2021 FRONTIERS IN PHYSICS EDITOR'S PICK

EDITED BY: Alex Hansen

PUBLISHED IN: Frontiers in Physics

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88971-076-8
DOI 10.3389/978-2-88971-076-8

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

2021 FRONTIERS IN PHYSICS EDITOR'S PICK

Topic Editor:
Alex Hansen, Norwegian University of Science and Technology, Norway

Citation: Hansen, A., ed. (2021). 2021 Frontiers in Physics Editor's pick. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88971-076-8

Table of Contents

frontiers in Physics

ORIGINAL RESEARCH
published: 24 January 2020
doi: 10.3389/fphy.2020.00004

Flow-Area Relations in Immiscible Two-Phase Flow in Porous Media

Subhadeep Roy[1]*, *Santanu Sinha*[2] *and Alex Hansen*[1,2]

[1] *PoreLab, Department of Physics, Norwegian University of Science and Technology, Trondheim, Norway,* [2] *Beijing Computational Science Research Center, Beijing, China*

We present a theoretical framework for immiscible incompressible two-phase flow in homogeneous porous media that connects the distribution of local fluid velocities to the average seepage velocities. By dividing the pore area along a cut transversal to the average flow direction up into differential areas associated with the local flow velocities, we construct a distribution function that allows us to not only re-establish existing relationships of between the seepage velocities of the immiscible fluids, but also to find new relations between their higher moments. We support and demonstrate the formalism through numerical simulations using a dynamic pore-network model for immiscible two-phase flow with two- and three-dimensional pore networks. Our numerical results are in agreement with the theoretical considerations.

Keywords: porous media, thermodynamic relations, seepage velocity, steady state flow, two phase flow

OPEN ACCESS

Edited by:
Daniel Bonamy,
Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA),
France

Reviewed by:
Wenzheng Yue,
China University of Petroleum, China
Eric Josef Ribeiro Parteli,
University of Cologne, Germany

****Correspondence:***
Subhadeep Roy
subhadeeproy03@gmail.com

Specialty section:
This article was submitted to Interdisciplinary Physics, a section of the journal Frontiers in Physics

Received: *27 November 2019*
Accepted: *06 January 2020*
Published: *24 January 2020*

Citation:
Roy S, Sinha S and Hansen A (2020) Flow-Area Relations in Immiscible Two-Phase Flow in Porous Media. Front. Phys. 8:4. doi: 10.3389/fphy.2020.00004

1. INTRODUCTION

When two immiscible fluids compete for the same pore space, we are dealing with immiscible two-phase flow in porous media [1]. A holy grail in porous media research is to find a proper description of immiscible two-phase flow at the continuum level, i.e., at scales where the porous medium may be treated as a continuum. Our understanding of immiscible two-phase flow at the pore level is increasing at a very high rate due to advances in experimental techniques combined with an explosive growth in computer power [2]. Still, the gap in scales between the physics at the pore level and a continuum description remains huge and the bridges that have been built so far across this gap are either complex to cross or rather rickety. To the latter class, we find the still dominating theory, first proposed by Wyckoff and Botset [3] and with an essential amendment by Leverett [4], namely relative permeability theory. The basic idea behind this theory is the following: Put yourself in the place of one of the two immiscible fluids. What does this fluid see? It sees a space in which it can flow limited by the solid matrix of the porous medium, but *also by the other fluid.* This reduces its mobility in the porous medium by a factor known as the relative permeability, which is a function of the how much space there is left for it. And here is the rickety part: this reduction of available space—expressed through the saturation—is the *only* parameter affecting the reduction factor or relative permeability. This is a very strong statement and clearly does not take into account that the distribution of immiscible fluid clusters will depend on how fast the fluids are flowing. Still, in the range of flow rates relevant for many industrial applications, this assumption works pretty well. It therefore, remains the essential work horse for practical applications.

Thermodynamically Constrained Averaging Theory (TCAT) [5–9] is built on the framework of relative permeability. However, it is based on a full analysis based on mechanical conservation laws, constitutive laws, e.g., for the motion of interfaces and contact lines, and on thermodynamics at the pore level. These are then scaled up using averaging theorems, which, loosely explained, consist of replacing derivatives of averages by averages of derivatives. In principle, this approach solves the up-scaling problem. However, as Gray and Miller point out in their book [9], each component of TCAT involves significant mathematical manipulations. The internal energy has contributions from the bulk liquids, the fluid-fluid and fluid-matrix interfaces at the pore level. The averaging process redefines the variables describing these contributions, but does not reduce their number. This accounts for a high level of complexity.

A further development somewhat along the same lines, based on non-equilibrium thermodynamics uses Euler homogeneity, more about this later, to define the up-scaled pressure. From this, Kjelstrup et al. derive constitutive equations for the flow [10, 11].

Another class of theories is based on detailed and specific assumptions concerning the physics involved. An example is Local Porosity Theory [12–17]. Another is DeProf theory which is a mechanical model combined with non-equilibrium statistical mechanics based on a classification scheme of fluid configurations at the pore level [18–20].

A recent work [21] has explored a new approach to immiscible two-phase flow in porous media based on elements borrowed from thermodynamics. That is, it is using the framework of thermodynamics, but without connecting it to processes involving heat. The spirit behind this approach is like that taken in Edwards and Oakshot's pseudo-thermodynamic theory of powders [22]. The approach consists in looking for general relations that transcends details of the physical processes involved. An example of such an approach in the field if immiscible two-phase flow in porous media is found in the Buckley-Leverett theory of invasion fronts [23]. The Buckley-Leverett equation is based solely on the principle of mass conservation and on the fractional flow rate being a function of the saturation. In the approach of Hansen et al. [21], equations are derived that originate from Euler homogeneity as in ordinary thermodynamics. These equations transcend the details of the physics involved in the same way that the equations of thermodynamics are universally applicable if a set of simple underlying conditions are met.

Thermodynamics is a theory that is valid on scales large enough so that the system it refers to may be regarded as a continuum. Statistical mechanics is then the theory that makes the connection between thermodynamics and the underlying atomistic picture.

It is the aim of this paper to formulate a description of immiscible two-phase flow in porous media that may form a link between the continuum-level approach of Hansen et al. [21] and the pore-level description of the problem—a sort of "statistical mechanics" from which the pseudo-thermodynamics may be derived, but which also describes the flow problem at the pore level.

After defining the system and the variables involved in section 2, we will in section 3 review the pseudo-thermodynamic approach [21]. The next section 4 we introduce the central object in the paper, the *differential transversal area distribution* which corresponds to the Boltzmann distribution in ordinary statistical mechanics, and relate it to the pseudo-thermodynamics relations. Then follows section 5 which then moves beyond the results of the pseudo-thermodynamics by focusing on fluctuations. In section 6 we use the dynamic network simulator [24] first introduced by Aker et al. [25] and then later refined [26–28] to verify the relations derived in the earlier sections. There is also a second goal behind this numerical work: the dynamic network model is a model at pore level and by its use, we show how the formalism developed here connect to the flow patterns at the pore level. Finally, we draw our conclusions in section 7.

2. SYSTEM DEFINITION

In two-phase flow, the steady state [29–31] is characterized by potentially strong fluctuations at the pore scale, but steady averages at the REV (Representative Elementary Volume) scale. As such they differ fundamentally from stationary states that are static at the pore scale as well. Steady states have much in common with ensembles in equilibrium statistical mechanics. They are also by implication assumed in the conventional descriptions of porous media flows that take the existence of an REV for granted.

Our REV is a block of homogeneous porous material of length L and area A. We prevent flow through the surfaces that are parallel to the L-direction which is the flow direction. The two remaining surfaces, each having an area A, act as inlet and outlet for the incompressible fluids that are injected and extracted from the REV. The porosity of the material is defined as

$$\phi = \frac{V_p}{V}, \tag{1}$$

where V_p is the pore volume and $V = AL$ is the volume of the REV. Due to the homogeneity of the porous medium, any cross section orthogonal to the axis along the L-direction will have a pore area that fluctuates around the value

$$A_p = \frac{V_p}{L} = \phi A. \tag{2}$$

There is also a solid matrix area fluctuating around

$$A_s = A - A_p = (1 - \phi)A. \tag{3}$$

The homogeneity assumption consists in the fluctuations being so small that they can be ignored.

There is a time averaged volumetric flow rate Q through the REV. The volumetric flow rate consists of two components, Q_w and Q_n, which are the volumetric flow rates of the more wetting (w for "wetting") and the less wetting (n for "non-wetting") fluids with respect to the porous medium. They are related through

$$Q = Q_w + Q_n. \tag{4}$$

In the porous medium, there is a volume V_w of the wetting fluid and a volume V_n of the non-wetting fluid so that $V_p = V_w + V_n$. We define the wetting and non-wetting saturations $S_w = V_w/V_p$ and $S_n = V_n/V_p$, so that $S_w + S_n = 1$.

We define the wetting and non-wetting transversal pore areas A_w and A_n as the parts of the transversal pore area A_p which occupied by the wetting or the non-wetting fluids, respectively. We have that

$$A_p = A_w + A_n. \tag{5}$$

As the porous medium is homogeneous, we will find the same averages A_w and A_n in any cross section through the porous medium orthogonal to the flow direction. We have therefore $A_w/A_p = (A_w L)/(A_p L) = V_w/V_p = S_w$, so that

$$A_w = S_w A_p. \tag{6}$$

Likewise,

$$A_n = S_n A_p = (1 - S_w)\, A_p. \tag{7}$$

We define the seepage velocities, i.e., the average flow velocities in the pores, for the two immiscible fluids, v_w and v_n as

$$v_w = \frac{Q_w}{A_w}, \tag{8}$$

and

$$v_n = \frac{Q_n}{A_n}. \tag{9}$$

The seepage velocity associated with the total flow rate Q is defined as

$$v_p = \frac{Q}{A_p}. \tag{10}$$

We may express Equation (4) in terms of the seepage velocities,

$$v_p = S_w v_w + S_n v_n. \tag{11}$$

3. PSEUDO-THERMODYNAMIC RELATIONS

Hansen et al. [21] derived a number of relations between the seepage velocities defined in (8)–(10) based on the volumetric flow rate being an Euler homogeneous function of order one with respect to the wetting and non-wetting transversal pore areas A_w and A_n. We present here a short review of the main results in that paper for completeness. The meaning of the statement that the volumetric flow rate is an Euler homogeneous function of order one is that it obeys the scaling relation

$$Q_p(\lambda A_w, \lambda A_n) = \lambda Q_p(A_w, A_n), \tag{12}$$

where λ is a scale factor. By taking the derivative of this equation with respect to λ and then setting $\lambda = 1$, we find

$$Q_p(A_w, A_n) = \left(\frac{\partial Q_p}{\partial A_w}\right)_{A_n} A_w + \left(\frac{\partial Q_p}{\partial A_n}\right)_{A_w} A_n. \tag{13}$$

By dividing this expression by the transversal pore area A_p and using Equations (5)–(7), we may write this equation as

$$v_p = S_w \left(\frac{\partial Q_p}{\partial A_w}\right)_{A_n} + S_n \left(\frac{\partial Q_p}{\partial A_n}\right)_{A_w}. \tag{14}$$

The two partial derivatives have the units of velocity, and Hansen et al. [21] name these velocity functions the *thermodynamic velocities,*

$$\hat{v}_w = \left(\frac{\partial Q}{\partial A_w}\right)_{A_n}, \tag{15}$$

and

$$\hat{v}_n = \left(\frac{\partial Q}{\partial A_n}\right)_{A_w}. \tag{16}$$

We use Equations (6) and (7) and the chain rule to derive

$$\begin{aligned}\left(\frac{\partial}{\partial A_w}\right)_{A_n} &= \left(\frac{\partial S_w}{\partial A_w}\right)_{A_n} \left(\frac{\partial}{\partial S_w}\right)_{A_p} \\ &+ \left(\frac{\partial A_p}{\partial A_w}\right)_{A_n} \left(\frac{\partial}{\partial A_p}\right)_{S_w} \\ &= \frac{S_n}{A_p} \left(\frac{\partial}{\partial S_w}\right)_{A_p} + \left(\frac{\partial}{\partial A_p}\right)_{S_w}.\end{aligned} \tag{17}$$

Likewise, we find

$$\left(\frac{\partial}{\partial A_n}\right)_{A_w} = -\frac{S_w}{A_p} \left(\frac{\partial}{\partial S_w}\right)_{A_p} + \left(\frac{\partial}{\partial A_p}\right)_{S_w}. \tag{18}$$

We now combine these two equations with the definitions (15) and (16), and use that $Q = A_p v_p$, i.e., Equation (10), to find

$$\hat{v}_w = v_p + S_n \frac{dv_p}{dS_w}, \tag{19}$$

and

$$\hat{v}_n = v_p - S_w \frac{dv_p}{dS_w}. \tag{20}$$

Combining the definitions (15) and (16) with Equation (14) gives

$$v_p = S_w \hat{v}_w + S_n \hat{v}_n, \tag{21}$$

which should be compared to Equation (11). We see that

$$S_w v_w + S_n v_n = S_w \hat{v}_w + S_n \hat{v}_n. \tag{22}$$

The seepage and thermodynamic velocities are related through a transformation $(v_w, v_n) \rightarrow (\hat{v}_w, \hat{v}_n)$ defining the *co-moving velocity* v_m,

$$\hat{v}_w = v_w + v_m S_n, \tag{23}$$

and

$$\hat{v}_n = v_n - v_m S_w. \tag{24}$$

We now calculate

$$\begin{aligned}\left(\frac{\partial Q}{\partial S_w}\right)_{A_p} = \left(\frac{\partial Q}{\partial A_w}\right)_{A_n} \left(\frac{\partial A_w}{\partial S_w}\right)_{A_p} + \\ \left(\frac{\partial Q}{\partial A_n}\right)_{A_w} \left(\frac{\partial A_n}{\partial S_w}\right)_{A_p}.\end{aligned} \tag{25}$$

Using Equations (6) and (7) together with Equations (19) and (20), we transform this equation into

$$\frac{dv_p}{dS_w} = \hat{v}_w - \hat{v}_n, \tag{26}$$

where we have used that $v_p = Q/A_p$, i.e., Equation (10). We now use Equation (21) to calculate

$$\frac{dv_p}{dS_w} = \hat{v}_w - \hat{v}_n + S_w \frac{d\hat{v}_w}{dS_w} + S_n \frac{d\hat{v}_n}{dS_w}. \tag{27}$$

Compare this equation to Equation (26) and we get an analog to the Gibbs-Duhem equation,

$$S_w \frac{d\hat{v}_w}{dS_w} + S_n \frac{d\hat{v}_n}{dS_w} = 0. \tag{28}$$

Using Equations (23) and (24), we find that the seepage velocities obey

$$\frac{dv_p}{dS_w} = v_w - v_n + v_m, \tag{29}$$

and

$$S_w \frac{dv_w}{dS_w} + S_n \frac{dv_n}{dS_w} = v_m, \tag{30}$$

where we have combined Equations (23) and (24) with Equation (28).

By combining Equations (15), (16), (23), and (24), one finds

$$v_w = v_p + S_n \left(\frac{dv_p}{dS_w} - v_m \right), \tag{31}$$

and

$$v_n = v_p - S_w \left(\frac{dv_p}{dS_w} - v_m \right). \tag{32}$$

These two equations, (31) and (32), may be seen as a transformation $(v_p, v_m) \rightarrow (v_w, v_n)$. The inverse of this transformation, i.e., $(v_w, v_n) \rightarrow (v_p, v_m)$ are given by Equations (11) and (29), i.e.,

$$\begin{aligned} v_p &= S_w v_w + S_n v_n, \\ v_m &= S_w v'_w + S_n v'_n, \end{aligned} \tag{33}$$

where $v'_w = dv_w/dS_w$ and $v'_n = dv_n/dS_w$.

But, what is the co-moving velocity v_m physically? We first need to understand the thermodynamic velocities $\hat{v}_w$ and $\hat{v}_n$. These are the velocities the two fluids would have had if they were miscible. Equation (26) then tells us that a change in the saturation S_w leads to a change in the average seepage velocity v_p which is the difference in seepage velocities of the two fluids. However, the two fluids are *not* miscible and they do get in each other's way. How much is dictated by the co-moving velocity through Equation (29).

From Equation (26) onwards to the end of this sections, none of the equations contain the size of the REV. If we now imagine a REV associated with each point in the porous medium, we have a continuum description. We may then add equations that transport the fluids between these points. Assuming that the fluids are incompressible, these equations are [1]

$$\phi \frac{\partial S_w}{\partial t} = \frac{\partial \phi S_w v_w}{\partial x}, \tag{34}$$

where t is the time coordinate and x is the spatial coordinate, and

$$\phi \frac{\partial S_n}{\partial t} = \frac{\partial \phi S_n v_n}{\partial x}. \tag{35}$$

We add the two equations and get

$$\frac{\partial}{\partial x} \phi v_p = 0. \tag{36}$$

The generalization to three dimensions is straight forward.

In order to connect the equations that now have been derived to a given porous medium, constitutive equations for v_p and v_m need to be supplied, linking the flow to the driving forces. These may in the simplest case be pressure gradient and saturation gradient.

4. DIFFERENTIAL TRANSVERSAL AREA DISTRIBUTIONS

In this section, we connect the pseudo-thermodynamic results of section 3 to the properties of an underlying ensemble distribution. This concept in the context of immiscible two-phase flow was first considered by Savani et al. [32]. Here we generalize this concept. In some sense, we introduce here a statistical mechanics from which the pseudo-thermodynamics ensue.

We define a *differential transversal pore area* $a_p = a_p(S_w, v)$ where v is a velocity such that $a_p dv$ is the pore area covered by fluid, wetting or non-wetting, that has a velocity in the range $[v, v + dv]$. Hence, a_p—and the other differential transversal pore areas that we will proceed to construct—are *statistical distributions of the pore level velocities.* The new idea we are introducing is that the velocity distribution is measured in terms of transversal pore areas. This makes it possible to make the connection between the flow at the pore level and the pseudo-thermodynamic theory reviewed in the previous section.

We must have that

$$A_p = \int_{-\infty}^{\infty} dv\, a_p, \tag{37}$$

where the integral runs over the entire range of negative and positive velocities since there may be local areas where the flow direction is opposite to the global flow. The total flow rate Q is given by

$$Q = \int_{-\infty}^{\infty} dv\, v\, a_p, \tag{38}$$

and the see page velocity defined in Equation (10) is then given by

$$v_p = \langle v \rangle_p = \frac{1}{A_p} \int_{-\infty}^{\infty} dv\, v\, a_p. \tag{39}$$

Likewise, we define a wetting differential pore area a_w and a non-wetting differential pore area a_n. They have the same properties except that they are restricted to the wetting or the non-wetting fluids only. That is, we have

$$A_w = \int_{-\infty}^{\infty} dv\, a_w, \tag{40}$$

and

$$A_n = \int_{-\infty}^{\infty} dv\, a_n. \tag{41}$$

They relate to the wetting and non-wetting seepage velocities defined in Equations (8) and (9) as

$$v_w = \langle v \rangle_w = \frac{1}{A_w} \int_{-\infty}^{\infty} dv\, v\, a_w, \tag{42}$$

and

$$v_n = \langle v \rangle_n = \frac{1}{A_n} \int_{-\infty}^{\infty} dv\, v\, a_n. \tag{43}$$

We have that

$$a_p = a_w + a_n. \tag{44}$$

We now combine this equation with Equation (39) to find

$$\begin{aligned} v_p &= \frac{1}{A_p} \int_{-\infty}^{\infty} dv\, v\, (a_w + a_n) \\ &= \left(\frac{A_w}{A_p}\right) \frac{1}{A_w} \int_{-\infty}^{\infty} dv\, v\, a_w \\ &+ \left(\frac{A_n}{A_p}\right) \frac{1}{A_n} \int_{-\infty}^{\infty} dv\, v\, a_n \\ &= S_w v_w + S_n v_n, \end{aligned} \tag{45}$$

which is Equation (11). We have here used Equations (6) and (7).

We may associate a differential area a_m to the co-moving velocity v_m defined in Equation (29). By using Equations (39), (42), and (43) in combination with Equation (29), we find

$$\begin{aligned} v_m &= \frac{dv_p}{dS_w} - v_w + v_n \\ &= \frac{1}{A_p} \int_{-\infty}^{\infty} dv\, v \left[\frac{\partial a_p}{\partial S_w} - \frac{a_w}{S_w} + \frac{a_n}{S_n}\right], \end{aligned} \tag{46}$$

so that

$$\begin{aligned} a_m &= \frac{\partial a_p}{\partial S_w} - \frac{a_w}{S_w} + \frac{a_n}{S_n} \\ &= \left(\frac{\partial a_w}{\partial S_w} - \frac{a_w}{S_w}\right) + \left(\frac{\partial a_n}{\partial S_w} + \frac{a_n}{S_n}\right), \end{aligned} \tag{47}$$

where we have used Equation (44). Equation (47) may be rewritten as

$$a_m = S_w \frac{\partial}{\partial S_w}\left(\frac{a_w}{S_w}\right) + S_n \frac{\partial}{\partial S_w}\left(\frac{a_n}{S_n}\right). \tag{48}$$

Averaging this equation over v and using Equations (42), (43), and (46) recovers Equation (30). Hence, we note that Equations (47) and (48) are the generalizations of Equations (29) and (30) to the differential transversal areas.

It follows that

$$A_m = \int_{-\infty}^{\infty} dv\, a_m = 0, \tag{49}$$

where A_m is the pore area associated with co-moving velocity v_m. This is to expected as the areas A_w, A_n, A_p and A_m are ways to partition the transversal pore area A_p; and we have that $A_w + A_n = A_p + 0$. This implies that there is no volumetric flow rate associated with the co-moving velocity since

$$Q_m = A_m v_m = 0. \tag{50}$$

Lastly, we may associate differential transversal areas to the thermodynamic velocities defined in Equations (19) and (20). We use Equations (23) and (24) to find

$$\hat{a}_w = a_w + S_n S_w\, a_m, \tag{51}$$

and

$$\hat{a}_n = a_n - S_w S_n\, a_m, \tag{52}$$

where a_m is given in Equation (48). The thermodynamic velocities are then given by

$$\hat{v}_w = \frac{1}{A_w} \int_{-\infty}^{\infty} dv\, v\, \hat{a}_w, \tag{53}$$

and

$$\hat{v}_n = \frac{1}{A_n} \int_{-\infty}^{\infty} dv\, v\, \hat{a}_n. \tag{54}$$

We find as expected that

$$\hat{A}_w = \int_{-\infty}^{\infty} dv\, \hat{a}_w = A_w, \tag{55}$$

and

$$\hat{A}_n = \int_{-\infty}^{\infty} dv\, \hat{a}_n = A_n. \tag{56}$$

Summing the two differential transversal areas for the thermodynamic areas gives

$$\hat{a}_w + \hat{a}_n = a_w + a_n = a_p. \tag{57}$$

This leads us to an important remark. The differential transversal areas are statistical velocity distributions at the pore level. We see that the differential transversal areas that are associated with the thermodynamic velocities are different from those associated with the seepage velocities. However, Equation (57) shows that the *combined* differential transversal area based upon the thermodynamic velocity distributions is the same as that based upon the distributions giving the seepage velocities. Hence, the two types of differential transversal areas represent a redistribution of the pore level velocities, but in such a way that A_w and A_n are preserved. We see the same from Equation (49) showing that A_m is zero and combining this Equations (51) and (52).

We see from Equation (47) that a_m is only zero if a_w and a_n are linear in S_w and S_n, respectively, i.e., $a_w = S_w b_w$ where b_w is independent of S_w and $a_n = S_n b_n$ where b_n is independent of S_n. Hence, this is the condition for the thermodynamic velocities to be equal to the seepage velocities.

5. MOMENTS AND FLUCTUATIONS

We define the qth moment of the seepage velocity distribution as

$$v_p^q = \langle v^q \rangle_p = \frac{1}{A_p} \int_{-\infty}^{\infty} dv\, v^q a_p. \tag{58}$$

By using Equation (44) we find immediately

$$v_p^q = v_w^q S_w + v_n^q S_n, \tag{59}$$

where we have defined

$$v_w^q = \langle v^q \rangle_w = \frac{1}{A_w} \int_{-\infty}^{\infty} dv\, v^q\, a_w, \tag{60}$$

and

$$v_n^q = \langle v^q \rangle_n = \frac{1}{A_n} \int_{-\infty}^{\infty} dv\, v^q\, a_n. \tag{61}$$

We may work out the moments of the co-moving velocity are given by

$$v_m^q = \frac{1}{A_p} \int_{-\infty}^{\infty} dv\, v^q\, a_m = \left[\frac{dv_p^q}{dS_w} - v_w^q + v_n^q \right], \tag{62}$$

where we have used (47).

The thermodynamic velocity moments may be defined as in a similar manner as the moments of the seepage velocities, (60) and (61),

$$\hat{v}_w^q = \langle \hat{v}^q \rangle_w = \frac{1}{A_w} \int_{-\infty}^{\infty} dv\, v^q\, \hat{a}_w, \tag{63}$$

and

$$\hat{v}_n^q = \langle \hat{v}^q \rangle_n = \frac{1}{A_n} \int_{-\infty}^{\infty} dv\, v^q\, \hat{a}_n. \tag{64}$$

and we find

$$\hat{v}_p^q = \hat{v}_w^q S_w + \hat{v}_n^q S_n, \tag{65}$$

where we have used Equations (52) and (55).

We may Fourier transform a_p, a_w, and a_n,

$$2\pi\, \tilde{a}_p(\omega) = A_p \langle e^{iv\omega} \rangle_p = \int_{-\infty}^{\infty} dv\, e^{iv\omega}\, a_p, \tag{66}$$

$$2\pi\, \tilde{a}_w(\omega) = A_w \langle e^{iv\omega} \rangle_w = \int_{-\infty}^{\infty} dv\, e^{iv\omega}\, a_w, \tag{67}$$

and

$$2\pi\, \tilde{a}_n(\omega) = A_n \langle e^{iv\omega} \rangle_n = \int_{-\infty}^{\infty} dv\, e^{iv\omega}\, a_n. \tag{68}$$

From Equation (44) we find

$$\tilde{a}_p(S_w, \omega) = \tilde{a}_w(S_w, \omega) + \tilde{a}_n(S_w, \omega), \tag{69}$$

and

$$\langle e^{iv\omega} \rangle_p = S_w \langle e^{iv\omega} \rangle_w + S_n \langle e^{iv\omega} \rangle_n. \tag{70}$$

We write $\langle \exp(iv\omega) \rangle_p$ as a cumulant expansion,

$$\langle e^{iv\omega} \rangle_p = \exp\left(\sum_{k=1}^{\infty} \frac{(i\omega)^k}{k!}\, C_p^k \right), \tag{71}$$

where C_p^k is the kth cumulant. We define the wetting and non-wetting velocity cumulants C_w^k and C_n^k in the same way. We also write $\langle \exp(iv\omega) \rangle_p$ as a moment expansion

$$\langle e^{iv\omega} \rangle_p = \sum_{m=0}^{\infty} \frac{(i\omega)^m}{m!}\, v_p^m. \tag{72}$$

By expanding the cumulant expression in Equation (71) and equating each power in $i\omega$ with the corresponding one in Equation (72), then repeating this for the wetting and non-wetting cumulants, and lastly combining them through Equation (70), we find for the term proportional to $(i\omega)^2$,

$$C_p^2 + (C_p^1)^2 = [C_w^2 + (C_w^1)^2] S_w + [C_n^2 + (C_n^1)^2] S_n. \tag{73}$$

Noting that $C_p^1 = v_p$, $C_w^1 = v_w$, and $C_n^1 = v_n$ and using that $\Delta v_p^2 = C_p^2$, $\Delta v_w^2 = C_w^2$, and $\Delta v_n^2 = C_n^2$, we find from this equation

$$\Delta v_p^2 = \Delta v_w^2 S_w + \Delta v_n^2 S_n + S_w S_n \left(v_w - v_n \right)^2. \tag{74}$$

We may follow this procedure for any of the cumulants.

The corresponding equation between the second cumulants of the thermodynamic velocities is

$$\Delta \hat{v}_p^2 = \Delta \hat{v}_w^2 S_w + \Delta \hat{v}_n^2 S_n + S_w S_n \left(\hat{v}_w - \hat{v}_n \right)^2. \tag{75}$$

6. NUMERICAL OBSERVATIONS

The relations presented in sections 4, 5 provide the bridge between the velocity distributions at the pore level and the pseudo-thermodynamic theory outlined in section 3. In order to test these relations, and to show how they may be used, we use a dynamic pore network simulator [24].

In pore network modeling, the porous medium is represented by a network of pores which transport two immiscible fluids. The pore-network model we consider here can be applied to regular networks such as a regular lattice with an artificial disorder as well as to irregular networks such as a reconstructed network from real samples. The flow of the two immiscible fluids is described in this model by keeping the track of all interface positions with time. This approach of pore network modeling was first introduced by Aker et al. [25] for drainage displacements in a regular network. Over the last two decades, new mechanisms have been developed to extend the model for the steady-state flow as well as for irregular networks. A detailed description of this model in its most recent form can be found in Gjennestad et al. [26, 27] and Sinha et al. [28] and we therefore describe it here only briefly.

The porous medium is represented by a network of links that are connected at nodes. All the pore space in this model is assigned to the links and, hence, the nodes do not contain any volume, they only represent the positions where the links meet. The flow rate q_j inside any link j of the network at any instant of time for fully developed viscous flow is obtained by [33, 34],

$$q_j = -\frac{g_j}{l_j \mu_j}\left[\Delta p_j - \sum p_{c,j}\right] \tag{76}$$

where Δp_j is the pressure drop across link, l_j is the link length and g_j is the link mobility which depends on the cross section of the link. The viscosity term μ_j is the saturation-weighted viscosity of the fluids inside the link given by $\mu_j = s_{j,w}\mu_w + s_{j,n}\mu_n$ where μ_w and μ_n are the wetting and non-wetting viscosities and $s_{j,w}$ and $s_{j,n}$ are the wetting and non-wetting fluid saturations inside the link, respectively. The term $\sum p_{c,j}$ corresponds to the sum of all the interfacial pressures inside the jth link. A pore typically consists of two wider pore bodies connected by a narrow pore throat. We model this by using hour-glass shaped links. The variation of the interfacial pressure with the interface position for such a link is modeled by [34],

$$|p_c(x)| = \frac{2\gamma \cos\theta}{r_j}\left[1 - \cos\left(\frac{2\pi x}{l_j}\right)\right] \tag{77}$$

where r_j is the average radius of the link and $x \in [0, l_j]$ is the position of the interface inside the link. Here γ is the surface tension between the fluids and θ is the contact angle between the interface and the pore wall. These two Equations (77) and (76), together with the Kirchhoff relations, that is, the sum of the net volume flux at every node at each time step will be zero, provide a set of linear equations. We solve these equations with conjugate gradient solver [35] to calculate the local flow rates. All the interfaces are then advanced accordingly with small time steps. In order to achieve steady-state flow, we apply periodic boundary conditions in the direction of flow.

We construct a diamond lattice with 64 × 64 links in two dimensions (2D) with link lengths $l_j = 1\,\mathrm{mm}$ for each link. Disorder is introduced by choosing the link radii r_j randomly from a uniform distribution in the range 0.1 mm and 0.4 mm. We use 10 different realizations of such network for our simulations in 2D. In three dimensions (3D), we use a network reconstructed from a $1.8 \times 1.8 \times 1.8\,\mathrm{mm}^3$ sample of Berea sandstone that contains 2,274 links and 1,163 nodes [36]. Simulations are performed under constant pressure drop ΔP across the network. For 2D, we have considered 3 different values for pressure drop such that, $\Delta P/L = 0.5$, 1.0, and 1.5 MPa/m. For the 3D network, values of $\Delta P/L$ are chosen as, 10, 20, 40, and 80 MPa/m. The values for surface tension γ are chosen to be 0.02, 0.03, and 0.04 N/m for both 2D and 3D. Three different values of viscosity ratios $M(= \mu_n/\mu_w) = 0.5$, 1.0, and 2.0 are considered. These values are chosen in such a way that the capillary number, defined as

$$\mathrm{Ca} = \frac{\mu_e Q}{\gamma A_p} \tag{78}$$

falls in a range of around 10^{-3} to 10^{-1}. Here μ_e is the saturation weighted effective viscosity of the system given by $\mu_e = S_w\mu_w + S_n\mu_n$. Specifically, we find Ca in the range of 0.004–0.074 for 2D and 0.001–0.271 for 3D in the steady state. As the simulations are performed under constant pressure drop, the capillary number fluctuates. Ca is therefore calculated as functions of time by measuring the total flow rate Q along any cross section of the network perpendicular to the applied pressure drop. For any set of parameters, saturations are varied in the steps of 0.05 from 0 to 1 which correspond to 21 saturation values.

The simulations are continued to the steady state which is defined by the global measurable quantities, such as the fractional flow or the total flow rate Q fluctuate around a steady average. In the steady state, we calculate the seepage velocities averaged over time. First we use direct measurements, where we measure the global flow rates (Q, Q_w, and Q_n) and the pore areas (A_p, A_w, and A_n) through any cross section orthogonal to the applied pressure drop and then use Equations (8)–(10) to calculate the seepage velocities. Next, we perform the measurements using the differential pore areas (a_p, a_w, and a_n) and calculate the seepage velocities using description given in section 4. We then compare the results from the two measurements and calculate the co-moving velocities. We then verify the relation between the seepage velocities and their higher moments.

For the direct measurements, imagine a cross section at any place of the network orthogonal to the overall direction of flow. For the regular diamond lattice in 2D, all the links have the same length. Different moments of the seepage velocities can therefore

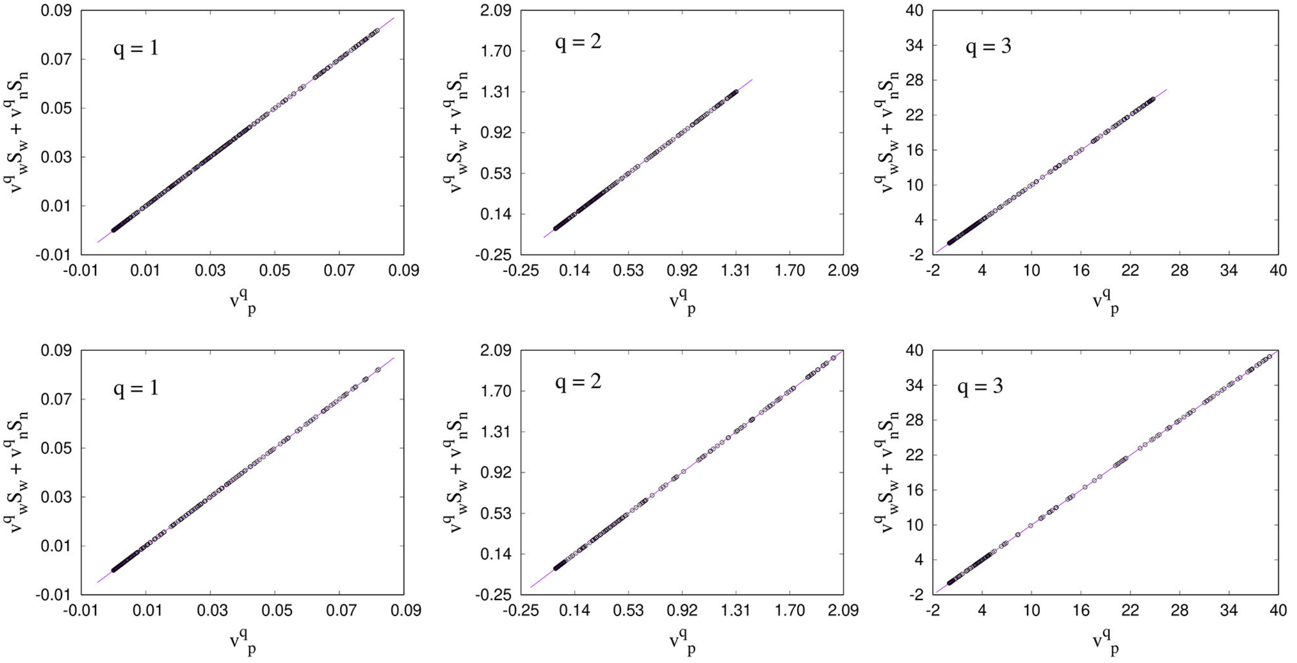

FIGURE 1 | Verification of the relations (11), (45), and (59) between the steady-state seepage velocities v_p, v_w, and v_n, and their higher moments for the 2D regular network. The top row represents the direct approach of measurements using Equations (79), (80), and (81). The bottom row corresponds to the velocities measured from the differential area distributions defined in Equations (58), (60), and (61). v_p has a unit mm/s. Subsequently, for $q = 2$ and 3, the units for v_p^q will be mm^2/s and mm^3/s, respectively.

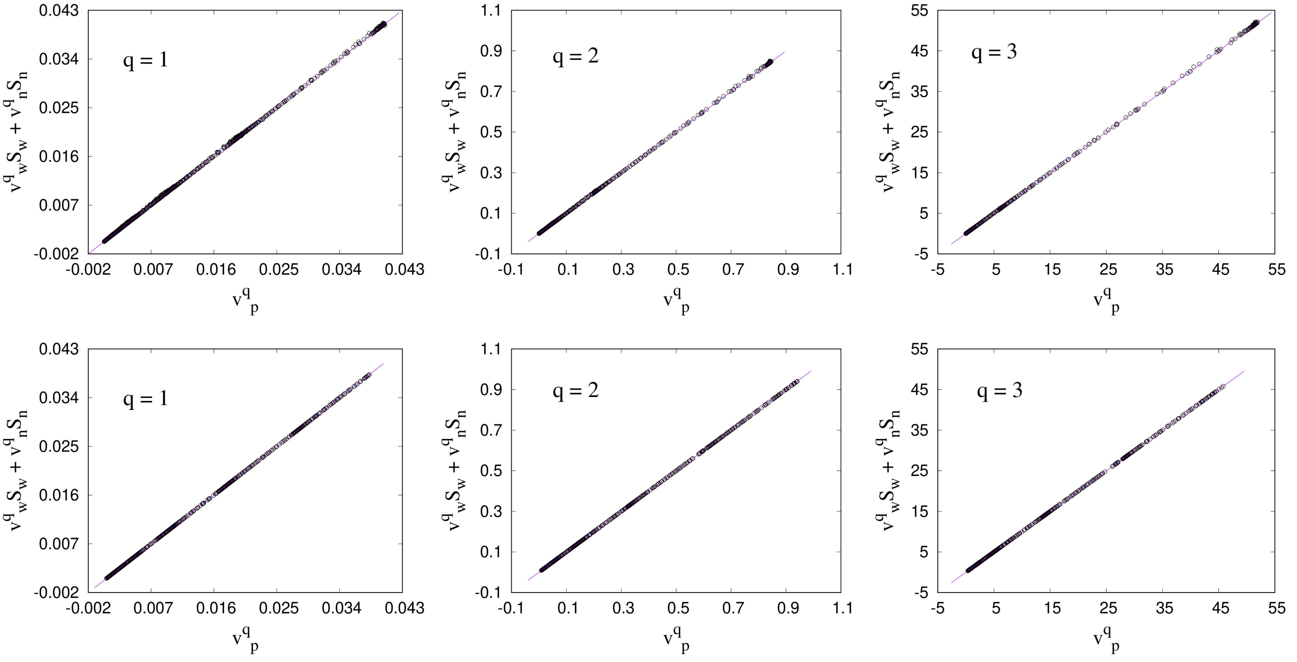

FIGURE 2 | Verification of the relations (11), (45), and (59) between the steady-state seepage velocities v_p, v_w, and v_n, and their higher moments for the 3D Berea network. The direct approach of measurements using Equations (82)–(84) are presented in the top row. The measurements using the differential area distributions defined in Equations (58), (60), and (61) are presented in the bottom row. v_p has a unit mm/s. Subsequently, for $q = 2$ and 3, the units for v_p^q will be mm^2/s and mm^3/s, respectively.

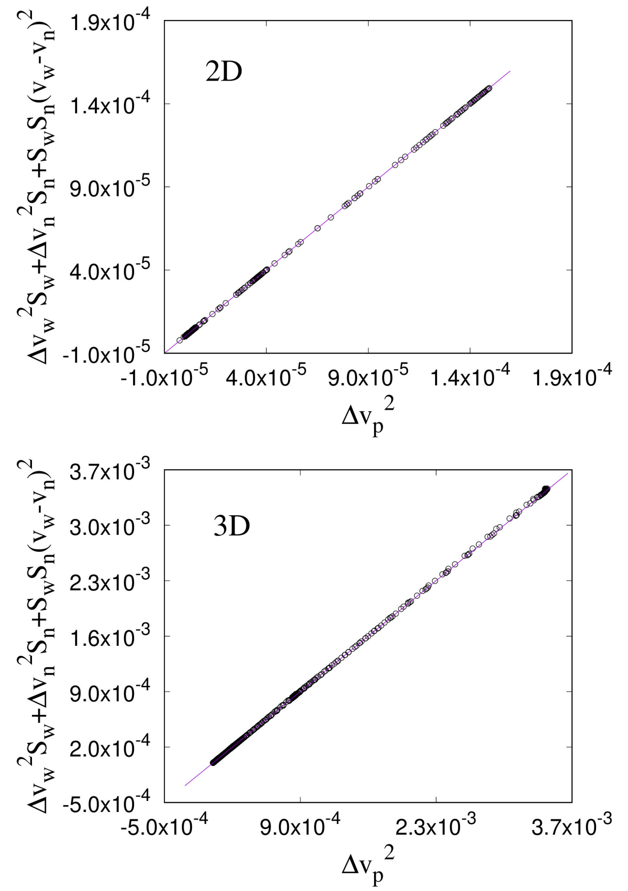

FIGURE 3 | Numerical verification of Equation (74) between the fluctuations in the seepage velocities. Δv_p^2 has a unit mm^2/s.

be calculated by

$$v_p^q = \frac{\sum_j \left(\frac{q_j}{a_j}\right)^q a_j}{\sum_j a_j}, \tag{79}$$

$$v_w^q = \frac{\sum_j \left(\frac{q_j}{a_j}\right)^q a_j S_{w,j}}{\sum_j a_j S_{w,j}}, \tag{80}$$

and

$$v_n^q = \frac{\sum_j \left(\frac{q_j}{a_j}\right)^q a_j S_{n,j}}{\sum_j a_j S_{n,j}}, \tag{81}$$

where a_j is the projection of the pore area of the jth link on the cross sectional plane. Here, all links have the same angle $\alpha = 45°$ with the direction of the overall flow. However, in case of the irregular network in 3D, Equations (79)–(81) need to be modified as the links have different lengths and orientations. In such case, the one can calculate the seepage velocities by [28],

$$v_p^q = \frac{\sum_j \left(\frac{q_j}{a_j}\right)^q a_j l_{x,j}}{\sum_j a_j l_{x,j}}, \tag{82}$$

$$v_w^q = \frac{\sum_j \left(\frac{q_j}{a_j}\right)^q a_j S_{w,j} l_{x,j}}{\sum_j a_j S_{w,j} l_{x,j}}, \tag{83}$$

and

$$v_n^q = \frac{\sum_j \left(\frac{q_j}{a_j}\right)^q a_j S_{n,j} l_{x,j}}{\sum_j a_j S_{n,j} l_{x,j}}, \tag{84}$$

where $l_{x,j} = l_j \cos \alpha_j$ is the projection of the link length (l_j) to the direction of the overall flow.

If we consider every link having the same length $l_j = l$ and same orientations $\alpha_j = \alpha$ in these equations, we retrieve the Equations (79)–(81). For the first moment ($q = 1$), the velocities v_p^q, v_w^q and v_n^q are equivalent to Q/A_p, Q/A_w, and Q/A_n, respectively, in both 2D and 3D.

For the second approach, we construct the distribution of differential transversal pore areas a_p, a_w, and a_n such that $a_p dv$, $a_w dv$, and $a_n dv$ express the transversal pore areas for the total, wetting and non-wetting fluids within the velocity range from v to $v + dv$, so that they satisfy Equations (38), (40), and (41). We therefore have,

$$\begin{aligned} a_p(v)dv &= \frac{1}{L}\sum_j a_j l_{x,j}, \\ a_w(v)dv &= \frac{1}{L}\sum_j a_j l_{x,j} S_{w,j}, \\ a_n(v)dv &= \frac{1}{L}\sum_j a_j l_{x,j} S_{n,j}, \end{aligned} \tag{85}$$

where j runs over all the sites satisfying the condition: $v < v_j < v + dv$, v_j being the local velocity of link j. In case of the 2D lattice, $l_{x,j}$s are same for any j and given by $l_{x,j} = l/\sqrt{2}$. With these, different moments of the seepage velocities are then calculated using Equations (58), (60), and (61), respectively.

For any saturation, the seepage velocities and their higher moments should follow the relations (11), (45), and (59). We plot our numerical measurements in **Figures 1**, **2** for 2D and 3D, respectively. The upper row in each figure corresponds to the direct measurements and the lower row correspond to the

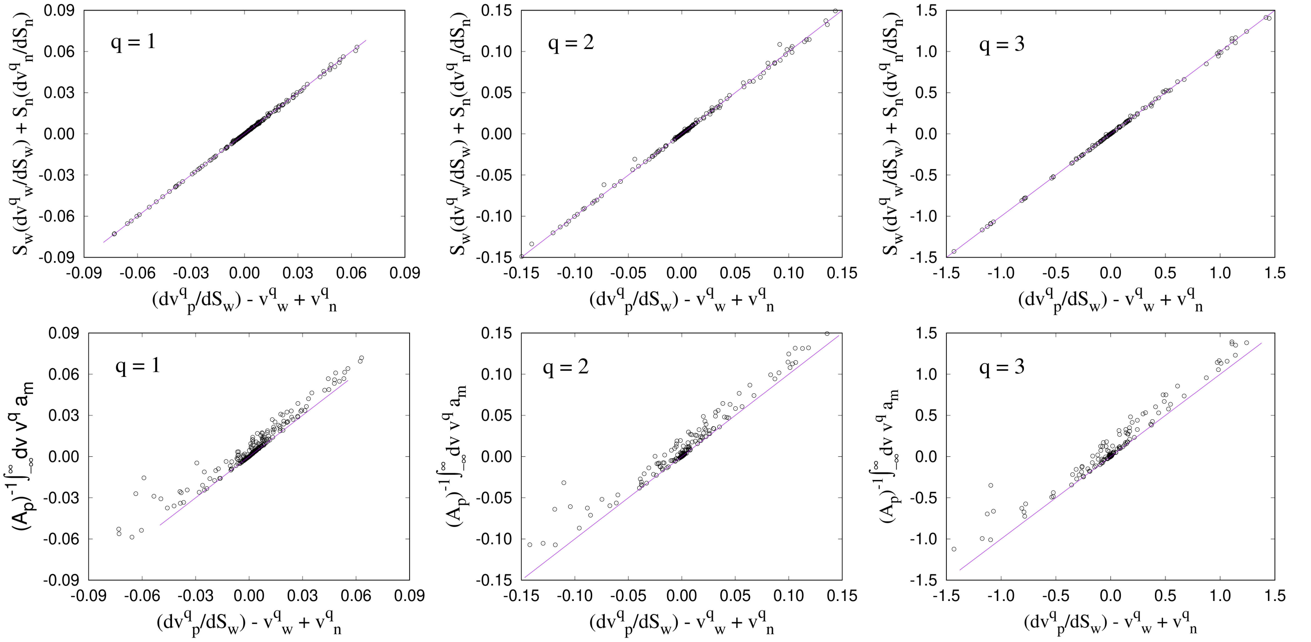

FIGURE 4 | Measurement of the co-moving velocity (v_m) and its higher moments for the 2D network. The top row corresponds to the calculations using Equations (29) and (30) with the direct measurements. The bottom row shows the measurements of v_m^q using the differential area distributions with Equation (46) and compared with the direct measurements where higher fluctuations are observed. v_m has a unit mm/s. Subsequently, for $q = 2$ and 3, the units for v_m^q will be mm^2/s and mm^3/s, respectively.

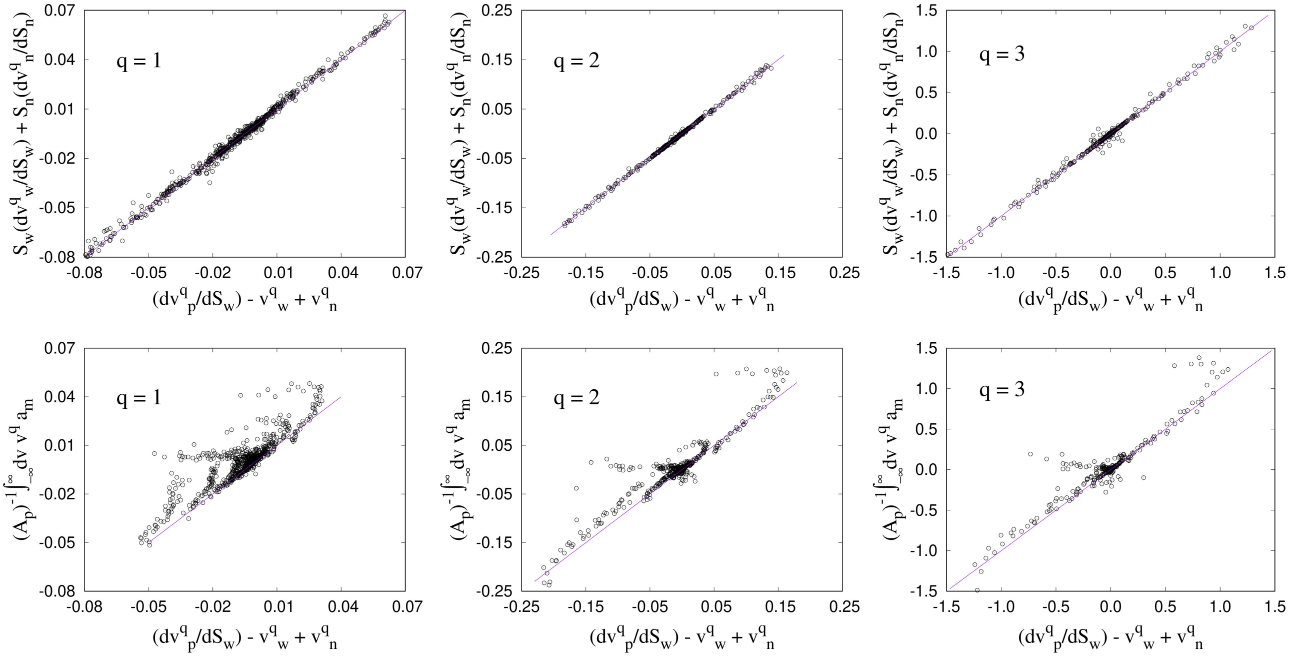

FIGURE 5 | Measurements of v_m^q for the 3D Berea network where the top row corresponds to the direct measurements using Equations (29) and (30), and the bottom row corresponds to the measurement from the differential area distributions using Equation (46). Here, larger fluctuations in the results calculated with the differential pore area are observed compared to the 2D network. v_m has a unit mm/s. Subsequently, for $q = 2$ and 3, the units for v_m^q will be mm^2/s and mm^3/s, respectively.

measurements from the differential area distribution. A good agreement with the relations can be observed for first as well as for the higher moments for both the networks.

Next we measure the fluctuations in the seepage velocities which obey Equation (74). Numerically, Δv_p^2, Δv_w^2, and Δv_n^2 are calculated from the knowledge of the 1st and 2nd moments by,

$$\begin{aligned}\Delta v_p^2 &= \langle v^2 \rangle_p - \langle v \rangle_p^2,\\ \Delta v_w^2 &= \langle v^2 \rangle_w - \langle v \rangle_w^2,\\ \Delta v_n^2 &= \langle v^2 \rangle_n - \langle v \rangle_n^2. \end{aligned} \tag{86}$$

In **Figure 3**, we plot these fluctuations for the two networks to compare with Equation (74) and good agreements is observed. There are some deviations in the results for the Berea network, since the results in 3D is based on only one network configuration whereas the results for 2D are averaged over 10 different configurations.

Finally, we verify the relations between seepage velocities and their higher moments while varying the fluid saturation as given by the Equations (29), (30), and (62). For this, we first calculated the co-moving velocity (v_m) and its higher moments from Equations (29) and (30) where we used the values of the seepage velocities measured with the direct approach. This is shown in the top rows of **Figures 4**, **5** for 2D and 3D, respectively, which show good agreements with Equations (29) and (30). We then compare these values of v_m^q with the measurements from the differential transversal areas using Equation (46). For this, we first constructed the histogram for the differential pore area a_m corresponding to the co-moving velocity from Equation (47) where we have used the variations of a_p, a_w, and a_n with the saturation S_w. For this purpose, we have considered 21 different values of saturations within 0 and 1 with an interval of 0.05. We then integrate a_m from $-\infty$ to ∞, weighted by the velocity and normalized by the total pore area to obtain the desired co-moving velocity with Equation (46). These results are plotted in the bottom row of **Figures 4**, **5** where they are compared with the results from direct measurements. The data points roughly follow the diagonal straight line showing satisfactory agreement with the theoretical formulations. However, we observe deviations in the results that is higher compared to the direct measurements. We believe this is due to the numerical errors that added up from several steps in the calculation such as the binning techniques while measuring the distributions, taking the derivatives and calculating the integrals. Moreover, the fluctuations for 3D are much higher compared to 2D, which is due to the lack of averaging over different samples as we have already mentioned earlier.

7. SUMMARY

The aim of this paper is to provide the link between the pseudo-thermodynamic theory at the continuum level developed in Hansen et al. [21] (see section 3) and the velocities occurring at the pore level during immiscible two-phase flow in porous media. This link is provided by defining the differential transversal pore areas defined in section 4, which essentially correspond to the statistical distributions of velocities at the pore level. The central quantities are the velocity differential transversal pore area a_p, the wetting fluid differential velocity transversal pore area a_w, the non-wetting fluid velocity differential transversal pore area a_n, and the co-moving velocity differential transversal pore area a_m. We also consider the thermodynamic velocity differential transversal pore areas $\hat{a}_w$ and $\hat{a}_n$. The relations found by Hansen et al. [21] for the average seepage velocities, the co-moving velocity and the thermodynamic velocities are generalized to the differential transversal areas here. In the following section 5, the relations are generalized to higher moments of the velocity distributions.

The theoretical derivations are then in section 6 validated by numerical simulations. We used dynamic pore-network modeling where an interface-tracking model is used to simulate steady-state two-phase flow. We used both regular pore networks and an irregular pore network reconstructed from a Berea sandstone for our simulations. By measuring the seepage velocities from the differential area distributions and comparing them with the direct measurements, we validated the essential predictions from the earlier theoretical sections.

Both Hansen et al. [21] and the present paper are to be seen as installments toward a theory for immiscible flow in porous media at the continuum scale. The structure of this theory reflects that found in thermodynamics: A set of general relations between the macroscopic variables based on energy conservation (i.e., the Gibbs relation) and Euler homogeneity. These general equations then have to be complemented by an equation of state which introduces the specifics of the system at hand. In the immiscible two-phase flow theory we are presenting here, Euler homogeneity and mass conservation provide the general equations that transcend the specifics of the porous medium. These general equations then have to be complemented by the constitutive equations for v_p and v_m, which provide the specifics of the porous medium.

The resulting set of equations may then be solved for structured porous media where the structure are associated with length scales larger than that set by the REV. This is e.g., seen in the explicit appearance of the porosity ϕ in Equations (34)–(36).

An open question, though, is what happens when there is non-trivial structure in the porous medium all the way from the pore scale to the continuum scale, see [37] and [38]—or when the saturation of the system is at a critical value, see [36]. The fundamental Euler scaling assumption (12) would then need to be modified, and with it, all the ensuing equations.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

SR did the numerical simulations and analysis. SS developed the codes and performed the 3D simulations. AH developed the theory.

FUNDING

SS was supported by the National Natural Science Foundation of China under grant number 11750110430. This work was partly supported by the Research Council of Norway through its Center of Excellence funding scheme, project number 262644.

ACKNOWLEDGMENTS

The authors thank Carl Fredrik Berg, M. Aa. Gjennestad, Knut Jørgen Måløy, Per Arne Slotte, Ole Torsæter, Morten Vassvik, and Mathias Winkler for interesting discussions. AH thanks the Beijing Computational Science Research Center CSRC for the hospitality and financial support.

REFERENCES

1. Bear J. *Dynamics of Fluids in Porous Media.* Mineola, NY: Dover (1988).
2. Blunt MJ. *Multiphase Flow in Permeable Media.* Cambridge: Cambridge University Press (2017).
3. Wyckoff RD, Botset HG. The flow of gas-liquid mixtures through unconsolidated sands. *Physics.* (1936) **7**:325–45. doi: 10.1063/1.1745402
4. Leverett MC. Capillary behavior in porous sands. *Trans AIMME.* (1940) **12**:152.
5. Hassanizadeh SM, Gray WG. Mechanics and thermodynamics of multiphase flow in porous media including interphase boundaries. *Adv Water Res.* (1990) **13**:169–86.
6. Hassanizadeh SM, Gray WG. Towards an improved description of the physics of two-phase flow. *Adv Water Res.* (1993) **16**:53–67.
7. Hassanizadeh SM, Gray WG. Thermodynamic basis of capillary pressure in porous media. *Water Resour Res.* (1993) **29**:3389–405.
8. Niessner J, Berg S, Hassanizadeh SM. Comparison of two-phase Darcy's law with a thermodynamically consistent approach. *Transp Por Med.* (2011) **88**:133–48. doi: 10.1007/s11242-011-9730-0
9. Gray WG, Miller CT. *Introduction to the Thermodynamically Constrained Averaging Theory for Porous Medium Systems*. Berlin: Springer Verlag (2014).
10. Kjelstrup S, Bedeaux D, Hansen A, Hafskjold B, Galteland O. Non-isothermal transport of multi-phase fluids in porous media. The entropy production. *Front Phys.* (2018) **6**:126. doi: 10.3389/fphy.2018.00126
11. Kjelstrup S, Bedeaux D, Hansen A, Hafskjold B, Galteland O. Non-isothermal transport of multi-phase fluids in porous media. Constitutive equations. *Front Phys*. (2019) **6**:150. doi: 10.3389/fphy.2018.00150
12. Hilfer R, Besserer H. Macroscopic two-phase flow in porous media. *Phys B*. (2000) **279**:125–9. doi: 10.1016/S0921-4526(99)00694-8
13. Hilfer R. Capillary pressure, hysteresis and residual saturation in porous media. *Phys A*. (2006) **359**:119–28. doi: 10.1016/j.physa.2005.05.086
14. Hilfer R. Macroscopic capillarity and hysteresis for flow in porous media. *Phys Rev E.* (2006) **73**:016307. doi: 10.1103/PhysRevE.73.016307
15. Hilfer R. Macroscopic capillarity without a constitutive capillary pressure function. *Phys A*. (2006) **371**:209–25. doi: 10.1016/j.physa.2006.04.051
16. Hilfer R, Döster F. Percolation as a basic concept for capillarity. *Transp Por Med*. (2010) **82**:507–19. doi: 10.1007/s11242-009-9395-0
17. Döster F, Hönig O, Hilfer R. Horizontal flow and capillarity-driven redistribution in porous media. *Phys Rev E.* (2012) **86**:016317. doi: 10.1103/PhysRevE.86.016317
18. Valavanides MS, Constantinides GN, Payatakes AC. Mechanistic model of steady-state two-phase flow in porous media based on Ganglion dynamics. *Transp Porous Media.* (1998) **30**:267–99. doi: 10.1023/A:1006558121674
19. Valavanides MS. Steady-state two-phase flow in porous media: review of progress in the development of the DeProF theory bridging pore-to statistical thermodynamics-scales. *Oil Gas Sci Technol.* (2012) **67**:787–804. doi: 10.2516/ogst/2012056
20. Valavanides MS. Review of steady-state two-phase flow in porous media: independent variables, universal energy efficiency map, critical flow conditions, effective characterization of flow and pore network. *Transp Porous Media.* (2018) **123**:45–99. doi: 10.1007/s11242-018-1026-1
21. Hansen A, Sinha S, Bedeaux D, Kjelstrup S, Gjennestad MA, Vassvik M. Relations between seepage velocities in immiscible, incompressible two-phase flow in porous media. *Transp Porous Media.* (2018) **125**:565–87. doi: 10.1007/s11242-018-1139-6
22. Edwards SF, Oakeshott RBS. Theory of powders. *Phys A*. (1989) **157**, 1080–90.
23. Buckley SE, Leverett MC. Mechanism of fluid displacements in sands. *Trans AIME.* (1942) **146**:107–17.
24. Joekar-Niasar V, Hassanizadeh SM. Analysis of fundamentals of two-phase flow in porous media using dynamic pore-network models: a review. *Crit Rev Environ Sci Technol*. (2012) **42**: 1895–76. doi: 10.1080/10643389.2011.574101
25. Aker E, Måløy KJ, Hansen A, Batrouni GG. A two-dimensional network simulator for two-phase flow in porous media. *Transp Porous Media.* (1998) **32**:163–86. doi: 10.1023/A:1006510106194
26. Gjennestad MA, Vassvik M, Kjelstrup S, Hansen A. Stable and efficient time integration of a dynamic pore network model for two-phase flow in porous media. *Front Phys*. (2018) **6**:56. doi: 10.3389/fphy.2018.00056
27. Gjennestad MA, Winkler M, Hansen A. Pore network modeling of the effects of viscosity ratio and pressure gradient on steady-state incompressible two-phase flow in porous media. *arXiv:1911.07490* (2019).
28. Sinha S, Gjennestad MA, Vassvik M, Hansen A. A dynamic network simulator for immiscible two-phase flow in porous media. *arXiv:1907.12842* (2019).
29. Tallakstad KT, Knudsen HA, Ramstad T, Løvoll G, Måløy KJ, Toussaint R, et al. Steady-state two-phase flow in porous media: statistics and transport properties. *Phys Rev Lett.* (2009) **102**:074502. doi: 10.1103/PhysRevLett.102.074502
30. Tallakstad KT, Løvoll G, Knudsen HA, Ramstad T, Flekkøy EG, Måløy KJ. Steady-state simultaneous two-phase flow in porous media: an experimental study. *Phys Rev E.* (2009) **80**:036308. doi: 10.1103/PhysRevE.80.036308
31. Aursjø O, Erpelding M, Tallakstad KT, Flekkøy EG, Hansen A, Måløy KJ. Film flow dominated simultaneous flow of two viscous incompressible fluids through a porous medium. *Front Phys.* (2014) **2**:63. doi: 10.3389/fphy.2014.00063
32. Savani I, Bedeaux D, Kjelstrup S, Sinha S, Vassvik M, Hansen A. Ensemble distribution for immiscible two-phase flow in porous media. *Phys Rev E.* (2017) **95**:023116. doi: 10.1103/PhysRevE.95.023116
33. Washburn EW. The dynamics of capillary flow. *Phys Rev*. (1921) **17**:273. doi: 10.1103/PhysRev.17.273
34. Sinha S, Hansen A, Bedeaux D, Kjelstrup S. Effective rheology of bubbles moving in a capillary tube. *Phys Rev E.*(2013) **87**:025001. doi: 10.1103/PhysRevE.87.025001
35. Batrouni GG, Hansen A. Fourier acceleration of iterative processes in disordered systems. *J Stat Phys.* (1988) **52**:747–73. doi: 10.1007/BF01019728
36. Ramstad T, Hansen A, Øren PE. Flux-dependent percolation transition in immiscible two-phase flow in porous media. *Phys Rev E.* (2009) **79**:036310. doi: 10.1103/PhysRevE.79.036310
37. Parteli EJR, da Silva LR, Andrade JS Jr. Self-organized percolation in multi-layered structures. *J Stat Mech.* (2010) **2010**:P03026. doi: 10.1088/1742-5468/2010/03/P03026
38. Hansen A, da Silva LR, Lucena L. Spatial correlations in permeability distributions due to extreme dynamics restructuring of unconsolidated sandstone. *Phys A*. (2011) **390**:553. doi: 10.1016/j.physa.2010.10.011

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 31 January 2020
doi: 10.3389/fphy.2020.00009

A Discrete Fracture Network Model With Stress-Driven Nucleation: Impact on Clustering, Connectivity, and Topology

Etienne Lavoine [1,2], Philippe Davy [1], Caroline Darcel [2] and Raymond Munier [3]*

[1] Univ Rennes, CNRS, Géosciences Rennes, UMR 6118, Rennes, France, [2] Itasca Consultants SAS, Écully, France, [3] Terra Mobile Consultants AB, Stockholm, Sweden

The realism of Discrete Fracture Network (DFN) models relies on the spatial organization of fractures, which is not issued by purely stochastic DFN models. In this study, we introduce correlations between fractures by enhancing the genetic model (UFM) of Davy et al. [1] based on simplified concepts of nucleation, growth and arrest with hierarchical rules. To do so, the nucleation of new fractures is correlated with the elastic strain energy of distortion stored in the matrix, which is a function of preexisting fractures. Discrete Fracture Networks so generated show multi-scale clustering effects with fractal dimensions below the topological dimension over a broad range of scales. The fractal dimension depends on the way one correlates the nucleation occurrence to the strain energy. Fracture clustering entails a spatial variability of the fracture density, which increases with the intensity of the coupling between stress and nucleation. The analysis of connected clusters density and of fracture intersections also highlights the differences between the UFM models and its equivalent Poisson model. We show that our stress-dependent nucleation model introduces some new fracture size-positions correlations, with small fractures tending to connect to the largest ones.

Keywords: discrete fracture networks (DFNs), nucleation, clustering, connectivity, topology

OPEN ACCESS

Edited by:
Eirik Grude Flekkøy,
University of Oslo, Norway

Reviewed by:
Srutarshi Pradhan,
Norwegian University of Science and Technology, Norway
Jonathan Bares,
Université de Montpellier, France

****Correspondence:***
Etienne Lavoine
e.lavoine@itasca.fr

Specialty section:
This article was submitted to Interdisciplinary Physics, a section of the journal Frontiers in Physics

Received: *11 October 2019*
Accepted: *08 January 2020*
Published: *31 January 2020*

Citation:
Lavoine E, Davy P, Darcel C and Munier R (2020) A Discrete Fracture Network Model With Stress-Driven Nucleation: Impact on Clustering, Connectivity, and Topology. Front. Phys. 8:9. doi: 10.3389/fphy.2020.00009

INTRODUCTION

Fractures are ubiquitous structures controlling both flows and rock mechanical strength in geological environments. Modeling the fracture network is thus a key prerequisite of forecasting modeling in many industrial applications such as managing groundwater/petroleum resources, assessing risks associated with geotechnical constructions or deep waste disposal, among others. In most of the cases, fractures cannot be modeled deterministically, because they cannot be observed in three-dimensions with sufficient resolution at all scales. Hence, the modeling must be stochastic, which consists of generating a 3D fractured medium statistically equivalent to measures and observations. Discrete Fracture Network modeling is one of the most convenient and used of the stochastic methods; it describes fractured rocks as a population of individual fractures, whose parameters (size, shape, orientation, aperture, and position) are drawn from statistical probability distributions derived from observation maps (mainly 2D outcrop and tunnels, 1D fracture intensity along wells, or 3D geophysics imagery) and models [see [2, 3] for reviews]. In its simplest form, the model (which we will refer as the Poisson model) consists of positioning fractures at random in the generation volume with a given density, and of assigning other fracture parameters independently by bootstrapping the parameter distributions of observations [4–9]. The method is easy to implement, but it neglects most of the complexity of the underlying fracturing process, in particular the correlations induced by fracture-to-fracture mechanical interaction [10–12]. This so-called Poisson model is thus a crude representation of geological fractures that can lead to large discrepancy between modeled and natural network in terms of network topology [13, 14], having dramatical impact on the estimated hydrological and mechanical behavior of the fractured rock mass [15–17]. A way to improve the realism of DFN models is the use of genetic models, in which the fracture hierarchy reproduce correlations between their different geometrical attributes. Nevertheless, a full mechanical description of the fracturing process [18–20] is not feasible when dealing with dense networks made of fractures having sizes from centimeter to tens of kilometers. The broad range of natural fracture size distributions and their power-law nature [11, 21–23] suggest that all scales matters. This is even more important for industries such as nuclear deep waste disposal where small fractures may have an important role at the nearfield of the repository, whose footprint extends to several kilometers. Recent papers [1, 24] have proposed a genetic model of DFN, called "Universal Fracture Model" (UFM model), using simplified fracturing-relevant rules for nucleation, growth and arrest of fractures to draw complex and dense networks. With simple kinematic rules that mimic the main mechanical processes, the model produces fracture size distributions and fracture intersections that are consistent with observations [24]. It results in less connected networks than the Poisson model and changes in the network topology [25], which has been proven to have an impact on flow properties, particularly decreasing effective permeability and increasing flow channeling [26]. Nonetheless, the random positioning of new fractures (nuclei) in this model is still too simplified to reproduce spatial variability and clustering effect observed in natural fracture network patterns [11, 23]. Indeed, nucleation is a complex process both controlled by the repartition of flaws in the rock matrix such as grain boundaries, pores or cleavage plans [27–29] and the stress distributions that make nuclei active or not [30–32]. This problem can be addressed as a quenched disordered process where flaws are initially present in the system and activate as the system evolves [33, 34], or as an annealed disorder where nuclei positioning is directly a function of the system evolution [35]. In this paper, we aim to improve the UFM model by better reproducing the complex feedback-loop process between the propagation of fractures and the emergence of new ones. Our model is also based on the nucleation, growth, and arrest scheme, but we propose to condition the positioning of new nuclei to the mechanical perturbation induced by existing fractures in a timewise manner. This perturbation is modeled as the superposition of stress redistributions induced by each fracture loaded by an allegedly known remote stress field. Such a pseudo-mechanical model does not aim to catch all the complexity of fracture mechanical processes, but this first order approximation of fracture interactions at the network scale may already change dramatically the topology and connectivity of the DFN models. Since larger stress perturbations are expected in the vicinity of fractures, with an intensity that depends on fracture size, we expect the stress-driven nucleation in the timewise process of the UFM model to increase fracture spatial correlations. In order to highlight spatial correlations of the DFNs so generated, we compute fracture positions correlation dimension, fracture density variability, and fractures intersection matrix, and compare results with equivalent Poisson model.

THE MODEL

The Discrete Fracture Network approach for modeling fractured rock masses refers to numerical models explicitly representing the geometry of each fractures forming the network. Generally, this geometry (positions, orientations, size…) is generated stochastically from data statistics. The simplest model considers fractures independent of each other; it will be referred to as the Poisson DFN model. The DFN model developed by Davy et al. [1] is based on basic mechanical concepts described in Davy et al. [24]. The fracturing process is divided in three main stages in a time-wise approach: nucleation, propagation and arrest of fractures. We first develop the model and how in its simplest form—i.e., with a Poisson distribution of nuclei—it controls the network size and intersection distributions. Then, we introduce a more complex nucleation model based on the stress perturbation of pre-existing fractures.

The Stress-Independent UFM Model

Nucleation is the fracture birth process, which is here defined by a nuclei size distribution $p_N\left(l\right)$ (that can be a power-law, exponential, etc…) and a rate $\dot{n}_N = dn_N/dt$ (with n_N the number of nuclei introduced in the system). Nuclei positions are assumed to be uniformly distributed in space here, we will refer to this model as the stress-independent UFM model.

Once created, fractures grow following a power law relationship to describe the crack tip velocity in the subcritical regime [36]:

$$v\left(l\right) = Cl^a$$

with l the fracture length, C the growth rate and a the growth exponent. If not arrested, nuclei size increases non-linearly with time and becomes infinite for a finite time t_∞ dependent on the initial nuclei size and the parameters C and a. For constant nucleation rate and no fracture arrest, even if fractures grow, there is a stationary solution for the fracture size distribution [1]:

$$n_G\left(l\right) = \frac{\dot{n}_N}{C} l^{-a}$$

The arrest rule is assumed to reflect the mechanical interaction between fractures. In this model, we consider these interactions as a binary law where fractures can only abut on larger ones, but the reverse is not likely to occur. It results in a large proportion of T-shape intersections that are consistent with field observation, and in a quasi-universal self-similar fracture size distribution:

$$n_A\left(l\right) = D\gamma^D l^{-D+1}$$

with D the topological dimension associated to fracture centers, and γ a geometrical parameter dependent on fracture orientations. Small fractures are statistically freely growing with the size distribution $n_G\left(l\right)$, while large fractures are statistically arrested and described by $n_A(l)$. The model thus results in a two-power-law size distribution, where the transition size between n_G and n_A is both the scale at which the network is connected and the average size of fracture blocks. The two power-law distribution is obtained for modeling time t_∞, when first fractures become infinite. For larger times, the number of arrested fractures increases, and the transition size decreases. We refer the reader to Davy et al. [1] for further information.

Stress-Driven Nucleation

In this section, we further develop the stress-independent UFM model by making nucleation dependent on stress redistributions caused by existing fractures. For this, we introduce a stress field based probabilistic sampling of nuclei locations. The stress field evolves over the whole domain as the fracturing process. Considering the system to be linear elastic, which may not be the case for highly damaged materials, we define the stress field $\overline{\overline{\sigma}}\left(\overline{x},t\right)$ at any position $\overline{x}$ in the domain at time t as the superposition of the remote stress field $\overline{\overline{\sigma^\infty}}(t)$ plus the contribution of every fracture $\overline{\overline{\sigma_f}}\left(\overline{x},t\ \right)$:

$$\overline{\overline{\sigma}}\left(\overline{x},t\right)=\overline{\overline{\sigma^\infty}}\left(t\right)+\sum_f\overline{\overline{\sigma_f}}\left(\overline{x},t\right)$$

New nuclei are progressively introduced in the system following a probability field $\mathbb{P}\left(\overline{x},t\right)$ derived of this stress field:

$$\mathbb{P}\left(\overline{x},t\right)\ \sim\left[\sigma_{VM}(\overline{x},t)\right]^m$$

where $\sigma_{VM}\left(\overline{x},t\right)$ is the Von Mises stress [37] and m a parameter that quantifies the coupling between nucleation occurrence and stress (thereafter called the selectivity parameter). The Von Mises stress is a scalar invariant measure of the deviatoric stress intensity and a measure of the elastic strain energy of distortion stored in the matrix. We then generate a scalar stress-intensity field that will serve as a basis to construct a discrete probability distribution for nuclei position sampling, without using any strength criteria. This stress-driven nucleation process is thus defined as an annealed disorder process [35] where nuclei positioning is directly a function of the system evolution. The model then needs two more parameters: the remote stress field tensor $\overline{\overline{\sigma^\infty}}$ and the selectivity parameter m. The latter quantifies the influence of the stress field heterogeneity on nucleation. For large m, nuclei tend to concentrate in regions with high stress intensity. In the following, we will refer to this model as the stress-driven UFM model. The case $m\ =\ 0$, where the nucleation is uniformly distributed in space, corresponds to the stress-independent UFM model.

For numerical implementation of the model, we use the same basic assumptions as Davy et al. [1]: fractures are modeled as interacting growing disks in a time-wise process. At each time step Δt, $\dot{n_N}\Delta t$ nuclei are introduced in the cubic system. Those who are not intersecting any existing fractures are kept in the system and grow following equation [1], until they cross a larger fracture or reach an infinite size, that we set to be twice the system size. Nuclei are not allowed to intersect pre-existing fractures so that the available space for new nuclei and the effective nucleation rate decrease with simulation time. The stress perturbation associated to each fracture can be approximated using the 3D tensorial analytical solutions of Fabrikant [38] for uniformly loaded freely-slipping penny-shaped cracks, considering traction and/or shearing. If the system is under compression, then only the shearing part is considered. Each fracture is assumed to be uniformly loaded by the remote stress field and generate a stress perturbation that depends on the fracture size, the input remote stress field intensity, and their relative orientations. To fasten calculations, we do not calculate the interaction terms between fractures and set them to zero. Although these terms may be non-negligible when fractures get close with each other [39–41], in particular when the fracture density increases, we have estimated that this approximation is consistent with the degree of simplification used for the different stages of the model. A more elaborate version is under development.

The stress field is computed over the whole domain, on a regular cartesian stress grid S_g of resolution r_{stress}. In order to obtain a dimensionless stress-intensity scalar field (**Figure 1**), each cell value is divided by the remote stress Von Mises value.

For each nucleation step, a cell is chosen from a discrete probability sampling over the whole stress grid S_g, where the probability for each cell c is defined by:

$$P\left(\overline{x}_c\right)=\frac{\left[\sigma_{VM}(\overline{x_c})\right]^m}{\sum_{c'\in S_g}\left[\sigma_{VM}\left(\overline{x_{c'}}\right)\right]^m}$$

Once the nucleus cell has been determined, the nucleus center position is randomly taken inside this cell. The number of

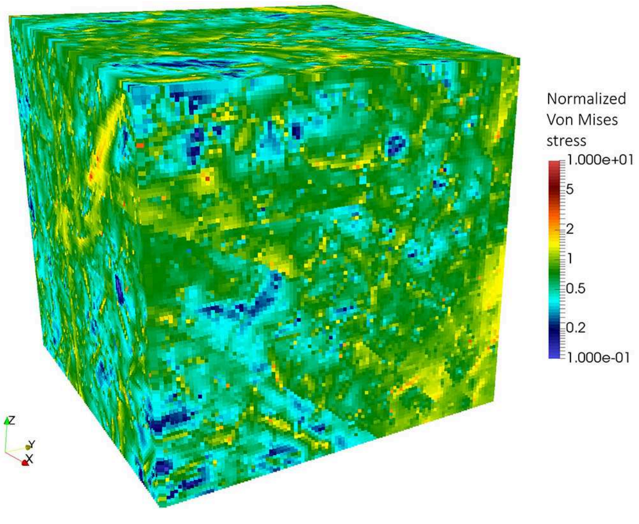

FIGURE 1 | Stress-intensity field generated on a regular grid of size **L = 1** and resolution **r = 0.01**.

computations is directly related to the nucleation rate, the stress grid resolution, and the increasing number of fractures already present in the medium, which can be large. Since the addition of a single small nucleus does not affect the stress field at the system size, a single stress grid can be used for several nucleation steps n_{step} in order to accelerate computations. This heterogeneous probabilistic point process of nuclei positions constitutes an improvement of the stress-independent UFM model, while keeping constant nucleation rate.

RESULTS

In this section, we focus on the spatial and topological analysis of fracture networks generated by this stress-driven UFM model. We analyze the evolution of the pattern complexity of this model with the selectivity parameter m, and compare the results with the stress-independent UFM ($m = 0$) and equivalent Poisson model (i.e., same population of fractures with random positions in the domain).

Numerical Simulations

For all models, we seed and let fractures grow in a domain of size $L = 1$ with a growth exponent $a = 3$, so that the power-law exponent of the dilute regime is -3, which is consistent with field data [24]. Nuclei appears in the system with a constant nucleation rate $\dot{n} = 20$, growth rate $C = 1$, and a size drawn from a narrow-ranged power-law distribution:

$$p_N\left(l\right) = \frac{\left(b-1\right)}{l_N}\left(\frac{l}{l_N}\right)^{-5}$$

l_N is the minimum nuclei size; its value has been set at 0.01 in order to cover two orders of magnitude in the resulting fracture size distribution. For each timestep Δt, we introduce $n_{step} = 200$

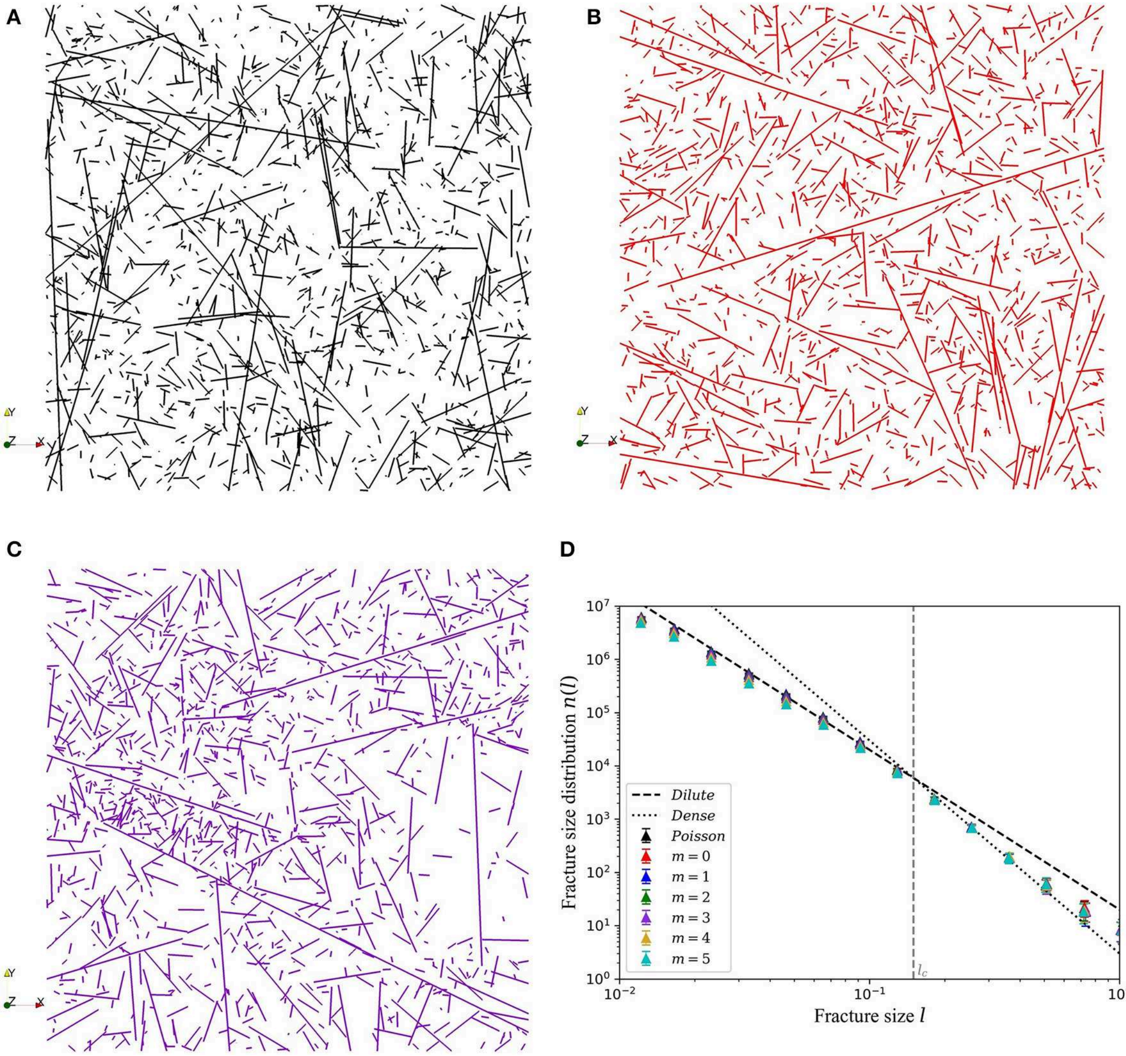

FIGURE 2 | 2D slices of three-dimensional **(A)** Poisson, **(B)** stress-independent UFM (**m** = **0**), and **(C)** stress-driven UFM (**m** = **3**) models, and **(D)** associated size distributions (**m** ∈ [**0**, **5**]).

new nuclei. Nuclei intersecting existing fractures are rejected, the effective nucleation rate is thus decreasing with time, since the available space for new fractures to form is also decreasing. We stop simulations when the time is close to the $t_{\infty}\left(l_N\right)$, i.e., the time necessary for the smallest nuclei to become infinite [1]. Networks are generated for different values of selectivity parameters $m = \{0, 1, 2, 3, 4, 5\}$ in order to quantify its impact on fracture clustering. The equivalent Poisson model is obtained by moving the fracture centers randomly in space, so that the size distribution remain identical but the correlations between fracture size and position are destroyed. We only show the Poisson model derived from the stress-independent UFM ($m = 0$). All stress-driven UFM are generated from the same constant and compressive remote stress field σ^{∞}, so that $\sigma_1 = \sigma_{xx} = -4$, $\sigma_2 = \sigma_{yy} = -2$, $\sigma_3 = \sigma_{zz} = -1$, and $\sigma_{xy} = \sigma_{xz} = \sigma_{yz} = 0$. The intensity and orientations of the principal stress components do not matter in this set of simulations since we consider that, for all generated DFNs, the orientation distribution is stress-decorrelated and is assumed uniform in order to minimize asymmetry due to the remote stress field. By doing so, we aim at focusing on the consequences of the stress field heterogeneity on spatial correlations only, but not on its spatialization. For this set of parameters, the simulation time for a stress-driven UFM model is about 60 times larger than for the stress-independent UFM model for which there is no stress. For all the models, we perform 50 realizations of each model for statistical analysis.

Figures 2A–C show 2D slices of generated three-dimensional Poisson, stress-independent UFM and stress-driven UFM ($m = 3$), respectively. Visually, the stress-driven nucleation process seems to increase the clustering effect of fractures positions. Simulations are stopped when $t = t_{\infty}(l_N)$, so that both the dilute and dense regime can be observed on the fracture size distribution (**Figure 2D**) and are consistent with equations [2] and [3], with a transitional length $l_c = 0.15$:

$$n\left(l\right) = \begin{cases} 20.l^{-3} & if\ l < l_c \\ 3.l^{-4} & if\ l > l_c \end{cases}$$

For all models, fracture size distributions are almost the same.

When increasing the selectivity parameter, new nuclei to form should be attracted to the tip of the largest existing fractures since the stress is high here, increasing the probability of rejection. Hence, for the same simulation time, the fracture density should decrease when increasing selectivity parameter m. We consider three-dimensional fracture densities here, such as defined by Dershowitz and Herda [42]: fracture number density p_{30} (number of fractures per unit volume), fracture intensity p_{32} (total fracture surface per unit volume), and percolation parameter p (total excluded volume around fractures per unit volume) that quantifies the network connectivity [43]. Considering the disk-shape assumption we made in our DFN models, we define these densities as:

$$p_{30} = \int n\left(l\right)\Pi\left(l, L\right)dl$$

$$p_{32} = \frac{\pi}{4}\int n\left(l\right)\Pi\left(l, L\right)l^2 dl$$

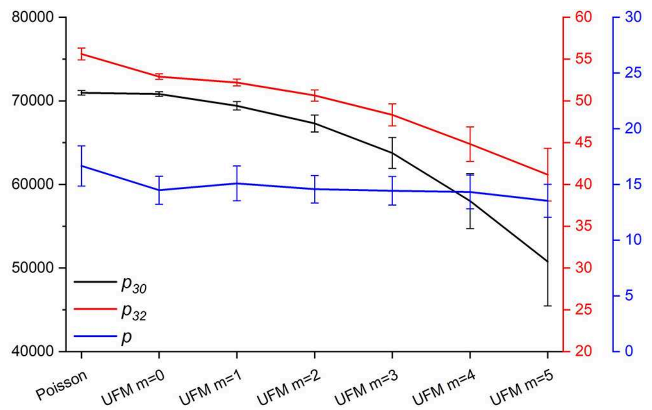

FIGURE 3 | Fracture density statistics of generated networks in cubic systems of volume $\mathbf{V} = \mathbf{1}$.

$$p = \frac{\pi^2}{8}\int n\left(l\right)\Pi\left(l, L\right)l^3 dl$$

$\Pi(l, L)$ is surface ratio of the fracture l included in the domain of size L. **Figure 3** summarize the density statistics of the generated networks.

One can notice a decreasing of both p_{30}, p_{32}, and p with selectivity parameter m due the increasing number of rejected nuclei. Moreover, the fracture density of the stress-independent UFM model is slightly slower than its equivalent Poisson model, because both models are subjected to different finite size effects.

Impact on Clustering

The spatial organization and topology of fractures in a network may have dramatical impact on its connectivity and hydraulic behavior. Quantifying the fractures organization in DFN models and comparing with natural networks is a key challenge to qualify the relevance of simplified models. Natural fracture networks have shown complex clustered patterns [11, 44] that has consequences on connectivity [21, 45, 46].

Considering the multiscale nature of fractures, and thus of the subsequent stress fluctuations, we expect fractal correlations to develop in our model. The full multifractal spectrum of fracture organization may be quantified using the box-counting method [47], computing the number of boxes to cover the network at different scales. Nevertheless, this technique has be shown to be strongly affected by finite size effects [11, 23, 48]. We then compute the 2-point correlation integral (or correlation pair function) to describe the spatial correlations of fractures positions. This method gives the probability for two fractures to belong to the same cluster. For a population of N_f fractures, the associated correlation pair function is defined by:

$$C_2\left(r\right) = \frac{2.N(r)}{N_f.(N_f - 1)}$$

with $N(r)$ the number of points whose mutual distance is less than r [47]. For a large population of points, this quantity tends to scale a power-law r^{D_2}, where D_2 is the correlation dimension of fracture centers. The correlation dimension can thus be obtained by computing the slope of the correlation pair function in a log-log plot. **Figure 4A** shows the evolution of the correlation pair function and its derivative with scale r. This function is affected by finite-size and resolution effects when the mutual distance r tends to domain finite size and nuclei size, respectively. We calculate a correlation dimension D_2 in the interval $[0.03, 0.08]$ that is not affected by size effects. Indeed, below the lower bound, the derivative of the correlation pair function increases dramatically because few fractures are so close to each other. This resolution effect is related to the no-intersecting assumption when introducing new nuclei in the UFM model. On the other hand, finite size effect due to the system size are already dramatic below the upper bound, as we should obtain a correlation dimension $D_2 = 3$, for the Poisson model. As expected, the correlation dimension of the 3D Poisson and stress-independent UFM models is $D_2 \sim 3$, which is consistent with a uniform distribution of fracture centers in space for both models. D_2 is smaller than 3 for the stress-driven UFM model and decreases when increasing the selectivity parameter m. For large values of m, nuclei concentrate in zones of high stress, mainly near the tips of the largest fractures, leading to a clustering of fractures. **Figure 4B** shows that correlations exist even at early simulation times. For large m values, the correlation dimension increases slightly with time, as the available space around fractures tips decreases.

The correlation dimension of fracture centers indicates how much a fracture network occupies its underlying metric space. Nevertheless, two networks can have the same correlation dimension but very different patterns. In order to describe the *texture* associated to a network, we use the concept of lacunarity [49]. Fundamentally, lacunarity is a dimensionless representation of the variance to mean ratio [50] defined here as:

$$\lambda_M(s) = \left[\frac{\sigma_M(s)}{\mu_M(s)}\right]^2$$

with $\sigma_M(s)$ and $\mu_M(s)$ the standard deviation and mean of a measure M at scale s. For any density measure M, if $\lambda_M(s) \to 0$, the pattern is perfectly homogeneous at scale s. Lacunarity is a scale dependent measure, whose analysis quantify the degree of clustering and anti-clustering [51], and potentially on different regimes [50, 52, 53], when analyzing lacunarity curves, showing $\lambda_M(s)$ evolution with scale s. Lacunarity analysis can then be used to analyze textural heterogeneity of fracture densities [54]. For three-dimensional fracture networks, we can define various measures M quantifying fracture density as defined by Dershowitz and Herda [42], or in section Numerical Simulations.

Figure 5 shows the lacunarity curves of the three-dimensional fracture densities defined in section Numerical Simulations. As expected, the p_{30} lacunarity curve is the same for the stress-independent UFM model and its equivalent Poisson

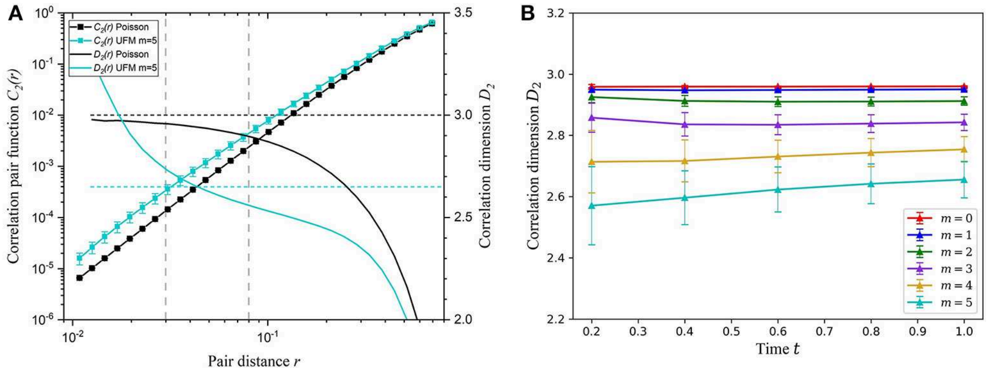

FIGURE 4 | (A) Correlation pair function, and **(B)** correlation dimension analysis for the Poisson, stress-independent UFM, and stress-driven UFM models.

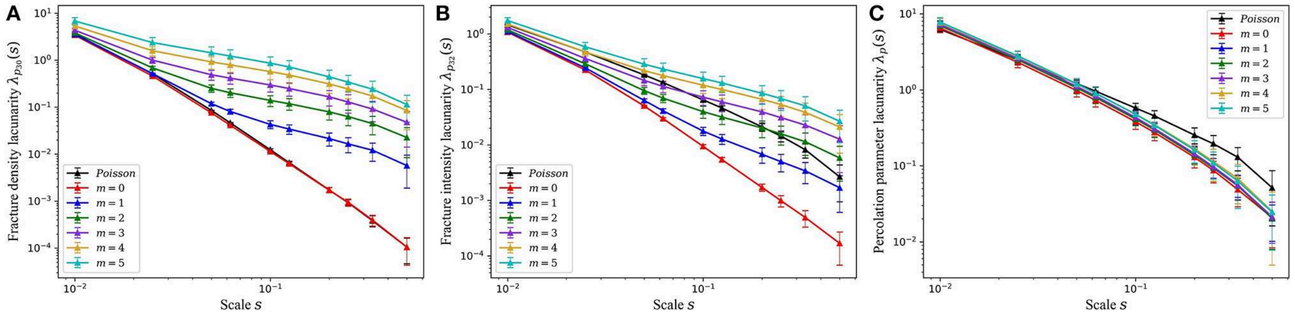

FIGURE 5 | Fracture density lacunarity curves for **(A)** $\mathbf{p_{30}}$, **(B)** $\mathbf{p_{32}}$, and **(C)** percolation parameters.

model (because both follow a homogeneous Poisson point process) and scales $\sim\ s^{-3}$. The p_{30} lacunarity of the stress-driven UFM models are much larger and evolves differently with scale, which emphasizes that fracture center positions are correlated. They all follow a scaling $As^{-\alpha}$, with A and α constant factors increasing with m. For fracture intensity p_{32}, the lacunarity of the stress-independent UFM model is smaller at all scales than that of the Poisson model. The p_{32} lacunarity decreases faster with scale for the stress-independent UFM model than for Poisson model. This reflects a fracture density much more homogeneous in space at all scales for the stress-independent UFM than for the Poisson case. Indeed, the UFM rule (a fracture cannot cross a larger one) tends to produce an interconnected network of blocks of size of the order the transition scale l_c [1]. The p_{32} lacunarity here quantifies the fracture position-size correlation induced by the UFM rule. The p_{32} lacunarity increases at all scales with the selectivity parameter m for the stress-driven UFM models,. Even for small values of m, the shift in lacunarity with the stress-independent UFM case is important, which points out the impact of the clustering of fracture positions on density variability. The lacunarity associated to the percolation parameter is smaller for all UFM models than for Poisson model, meaning that the connectivity of the network is more homogeneous for UFM models. All percolation lacunarity curves are similar whatever the value of m, suggesting that the percolation parameter is more sensitive to the fracture position-size correlations induced by the UFM rule, than the fracture centers correlations.

Consequences on Connectivity and Topology

Fracture correlation is likely changing the connectivity of the overall network. Maillot et al. [26] show that stress-independent UFM networks (which they refer as a "kinematic" model) have permeabilities 1.5–10 times smaller than the equivalent Poisson model, and a higher channeling (a higher portion of the total fracture surface where the flow is significant). Some studies [46, 55] show that, for networks with a power-law size distribution, the evolution of connectivity with scale is strongly dependent on the power-law exponent a and on the fractal dimension of fracture centers D_2. In our case, because fracture centers tend to concentrate in clusters around large fracture tips, the fracture interconnectivity may also increase. Increasing the connectivity between fractures should increase the backbone density, defined as the structure carrying flow in the network [56]. We here define the backbone by removing iteratively fractures having only one connection with the network or the boundary, keeping only the connected clusters without flow dead-ends. **Figure 6** shows that the percentage of fractures in number and in total area (p_{30} and p_{32}) involved in the backbone is smaller for any kind of UFM model (stress-independent or stress-driven) than for corresponding Poisson model. Moreover, for the UFM models, even if $<\ 25\%$ of fractures are part of the backbone, this represents more than 65% of the backbone surface, which shows that connectivity is mostly ensured by large structures [21, 45]. Finally, as the number of fractures involved in the backbone structure increases more than the total area with m, we can conclude that we tend to connect mostly small fractures to the backbone with this stress-driven nucleation process.

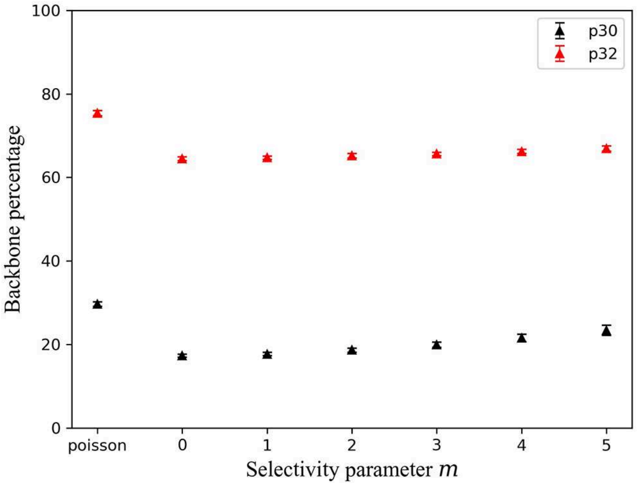

FIGURE 6 | Evolution of backbone percentage with selectivity parameter **m** (**m** = **Poisson** refers to Poisson model).

The number of intersections per fracture is a good indicator of fracture connectivity. Maillot et al. [26] showed that the number of intersections per fracture is a function of fracture size for both Poisson and UFM models. Moreover, they showed that whatever the fracture size, the number of intersections is about two times lower for UFM model than for equivalent Poisson model. We here push further this topological analysis computing the fracture intersection matrix P_I so that for n_i and n_j fractures of size l_i and l_j, respectively, $P_I[i,j]$ gives the number of fractures of size l_i intersecting fractures of size l_j, divided by n_in_j. **Figures 7A–C** show the mean intersection matrix for all generated Poisson, stress-independent UFM ($m = 0$), and stress-driven UFM ($m = 5$) models, respectively. As expected, in any case, the probability of intersection increases with fractures sizes. **Figures 7D,E** show the mean intersection matrix for stress-independent UFM ($m = 0$), and stress-driven UFM ($m = 5$) models, normalized by the mean equivalent Poisson's model intersection matrix, in order to highlight differences between UFM and Poisson models. Our analysis shows that the number of fractures intersections is smaller for UFM models than their equivalent Poisson. We can also notice that this number is much smaller for fractures of the same size, which is a consequence of the UFM rule assuming that a fracture cannot cross a larger one. Finally, **Figure 7F** shows the stress-dependent UFM ($m = 5$) model intersection matrix, normalized by the one of the stress-independent UFM ($m = 0$) model, showing that our stress-driven nucleation process tend to increase the connectivity of small fractures with the smallest and the largest fractures. Indeed, new fractures tend to develop at the tip of the largest existing fractures, increasing connectivity between both.

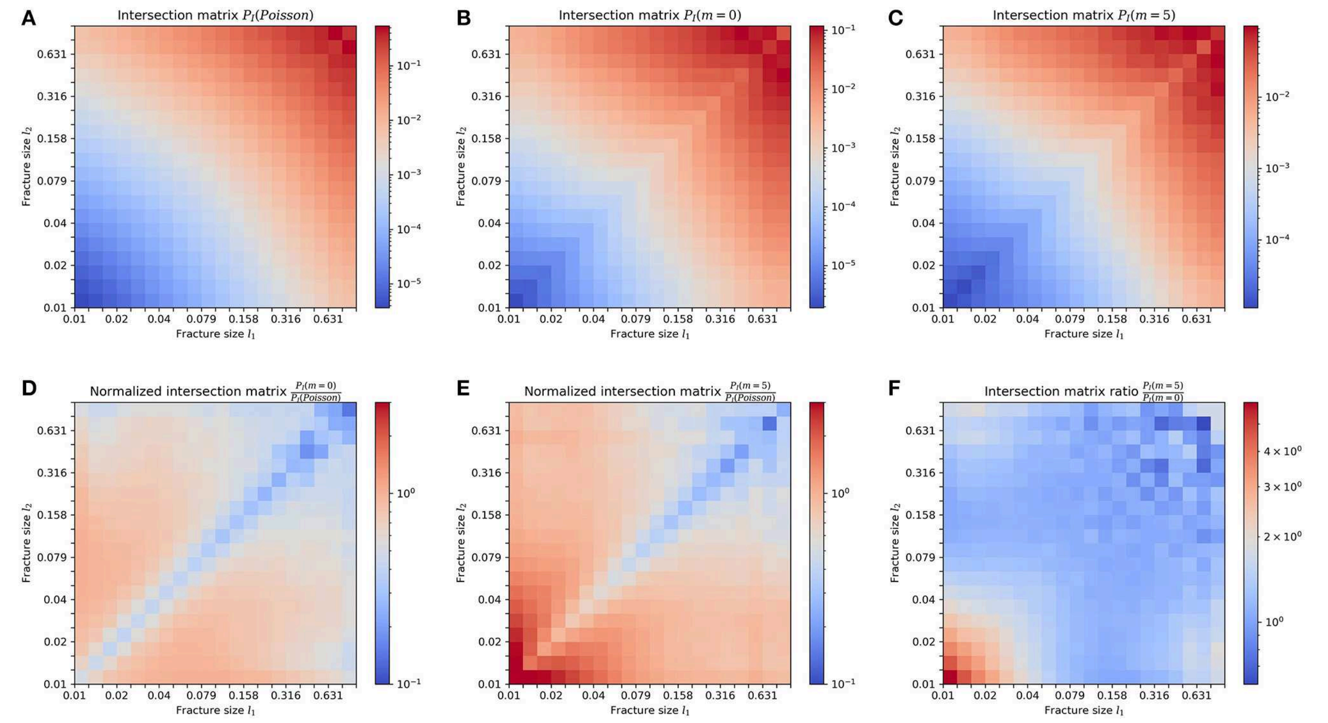

FIGURE 7 | Intersection matrix for **(A)** Poisson, **(B)** stress-independent UFM, and **(C)** stress-driven UFM models. **(D)** Normalized intersection matrix $\frac{\mathbf{P_I(m=0)}}{\mathbf{P_I(Poisson)}}$ for the stress-independent UFM models, **(E)** Normalized intersection matrix $\frac{\mathbf{P_I(m=5)}}{\mathbf{P_I(Poisson)}}$ for the stress-driven UFM (**m = 5**) models, **(F)** Intersection matrix ratio $\frac{\mathbf{P_I(m=5)}}{\mathbf{P_I(m=0\ )}}$.

CONCLUSION

The genetic UFM model developed by Davy et al. [1], describing the fracturing process as a combination of simplified nucleation, growth and arrest laws, introduces a fracture size-position correlation in DFN modeling, that does not exist in equivalent Poisson model. It results in two distinct power-law fracture size distributions and a number of T-intersections that are consistent with field data [24]. Nevertheless, the model does not take into account the mechanical feedback loop between fracture growth and birth, therefore neglecting fracture-to-fracture positioning correlations. In this paper, we pushed further the model by improving the nucleation process, conditioning the position of newly created fractures by the stress perturbation induced by preexisting ones, in a timewise process. This stress perturbation is a function of fractures geometry (size and orientation), and the applied remote stress (orientation and intensity). This results in more correlated networks, showing fractal positioning, and a higher variability of fracture densities. We introduce a selectivity parameter m that quantifies the dependency of nucleation with the stress field. When nucleation is stress-independent ($m = 0$, uniform positions), we show that fracture density variability associated to UFM networks is much smaller than equivalent Poisson model. This means that the UFM rule, imposing that a fracture cannot cross a larger one, tends to organize networks into more homogeneous patterns than if fractures were positioned at random. Nonetheless, when nucleation is stress-driven ($m \neq 0$), the higher m, the lower the correlation dimension of fracture positions, and the higher the spatial variability of fracture densities. This effect is dependent on the density measure, i.e., on the dependency of the density with fracture size. It is higher for the number of fractures per unit volume than for the percolation parameter. Moreover, our connectivity analysis brings up that the UFM rule tends to create a hierarchy between fractures, so that fractures of the same size order are less likely to cross one each other. The stress-driven nucleation process we propose tends to connect small fractures all together with the largest ones, that are responsible for the main stress perturbations. We also show that UFM models have lower percentage of fractures involved in the backbone than their equivalent Poisson model [26], whatever the selectivity parameter.

Finally, constraining fractures orientation according to the computed local stress field, in order to account for fracture position-orientation correlations, would constitute a huge improvement of the model. Once a full simplified mechanical description of the fracturing process is performed, this would allow us to perform real case studies, and compare our analysis results (clustering, connectivity…) between 2D numerical outcrops from generated DFNs, with real ones.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

EL, PD, CD, and RM conceived the idea of this study. EL and PD contributed in developing the model. EL ran network analysis and took the lead in writing manuscript. All authors provided critical feedback.

FUNDING

This work was funded by the CIFRE grant ITASCA Consultants SAS/ANRT (Association Nationale Recherche Technologie, France) 2016/1002, and partially supported by the Swedish Nuclear Fuel and Waste Management Co. (SKB).

REFERENCES

1. Davy P, Le Goc R, Darcel C. A model of fracture nucleation, growth and arrest, and consequences for fracture density and scaling. *J Geophys Res.* (2013) **118**:1393–407. doi: 10.1002/jgrb.50120
2. Jing L. A review of techniques, advances and outstanding issues in numerical modelling for rock mechanics and rock engineering. *Int J Rock Mech Mining Sci.* (2003) **40**:283–353. doi: 10.1016/S1365-1609(03)00013-3
3. Lei Q, Latham J-P, Tsang C-F. The use of discrete fracture networks for modelling coupled geomechanical and hydrological behaviour of fractured rocks. *Comput Geotech.* (2017) **85**:151–76. doi: 10.1016/j.compgeo.2016.12.024
4. Long J, Remer J, Wilson C, Witherspoon P. Porous media equivalents for networks of discontinuous fractures. *Water Resour Res.* (1982) **18**:645–58. doi: 10.1029/WR018i003p00645
5. Baecher GB. Statistical analysis of rock mass fracturing. *J Int Assoc Math Geol.* (1983) **15**:329–48. doi: 10.1007/BF01036074
6. Andersson J, Shapiro AM, Bear J. A stochastic model of a fractured rock conditioned by measured information. *Water Resour Res.* (1984) **20**:79–88. doi: 10.1029/WR020i001p00079
7. Endo H, Long J, Wilson C, Witherspoon P. A model for investigating mechanical transport in fracture networks. *Water Resour Res.* (1984) **20**:1390–400. doi: 10.1029/WR020i010p01390
8. Long J, Gilmour P, Witherspoon PA. A model for steady fluid flow in random three-dimensional networks of disc-shaped fractures. *Water Resour Res.* (1985) **21**:1105–15. doi: 10.1029/WR021i008p01105
9. Andersson J, Dverstorp B. Conditional simulations of fluid flow in three-dimensional networks of discrete fractures. *Water Resour Res.* (1987) **23**:1876–86. doi: 10.1029/WR023i010p01876
10. Ackermann RV, Schlische RW. Anticlustering of small normal faults around larger faults. *Geology.* (1997) **25**:1127–30. doi: 10.1130/0091-7613(1997)025<1127:AOSNFA>2.3.CO;2
11. Bonnet E, Bour O, Odling NE, Davy P, Main I, Cowie P, et al. Scaling of fracture systems in geological media. *Rev Geophys.* (2001) **39**:347–83. doi: 10.1029/1999RG000074
12. Du Bernard X, Labaume P, Darcel C, Davy P, Bour O. Cataclastic slip band distribution in normal fault damage zones, Nubian sandstones, Suez rift. *J Geophys Res.* (2002) **107**:ETG 6-1-ETG 6-12. doi: 10.1029/2001JB000493
13. Andresen CA, Hansen A, Le Goc R, Davy P, Hope SM. Topology of fracture networks. *Front Phys.* (2013) **1**:7. doi: 10.3389/fphy.2013.00007
14. Sanderson DJ, Nixon CW. The use of topology in fracture network characterization. *J Struct Geol.* (2015) **72**:55–66. doi: 10.1016/j.jsg.2015.01.005
15. Odling NE, Webman I. A "conductance" mesh approach to the permeability of natural and simulated fracture patterns. *Water Resour Res.* (1991) **27**:2633–43. doi: 10.1029/91WR01382
16. Berkowitz B, Hadad A. Fractal and multifractal measures of natural and synthetic fracture networks. *J Geophys Res.* (1997) **102**:12205–18. doi: 10.1029/97JB00304
17. Lei Q, Latham J-P, Xiang J, Tsang C-F, Lang P, Guo L. Effects of geomechanical changes on the validity of a discrete fracture network representation of a realistic two-dimensional fractured rock. *Int J Rock Mech Mining Sci.* (2014) **70**:507–23. doi: 10.1016/j.ijrmms.2014.06.001
18. Moës N, Dolbow J, Belytschko T. A finite element method for crack growth without remeshing. *Int J Numer Methods Eng.* (1999) **46**:131–150. doi: 10.1002/(SICI)1097-0207(19990910)46:1<131::AID-NME726>3.0.CO;2-J
19. Paluszny A, Matthäi SK. Numerical modeling of discrete multi-crack growth applied to pattern formation in geological brittle media. *Int J Solids Struct.* (2009) **46**:3383–97. doi: 10.1016/j.ijsolstr.2009.05.007
20. Paluszny A, Zimmerman RW. Numerical simulation of multiple 3D fracture propagation using arbitrary meshes. *Comput Methods Appl Mech Eng.* (2011) **200**:953–66. doi: 10.1016/j.cma.2010.11.013
21. Odling NE. Scaling and connectivity of joint systems in sandstones from western Norway. *J Struct Geol.* (1997) **19**:1257–71. doi: 10.1016/S0191-8141(97)00041-2
22. Bour O, Davy P. Clustering and size distributions of fault patterns: theory and measurements. *Geophys Res Lett.* (1999) **26**:2001–4. doi: 10.1029/1999GL900419
23. Bour O. A statistical scaling model for fracture network geometry, with validation on a multiscale mapping of a joint network (Hornelen Basin, Norway). *J Geophys Res.* (2002) **107**:ETG 4-1-ETG 4-12. doi: 10.1029/2001JB000176
24. Davy P, Le Goc R, Darcel C, Bour O, de Dreuzy JR, Munier R. A likely universal model of fracture scaling and its consequence for crustal hydromechanics. *J Geophys Res.* (2010) **115**:B10411. doi: 10.1029/2009JB007043
25. Hope SM, Davy P, Maillot J, Le Goc R, Hansen A. Topological impact of constrained fracture growth. *Front Phys.* (2015) 3:75. doi: 10.3389/fphy.2015.00075
26. Maillot J, Davy P, Le Goc R, Darcel C, De Dreuzy J-R. Connectivity, permeability, and channeling in randomly distributed and kinematically defined discrete fracture network models. *Water Resour Res.* (2016) **52**:8526–45. doi: 10.1002/2016WR018973
27. Atkinson BK, Meredith PG. The theory of subcritical crack growth with applications to minerals and rocks. *Fracture mechanics of rock.* (1987) 2:111–66. doi: 10.1016/B978-0-12-066266-1.50009-0
28. Engelder T. Joints and shear fractures in rock. *Fract Mech Rock.* (1987) 27–69. doi: 10.1016/B978-0-12-066266-1.50007-7
29. Pollard DD, Aydin A. Progress in understanding jointing over the past century. *Geol Soc Am Bull.* (1988) **100**:1181–204. doi: 10.1130/0016-7606(1988)100<1181:PIUJOT>2.3.CO;2
30. Tapponnier P, Brace W. Development of stress-induced microcracks in Westerly granite. *Int J Rock Mech Min Sci Geomech Abstr.* (1976) **13**:103–12. doi: 10.1016/0148-9062(76)91937-9
31. Ingraffea AR. Theory of crack initiation and propagation in rock. *Fract Mech Rock.* (1987) **10**:93–4. doi: 10.1016/B978-0-12-066266-1.50008-9
32. Betekhtin V, Kadomtsev A. Evolution of microscopic cracks and pores in solids under loading. *Phys Solid State.* (2005) **47**:825–31. doi: 10.1134/1.1924839
33. Renshaw CE, Pollard DD. Numerical simulation of fracture set formation: a fracture mechanics model consistent with experimental observations. *J Geophys Res.* (1994) **99**:9359–72. doi: 10.1029/94JB00139
34. Olson JE. Predicting fracture swarms—The influence of subcritical crack growth and the crack-tip process zone on joint spacing in rock. *Geol Soc Lond Spec Publ.* (2004) **231**:73–88. doi: 10.1144/GSL.SP.2004.231.01.05
35. Bonneau F, Caumon G, Renard P. Impact of a stochastic sequential initiation of fractures on the spatial correlations and connectivity of discrete fracture networks. *J Geophys Res.* (2016) **121**:5641–58. doi: 10.1002/2015JB012451
36. Charles RJ. Dynamic fatigue of glass. *J Appl Phys.* (1958) **29**:1657–62. doi: 10.1063/1.1723019
37. Mises RV. Mechanik der festen Körper im plastisch-deformablen Zustand. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse.* (1913) **1**:582–92.

38. Fabrikant VI. Complete solutions to some mixed boundary value problems in elasticity. *Adv Appl Mech.* (1989) **27**:153–223. doi: 10.1016/S0065-2156(08)70196-0
39. Kachanov M. Three-dimensional problems of strongly interacting arbitrarily located penny-shaped cracks. *Int J Fract*. (1989) **41**:289–313. doi: 10.1007/BF00018861
40. Kachanov M. On the problems of crack interactions and crack coalescence. *Int J Fract.* (2003) **120**:537–43. doi: 10.1023/A:1025448314409
41. Thomas RN, Paluszny A, Zimmerman RW. Quantification of fracture interaction using stress intensity factor variation maps. *J Geophys Res.* (2017) **122**:7698–717. doi: 10.1002/2017JB014234
42. Dershowitz WS, Herda HH. Interpretation of fracture spacing and intensity. In: *The 33th US Symposium on Rock Mechanics (USRMS)* Santa Fe, NM: American Rock Mechanics Association (1992).
43. De Dreuzy JR, Davy P, Bour O. Percolation parameter and percolation-threshold estimates for three-dimensional random ellipses with widely scattered distributions of eccentricity and size. *Phys Rev E.* (2000) **62**:5948–52. doi: 10.1103/PhysRevE.62.5948
44. Davy P. On the frequency-length distribution of the San Andreas fault system. *J Geophys Res.* (1993) **98**:12141–51. doi: 10.1029/93JB00372
45. Bour O, Davy P. On the connectivity of three-dimensional fault networks. *Water Resour Res.* (1998) **34**:2611–22. doi: 10.1029/98WR01861
46. Darcel BO, Davy P, de Dreuzy JR. Connectivity properties of two-dimensional fracture networks with stochastic fractal correlation. *Water Resour Res.* (2003) **39**:1272. doi: 10.1029/2002WR001628
47. Hentschel H, Procaccia I. The infinite number of generalized dimensions of fractals and strange attractors. *Phys D.* (1983) **8**:435–44. doi: 10.1016/0167-2789(83)90235-X
48. Moein MJA, Valley B, Evans KF. Scaling of fracture patterns in three deep boreholes and implications for constraining fractal discrete fracture network models. *Rock Mech Rock Eng.* (2019) **52**:1723–43. doi: 10.1007/s00603-019-1739-7
49. Mandelbrot BB, Pignoni R. *The Fractal Geometry of Nature.* New York, NY: WH Freeman (1983). doi: 10.1119/1.13295
50. Plotnick RE, Gardner RH, Hargrove WW, Prestegaard K, Perlmutter M. Lacunarity analysis: a general technique for the analysis of spatial patterns. *Phys Rev E.* (1996) **53**:5461. doi: 10.1103/PhysRevE.53.5461
51. Kaye BH. *A Random Walk Through Fractal Dimensions*. John Wiley & Sons (2008).
52. Roy A, Perfect E, Dunne WM, Odling N, Kim J-W. Lacunarity analysis of fracture networks: evidence for scale-dependent clustering. *J Struct Geol.* (2010) **32**:1444–9. doi: 10.1016/j.jsg.2010.08.010
53. Roy A, Perfect E, Dunne WM, McKay LD. A technique for revealing scale-dependent patterns in fracture spacing data. *J Geophys Res.* (2014) **119**:5979–86. doi: 10.1002/2013JB010647
54. Lavoine E, Davy P, Darcel C, Le Goc R. On the density variability of poissonian discrete fracture networks, with application to power-law fracture size distributions. *Adv Geosci.* (2019) **49**:77–83. doi: 10.5194/adgeo-49-77-2019
55. Bour O, Davy P. Connectivity of random fault networks following a power law fault length distribution. *Water Resour Res.* (1997) **33**:1567–83. doi: 10.1029/96WR00433
56. Stauffer D, Aharony A. *Introduction to Percolation Theory: Revised.* 2nd Edition. CRC Press (2014).

Conflict of Interest: RM was employed by the company Svensk Kärnbränslehantering AB (SKB).

The authors declare that this study received funding from SKB. The funder had the following involvement with the study: formulation of some of the problems addressed and review of early drafts of the manuscript.

frontiers in Physics

ORIGINAL RESEARCH
published: 18 February 2020
doi: 10.3389/fphy.2020.00024

Characterization and Control of an Ion-Acoustic Plasma Instability Downstream of a Diverging Magnetic Nozzle

Scott J. Doyle[1]*, Alex Bennet[2], Dimitrios Tsifakis[2], James P. Dedrick[1], Rod W. Boswell[2] and Christine Charles[2]

[1] Department of Physics, York Plasma Institute, University of York, York, United Kingdom, [2] Space Plasma, Power and Propulsion Laboratory, Research School of Physics and Engineering, The Australian National University, Canberra, ACT, Australia

The study and control of resonant instabilities in magnetized plasmas is of fundamental interest over a wide range of applications from industrially relevant plasmas to plasma sources for spacecraft propulsion. In this work electrostatic probes were employed to measure a 4–20 kHz instability in the ion saturation current downstream of an electric double layer (DL) in an expanding helicon plasma source. The amplitude and frequency of the instability were found to vary in inverse proportion to the operating argon gas pressure (0.2–0.6 mTorr) and in direct proportion to the applied rf power (100–600 W) and applied solenoid current (3–8 A). A spatially resolved characterization of the maximum instability amplitude determined two radial maxima, corresponding to the locations of most positive radial ion density gradient. Control and inhibition of the instability were achieved through the application of a kHz voltage amplitude modulation to the 13.56 MHz radio-frequency (rf) power supplied to the helicon antenna. Through the application of voltage amplitude modulations in the frequency range 2–12 kHz the instability was reduced by up to 65%, exhibiting a greater reduction at higher applied modulation frequencies. This effect is described through a variation in the radial ion density gradient via asymmetrically attenuated ion acoustic density perturbations induced by the applied voltage modulation. The application of voltage amplitude modulations has been demonstrated as a potential control mechanism for density gradient driven instabilities in magnetized plasmas.

Keywords: helicon, magnetized plasmas, double-layer, ion-acoustic instability, radio-frequency

OPEN ACCESS

Edited by:
Ashild Fredriksen,
Arctic University of Norway, Norway

Reviewed by:
Bill Amatucci,
United States Naval Research Laboratory, United States
Saikat Chakraborty Thakur,
University of California, San Diego, United States

****Correspondence:***
Scott J. Doyle
Scott.Doyle@Physics.org

Specialty section:
This article was submitted to Plasma Physics, a section of the journal Frontiers in Physics

Received: *20 November 2019*
Accepted: *27 January 2020*
Published: *18 February 2020*

Citation:
Doyle SJ, Bennet A, Tsifakis D, Dedrick JP, Boswell RW and Charles C (2020) Characterization and Control of an Ion-Acoustic Plasma Instability Downstream of a Diverging Magnetic Nozzle. Front. Phys. 8:24. doi: 10.3389/fphy.2020.00024

1. INTRODUCTION

Electromagnetic propulsion devices present a growing alternative to chemical propulsion sources as they are capable of providing higher specific impulses and employ less volatile propellants [1–4]. Within this catagory, radio-frequency (rf), electrodeless plasma thrusters are of particular interest as they could provide extended operational lifetimes as compared to more commonly employed Hall and gridded ion thrusters. The Helicon Double Layer Thruster (HDLT) is an example of a rf, electrodeless, neutraliser-free plasma thruster and employs a current-free double layer (CFDL) to accelerate an ion beam to velocities beyond the local ion sound speed, producing variable thrust in the mN range [5–9].

Over the last two decades, a number of expanding plasma devices have been used to explore the physics of low pressure plasmas flowing through diverging magnetic nozzles for conditions relevant to HDLT-type thrusters. These devices typically consist of a rf plasma source surrounded by a set of solenoids and contiguously attached to a larger radius expansion chamber. Much of the previous work conducted in these devices has focused on characterizing the axial DL [9–13] and subsequent ion beam [14–16] as well as a region of high density plasma located off-axis in the expansion chamber known as the high-density conics [17–22]. The conics are ionized locally by hot electron populations, which stream along the most radial magnetic field lines to escape the source region, and result in a hollow radial plasma density profile in the expansion region. The downstream region of these expanding plasma devices can therefore be described by an axial supersonic ion beam radially surrounded by the stationary high density conics.

Strong gradients in magnetic field strength, plasma potential, density, and electron temperature are inherent to the expanding plasma devices in which DLs, ion beams and conics have been observed [23, 24]. Ion acoustic instabilities generated by strong gradients in these plasma properties have been the subject of investigation for many years [25–28], and extensive work has been preformed on density gradient driven, kHz range, drift wave instabilities in helicon sources [29–34] and turbulent or anomalous transport in other types of electric propulsion devices, e.g., spoke instabilities in Hall thrusters [35–37].

In an experimental study on the HDLT, Aanesland et al. showed the presence of a 10–20 kHz upstream ionization instability in the source region of the Chi Kung reactor at the Australian National University (ANU) and investigated its source [38, 39]. The authors of that study theorized that the instability was an ionization instability caused by an energetic electron population accelerated into the plasma source from the expansion region by the CFDL. However, measurements of the axial upstream Electron energy probability function (EEPF) taken by Takahashi et al. in the same reactor did not show the presence of an accelerated electron beam in the source region [40]. The source of the instability presented in the previous work is yet to be fully understood and further investigation is necessary to understand the particle dynamics in these expanding plasma devices.

In this work, an instability found at similar frequencies to that presented in the previous work by Aanesland et al. is detected in the downstream region of the Chi Kung reactor. Control of the instability is achieved via the application of variable frequency voltage amplitude modulations to the rf power supplied to the rf antenna surrounding the source region. Here, ion acoustic waves, i.e., ion density perturbations, are employed to control and reduce the instability via modification of the radial ion density gradient. An overview of the Chi-Kung reactor and the electrostatic diagnostics employed is given in section 2. The instability amplitude and frequency are spatially characterized in section 3, revealing the apparent source of the instability and the dependence on operating pressure, applied rf antenna power and solenoid current. The voltage amplitude modulation technique is detailed in section 4.1 and control of the instability is demonstrated for varying voltage amplitude modulation frequencies in section 4.2 including a mathematical description of the proposed control mechanism.

2. DESCRIPTION OF THE EXPERIMENTAL SETUP

The measurements presented in this study were taken in the Chi-Kung helicon plasma reactor, shown in **Figure 1**. The operation of Chi Kung is described in detail in Charles and Boswell [8] and is briefly outlined here for completeness. The Chi Kung plasma reactor source chamber consists of a 31 cm long, 0.65 cm thick, 13.7 cm inner diameter cylindrical Pyrex tube, contiguously connected to a 30 cm long, 32 cm inner diameter grounded aluminum expansion chamber. The axial coordinate system Z is zeroed at the interface between the source chamber and the expansion region, as shown in **Figure 1**. To compare with previous measurements of the instability in Aanesland et al. [39], an insulating glass plate is positioned on the upstream side of the source chamber, illustrated in **Figure 1** by a light blue section at $Z = -31$ cm. The presence of an insulating backplate ensures that all walls in the source region are floating, enforcing an insulated boundary condition upstream [41, 42] leading to enhanced electron fluxes downstream. A pumping system consisting of a turbomolecular and rotary pump is used to maintain a base chamber pressure of 5×10^{-6} Torr. Argon gas is introduced into the system through a vacuum feedthrough at $Z = -31$ cm and is regulated using a mass flow controller. Experiments in this study are conducted at operating argon gas pressures between 0.2 and 0.6 mTorr, as measured by an MDC P/N 432025 pressure gauge. Two solenoids in a Helmholtz pair configuration, positioned at $Z = -28.6$ cm, and $Z = -9$ cm,

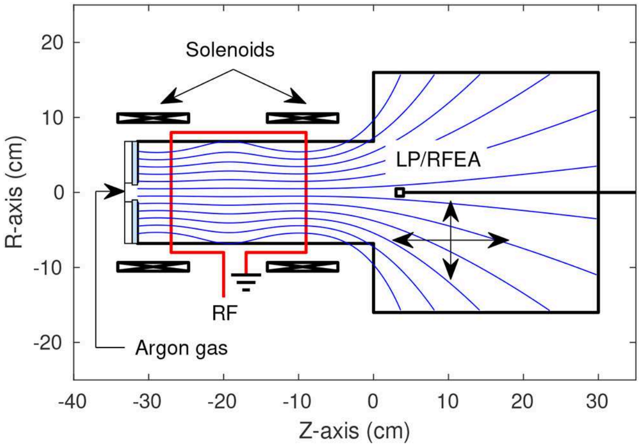

FIGURE 1 | Schematic of the Chi Kung reactor showing the locations of the solenoids, RF antenna, gas inlet, insulating glass plate in the source region and diagnostic probes. The magnetic field lines generated by the solenoids are pictured as blue lines. The Langmuir probe (LP) and retarding field energy analyzer (RFEA) possessed an axial range of $-20 \le Z \le 30$ cm and a radial range of $-14 \le R \le 14$ cm.

provide a near-parallel magnetic field in the source region and a diverging magnetic field in the downstream expansion chamber. A 3–8 A DC current is supplied to both solenoids, resulting in an on-axis magnetic field strength of between ∼48 and 193 G. Power is coupled to the plasma at 13.56 MHz through an 18 cm long double saddle antenna, centered on $Z = -20$ cm, and supplied by an ENI OEM-25, 3 kW solid state power supply via an impedance matching network.

A Langmuir probe (LP) was used to measure the ion saturation current (probe biased to –67 V) downstream of the double layer, located just upstream of the source aperture between –9 and 0 cm. The probe consisted of a one-sided 1.5 mm radius nickel disk installed at the end of a 5 cm long piece of alumina tube, itself fixed to a grounded probe shaft. The probe was free to move both axially and radially without breaking vacuum, allowing for an axial range of $-20 \leq Z \leq 30$ cm and a radial range of $-14 \leq R \leq 14$ cm. Ion saturation currents were acquired and fast Fourier transformed by a National Instruments NI PXI-5122 high-speed digitizer (100 MHz sampling rate), prior to further analysis.

Ion densities downstream of the double layer were obtained employing a retarding field energy analyser (RFEA) operating in ion collection mode. The RFEA follows a design published previously in Charles and Boswell [8] and consists of a stack of biased nickel grids positioned in front of a nickel collector plate. The grids act as a directional energy filter for ions accelerated by the grounded sheath surrounding the RFEA and entering the probe orifice. In ion mode operation, the RFEA repeller grid is aligned perpendicular to the ion beam and biased to –80 V and the applied discriminator voltage V_d is scanned between 0 and 80 V. Those ions with sufficient kinetic energy to overcome the potential barrier provided by the discriminator grid are measured as an integrated ion current at the collecting electrode, $I_c(V_d)$.

As V_d is swept from 0 to 80 V, the increasing potential barrier between the discriminator grid and the collecting plate results in a decrease in the current incident upon the collector plate. Eventually, V_d becomes too large and no ions are detected at the collecting electrode. The total stationary ion current can then be extracted by applying a Gaussian fit to the first derivative of RFEA I-V trace, and converted into the stationary ion density via comparison to the stationary ion current measured via a Langmuir probe, as performed and described in Bennet et al. [22].

3. IDENTIFICATION AND CHARACTERIZATION OF A kHz INSTABILITY

An instability in the ion saturation current has been observed in the expansion chamber of the Chi Kung reactor under conditions known to support an ion beam [21]. A representative example of the instability measured in a 13.56 MHz, 300 W, 0.25 mTorr argon discharge is shown in **Figure 2A**, indicating the maximum amplitude ξ and central frequency ν_ξ. The radial distribution in the maximum instability amplitude and the radial ion density gradient $\frac{\delta n_i}{\delta R}$ are shown in **Figure 2B**, as measured downstream of the DL ($Z = 2$ cm) for the same operating conditions. Ion density measurements were obtained between $0 \leq R \leq 10$ cm employing identical conditions for a 0.30 mTorr operating pressure.

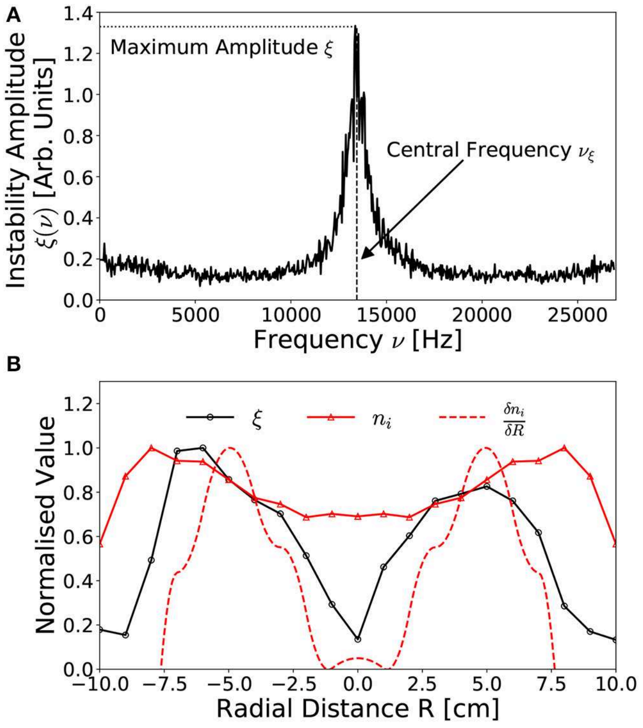

FIGURE 2 | Representative examples of **(A)** the fourier transformed ion saturation current measured off-axis (R, Z = –6 cm, 2 cm) downstream of the DL, exhibiting an anomalous instability centered on $\nu_\xi = 13.4$ kHz with a 2 kHz FWHM bandwidth and **(B)** the radial distribution (at $Z = 2$ cm) in the maximum instability amplitude ξ, the ion density n_i and the radial ion density gradient $\frac{\delta n_i}{\delta R}$. Solid lines added to guide the eye. Operating conditions: helicon antenna supplied $P_{rf} = 300$ W at $\nu_{rf} = 13.56$ MHz employing 0.25–0.30 mTorr argon with $I_{1,2} = 6$ A solenoid current.

The instability shown in **Figure 2A** represents an ion acoustic wave at $\nu_\xi = 13.4$ kHz propagating downstream of the DL (R, Z = –6 cm, 2 cm). Notably, this frequency does not correspond to any harmonic or alias of the driving 13.56 MHz frequency nor to the ion cyclotron resonance frequency, instead arising from anomalous periodic interactions within the plasma. The relative amplitude and frequency range of the instability agree with previously measured anomalous signals observed under similar operating conditions, presented in Aanesland et al. [39].

The spatial distribution in the amplitude of the instability, shown in **Figure 2B** for $-10 \leq R \leq 10$ cm at $Z = 2$ cm, exhibits two approximately symmetric maxima located R∼5–6 cm off-axis, reducing on-axis and adjacent to the expansion chamber walls. Additionally **Figure 2B** contains the normalized radial ion density profile measured between $0 \leq R \leq 10$ cm employing an axially aligned RFEA positioned at the same axial location, where ion density data is obtained from Bennet et al. [22]. The radial density profile shows the high-density conics as locations of increased ion density off axis centered around

R ~7–8 cm [15, 22]. The amplitude of the instability drops rapidly toward the regions of peak ion density, limiting its radial extent to within the conics. In fact, the radial distribution of the instability demonstrates maxima corresponding with the regions of highest increasing radial ion density gradient. To illustrate this, the radial ion density gradient $\frac{\delta n_i}{\delta R}$, obtained from a differentiated smoothing spline fit of the radial ion density, has been plotted on **Figure 2B**, where $\frac{\delta n_i}{\delta R}$ features two peaks located off-axis which align closely with the peaks of the instability amplitude. This indicates that the observed instability could be caused by the radial density gradient in the downstream generated by the conics. Note that the ion density in **Figure 2B** was measured only on the +R side, with the ion density and density gradient for the –R side presented as a symmetric mapping.

To further investigate the dependencies between the spatial distribution of the instability and the radial ion density gradients, LP measurements of the instability magnitude and central frequency were performed with a 1 cm radial and axial resolution across the mid-plane downstream of the Chi-Kung source. The central frequency of the instability downstream of the DL was found to be approximately spatially invariant, however the amplitude of the instability exhibited both a radial and axial spatial variation. A 2D map of the maximum amplitude of the instability is shown in **Figure 3A** for a 13.56 MHz, 300 W, argon discharge at 0.25 mTorr. **Figure 3B** shows a 2D map of the normalized radial ion density gradient, again employing smoothing splines allowing for calculation of $\frac{\delta n_i}{\delta R}$ throughout the measured area.

As observed previously in **Figure 2**, a similarity exists between the topography of the instability amplitude and the most positive $\frac{\delta n_i}{\delta R}$, exhibited in **Figures 3A,B**. Both the amplitude of the instability and the radial ion density gradient exhibit maxima coinciding with the beam-conic interface, denoted by the white dashed lines. The beam-conic interface marks the transition from the low density, high ion energy conditions on-axis into the high density, low ion energy conics.

In addition, there exists a "global" radial asymmetry in the topology of the instability, with the highest instability amplitudes measured for the –R side of the Chi-Kung reactor. This global asymmetry likely arises from the rf antenna geometry, where higher average instability amplitudes coincide with the "hot" –R ($R < 0$ cm) side of the antenna, while the lower average instability amplitudes align with the "cold" +R ($R > 0$ cm) grounded end of the antenna. Note that due to the radial mirroring of the ion density measurements, performed on the "cold" +R side of the source, this asymmetry is not observed in **Figure 3B**. Previous work employing similar operational conditions have observed an asymmetric ion density distribution about the central axis, with the higher ion density localized to the –R side of the reactor [22]. From the topological comparisons of the instability to the ion density gradients presented in **Figures 2B**, **3** it is likely that the instability forms either as a result of ion acoustic interactions across the beam-conic interface [23, 24], or as a density gradient driven drift wave instability [32, 33].

Before attempting to modulate the radial ion density gradients via the introduction of a voltage waveform amplitude modulation, the behavior of the instability was first investigated with respect to the applied rf power P_{rf}, the DC solenoid current $I_{1,2}$ and the operating argon pressure. These findings are presented in section 3.1, where measurements of the ion saturation current were performed as described previously with reference to **Figure 3**, employing a radial scan across $-10 \leq R \leq 10$ cm at $Z = 2$ cm downstream of the DL.

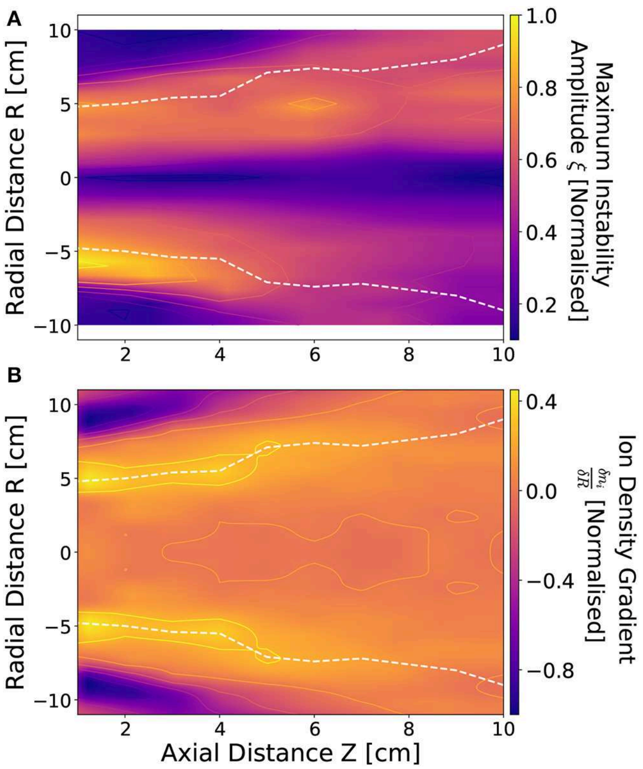

FIGURE 3 | Normalized **(A)** maximum instability amplitude ξ and **(B)** radial ion density gradient $\frac{\delta n_i}{\delta R}$ downstream of the DL between $1 \leq Z \leq 10$ cm and $-10 \leq R \leq 10$ cm. White dashed lines denote the approximate location of beam-conic interface. Operating conditions: helicon antenna supplied $P_{rf} = 300$ W at $\nu_{rf} = 13.56$ MHz employing 0.25 mTorr argon with $I_{1,2} = 6$ A solenoid current.

3.1. Power, Operating Pressure, and Solenoid Current Dependence

The instability amplitude and central frequency are shown with respect to varying applied power for a 0.25 mTorr pressure, 6 A solenoid current discharge in **Figures 4A,D**, respectively, with respect to varying solenoid current for a 300 W, 13.56 MHz, 0.25 mTorr pressure discharge in **Figures 4B,E**, respectively and with respect to varying operating pressure for a 300 W, 6 A solenoid current discharge in **Figures 4C,F**, respectively. Measurements were taken by employing a LP downstream (Z - 2 cm) of the Chi-Kung source, positioned on-axis ($R = 0.0$ cm) and at the radial beam-conic interfaces ($R = \pm 5$ cm).

Increasing the applied antenna power results in a proportional increase in the amplitude of the instability, shown in **Figure 4A**, where the constant of proportionality varies with radial location. Generally however, a factor of three increase in the rf power

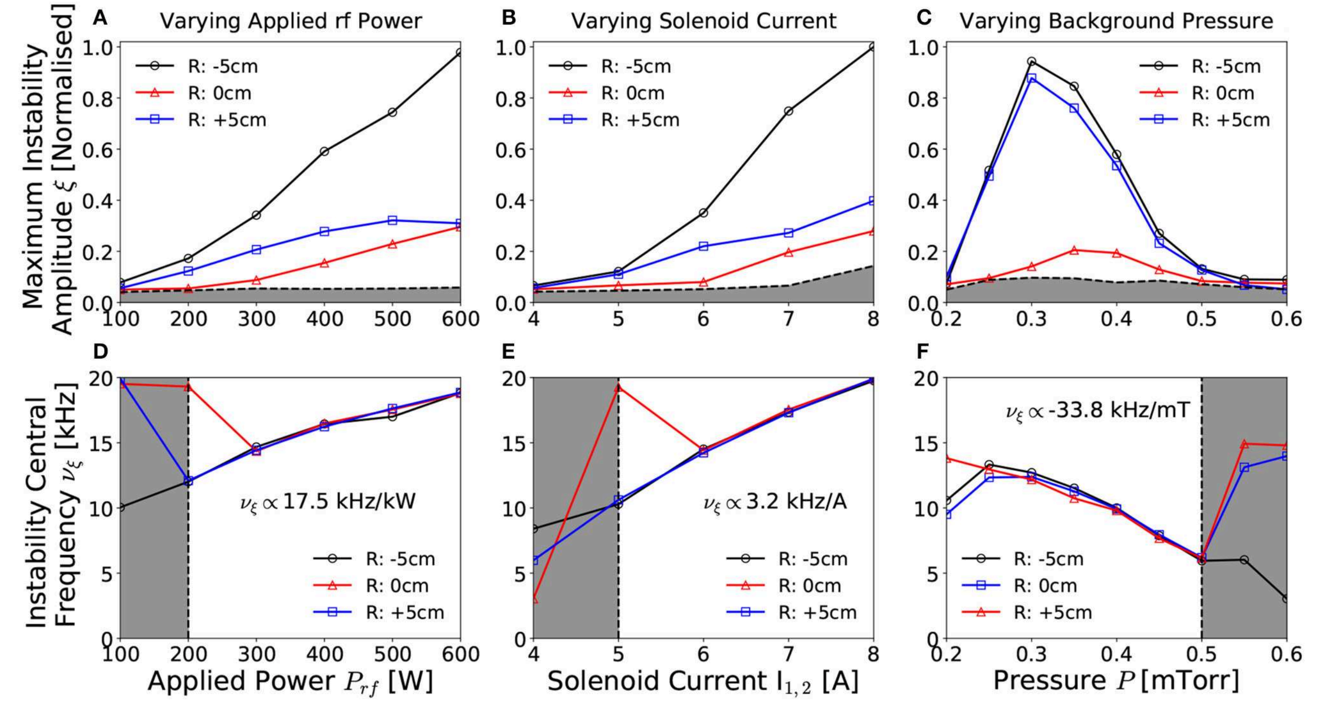

FIGURE 4 | Normalized instability amplitude and frequency with respect to **(A,D)** rf power P_{rf}, **(B,E)** solenoid current $I_{1,2}$ and **(C,F)** operating argon pressure respectively, as measured on-axis ($R = 0.0$ cm) and off-axis ($R = \pm 5$ cm), $Z = 2$ cm downstream of the DL. Off-axis measurements correspond to the beam-conic interface, correlating with the region of highest radial ion density gradient. The horizontal gray shaded regions in **(A–C)** denote the background noise level, while the vertical gray shaded regions in **(E–F)** denote measurements for which SNR $\leq$ 2. Solid lines added to guide the eye. Operating conditions: helicon antenna supplied $P_{rf} = 100 - 600$ W at $\nu_{rf} = 13.56$ MHz employing 0.2–0.6 mTorr argon with $I_{1,2} = 4 - 8$ A solenoid current.

relates to an order of magnitude increase in the amplitude of the instability. The spatial distribution continues to exhibit two maxima coinciding with the ±R beam-conic interfaces, as observed previously for **Figure 3A**. Note that the ratio between the maximum amplitude measured at the "hot" –R and "cold" +R sides of the chamber increases with increasing applied rf power. While this is likely due to an increasing asymmetry between the –R and +R ion density gradients, it should be noted that other plasma conditions, such as the localized electric field adjacent to the antenna, may have an influence on the plasma stability. Note however, in contrast to the instability amplitude, the frequency of the instability in **Figure 4D** remains spatially invariant and exhibits a linear dependence on the applied rf power, increasing at a rate of 17.5 kHz kW^{-1}, corresponding to a doubling in central frequency for a factor three increase in rf power.

Previous studies of the Chi-Kung reactor and similar inductively coupled rf coupled plasma sources observed an approximately linear relationship between the applied power and the maximum plasma density [5, 21, 43]. This increase is typically not spatially homogeneous, but instead the maximum density increase is localized on-axis, leading to enhanced radial density gradients [22]. The observed correlation between these proportionalities with the topological similarities observed previously between **Figures 3A,C** supports the hypothesis that the instability is indeed primarily influenced by the ion density distribution across the beam-conic interface.

The axial magnetic field strength was controlled by altering the DC current supplied to the solenoids, where a symmetrical current $I_{1,2} = I_1 = I_2$ was supplied for all cases in **Figure 4**. The instability is first observed above background noise in **Figure 4B** for $I_{1,2} \geq 5$ A (≈ 125 G on-axis) agreeing with previously determined minimum solenoid currents required for ion beam formation [44]. The amplitude and frequency of the instability increase in proportion to the solenoid current, shown in **Figures 4B,E**, exhibiting an order of magnitude increase in the maximum amplitude for a doubling of the applied solenoid current. Such behavior is to be expected due to the increased radial confinement giving rise to enhanced axial and radial ion density gradients within the source [11, 22]. Increasing the radial confinement also amplifies any inherent radial asymmetry in the ion density arising from the antenna geometry, again resulting in higher maximum instability amplitudes on the "hot" –R side of the chamber as compared to the "cold" +R side.

The central frequency of the instability exhibits an approximately linear 3.1 kHz A^{-1} proportionality to the applied solenoid current, shown in **Figure 4E**, and remains spatially invariant for currents known to produce an ion beam. Here, a doubling of the solenoid current exhibits a factor of four increase in the frequency of the instability. Note also that reversing the orientation of the magnetic field had no effect on the frequency, amplitude or radial asymmetry of the instability.

Both the amplitude and frequency of the instability exhibited the greatest response to variations in the operating pressure, shown in **Figures 4C,F**, respectively. Here, the instability is observed for pressures between 0.25 and 0.50 mTorr, agreeing with the measured range in Aanesland et al. [38], and the lower limit of which corresponds to the minimum pressure for ion beam formation [8, 15, 21]. The rapid growth in the amplitude of the instability (off-axis) for increasing pressure between 0.20 and 0.30 mTorr suggests that the presence of an ion beam may be necessary for the formation of the instability. Beyond 0.30 mTorr the amplitude of the instability exhibits an inversely proportional relationship to the operating pressure, falling below the background noise for pressures above 0.50 mTorr. Note that the global asymmetry in the amplitude of the instability between the "hot" –R and "cold" +R sides of the reactor is less pronounced with changes in operating pressure than with power or solenoid current.

The frequency of the instability, shown in **Figure 4F**, exhibits an inversely proportional relationship with the operating pressure, varying by –33.8 kHz mTorr^{-1} between 0.25 and 0.50 mTorr, or a factor three reduction in the central frequency for a doubling of the pressure. Increasing the operating pressure results in an increased ionization rate, reducing the on-axis confinement, potentially leading to reduced radial ion density gradients. In addition to altering the radial ion density gradient, previous work has shown that the potential drop across the DL, and hence the ion beam energy and flux, reduces with increasing operating pressure [21]. Therefore, a reduction in the ion beam energy, coupled with the reduced radial ion density gradient, reduces the instability amplitude and frequency, as observed in **Figures 4C,F**, resulting in a greater influence than that exerted by varying the applied power or solenoid current.

4. CONTROL OF kHz INSTABILITY VIA AN IMPOSED VOLTAGE AMPLITUDE MODULATION

The characterization performed in section 3 indicates a relationship between the radial ion density gradient and the amplitude and frequency of the instability. Further, the appearance of the instability at operating pressures corresponding to the limiting conditions for ion beam formation suggest a link between the two phenomena. Control and reduction of the instability may therefore impact the ion beam parameters. Such control may be achieved through altering the radial ion density gradient, implemented in this work through the application of voltage waveform amplitude modulations.

4.1. Voltage Waveform Amplitude Modulation Technique

Voltage waveform amplitude modulation was employed to produce variable frequency and amplitude ion acoustic waves, i.e., time varying ion density fluctuations, within the Chi-Kung source. Voltage waveform amplitude modulation involves the superposition of two non-harmonic sinusoidal voltage waveforms; consisting of a high frequency "carrier" waveform, facilitating power coupling to the plasma, and a low frequency "envelope" waveform, introducing the desired ion acoustic wave but otherwise not significantly affecting the average power deposition. The resulting voltage waveform $\phi_{rf}(t)$ is described by Equation (1):

$$\phi_{rf}(t) = \big(V_{rf}\sin(2\pi\nu_{rf}t)\big)\cdot\big(V_{mod}\sin(2\pi\nu_{mod}t)\big) \qquad (1)$$

where, V_{rf} and V_{mod} are the carrier and modulation voltage amplitudes, respectively and ν_{rf} and ν_{mod} represent the carrier and modulation frequencies, respectively. As the modulation voltage is applied symmetrically about the mean carrier voltage, the power deposited into the source remains approximately constant so long as $V_{mod} \leq V_{rf}$. Modulation frequencies in the range $1 \leq \nu_{mod} \leq 12000$ Hz were employed, while the carrier frequency was maintained at $\nu_{rf} = 13.56$ MHz. For ease of discussion, the extent to which the carrier waveform is modulated is discussed in terms of the modulation fraction ϕ_{mod}, defined as the fraction of the modulation amplitude with respect to the carrier waveform ϕ_{rf} within a single modulation phase cycle τ_{mod}.

In practice, voltage amplitude modulation was achieved through applying a seed envelope voltage waveform, supplied by a Siglent SDG 5162 waveform generator, to the "volume" input on the ENI OEM-25 power supply, directly varying V_{mod}. An example of a modulated voltage waveform generated in this way is shown in **Figure 5**.

The voltage waveform in **Figure 5** consists of a base 13.56 MHz voltage frequency, modulated at $\nu_{mod} = 8$ kHz by a modulation fraction of $\phi_{mod} = 0.16\ \phi_{rf}$. Ions within the source chamber, being heavier and less mobile than the electrons, will respond preferentially to the low-frequency component of the modulated waveform. Ion acoustic waves induced by the kHz modulation voltage are then free to propagate through the source chamber and into the downstream region. Altering the modulation amplitude and frequency therefore provides a means of selectively influencing the ion dynamics, providing a control

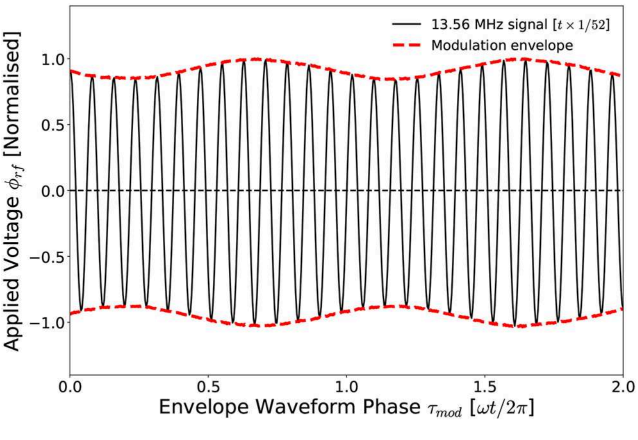

FIGURE 5 | Example of rf amplitude modulation where the base frequency of $\nu_{rf} = 13.56$ MHz is modulated at $\nu_{mod} = 8$ kHz by a modulation fraction of $\phi_{mod} = 0.16$. For clarity the 13.56 MHz signal is plotted at $\tau_{rf} = \frac{1}{52}\ \tau_{mod}$ temporal scale.

mechanism for the ion density gradients across the ion beam-conic interface. The application of voltage amplitude modulation can therefore be used to investigate and control the instability characterized in section 3.

4.2. Control of Instability Frequency and Amplitude

As an initial proof of concept, a voltage waveform amplitude modulation of frequency $\nu_{mod} = 1$ Hz and modulation fraction $\phi_{mod} = 0.5\,\phi_{rf}$ was applied to the 13.56 MHz carrier waveform, as described in section 4.1, for a $P_{rf} = 300$ W, $I_{1,2} = 6$ A, 0.25 mTorr argon discharge. The resulting temporally-averaged and temporally-resolved LP ion saturation current measurements of the instability amplitudes measured off-axis downstream of the DL (R, Z = –6 cm, 2 cm) are shown in **Figures 6A,B**, respectively.

Figure 6A indicates that, in the absence of an applied amplitude modulation, the instability exhibits an approximately Gaussian frequency space profile centered at 14.2 kHz, agreeing with previous observations. Upon application of a 1 Hz voltage modulation the instability profile splits into a bi-modal distribution, consisting of a low amplitude, low frequency ($\approx$ 12.6 kHz) component and a high amplitude, high frequency ($\approx$ 15.7 kHz) component. This "smearing" of the instability over a wider frequency range arises due to the periodic alteration in the applied rf power, following the proportionalities exhibited in **Figures 4A,B**. Varying the power sinusoidally therefore produces a low amplitude, low frequency peak corresponding to the minima of the envelope waveform and a high amplitude, high frequency peak corresponding to the maxima of the envelope waveform. This is more clearly visualized in the temporally resolved instability amplitude, shown in **Figure 6B**, illustrating approximately 12 voltage amplitude modulation cycles. Note that the LP measurement timescale and the $\nu_{mod} = 1$ Hz amplitude modulation timescale are not synchronized, resulting in a beat-effect where the instability is measured at different phases of the amplitude modulation.

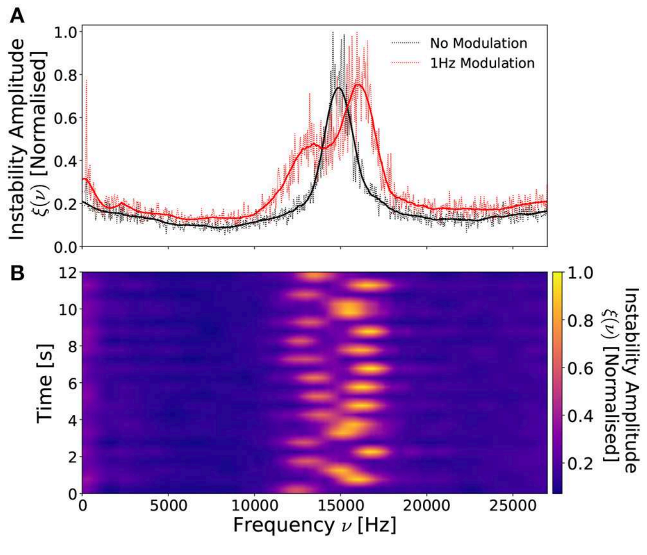

FIGURE 6 | Normalized **(A)** time-integrated instability amplitude profiles with (red curve) and without (black curve) a $\nu_{mod} = 1$ Hz, $\phi_{mod} = 0.5\phi_{rf}$ voltage amplitude modulation and **(B)** time-resolved instability profiles employing the same voltage amplitude modulation. Instability measured at the beam-conic interface ($R = -6$ cm) downstream of the DL ($Z = 2$ cm). Solid lines in **(A)** were smoothed employing a SavitzkyGolay filter. Operating conditions: helicon antenna supplied $P_{rf} = 300$ W at $\nu_{rf} = 13.56$ MHz employing 0.25 mTorr argon with $I_{1,2} = 6$ A solenoid current.

Recalling that the amplitude of the instability varies with respect to the radial ion density gradient across the beam-conic interface, see **Figures 3A,B**, it is plausible that effects arising from the voltage amplitude modulation in this "low-frequency" regime are influenced not only by the ion acoustic wave density perturbations $\delta n_i(\tau_{mod})$, but also by variations in the stationary ion density gradients $\left|\frac{\delta n_i}{\delta R}\right|_0(\tau_{mod})$, the latter of which follow trends observed in **Figures 4A,B** described by Equation (2):

$$\frac{\delta n_i}{\delta R}(\tau_{mod}) \propto \left|\frac{\delta n_i}{\delta R}\right|_0 (\tau_{mod}) \pm \delta n_i(\tau_{mod}) \tag{2}$$

Such behavior is indicative of a modulation timescale $\tau_{mod} = \nu_{mod}^{-1}$ low enough such that the stationary ion density is capable of responding. Increasing the applied modulation frequency results in a reduced "frequency smearing" effect, previously observed in **Figure 6B**, indicating a decreasing temporal variation in the stationary ion density gradient. For modulation frequencies in excess of $\nu_{mod} \geq 2$ kHz the instability exhibits no evidence of a frequency smearing effect, indicating a constant stationary radial ion density gradient. The time-resolved radial ion density gradient, described by Equation (3), can therefore be expected to consist of a time-invariant stationary density $\left|\frac{\delta n_i}{\delta R}\right|_0$ superimposed with a time-varying ion density perturbation, induced by the modulation voltage:

$$\frac{\delta n_i}{\delta R}(\tau_{mod}) \propto \left|\frac{\delta n_i}{\delta R}\right|_0 \pm \delta n_i(\tau_{mod}) \tag{3}$$

This "high-frequency" regime presents the capability to influence the instability through specifically varying the amplitude and frequency of the ion density perturbations. The remainder of this study addresses the degree to which the instability can be controlled within the high-frequency modulation regime.

The frequency distribution of the instability for three applied $\nu_{mod} = 2$ kHz voltage amplitude modulations in a $P_{rf} = 300$ W, $I_{1,2} = 6$ A, $\nu_{rf} = 13.56$ MHz, 0.42 mTorr discharge are shown in **Figure 7A**, with the corresponding maximum instability amplitude shown with respect to varying amplitude modulation fraction between 0.0 $\phi_{rf} \leq \phi_{mod} \leq 0.74\,\phi_{rf}$ presented in **Figure 7B**.

The unmodulated ($\phi_{mod} = 0.0$) instability in **Figure 7A** exhibits an approximately Gaussian frequency distribution with a central frequency of 9.2 kHz, lower than that observed in **Figure 2A**, due to the increased operating pressure. Applying a $\nu_{mod} = 2$ kHz, $\phi_{mod} = 0.38\,\phi_{rf}$ voltage amplitude modulation has no significant affect on the instability, altering

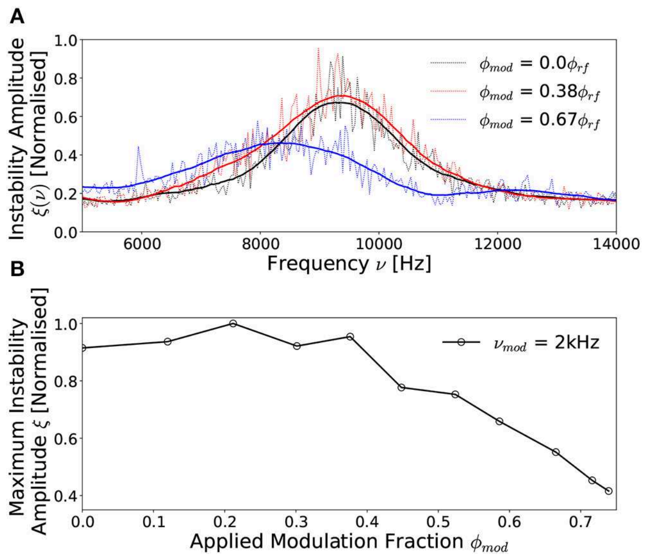

FIGURE 7 | Normalized instability amplitude profiles as measured at the beam-conic interface downstream of the DL (R, Z = –6 cm, 2 cm) with a $\nu_{mod} = 2$ kHz voltage amplitude modulation for **(A)** $\phi_{mod} = 0.0\ \phi_{rf}$, $\phi_{mod} = 0.38\ \phi_{rf}$ and $\phi_{mod} = 0.67\ \phi_{rf}$ and **(B)** the variation in peak instability amplitude with respect to increasing ϕ_{mod}. Solid lines in **(A)** were smoothed employing a SavitzkyGolay filter and included in **(B)** to guide the eye. Operating conditions: helicon antenna supplied $P_{rf} = 300$ W at $\nu_{rf} = 13.56$ MHz employing 0.42 mTorr argon with $I_{1,2} = 6$ A solenoid current.

neither the amplitude nor the central frequency. The preservation of a Gaussian frequency profile differs from that observed previously at $\nu_{mod} = 1$ Hz in **Figure 6A** demonstrating independence from the time-varying rf power indicating a high-frequency modulation regime. It should be noted that the transition between the low-frequency to high-frequency regimes is relatively gradual with increasing modulation frequency. As such $\nu_{mod} = 2$ kHz represents an empirically chosen frequency marking this transition as it is the lowest modulation frequency for which evidence of a bi-modal distribution is on-par with the background noise level. Increasing the modulation fraction beyond $\phi_{mod} \geq 0.4\ \phi_{rf}$ results in a significant reduction in both the amplitude and central frequency of the instability, albeit also extending the frequency range. The spreading of the instability in frequency space may arise from non-linear coupling between the instability and the modulation frequency [29]. Evidence of this may be observed from a comparison of **Figures 6**, **7**, where for a $\nu_{mod} = 1$ Hz modulation, the highest amplitude component of the instability is up-modulated from 14.2 to 15.7 kHz. In comparison, for a $\nu_{mod} = 2$ kHz modulation, the central frequency of the instability is down-modulated from 9.2 to 8.1 kHz. While these demonstrate the potential for coupling between the modulation frequency and the instability frequency, a full analysis of this phenomena is beyond the scope of this work.

The amplitude of the instability is shown with respect to applied modulation fraction in **Figure 7B**, exhibiting a limited variation in the amplitude of the instability between $0.0\ \phi_{rf} \leq \phi_{mod} \leq 0.4\ \phi_{rf}$, followed by a significant reduction beyond $\phi_{mod} \approx 0.4\ \phi_{rf}$. The inversely proportional relationship follows an approximately linear trend between $0.4\ \phi_{rf} \leq \phi_{mod} \leq 0.74\ \phi_{rf}$, with ξ reducing by 50%, at which point the voltage amplitude modulation fraction was limited by the internal response time of the rf power supply.

The observed variation in the amplitude of the instability with respect to increasing modulation fraction suggests that the ion acoustic waves are altering the time-averaged radial ion density gradient across the beam-conic interface. Assuming a time-invariant stationary ion density and time-varying ion density perturbation from Equation (3), the resulting time-averaged radial ion density gradient can be described by Equation (4).

$$\frac{\delta n_i}{\delta R} = \left|\frac{\delta n_i}{\delta R}\right|_0 \pm \int_0^{\tau_{mod}} \delta n_i(\tau_{mod}) d\tau_{mod} \qquad (4)$$

Here, the ion density perturbation varies in proportion to the applied modulation voltage waveform $\delta n_i(\tau_{mod}) \propto sin(2\pi \nu_{mod} t)$, via Equation (1). Ignoring wave damping and assuming a homogeneous stationary ion density, this would result in a temporally symmetric modification of the radial density gradient, leading to a zero time-averaged contribution from the perturbation.

$$\int_0^{\tau_{mod}} \delta n_i(\tau_{mod}) d\tau_{mod} \propto \int_0^t \sin(2\pi \nu_{mod} t) dt = 0 \qquad (5)$$

However for ion acoustic waves traveling through an inhomogeneous medium, as is the case here, the amplitude of the acoustic wave is attenuated with respect to the density of the background medium. Notably, ion acoustic waves propagating radially within the downstream region experience an increasing stationary ion density, see **Figure 3B**. The positive amplitude (compressive) phase of the wave is therefore likely to experience a higher energy loss than the negative amplitude (rarefaction) phase leading to an asymmetric attenuation of the ion acoustic waveform. The resulting ion density perturbation would therefore be expected to exhibit an amplitude asymmetry, represented in **Figure 8**.

The symmetrically attenuated waveform in **Figure 8** exhibits a zero time-averaged value, agreeing with Equation (5). In contrast, the asymmetrically attenuated ion acoustic wave exhibits a non-zero time-averaged ion density perturbation, such that:

$$\int_0^{\tau_{mod}} \delta n_i(\tau_{mod}) d\tau_{mod} \neq 0 \qquad (6)$$

Ion acoustic waves possessing a negative amplitude asymmetry, such as those shown in **Figure 8** where the negative amplitude of the wave exceeds the positive amplitude, are therefore expected to reduce the time-averaged ion density gradient, via Equation (4), leading to a reduction in the amplitude of the instability. These expected outcomes concur with the measured direction of the ion density gradient in **Figure 2B** and the associated reduction in the amplitude of the instability in **Figures 7A,B**. In addition, note that the

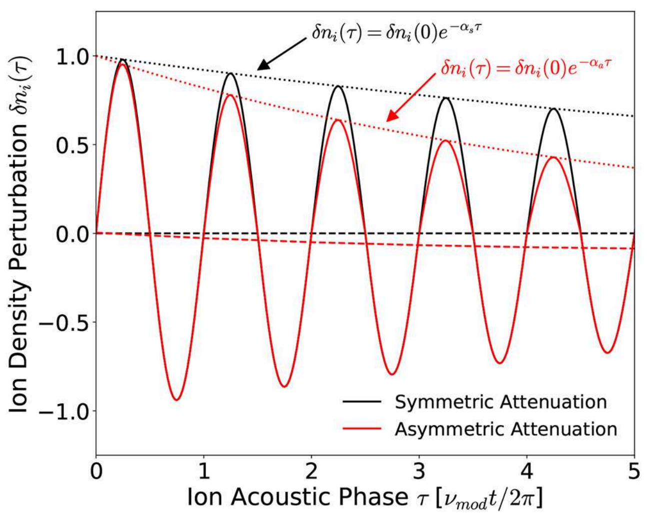

FIGURE 8 | Representative ion acoustic waveforms exhibiting symmetric α_s and asymmetric α_a amplitude attenuations. The time-averaged values and positive amplitudes of the waveforms are denoted by the dashed and dotted lines, respectively.

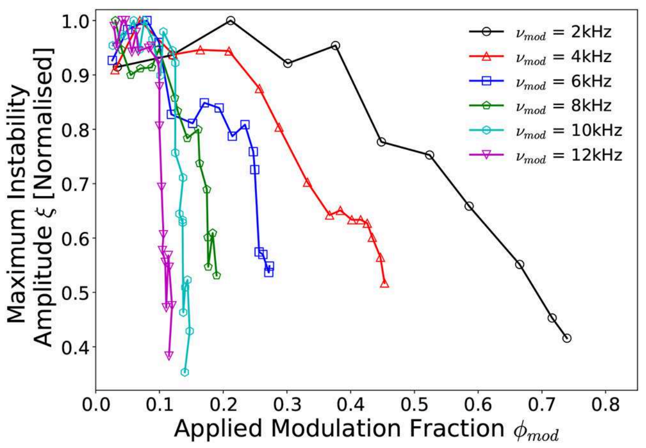

FIGURE 9 | Normalized instability amplitude profiles as measured at the beam-conic interface downstream of the DL (R, Z = –6 cm, 2 cm) with respect to voltage amplitude modulation frequency 2 kHz $\leq \nu_{mod} \leq$ 12 kHz and modulation fraction 0.0 $\leq \phi_{mod} \leq$ 0.74. Solid lines added to guide the eye. Operating conditions: helicon antenna supplied P_{rf} = 300 W at ν_{rf} = 13.56 MHz employing 0.42 mTorr argon with $I_{1,2}$ = 6 A solenoid current.

magnitude of the time-averaged asymmetry increases with the number of ion acoustic phase cycles, i.e., for a fixed source dimension the attenuation of an sub-cyclotronic ion acoustic wave increases in proportion to the applied frequency [45]. Therefore, increasing the voltage modulation frequency would be expected to enhance the effects observed for ν_{mod} = 2 kHz, further reducing the magnitude of the instability.

To investigate the effects of increasing the applied modulation frequency, **Figure 9** shows the maximum instability amplitude with respect to modulation fraction for modulation frequencies in the range 2 $\leq \nu_{mod} \leq$ 12 kHz for a 300 W, 13.56 MHz, $I_{1,2}$ = 6 A, 0.42 mTorr discharge. The instability amplitudes are measured R = –6 cm off-axis, Z = 2 cm downstream of the DL corresponding to the location of peak instability amplitude, see **Figure 3A**.

The instability amplitudes in **Figure 9** were found to be inversely proportional to the frequency of the voltage amplitude modulation, i.e., increasing the modulation frequency results in a substantial reduction in the amplitude of the instability for the same modulation fraction. Quantifying this observation employing a least-squares exponential fit yields that for a fixed reduction in instability amplitude, the required modulation fraction varies as: $\phi_{mod} \propto exp^{-0.2\nu_{mod}}$. This effect saturates as the modulation frequency is increased, trending toward a minimum required modulation fraction of $\phi_{mod} \approx 0.1\phi_{rf}$. These findings are concurrent with the hypothesis that asymmetrically attenuated ion acoustic waves are responsible for a reduction in the radial ion density gradient, via Equation (4), and subsequently also the amplitude of the instability.

Employing a ν_{mod} = 12 kHz, ϕ_{mod} = 0.11 ϕ_{rf} voltage amplitude modulation, the maximum amplitude of the instability was reduced by up to 65%. The central frequency of the instability (not shown) remained approximately independent of the applied modulation frequency, varying by at most 13% (1.2 kHz) over the range 0.0 $\leq \phi_{mod} \leq$ 0.74 for the ν_{mod} = 2 kHz modulation case. It should be noted that these trends are representative of the behavior of the instability obtained from a single spatial location (R, Z = –6 cm, 2 cm) downstream of the DL. However, as the induced ion density perturbations are not limited to any particular region of the source or expansion region, reduction of the instability amplitude is expected to be observed over the full downstream region.

Figure 10 shows the radially resolved maximum instability amplitudes as measured Z = 2 cm downstream of the DL between −10 cm $\leq R \leq$ 10 cm, for a 300 W, 13.56 MHz, $I_{1,2}$ = 6 A, 0.42 mTorr argon discharge employing a ν_{mod} = 12 kHz voltage amplitude modulation where $\phi_{mod} = 0.0\phi_{rf}$ (i.e. no modulation) and $\phi_{mod} = 0.11\phi_{rf}$.

In the absence of an applied voltage amplitude modulation the spatial distribution of the instability in **Figure 10** agrees with that previously measured in **Figure 3A**. The application of a ν_{mod} = 12 kHz, ϕ_{mod} = 0.11 ϕ_{rf} voltage amplitude modulation results in reduced instability amplitudes at all radii. The effect is most pronounced at the beam-conic interfaces (R = −5 cm, R = +6 cm), reducing on-axis and toward the downstream chamber walls. Greater inhibition of the instability adjacent to the beam-conic interface suggests a spatially varying modification of the radial ion density gradient arising from the voltage amplitude modulation. While this may arise from a number of effects, it is consistent with an increasing δn_i contribution arising from increasingly asymmetric ion acoustic waves attenuated through a radially varying ion density gradient. The trends presented in **Figures 7**, **9**, **10**, supported by the mathematical framework introduced in section 4.2, demonstrate voltage amplitude modulation as an effective

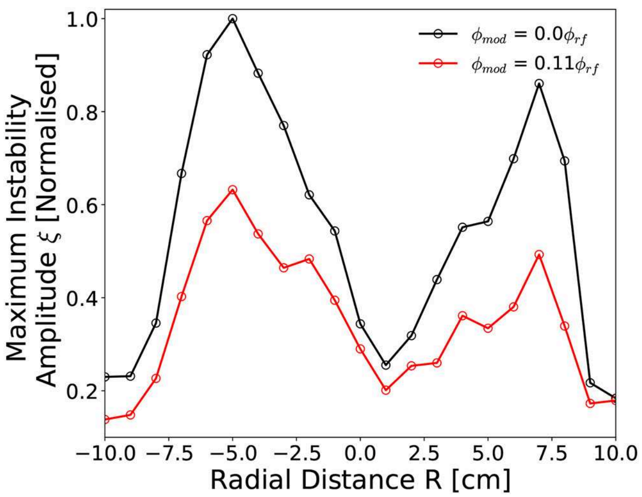

FIGURE 10 | Radially resolved (-10 cm $\leq R \leq$ 10 cm) instability amplitude downstream of DL ($Z = 2$ cm) employing a with a variable 12 kHz voltage amplitude modulation fraction. Solid lines added to guide the eye. Operating conditions: helicon antenna supplied $P_{rf} = 300$ W at $\nu_{rf} = 13.56$ MHz employing 0.42 mTorr argon with $I_{1,2} = 6$ A solenoid current.

technique for the reduction of ion density driven instabilities in magnetized plasmas.

5. CONCLUSIONS

A kHz instability in the ion saturation current downstream of a current free double layer has been spatially characterized with respect to the operating pressure, applied rf power and on-axis magnetic field strength. The amplitude of the instability was found to vary in proportion to the radial ion density gradient across the ion beam-conic interface. Reduction of the instability was achieved through the application of a kHz frequency voltage amplitude modulation ν_{mod} to the "carrier" 13.56 MHz voltage waveform ϕ_{rf}. Two amplitude modulation regimes were identified: a "low-frequency" regime (1 Hz $\leq \nu_{mod} \leq$ 2 kHz), characterized by a temporal "smearing" effect, where the instability is primarily influenced through variations in the time-averaged stationary ion density and a 'high-frequency' regime ($\nu_{mod} \geq$ 2 kHz) where the instability is primarily influenced via induced ion acoustic wave density perturbations. Employing a ν_{mod} = 12 kHz, ϕ_{mod} = 0.11 ϕ_{rf} voltage amplitude modulation, the amplitude of the instability was reduced by up to 65%. Measurements of the ion energy distribution downstream of the double layer indicated an increased proportion of high energy ions within the ion beam, coinciding with a reduction in the instability amplitude. The application of voltage amplitude modulation has the potential to be employed as an effective technique for the reduction of ion density driven instabilities in magnetized plasmas.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

SD: measurement of data, analysis of data, and preparation of manuscript. AB: lab assistance, preparation of manuscript introduction, and editing. DT and RB: lab assistance and editing. JD: editing. CC: project coordinator, lab facilities host, lab assistance, and editing.

FUNDING

This work presented herein was carried out within the SP3 lab at the Australian National University and funded by the Engineering and Physical Sciences Research Council (EPSRC), grant reference number: EP/m508196/1.

ACKNOWLEDGMENTS

The authors wish to thank Peter Alexander for technical support and gratefully acknowledge support from the Norma Ann Christie Scholarship in Low Temperature Plasma Physics.

REFERENCES

1. Adamovich I, Baalrud SD, Bogaerts A, Bruggeman PJ, Cappelli M, Colombo V, et al. The 2017 Plasma Roadmap: low temperature plasma science and technology. *J Phys D Appl Phys*. (2017) **50**:323001. doi: 10.1088/1361-6463/aa76f5
2. Mazouffre S. Electric propulsion for satellites and spacecraft: established technologies and novel approaches. *Plasma Source Sci Technol*. (2016) **25**:033002. doi: 10.1088/0963-0252/25/3/033002
3. Charles C. Grand challenges in low-temperature plasma physics. *Front Phys*. (2014) **2**:39. doi: 10.3389/fphy.2014.00039
4. Charles C. Plasmas for spacecraft propulsion. *J Phys D Appl Phys*. (2009) **42**:163001. doi: 10.1088/0022-3727/42/16/163001
5. Lafleur T, Takahashi K, Charles C, Boswell RW. Direct thrust measurements and modelling of a radio-frequency expanding plasma thruster. *Phys Plasmas*. (2011) **18**:1–4. doi: 10.1063/1.3610570
6. West MD, Charles C, Boswell RW. Testing a helicon double layer thruster immersed in a space-simulation chamber. *J Propul Power*. (2008) **24**:134–41. doi: 10.2514/1.31414
7. Charles C, Boswell RW. Effect of exhaust magnetic field in a helicon double-layer thruster operating in Xenon. *IEEE Trans Plasma Sci*. (2008) **36**:2141–6. doi: 10.1109/TPS.2008.2004233
8. Charles C, Boswell RW. Laboratory evidence of supersonic ion beam generated by a current-free "helicon" double-layer. *Phys Plasmas*. (2004) **11**:1706–14. doi: 10.1063/1.1652058
9. Charles C, Boswell R. Current-free double-layer formation in a high-density helicon discharge. *Appl Phys Lett*. (2003) **82**:1356–8. doi: 10.1063/1.1557319
10. Zhang Y, Charles C, Boswell R. Effect of radial plasma transport at the magnetic throat on axial ion beam formation. *Phys Plasmas*. (2016) **23**:083515. doi: 10.1063/1.4960828

11. Lafleur T, Charles C, Boswell RW. Ion beam formation in a very low magnetic field expanding helicon discharge. *Phys Plasmas*. (2010) **17**:1–6. doi: 10.1063/1.3381093
12. Byhring HS, Charles C, Fredriksen A, Boswell RW. Double layer in an expanding plasma: simultaneous upstream and downstream measurements. *Phys Plasmas*. (2008) **15**:102113. doi: 10.1063/1.3002396
13. Plihon N, Chabert P, Corr CS. Experimental investigation of double layers in expanding plasmas. *Phys Plasmas*. (2007) **14**:013506. doi: 10.1063/1.2424429
14. Sun X, Keesee AM, Biloiu C, Scime EE, Meige A, Charles C, et al. Observations of ion-beam formation in a current-free double layer. *Phys Rev Lett*. (2005) **95**:025004. doi: 10.1103/PhysRevLett.95.025004
15. Cox W, Charles C, Boswell RW, Hawkins R. Spatial retarding field energy analyzer measurements downstream of a helicon double layer plasma. *Appl Phys Lett*. (2008) **93**:2006–9. doi: 10.1063/1.2965866
16. Ahedo E, Merino M. Two-dimensional plasma expansion in a magnetic nozzle : separation due to electron inertia. *Phys Plasmas*. (2012) **19**:083501. doi: 10.1063/1.4739791
17. Charles C. High density conics in a magnetically expanding helicon plasma. *Appl Phys Lett*. (2010) **96**:051502. doi: 10.1063/1.3309668
18. Saha SK, Chowdhury S, Janaki MS, Ghosh A, Hui AK, Raychaudhuri S. Plasma density accumulation on a conical surface for diffusion along a diverging magnetic field. *Phys Plasmas*. (2014) **21**:043502. doi: 10.1063/1.4870758
19. Takahashi K, Akahoshi H, Charles C, Boswell RW, Ando A. High temperature electrons exhausted from rf plasma sources along a magnetic nozzle a magnetic nozzle. *Phys Plasmas*. (2017) **24**:084503. doi: 10.1063/1.4990110
20. Gulbrandsen N, Fredriksen Å. RFEA measurements of high-energy electrons in a helicon plasma device with expanding magnetic field. *Front Phys*. (2017) **5**:2. doi: 10.3389/fphy.2017.00002
21. Bennet A, Charles C, Boswell R. *In situ* electrostatic characterisation of ion beams in the region of ion acceleration. *Phys Plasmas*. (2018) **25**:023516. doi: 10.1063/1.5017049
22. Bennet A, Charles C, Boswell R. Separating the location of geometric and magnetic expansions in low-pressure expanding plasmas. *Plasma Sources Sci Technol*. (2018) **27**:075003. doi: 10.1088/1361-6595/aacd6d
23. Thakur SC, Harvey Z, Biloiu IA, Hansen A, Hardin RA, Przybysz WS, et al. Increased upstream ionization due to formation of a double layer. *Phys Rev Lett*. (2009) **102**:1–4. doi: 10.1103/PhysRevLett.102.035004
24. Thakur SC, Hansen A, Scime EE. Threshold for formation of a stable double layer in an expanding helicon plasma. *Plasma Sources Sci Technol*. (2010) **19**:025008. doi: 10.1088/0963-0252/19/2/025008
25. Allan W, Sanderson JJ. Temperature gradient drive ion acoustic instability. *Plasma Phys*. (1974) **16**:753. doi: 10.1088/0032-1028/16/8/005
26. Priest ER, Sanderson JJ. Ion acoustic instability in collisionless shocks. *Plasma Phys*. (1972) **14**:951. doi: 10.1088/0032-1028/14/10/005
27. Jassby DL. Transverse velocity shear instabilities within a magnetically confined plasma. *Phys Fluids*. (1972) **15**:1590–604. doi: 10.1063/1.1694135
28. Keen BE, Fletcher WHW. Suppression and enhancement of an ion-sound instability by nonlinear resonance effects in a plasma. *Phys Rev Lett*. (1969) **23**:760–3. doi: 10.1103/PhysRevLett.23.760
29. Thakur SC, Brandt C, Cui L, Gosselin JJ, Light AD, Tynan GR. Multi-instability plasma dynamics during the route to fully developed turbulence in a helicon plasma. *Plasma Sources Sci Technol*. (2014) **23**:044006. doi: 10.1088/0963-0252/23/4/044006
30. Yamada T, Itoh SI, Maruta T, Kasuya N, Nagashima Y, Shinohara S, et al. Anatomy of plasma turbulence. *Nat Phys*. (2008) **4**:721–5. doi: 10.1038/nphys1029
31. Burin MJ, Tynan GR, Antar GY, Crocker NA, Holland C. On the transition to drift turbulence in a magnetized plasma column. *Phys Plasmas*. (2005) **12**:1–14. doi: 10.1063/1.1889443
32. Schröder C, Grulke O, Klinger T, Naulin V. Spatial mode structures of electrostatic drift waves in a collisional cylindrical helicon plasma. *Phys Plasmas*. (2004) **11**:4249–53. doi: 10.1063/1.1779225
33. Schröder C, Grulke O, Klinger T, Naulin V. Drift waves in a high-density cylindrical helicon discharge. *Phys Plasmas*. (2005) **12**:1–6. doi: 10.1063/1.1864076
34. Light M, Chen FF, Colestock PL. Low frequency electrostatic instability in a helicon plasma. *Phys Plasmas*. (2001) **8**:4675–89. doi: 10.1063/1.1403415
35. Croes V, Lafleur T, Bonaventura Z, Bourdon A, Chabert P. 2D particle-in-cell simulations of the electron drift instability and associated anomalous electron transport in Hall-effect thrusters. *Plasma Sources Sci Technol*. (2017) **26**:034001. doi: 10.1088/1361-6595/aa550f
36. McDonald MS, Gallimore AD. Rotating spoke instabilities in hall thrusters. *IEEE Trans Plasma Sci*. (2011) **39**:2952–3. doi: 10.1109/TPS.2011.2161343
37. Zhurin VV, Kaufman HR, Robinson RS. Physics of closed drift thrusters. *Plasma Sources Sci Technol*. (1999) **8**:R1–20. doi: 10.1088/0963-0252/8/1/021
38. Aanesland A, Charles C, Lieberman MA, Boswell RW. Upstream ionization instability associated with a current-free double layer. *Phys Rev Lett*. (2006) **97**:1–4. doi: 10.1103/PhysRevLett.97.075003
39. Aanesland A, Lieberman MA, Charles C, Boswell RW. Experiments and theory of an upstream ionization instability excited by an accelerated electron beam through a current-free double layer. *Phys Plasmas*. (2006) **13**:122101. doi: 10.1063/1.2398929
40. Takahashi K, Charles C, Boswell R, Hatakeyama R. Radial characterization of the electron energy distribution in a helicon source terminated by a double layer. *Phys Plasmas*. (2008) **15**:074505. doi: 10.1063/1.2959137
41. Thakur SC, Xu M, Manz P, Fedorczak N, Holland C, Tynan GR. Suppression of drift wave turbulence and zonal flow formation by changing axial boundary conditions in a cylindrical magnetized plasma device. *Phys Plasmas*. (2013) **20**:012304. doi: 10.1063/1.4775775
42. Vaezi P, Holland C, Thakur SC, Tynan GR. Understanding the impact of insulating and conducting endplate boundary conditions on turbulence in CSDX through nonlocal simulations. *Phys Plasmas*. (2017) **24**:042306. doi: 10.1063/1.4980843
43. Lieberman MA, Lichtenberg AJ. *Principles of Plasma Discharges and Materials Processing*. 2nd ed. Hoboken, NJ: John Wiley & Sons (2005).
44. Charles C, Boswell RW, Cox W, Laine R, MacLellan P. Magnetic steering of a helicon double layer thruster. *Appl Phys Lett*. (2008) **93**:10–3. doi: 10.1063/1.3033201
45. Ceglio NM, Lidsky LM. Ion acoustic wave propagation near the ion cyclotron frequency. *Phys Fluids*. (1970) **13**:1108. doi: 10.1063/1.1693018

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Physics

ORIGINAL RESEARCH
published: 19 February 2020
doi: 10.3389/fphy.2020.00019

Numerical Performance Analysis of Terahertz Spectroscopy Using an Ultra-Sensitive Resonance-Based Sensor

Sajad Niknam [1], Mehran Yazdi [2], Salman Behboudi Amlashi [3]*† and Mohsen Khalily [3]†*

[1] Department of Computer Science and Engineering and Information Technology, Shiraz University, Shiraz, Iran, [2] Department of Communications and Electronics Engineering Technology, Shiraz University, Shiraz, Iran, [3] Institute for Communication Systems, Home of 5G Innovation Centre, University of Surrey, Guildford, United Kingdom

OPEN ACCESS

Edited by:
Stephen Watts,
University of Manchester,
United Kingdom

Reviewed by:
Sally Seidel,
University of New Mexico,
United States
Peter R. Hobson,
Queen Mary University of London,
United Kingdom

****Correspondence:***
Mehran Yazdi
yazdi@shirazu.ac.ir
Salman Behboudi Amlashi
s.behboudiamlashi@surrey.ac.uk

† These authors share senior authorship

Specialty section:
This article was submitted to Radiation Detectors and Imaging, a section of the journal Frontiers in Physics

Received: *27 July 2019*
Accepted: *23 January 2020*
Published: *19 February 2020*

Citation:
Niknam S, Yazdi M, Behboudi Amlashi S and Khalily M (2020) Numerical Performance Analysis of Terahertz Spectroscopy Using an Ultra-Sensitive Resonance-Based Sensor. Front. Phys. 8:19. doi: 10.3389/fphy.2020.00019

A terahertz sensor structure is proposed that can sense any variations in analyte permittivity. The sensor essentially works according to the shifts in the resonance frequencies of its propagated spoof surface plasmonic modes. The proposed structure shows great support for surface plasmon oscillations, which is proved by the calculated dispersion diagram. To achieve this in terahertz frequencies, a metamaterial structure is presented in the form of a structure with two-dimensional periodic elements. Afterward, it is shown that the performance of the sensor can be affected by different parameters such as metal stripe thickness, length of metal stripe, and width of metal stripe as the most influential parameters. Each of the parameters mentioned can directly influence on the electric field confinement in the metal structure as well as the strength of propagation modes. Therefore, two propagation modes are compared, and the stronger mode is chosen for sensing purposes. The primary results proved that the quality factors of the resonances are substantially dependent on certain physical parameters. To illustrate this, a numerical parametric sweep on the thickness of the metal stripe is performed, and the output shows that only for some specific dimensions the electromagnetic local field binds strongly with the metal part. In a similar way, a sweeping analysis is run to reveal the outcome of the variation in analyte permittivity. In this section, the sensor demonstrates an average sensitivity value, ∼1,550 GHz/Permittivity unit, for a permittivity range between 1 and 2.2, which includes the permittivity of many biological tissues in the terahertz spectrum. Following this, an analysis is presented, in the form of two contour plots, for two electrical parameters, maximum electric field and maximum surface current, based on 24 different paired values of metal thickness and metal width as the two most critical physical parameters. Using the plotted contour diagrams, which are estimated using the bi-harmonic fitting function, the best physical dimension for the maximum capability of the proposed sensor is achieved. As mentioned previously, the proposed sensor can be applied for biological sensing due to the simplicity of its fabrication and its performance.

Keywords: terahertz spectroscopy, terahertz sensing, Terahertz resonance sensor, metamaterial, spoof surface plasmon, surface wave

INTRODUCTION

Terahertz science has shown great potential in different applications such as security, medical imaging, and communications [1–6]. One of the main reasons for such interest in this field is the exclusive interaction of terahertz waves with some specific molecules [7–9]. Additionally, due to the low photon energy of terahertz radiation, it is a good candidate for non-destructive test (NDT) applications. Based on these interesting features of terahertz wave, many types of research have been done in the field of terahertz spectroscopy [10–15]. Due to the special and safe interaction of THz wave with biomolecules, THz spectroscopy for biomedical applications has become a rapidly evolving research area [16, 17]. Therefore, many researchers have presented various spectroscopy techniques in the THz regime. Terahertz spectroscopy methods are based on both time domain and frequency domain data acquisition [18]. Resonance Terahertz sensors work according to the high-quality resonance frequencies of a structure that occur in response to an incident THz wave. The Q-factor is one of the parameters that can be used for the evaluation of a resonance sensor. Terahertz resonance sensors can distinguish any changes in the refractive index of an analyte that is put on the surface of the sensor. However, the minimum detectable amount of variation in the refractive index of an analyte depends on the sensitivity parameters of the structure. Accordingly, having sharp resonances in terahertz frequencies can be regarded as a merit for a sensor structure. However, in THz frequencies, the loss of such structures is significant. To address this issue, different metal structures have been proposed to enhance the field confinement in the THz spectrum [19–21]. By increasing the confinement of the electromagnetic field in the structure, sharper resonances occur. One of the best solutions for this problem is using surface plasmon polaritons (SPPs) for better local electric field confinement. Although, in a normal state, sensing with SPPs is not possible in the THz spectrum because of the higher plasmonic frequencies of metals, which reach at least to the visible frequencies. However, some studies have proved that some specific forms of metal structures can bind the electromagnetic field in their surface to act like surface plasmon waves in higher frequencies. This confined surface wave that mimics the SPPs is called a Spoof Surface Plasmon (SSP) and can be generated in THz frequency range [22–25]. This type of THz sensor, which supports SSPs, employs corrugated metal structures in a two-dimensional metamaterial lattice. The metal structures can support surface waves in the THz spectrum using metamaterial techniques. The metamaterial structure can be manipulated to the extent that it gives the desired characteristics in a specific frequency. Thanks to the advantageous features of metamaterial techniques, THz sensors can be stimulated for very sharp resonances that can cause high Q-factor structures. Sensing with a THz resonance sensor can be achieved by the observation of any change in resonance frequency, phase, amplitude, etc. Among these parameters, interpreting the shifted resonances is a powerful tool for sensing purposes when a change occurs in the refractive index (Permittivity) of an analyte. Specifically, any variation in the Permittivity of the superimposed sample causes a redshift in the resonance frequency of the structure. In this paper, a metamaterial THz sensor is proposed and analyzed for its sensitivity parameters, and its ability to perform THz spectroscopy is then investigated. The results show that the structure can support the spoof surface wave in the THz frequency range. Additionally, the sensitivity of the structure is improved in comparison to previous works with lower manufacturing complexity. It is worth noting that the proposed structure is designed in a simple form that lessens the manufacturing difficulties substantially.

MATERIALS AND METHODS

As can be seen in **Figure 1**, the metamaterial structure consists of 2-D periodic metallic elements with periodicity T_x and T_y in the two directions of x and y. To measure the resonance frequencies of the structure, a transmission diagram should be calculated. Hence, a plane wave is illuminated from above the structure, the analyte is laid over the sensor, and afterward, the transmitted wave should be compared with the illuminated plane wave. Moreover, a substrate of polyimide is placed beneath the metal part. In terahertz frequencies, the polyimide has a very low absorption coefficient and, as a result, it can be assumed to be transparent for THz radiation. Besides, it has good mechanical properties as a substrate material for such structures. A dispersion diagram analysis has been run on CST Studio Suite 2018 to investigate the relationship between frequency (f) and wave number (K) [26]. To plot the dispersion curve, the boundary condition in the x and y directions is considered as periodic, and the electric boundary condition ($Et = 0$) is applied on the z-direction. The phase variation in the x-direction is then swept from 0 to 180 degrees, based on the Brillouin zone definition of a 2D-periodic structure. The output of the CST simulation gives the relation between frequency and phase, which is converted to an f-k dispersion diagram. The asymptotic behavior of the dispersion diagram proofs the support of the surface wave in the proposed metallic structure. Following this, the transmission curves are obtained from the response of the structure to an illuminated electromagnetic plane wave while the excitation wave is considered as Floquet port mode with two essential propagation modes. As is depicted, the transmitted THz wave must be measured for any shift in its resonance frequencies. In **Figure 1C**, it is presented that the structure can support surface waves in different and tunable frequencies according to the size of the structure. The most important parameter of the unit cell, which can affect the resonance frequency of the surface plasmon, is $d_{y,}$ the length of the metal stripe. The value of d_y is equal to 26 μm through this work, but, for different values, there are different resonance frequencies. In **Figure 1C**, it can be clearly seen that for $d_y = 20\,\mu m$, the resonance frequency is just above 3.5 THz, for $d_y = 26\,\mu m$, it is about 3 THz, and when $d_y = 30\,\mu m$, the resonance frequency is almost 2.2 THz. Actually, the frequency should be designed with respect to the application and the target analyte. Besides, the parameter d_x has a value of 10 μm as the best value for the width of the metal structure, which will be discussed later. Additional physical parameters are listed in

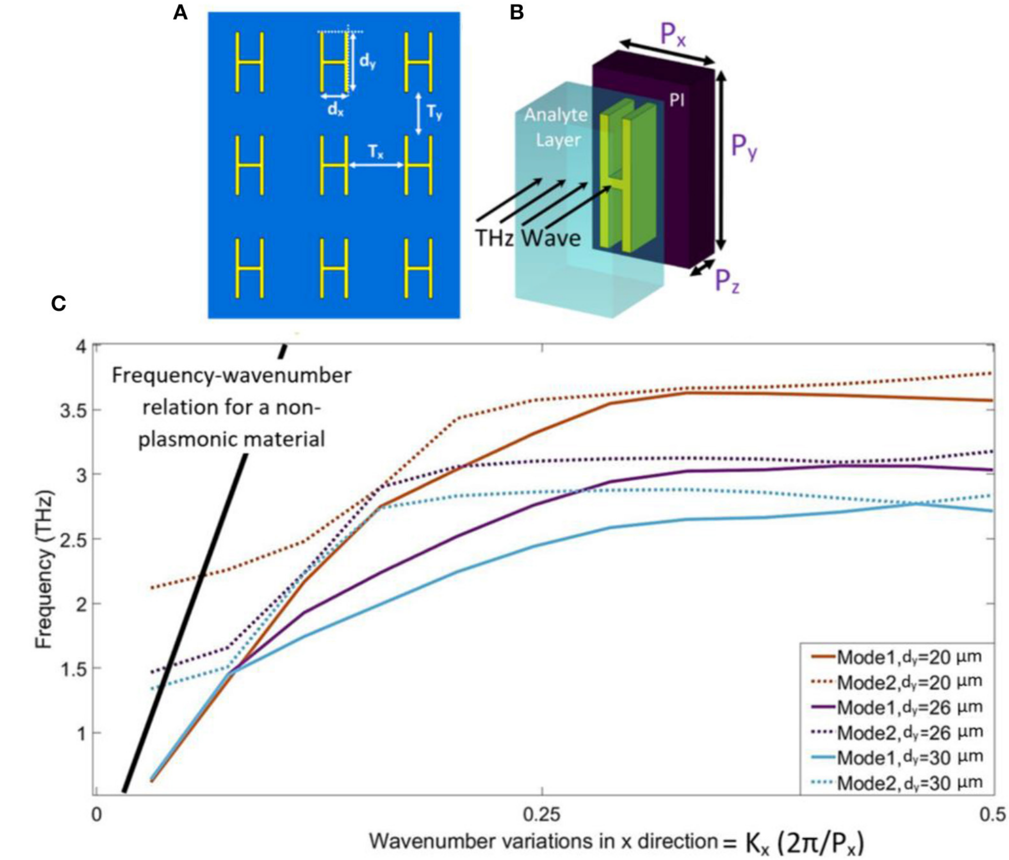

FIGURE 1 | The structure of the proposed THz sensor. A THz wave must be illuminated directly and transmits through the analyte and sensor structure. **(A)** Structure and the configuration of metal elements in the terahertz metamaterial structure. **(B)** Unit cell of the metamaterial periodic structure. A layer of polyimide is laid under the metal stripe. **(C)** Dispersion diagram of the structure. According to the dispersion relation between frequency and wave number, the graph illustrates that the surface plasmon is supported by the structure. There are two main propagation modes in this structure with different plasmon frequencies. The variation in the length of the metal (d_y) can tune the resonance frequency for the desired value. It can be clearly seen that for $d_y = 20\,\mu$m, the resonance frequency is just above 3.5 THz, for $d_y = 26\,\mu$m, it is about 3 THz, and when $d_y = 30\,\mu$m, the resonance frequency is almost 2.2 THz.

Table 1. Parameters such as P_x, P_y, and P_z are the dimensions of the substrate of the unit cell and have been defined as equal to the periodicity of the unit cells in such a way that no gap exists between the unit cells. It should be added that the parameters of P_x and P_y are optimized for the best characteristics of the metamaterial structure. In this paper, gold is considered as the material for the metal stripe because of its high conductivity in this range of frequency.

TABLE 1 | The optimized physical dimensions of the proposed structure and the periodicity of the periodic elements.

Parameter	Value (μm)
T_x	22
T_y	36
P_x	22
P_y	36
P_z	10
Metal stripe thickness	2

RESULTS AND DISCUSSION

The structure, as described in the previous section, can support spoof surface plasmon oscillations, which can produce strong confinement for the local electric field on the boundary of the metal-dielectric intersection. In this way, SSP modes can oscillate in a subwavelength scale in the THz frequency range, similar to what occurs at visible light frequencies. Based on these oscillations, multiple resonance frequencies are activated, which are very sensitive to any change in the permittivity of the analyte above the metal structure. It should be noted that, in this paper, only the redshifts due to change in the dielectric constant of the analyte are considered, but there may be other alterations such as the amplitude, phase, etc. of the resonance frequencies.

The main resonance frequencies can be determined using the simulation of the unit cell. As depicted in **Figure 2**, multiple

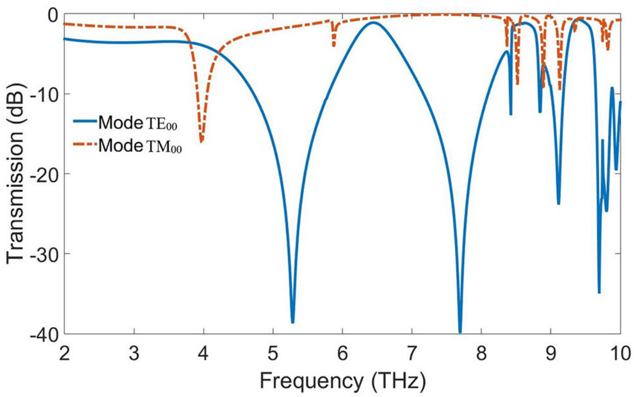

FIGURE 2 | Comparison between two dominant resonances for different propagation modes. Mode 1, which corresponds to the lower propagation mode, shows greater Q-factors than second resonance mode. Multiple resonances are activated for different propagation modes of the structure. One of the resonance frequencies corresponds to mode TM_{00}, and the other one denotes TE_{00}. This diagram depicts that the resonances for mode TE_{00} are more suitable for the sensing purpose thanks to the higher Q-factor of just above −40 dB.

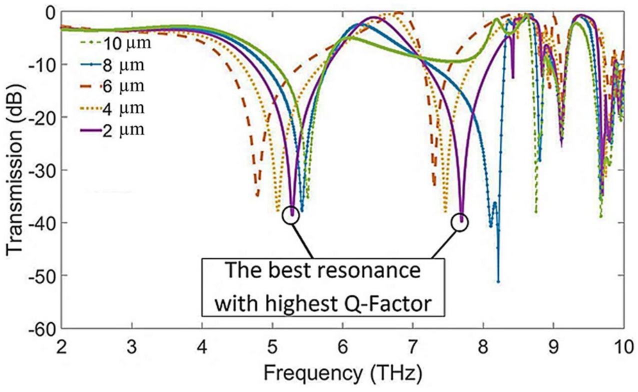

FIGURE 3 | Different resonances calculated by variation in the thickness of the metal structure. The plotted curves reveal the best value of the metal thickness for the highest Q-factor.

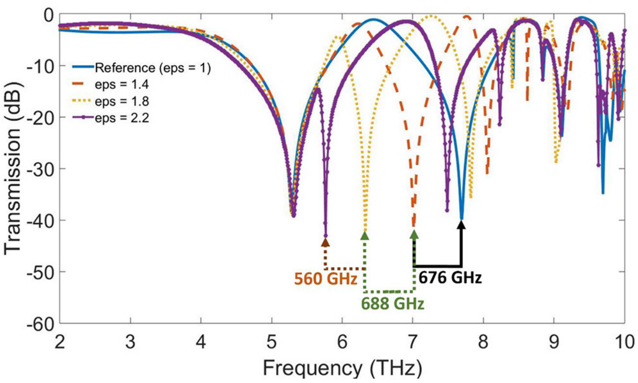

FIGURE 4 | Shifted resonances according to the change in the dielectric constant of the analyte. The range of investigated permittivities is between 1 and 2.2 for the analyte. The permittivity of $\varepsilon_r = 1$ is considered as the reference resonance curve. For $\varepsilon_r = 2.4$, the resonances reach their saturation state, and there are no further resonance frequency shifts for higher permittivity.

resonances are activated for different propagation modes of the structure. In **Figure 2**, one of the resonance frequencies corresponds to mode TM_{00} and the other denotes TE_{00}. As can be clearly seen, the resonances for mode TM_{00} show smaller Q-factors than the activated resonances for mode TE_{00}. The diagram illustrates that mode TE_{00} is more reliable for sensing purposes because of its higher Q-factor and sharper dip. Both of the resonances of TE_{00} mode have a perfect Fabry-Perot resonance form, which is very applicable for practical spectroscopy. Looking at mode TE_{00}, there are two strongly activated resonances with Q-factors of 110 and 296 at frequencies 5.28 THz and 7.696 THz, respectively. The calculated Q-factors demonstrate very strong local plasmonic oscillations, which bring about strong sensing capabilities. In fact, such characteristics ensure that the structure can be used for spectrometry and for sensing the variations in the permittivity of an analyte that is laid over the sensor structure. It is worth noting that the resonances could be changed due to different parameter values in the structure. The thickness of the metal stripe is one of the important parameters and can influence the sharpness of the resonance dips. For this purpose, the best thickness for the metal part is calculated and depicted in **Figure 3**. In the figure, it is obvious that for different thicknesses of the metal stripe, there are multiple resonances with different Q-factors as well as different frequencies. For the proposed structure, the thickness of 2 μm has the highest QF of the thicknesses considered, reaching 296 at 7.696 THz. Interestingly, for a metal stripe thickness of 8 μm, the second resonance of the TE_{00} mode has a much lower resonance value of transmission spectra, below −50 dB, than the second resonance of the metal stripe thickness of 2 μm. However, this resonance cannot be applied for sensing due to its partially Fabry-Perot form. Comparing the forms of these resonances, it can be seen that the transmission diagram for the thickness of 8 μm does not have a perfect Fabry-Perot resonance shape, and the simulation results proved that there would be no frequency shift for any change in the permittivity of the analyte. Hence, the resonances of the transmission diagram for the metal stripe with a thickness of 2 μm are considered to be the best achievable results.

The sensing abilities of the proposed THz sensor were investigated by carrying out a simulation study for an analyte with different permittivities with the frequency-domain solver of CST Suite Studio. In this calculation, an analyte layer is superimposed on the sensor structure. Afterward, the dielectric constant (ε_r = eps) of the analyte is swept between 1 and 2.4. Actually, this range of permittivity includes the permittivities of most biological tissues and is therefore very important for medical sensing applications [27–29]. **Figure 4** is plotted based on these calculations and is investigated for frequency shifts due to variations in the dielectric constant of the analyte. The main resonance curve is obtained for $\varepsilon_r = 1$ as the reference for other values of permittivity. Thus, the solid blue line is used as the reference resonance for the resonance shifts. It should be noted that each resonance curve has two Fabry-Perot form resonances, as previously pointed out, and that the second resonance is chosen for sensing purposes. As **Figure 4** shows, the first resonance does not have sensing capability because it does not experience any redshifts when there is a change in the permittivity of the analyte. On the other hand, the second resonance reacts properly by significant redshifts according to any variation in the dielectric constant of the analyte. By increasing the permittivity of the analyte, the redshifts approach a saturation state. Here, the saturation permittivity is defined as a value at which, for higher values, any change in the permittivity of the analyte cannot cause a shift in the resonance frequency. According to the simulation results, these redshifts reach the saturation state with eps = 2.4, which means that this is the maximum sensible permittivity for the dielectric constant of an analyte with this sensor. As can be calculated, the maximum achievable sensitivity for this sensor is 1.705 THz/Permittivity unit, which is obtained based

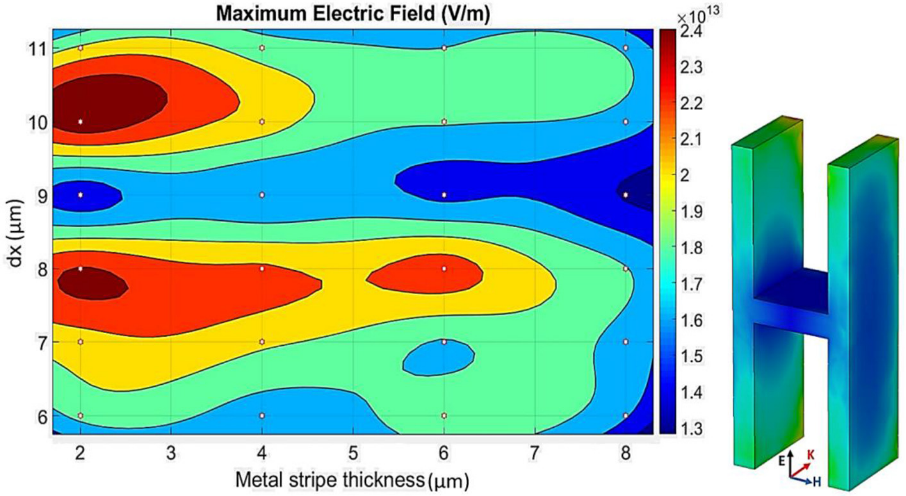

FIGURE 5 | Maximum electric field of the structure. This parameter is an indicator of the field confinement in the metal stripe. A fitting tool is applied to estimate the best physical dimensions for the highest possible electric field. Therefore, for 24 different combinations of metal thickness and d_x (red stars), the electric field is calculated, and then, using a biharmonic fitting function, a 2-D contour plot is generated. According to the plot, the highest electric field confinement can be achieved at metal stripe thickness= 2 µm and d_x = 8, 10 µm.

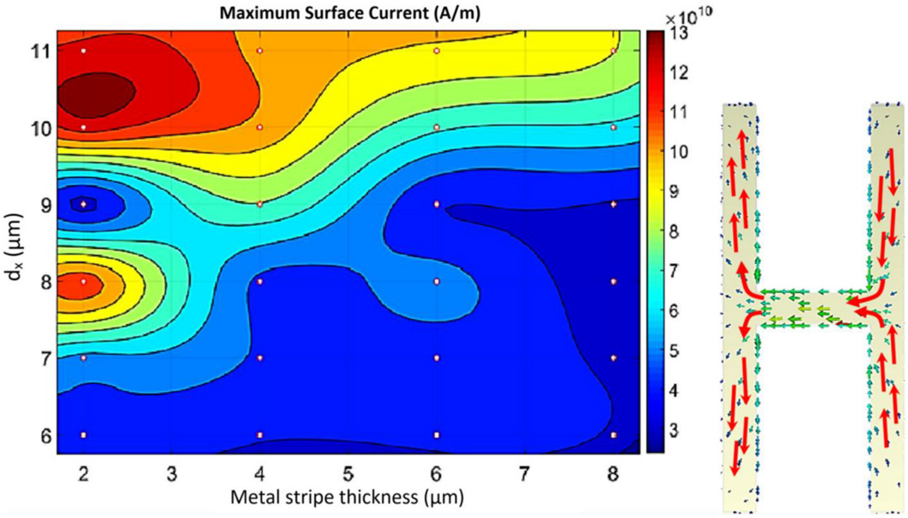

FIGURE 6 | Maximum surface current of the structure. In a similar way, a higher surface current can cause a sharper resonance dip. Consequently, investigating the best possible physical dimensions can increase the sensing capabilities. Similar to what has been done for the maximum electric field, 24 simulated data points are used for the estimation of all possible combinations. Thus, a biharmonic function is considered as a fitting tool to estimate the surface current for all combinations. As can be predicted, the same dimension values are obtained for the maximum surface current as have been calculated for the electric field.

on a 0.8-unit change in the permittivity of the analyte and corresponding 1.364 THz frequency shifts at the resonance frequency. The achieved sensitivity is very comparable with previous works [30, 31]. The results clearly prove that the proposed THz sensor can resonate in specified and tunable THz frequencies that can be applied for medical diagnostic applications. Many malignant biological tissues show variations in their permittivities in the range of frequencies investigated in this paper.

To take a closer look at the resonance quality of the proposed sensor, the electric field and the surface current of the structure are investigated numerically. As pointed out earlier, the length of the metal structure (d_y) can determine the resonance frequency of the sensor, but the results show that parameters such as the width of the metal structure (d_x) and metal stripe thickness can strongly affect the strength of the resonances. To make this clearer, numerical analyses of these two parameters have been done. In the first analysis, the maximum electric field of the structure has been calculated for different combinations of the two mentioned parameters. Accordingly, a 2-D contour plot of the maximum electric field has been estimated by a biharmonic interpolation function, as shown in **Figure 5**. The red stars in the figure are obtained from several simulations for the electric field of the structure and are used as inputs for the biharmonic surface fitting function. As the figure shows, it is possible to pinpoint the best values for the two physical parameters of d_x and metal stripe thickness. Based on the generated plot, a stripe thickness of 2 μm and $d_x = 8, 10\ \mu$m are the best choices for the proposed structure. For these values, the maximum electric field can have the best state, and therefore the structure can experience the maximum field confinement.

In a similar way, another contour plot is produced for the maximum surface current on the metal stripe. The surface current distribution on the structure is shown in **Figure 6**. The red arrows on the structure demonstrate that the surface currents on each side of the structure have opposite directions. Due to this, the resonances form sharper dips and have stronger sensing capabilities thanks to the dipole-dipole resonance mode. The contour plot in **Figure 6** can be easily used for better prediction of the best dimensional values of the structure. Very similar to the previous contour plot, the surface current contour plot also proves that a metal stripe with a thickness of 2 μm and $d_x = 8, 10\ \mu$m can have approximately the best electromagnetic field confinement and consequently better sensing strength. To summarize, the proposed terahertz sensor works as a resonance-based structure and, in this paper, is analyzed numerically in terms of physical dimensions and electric parameters. The designed structure benefits from a very simple designation, which reduces the complexity of fabrication in this range of wavelengths. Moreover, this sensor can be applied for medical sensing purposes for different biological tissues.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/supplementary material.

AUTHOR CONTRIBUTIONS

SN contributed to the concept of the work, analysis of results and simulations, and the draft of the manuscript. MY contributed to the concept of the work, supervising the analysis, and revising and finalizing the manuscript. SB contributed to the concept of the work, simulations and analysis of results, and revising the manuscript. MK contributed to the concept of the work and scientifically revising the manuscript.

REFERENCES

1. Amlashi SB, Araghi A, Dadashzadeh G. Design of a photoconductive antenna for pulsed-terahertz spectroscopy with polarization diversity. In: *2018 International Symposium on Networks, Computers and Communications (ISNCC)*. Rome: IEEE (2018), p. 1–5. doi: 10.1109/ISNCC.2018.8530985
2. Choudhury B, Menon A, Jha RM. Active terahertz metamaterial for biomedical applications. Singapore: Springer (2016), p. 1–41. doi: 10.1007/978-981-287-793-2_1
3. Niknam S, Yazdi M, Amlashi SB. Design of a pulsed-terahertz photoconductive antenna for spectroscopy applications. In: *2018 Fifth International Conference on Millimeter-Wave and Terahertz Technologies (MMWaTT)*. Tehran: IEEE (2018), p. 16–19. doi: 10.1109/MMWaTT.2018.8661244
4. Wang B-X, Zhai X, Wang G-Z, Huang W-Q, Wang L-L. A novel dual-band terahertz metamaterial absorber for a sensor application. *J Appl Phys.* (2015) **117**:014504. doi: 10.1063/1.4905261
5. Yang X, Zhao X, Yang K, Liu Y, Liu Y, Fu W, et al. Biomedical applications of terahertz spectroscopy and imaging. *Trends Biotechnol.* (2016) **34**:810–24. doi: 10.1016/J.TIBTECH.2016.04.008.
6. Zhou R, Wang C, Xu W, Xie L. Biological applications of terahertz technology based on nanomaterials and nanostructures. *Nanoscale.* (2019) **11**:3445–57. doi: 10.1039/C8NR08676A
7. Nagai N, Imai T, Fukasawa R, Kato K, Yamauchi K. Analysis of the intermolecular interaction of nanocomposites by THz spectroscopy. *Appl Phys Lett.* (2004) **85**:4010–2. doi: 10.1063/1.1811795
8. Zalden P, Song L, Wu X, Huang H, Ahr F, Mücke OD, et al. Molecular polarizability anisotropy of liquid water revealed by terahertz-induced transient orientation. *Nat Commun.* (2018) **9**:2142. doi: 10.1038/s41467-018-04481-5
9. Zhao Y, Li Z, Liu J, Hu C, Zhang H, Qin B, et al. Intermolecular vibrational modes and H-bond interactions in crystalline urea investigated by terahertz spectroscopy and theoretical calculation. *Spectrochim Acta Part A Mol Biomol Spectrosc.* (2018) **189**:528–34. doi: 10.1016/J.SAA.2017.08.041
10. Aytekin YS, Köktürk M, Esenturk O. Analysis of active pharmaceutical ingredients by terahertz spectroscopy. Dordrecht: Springer. (2017),. p. 69–73. doi: 10.1007/978-94-024-1093-8_10
11. Choi G, Hong SJ, Bahk Y-M. Graphene-assisted biosensing based on terahertz nanoslot antennas. *Sci Rep.* (2019) **9**:9749. doi: 10.1038/s41598-019-46095-x
12. Choi WJ, Cheng G, Huang Z, Zhang S, Norris TB, Kotov NA. Terahertz circular dichroism spectroscopy of biomaterials enabled by kirigami polarization modulators. *Nat Mater.* (2019) **18**:1. doi: 10.1038/s41563-019-0404-6
13. Ma Y, Huang H, Hao S, Qiu K, Gao H, Gao L, et al. Insights into the water status in hydrous minerals using terahertz time-domain spectroscopy. *Sci Rep.* (2019) **9**:9265. doi: 10.1038/s41598-019-45739-2
14. Ray S, Pesala B, Dash J, Devi N, Sasmal S. Understanding the effect of nanosilica incorporation on dicalcium silicate hydration using terahertz spectroscopy. In: Sadwick LP and Yang T, editors. *Terahertz, RF, Millimeter, and Submillimeter-Wave Technology and Applications XI*. San Francisco, CA: SPIE (2018), p. 53. doi: 10.1117/12.2289672

15. Zhang XC, Shkurinov A, Zhang Y. Extreme terahertz science. *Nat Photonics.* (2017) **11**:16–8. doi: 10.1038/nphoton.2016.249
16. Danciu M, Alexa-Stratulat T, Stefanescu C, Dodi G, Tamba BI, Mihai CT, et al. Terahertz spectroscopy and imaging: a cutting-edge method for diagnosing digestive cancers. *Materials.* (2019) **12**:1519. doi: 10.3390/ma12091519
17. Schmuttenmaer CA. Two decades of terahertz transient photoconductivity spectroscopy: where do we stand and where are we going? In: *2018 43rd International Conference on Infrared, Millimeter, and Terahertz Waves (IRMMW-THz).* Nagoya: IEEE (2018), p. 1. doi: 10.1109/IRMMW-THz.2018.8510028
18. Peiponen K-E, Zeitler A, Kuwata-Gonokami M. *Terahertz Spectroscopy and Imaging.* Berlin; Heidelberg: Springer Berlin Heidelberg (2013). doi: 10.1007/978-3-642-29564-5
19. Al-Naib I. Thin-film sensing via fano resonance excitation in symmetric terahertz metamaterials. *J Infrared Millimeter Terahertz Waves.* (2018) **39**:1–5. doi: 10.1007/s10762-017-0448-0
20. Niknam S, Yazdi M, Behboudi Amlashi S. Enhanced ultra-sensitive metamaterial resonance sensor based on double corrugated metal stripe for terahertz sensing. *Sci Rep.* (2019) **9**:7516. doi: 10.1038/s41598-019-44026-4
21. Pal BP, Chowdhury DR, Rao SJM, Islam M, Kumar G. Single split gap resonator based terahertz metamaterials for refractive index sensing. In Sadwick LP and Yang T, editors. *Terahertz, RF, Millimeter, and Submillimeter-Wave Technology and Applications XI,* SPIE (2018), p. 58. doi: 10.1117/12.2287320.
22. Cheng D, Zhang B, Liu G, Wang J, Luo Y. Terahertz ultrasensitive biosensing metamaterial and metasurface based on spoof surface plasmon polaritons. *Int J Numer Model Electron Netw Devices Fields.* (2018) e2529. doi: 10.1002/jnm.2529
23. Huang T-J, Liu J-Y, Yin L-Z, Han F-Y, Liu P-K. Superfocusing of terahertz wave through spoof surface plasmons. *Opt Express.* (2018) **26**:22722. doi: 10.1364/OE.26.022722
24. Unutmaz MA, Unlu M. Terahertz spoof surface plasmon polariton waveguides: a comprehensive model with experimental verification. *Sci Rep.* (2019) **9**:7616. doi: 10.1038/s41598-019-44029-1
25. Zhang Y, Xu Y, Tian C, Xu Q, Zhang X, Li Y, et al. Terahertz spoof surface-plasmon-polariton subwavelength waveguide. *Photonics Res.* (2018) **6**:18. doi: 10.1364/PRJ.6.000018
26. Davidson DB. *Computational Electromagnetics for RF and Microwave Engineering.* Cambridge: Cambridge University Press (2005). doi: 10.1017/CBO9780511611575
27. Labrou NE, Walker JM, Chen P, Levy DL, Good MC, Heald R, et al. Cell refractive index for cell biology and disease diagnosis: past, present and future. *Lab Chip.* (2018) 16:634–44. doi: 10.1039/C5LC01445J
28. Matcher SJ, Cope M, Delpy DT. Use of the water absorption spectrum to quantify tissue chromophore concentration changes in near-infrared spectroscopy. *Phys Med Biol.* (1994) **39**:177–96. doi: 10.1088/0031-9155/39/1/011
29. Pogue BW, Patterson MS. Review of tissue simulating phantoms for optical spectroscopy, imaging and dosimetry. *J Biomed Opt.* (2006) **11**:041102. doi: 10.1117/1.2335429
30. Chen X, Fan W. Ultrasensitive terahertz metamaterial sensor based on spoof surface plasmon. *Sci Rep.* (2017) **7**:2092. doi: 10.1038/s41598-017-01781-6
31. Singh R, Cao W, Al-Naib I, Cong L, Withayachumnankul W, Zhang W. Ultrasensitive terahertz sensing with high- *Q* Fano resonances in metasurfaces. *Appl Phys Lett.* (2014) **105**:171101. doi: 10.1063/1.4895595

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 21 February 2020
doi: 10.3389/fphy.2020.00034

An Inductively-Coupled Plasma Electrothermal Radiofrequency Thruster

Dimitrios Tsifakis, Christine Charles and Rod Boswell*

Space Plasma, Power and Propulsion Laboratory, Research School of Physics, The Australian National University, Canberra, ACT, Australia

The "cubesat" form factor has been adopted as the defacto standard for a cost effective and modular, nano-satellite platform. Many commercial options exist for nearly all components required to build such satellite; however, there is a limited range of thruster options that suit the power and size restrictions of a cubesat. Based on the prior work on the "Pocket Rocket" electro-thermal capacitively-coupled radiofrequency (RF) plasma thruster operating at 13.56 MHz, a new design is proposed which is based on an inductively-coupled radiofrequency plasma system operating at 40.68 MHz. The new thruster design, including a compact and efficient radiofrequency matching network adjacent to the plasma cavity, is presented, together with the first direct thrust measurements using argon as the propellant with the thruster immersed in vacuum and attached to a calibrated thrust balance. These initial results of the unoptimized first proof of concept indicate an up to 40% instantaneous thrust gain from the plasma compared to the cold gas thrust: typical total thrust at 100 SCCM of argon and 50 W RF power is ∼1.1 mN.

Keywords: cubesat propulsion, electrothermal thruster, inductively coupled plasma thruster, cubesat thruster, direct thrust measurement

OPEN ACCESS

Edited by:
Gianpiero Colonna,
Italian National Research Council, Italy

Reviewed by:
Andrea Cristofolini,
University of Bologna, Italy
Stephen Bernard Gabriel,
University of Southampton,
United Kingdom

****Correspondence:***
Dimitrios Tsifakis
dimitrios.tsifakis@anu.edu.au

Specialty section:
This article was submitted to
Plasma Physics,
a section of the journal
Frontiers in Physics

Received: *12 August 2019*
Accepted: *05 February 2020*
Published: *21 February 2020*

Citation:
Tsifakis D, Charles C and Boswell R (2020) An Inductively-Coupled Plasma Electrothermal Radiofrequency Thruster. Front. Phys. 8:34. doi: 10.3389/fphy.2020.00034

1. INTRODUCTION

Electric propulsion has been used by satellite operators for station keeping and orbit modification since the 1960's [1]. There are many thruster designs such as gridded ion thrusters and Hall effect thrusters with a proven track record. These systems are complex and are generally designed for larger satellites with the ability to carry large thruster and propellant mass and use high power in the order of kilowatts to achieve their orbit modification goals. In the recent years, there has been a disruption in satellite design by the creation of the cubesat form factor unit with $10 \times 10 \times 10$ cm^3 dimensions and up to 1.33 kg of mass [2]. At the time of writing, the online Nanosat database [3] reports over 1,300 nanosats and cubesats that have already been placed in orbit. The majority do not have a propulsion system. There have been many proposed propulsion system designs for this type of satellite [4–11] but most are laboratory designs and very little has been demonstrated or reported in "real" space missions. One reason for this is the many restrictions imposed on the thruster designer by the small satellite form factor with the most important ones being size (∼1/3 of the satellite volume and mass) and average power available to the payload (∼1 W).

The original Pocket Rocket was proposed in 2012 [12] as a thruster candidate that suits the requirements of a propulsion system of a small satellite. It is an asymmetric, capacitively-coupled, collisional (∼1 Torr) RF plasma device and falls under the electrothermal thruster category, together with resistojets and arcjets [13]. The inductive Pocket Rocket (PR) described in this paper is derived from the original capacitive PR, in an attempt to improve its performance and gain further insight into the gas heating and thrust mechanism. In a capacitively coupled plasma system, energy is transferred to the electrons by the electric component of the RF field while in an inductively coupled plasma system it is the magnetic component that serves this role. The switch from a capacitive to an inductive system has potential advantages in both performance by producing an increased ion density [14] and consequently increased propellant heating, and engineering by allowing a simpler, capacitor-only, RF matching network [15].

In order to accommodate the low average power available for the generation of RF, a thruster of the PR type is envisaged to be operated in a low duty cycle. An example of this type of operation is a Hohmann transfer consisting of two, 1-min burns per orbit. In the typical LEO 90-min period orbit, this allows enough time to recharge the batteries between thruster operations. Using this operation scheme and the performance obtained by the inductive PR proof of concept presented in this study, a 2 kg cubesat in a circular 400 km orbit will gain roughly 100 m of altitude per orbit using the thruster at 100 SCCM and 50 W of RF. This can be repeated over a number of orbits in order to achieve the desirable altitude. In terms of power requirement to operate the thruster, two 1-min burns at 50 W RF power need 1.67 Wh of energy. This energy needs to be collected from the solar panels in a period of 90 min (400 km orbit), implying a requirement of approximately 1.11 W of average power, probably closer to 1.5 W if RF amplifier and charging efficiencies are taken into account. This is within the capabilities of a 2-U satellite with extendable solar panels or larger cubesat.

The plasma coupling mode best suited to thrust generation in the capacitive PR is the Gamma mode [16] and not the very low plasma density alpha mode [17, 18]. In the Gamma mode, the electron density increases linearly with RF power and is a result of ion-induced secondary electron emission from the plasma cavity walls surrounded by the RF annulus electrode; in such capacitively coupled thruster, a self-bias generates as a result of the asymmetry between the small powered electrode and the large grounded walls/electrodes. The plasma thrust gain results from two terms, a bulk plasma propellant heating thrust term (resulting from ion-neutral charge exchange collisions) which immediately takes place at turn on (first 100 ms) and a wall propellant heating thrust term (resistojet effect) which increases with burn time as a result of ion bombardment of the radial plasma cavity walls. These two thrust terms have never been directly demonstrated and quantified experimentally. They have been indirectly deduced via gas temperature measurements [19]. Due to the small aperture area and large neutral density compared to the ion density, any thrust term from ion acceleration would remain small.

In terms of power coupling to the plasma, the inductive PR resembles the well documented low pressure (∼ *mTorr*) Helicon/inductive plasma sources and thrusters [20] which exhibit a capacitive coupling mode at low RF power and an inductive mode at higher power. For the mm size (diameter) and cm size (length) inductive PR it is necessary to operate at pressures around a few Torr to couple the plasma and the capacitive-inductive transition occurs at about 20 W. The present study focuses on the inductive mode obtained in the 20 W to 50 W power range. In terms of thrust generation, the main source of thrust expected from the inductive PR should be similar to that of the capacitive PR due to the plasma source scaling, operating pressure range and input power range or a few tens of watts. Basic thrust measurements have been previously performed for a larger inductively coupled thruster (i.e., 5 cm diameter) operating with input power ranges of a few hundred watts [21, 22] and at operating pressures of a few mTorr. It was shown that a typical power input of 100 W would result in about 1 mN of thrust, the maximum electron pressure in the plasma cavity (experimentally measured and also the result of a basic global model including particle balance and power balance, as described in chapter 10 of [14]) is converted into ion momentum along the expanding plasma as predicted by Fruchtman [23]. Fruchtman [24] also discussed the complexity of thrust imparted by low pressure and high pressure expanding plasma sources. For capacitive PR, the basic understanding of the plasma-generated thrust increase at 13.56 MHz from the cold gas thrust has been validated using CFD Ace+ [25] and HEMP fluid-plasma transient simulations [21]. In the present study, we demonstrate direct thrust measurement of a mm size inductively coupled plasma source using a vacuum compatible miniaturized impedance matching system at 40.68 MHz. Since this is simply a proof of concept, it is premature to directly compare with other technologies. Instead we provide additional references cases of a cold gas system and of a filament heated resistojet system and some comparative discussion with larger low pressure inductively coupled thruster.

2. PHYSICAL DESCRIPTION

2.1. First Experimental Configuration

Two experimental configurations are used to develop and characterize the inductive PR: a first configuration where the prototype is directly mounted onto a small size vacuum chamber ("Chi-Kung") and a second configuration where the prototype is immersed in a much larger vacuum chamber ("Wombat"). A photo of the inductive PR mounted onto the previously described Chi-Kung vacuum chamber [26] is shown in **Figure 1** and a simplified schematic is shown in **Figure 2**. The inductive PR consists of a plenum (a grounded cylindrical cavity into which gas is introduced) contiguously connected to a 5.5 mm outer diameter and 4.0 mm inner diameter ceramic (alumina) tube which forms the plasma cavity. The length of the alumina tube is approximately 6 cm. The ceramic tube is surrounded by a multiple turn RF loop antenna. The plenum (left exit of the plasma cavity on **Figure 1**) is made with aluminum and has a cylindrical shape with 25 mm diameter and 15 mm depth resulting in ∼7 cm^3 of volume. Viton o-rings are used to ensure a good vacuum seal between the ceramic and the aluminum. The o-rings are compressed with aluminum clamps which are tightened in place with screws. The exhaust end of the ceramic tube is connected via an appropriate aluminum flange to the 30 cm long 15 cm diameter pyrex tube of the Chi-Kung vacuum system, simulating the vacuum conditions of space.

Gas is injected to the plenum via 1/4″ flexible hose (**Figure 1**). For the Chi-Kung testing configuration, an MKS 626B Baratron pressure sensor is connected to the plenum via a short, rigid 1/4″ pipe and KF25 vacuum fittings in order to provide pressure measurements with the plasma "off" or "on," a parameter previously identified as a measure of neutral gas heating [16, 27]. The gas supply consists of a centrally located argon cylinder at high pressure, a ∼60 PSI regulator and an MKS Type 247 display and MKS Mass-Flo gas flow controller which allows an accurate selection of gas flow up to 140 SCCM. The plenum pressure varies with the cold gas flow rate from a fraction of a Torr at a few SCCM of gas flow to ∼3 Torr, at 140 SCCM. The Chi-Kung expansion chamber is pumped by a rotary vacuum pump and a secondary turbomolecular pump. The vacuum base pressure in the chamber is typically $\sim 8\times10^{-4}$ Torr when the gas supply is off and rises up to $\sim 10^{-2}$ Torr with the gas supply at the maximum setting of 140 SCCM. The pressure in the plenum and plasma cavity is greater by a factor of >100 to that in the vacuum chamber, ensuring the system is operating in choked flow conditions [25, 28]. This is checked by confirming that the plenum pressure is not affected by the downstream pressure.

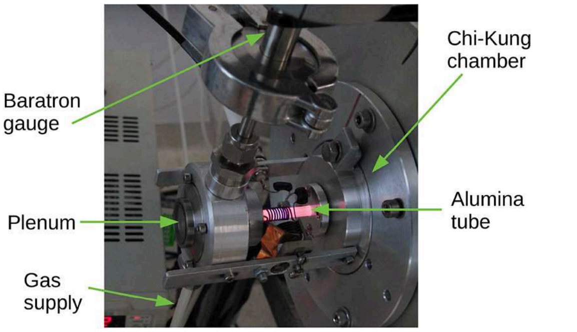

FIGURE 1 | Photo of the inductive PR (center left) connected to the Chi-Kung vacuum chamber (right). The baratron vacuum gauge is connected to the top of the plenum and the gas supply to the bottom. The inductive PR is operating with argon at 100 SCCM and 30 W of RF power at 40.68 MHz. Part of the RF matching network is also visible, in this instance consisting of silver mica capacitors and a variable capacitor which was used for fine tuning.

A 10-turn, 1 mm diameter, ∼15 mm long along the thruster main axis, copper wire inductor forming the RF antenna is wound on the ceramic tube as seen in **Figure 2**. It is not in contact with the plasma making this system an electrodeless thruster. The plenum end of the inductor is grounded and the exhaust end is connected to the RF matching circuit (described in detail in section RF Circuit Description). The matching circuit matches the inductive load formed by the antenna and the plasma system to 50 Ω to allow easy connection via any length 50 Ω coaxial cable to a laboratory grade wideband RF amplifier (Mini Circuits ZHL-100W-GAN+). The amplifier is driven by a 40.68 MHz sine wave signal from a Keysight 33600A arbitrary waveform generator which is used to control the output power and, when pulsed, the duty cycle of the RF supplied to the thruster (a feature not used in the present study). The RF circuit also includes a cross-needle Daiwa CN-801 SWR meter which is used to confirm power and matching quality. A high-flow cooling fan can be used to cool the inductor to ensure the Viton o-rings do not overheat when operated at high power and for long duration. With the cooling supplied from the external fan, the prototype can run continuously at 40 W reaching a maximum temperature near the center of the coil not exceeding 200°C. In these conditions, thermal equilibrium is achieved within a minute of RF being switched on. The inductive PR Chi-Kung setup can be used not only to test and optimize the RF circuitry but also the RF power transfer into the plasma by allowing plenum pressure measurements.

2.2. Second Experimental Configuration

The main aim of this study is to demonstrate thrust gain from gas heating during plasma operation with the newly designed 40.68 MHz compact impedance matching system. To this effect a second configuration is used where the inductive PR with matching network is attached to a thrust balance and fully immersed within the previously described, larger Wombat

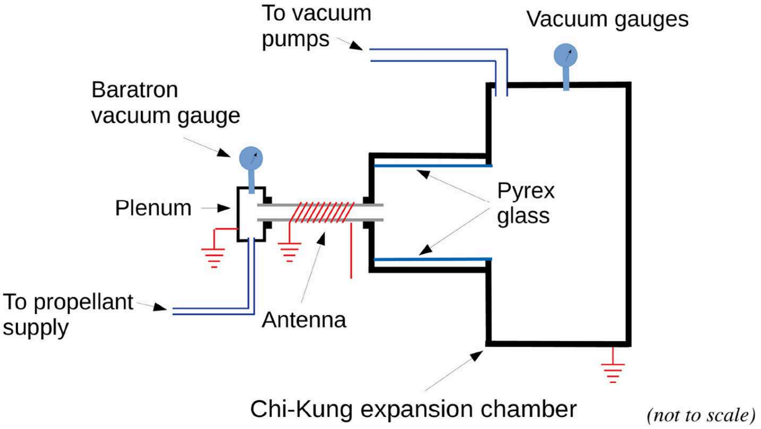

FIGURE 2 | A simplified sketch of the inductive PR, shown here connected to the Chi-Kung expansion chamber at lower pressure (as seen in **Figure 1**). The Pyrex glass inside the chamber is part of a different experiment also run on Chi-Kung and bears no significance in the experiments described in this paper.

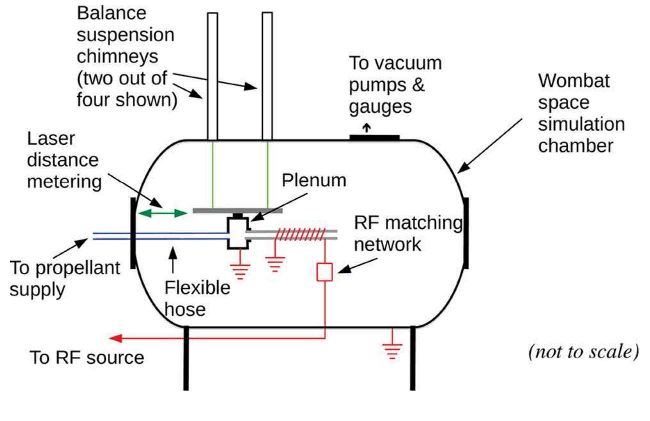

FIGURE 3 | A simplified sketch of the inductive PR, shown here installed in the Wombat vacuum chamber equipped with a thrust balance.

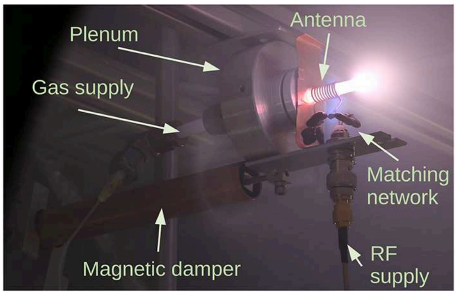

FIGURE 4 | Photo of the inductive PR mounted on the thrust balance and shown operating inside the Wombat chamber. The magnetic damper consisting of a copper tube attached to the chamber and a set of permanent magnets attached to the thruster can be seen below the thruster, to the left hand side. The RF matching network can also be seen, in this instance consisting of fixed value silver mica capacitors only.

vacuum chamber [15, 29]. Major components of this assembly are shown in **Figure 3**; a photo of the prototype operating in the Wombat chamber is shown in **Figure 4**. In this set up the RF antenna is located at a distance of ∼8 mm from the exhaust plane of the plasma cavity. The length of the ceramic tube is reduced to a length of ∼4 cm compared to the Chi-Kung setup. Similar measurements have been conducted in this chamber in the past for the capacitive PR mounted on the large, four-arm, thrust balance [29]. Here essential improvements to the thrust measurement procedure have been carried out as follows: the propellant is supplied via a flexible PTFE hose with an outer diameter of 1.5 mm and inner diameter of 0.8 mm, and RF power is supplied via a flexible RF cable (RG316). Both lines are attached to suitable vacuum feed through connectors to allow connection with the gas controller (Alicat scientific) and RF power supply (Oregon Physics VRG1000A), both located outside the chamber. Care has been taken to ensure both gas line and RF cable are mounted in a way that is not influencing the measurements. The lack of influence in the displacement from the gas line and RF cable was confirmed by the repeatability of the results in a large number of cold gas operations and by the observation that there is no displacement change when RF was switched on with no gas. When the thruster operates, either in cold gas or plasma mode, the balance moves in reaction to the applied force and the displacement D is measured with a previously described laser-sensor system [29]. This system is based on a laser triangulation displacement sensor (ILD1700) and has a resolution of 0.1 μm. The thrust to displacement calibration factor is 0.044 mN/μm and is obtained by using a calibration system consisting of a set of well known weights on a string connected via a pulley to the balance. The application of the weights is controlled by a stepper motor, installed in the chamber. Vacuum in the chamber is produced by three pumps: a Neovac SS120W roughing pump achieves $\sim 10^{-2}$ Torr with no gas flow and a large Varian V1800A turbomolecular pump with a pumping speed of 1600 l/s (N_2) improves the vacuum to $\sim 10^{-6}$ Torr. A cryopump is also available but was not used in these measurements.

To avoid putting the balance in an undamped oscillation every time there is a thrust change or external stimulus (mechanical vibration) and to quickly return to a baseline position, a custom-made magnetic damper is installed. The damper, seen in **Figure 4**, uses permanent magnets attached to the balance and inserted into a copper tube which is attached to the chamber. The eddy currents produced in the copper by the movement of the magnets result in a force opposing the movement and as a result have a damping effect. The time constant of this damping mechanism is in the order of a few seconds. This damping effect is further augmented by filtering done at the raw data level resulting in an improved sensitivity system. The thrust measurement campaign aims at assessing the cold gas thrust and the thrust gain when the plasma is turned on for varying gas flow rate and varying RF power. There is presently no cooling system and plasma runs (burns) are kept to a minute or less to avoid overheating damage to the Viton o-ring. No thermal drift of the balance was observed due to the short plasma burns.

2.3. RF Circuit Description

2.3.1. Frequency Selection of 40.68 MHz

The selection of frequency of operation for the inductive PR is based on two distinct requirements. The first one is the requirement to operate on a frequency that is not likely to cause interference to any other user of the radio spectrum. This requirement must be satisfied for all areas the satellite will be flying over, as determined by its orbit. The second one is to select a frequency which provides optimal performance of the thruster and is compatible with the cubesat architecture restrictions stated earlier in this paper.

To facilitate the operation of RF systems of Industrial, Scientific and Medical (ISM) nature (such as the cubesat thruster), the International Telecommunications Union (ITU) has determined a number of bands which are agreed by all member states to be used for such applications in a global scale. These bands are listed in the 2016 edition of the ITU Radio Regulations document (footnote 5.150) and are 13.56 MHz, 27.12 MHz and 40.68 MHz in the HF/VHF part of the spectrum, followed by 2.4 GHz, 5.8 GHz, and 24 GHz in the microwave part of the spectrum. There are other bands between VHF and microwaves allocated for similar applications by local administrations in many countries, however they do not enjoy the same global recognition as the ISM bands do. The technology to build suitable RF sources for the three lower ISM bands has been demonstrated and includes novel amplifier designs such as the Class-E [30]. Class-E amplifiers for the HF/VHF part of the spectrum have been constructed with 90% efficiency or better [31, 32]. While the inductive PR thruster could in theory be designed to be powered by a microwave source, the strict restrictions of the cubesat platform make this choice a more challenging one due to generally lower efficiency achievable on these frequencies. A typical efficiency value of high-efficiency RF sources for 2.4 GHz systems is ∼50% which indicates not only a higher input power requirement for a given RF output power, but also an exacerbated semiconductor heat dissipation problem.

In order to select one of the three lower ISM bands, it is important to explore the impedance characteristics of an inductively coupled plasma device like the inductive PR. The impedance Z at the antenna with the plasma ignited can be represented as

$$Z = (R_p + R_l) + j\omega L$$

where R_p is the resistance component and ωL is the inductive reactance of the antenna. The angular frequency, ω, relates to the RF frequency, f, according to the formula $\omega = 2\pi f$. The inductance, L, is mostly dependent on the antenna length, diameter and number of turns. The R_p component of the impedance is due to the plasma absorbing energy and is not dependent on frequency. The power deposited in the plasma is $P_p = I^2 R_p$ where I is the antenna current. For a given current (or plasma deposited power), the voltage across the antenna V_{ant} is determined by the reactance and not the resistance as, for the typical inductive PR operating conditions, $R \ll \omega L$ and is $V_{ant} \approx \omega LI$. The higher voltage offered by the higher frequency is an important advantage to consider in the thruster design as it facilitates a quicker and more predictable striking of the plasma. R_l represents the loss component and is mostly due to the skin effect influenced AC resistance of the copper wire. In the case of the inductive PR, $R_l \ll R_p$ for all three ISM bands. Typical values at room temperature for the highest frequency of 40.68 MHz and copper conductors are $R_l = \sim 0.1\ \Omega$ and $R_p = \sim 3.1\ \Omega$. R_l increases with frequency but $R_l \ll R_p$ still holds at the expected operating frequencies of the prototype. The losses can be further improved by using silver plated conductors instead of copper. Based on the above the frequency of operation chosen for the present development of the prototype is the ISM band of 40.68 MHz, a change from the 13.56 MHz previously used for the capacitive PR [12].

2.3.2. Impedance Matching Network

The plasma producing inductor appears as an impedance of $\sim 3.2 + j56\ \Omega$ (at 40.68 MHz), obtained by direct voltage and current measurements and calculations based on the matching network component values. Small variations in this impedance may be observed and are due to parasitic impedances in the space immediately around the inductor as well as the gas flow rate and copper temperature, with the last two affecting the real part of the impedance. There are multiple circuits that will match the ignited plasma impedance to 50 Ω (the output impedance of commercial RF sources used in the laboratory) consisting of inductors and capacitors. In general, inductors introduce added losses due to ohmic heating, can be affected by objects in their surrounding and take more space. While innovative solutions for manufacturing inductors have been proposed [33], a capacitor-only matching network is preferred if possible [15]. The two simplest capacitor-only networks are shown in **Figure 5**, together with the Smith chart solutions.

Out of these two matching circuits, the one using C1/C2 was selected because it results in lower value capacitors which are generally easier to obtain with very low or zero temperature coefficient (C0G/NP0 types). In the prototype C1 is implemented with a fixed capacitor network resulting in a total capacitance of ∼53 pF and C2 is a fixed ∼18 pF capacitor. Another advantage of the C1/C2 matching circuit over the C3/C4 one is avoiding the relatively high current that will flow on C4.

Before plasma ignition, the matching network does not provide a good match to 50 Ω due to the lack of the plasma resistance. In this case, circuit simulations show that there is a higher current I flowing on the inductor, which results in a higher voltage V across it. This is beneficial for the successful ignition of the plasma. If the RF power is decreased below the design range of 20 W to 50 W, the plasma switches to a capacitive mode. This is visually observed from the brightness of the plasma being asymmetric and much brighter at the "hot" (not grounded) end of the inductor. If the power is increased to the design range, the discharge becomes inductively coupled and the brightest spot moves to the center of the coil. The voltage across the inductor is $\sim 250\ V_{peak}$ for 40 W RF power.

It is convenient to have a 50 Ω system in the laboratory. However, this is not a requirement for the final design that will operate on a cubesat. In this case, the amplifier can be designed with an output impedance of 3.2 Ω and the matching network can be reduced to a single capacitor with enough capacitive reactance to tune out the $+j56\ \Omega$ inductive reactance of the load. The impedance was found not to vary significantly when the thruster is operated within its design range of power (20 W to 50 W) and argon gas flow (20 SCCM to 100 SCCM) At 40.68 MHz, this results in a ∼70 pF capacitor. Such direct impedance match has been successfully demonstrated for capacitive PR [31].

3. EXPERIMENTAL RESULTS

3.1. Direct Thrust Measurements: Cold Gas Thrust

In an electrothermal RF plasma thruster the total thrust (F_T) consists of two main components, the cold gas thrust (F_{cg}) and the plasma thrust (F_p):

$$F_T = F_{cg} + F_p$$

Figure 6 shows the raw displacement D (from the thrust balance laser-sensor system) over time in cold gas operation for increasing argon flow rate (0 SCCM to 200 SCCM in 20 SCCM incremental steps) and decreasing flow rate (200 SCCM to 0 SCCM in 50 SCCM incremental steps). Each step is ∼20 s long. The data sampling rate is 312.5 Hz with a moving average

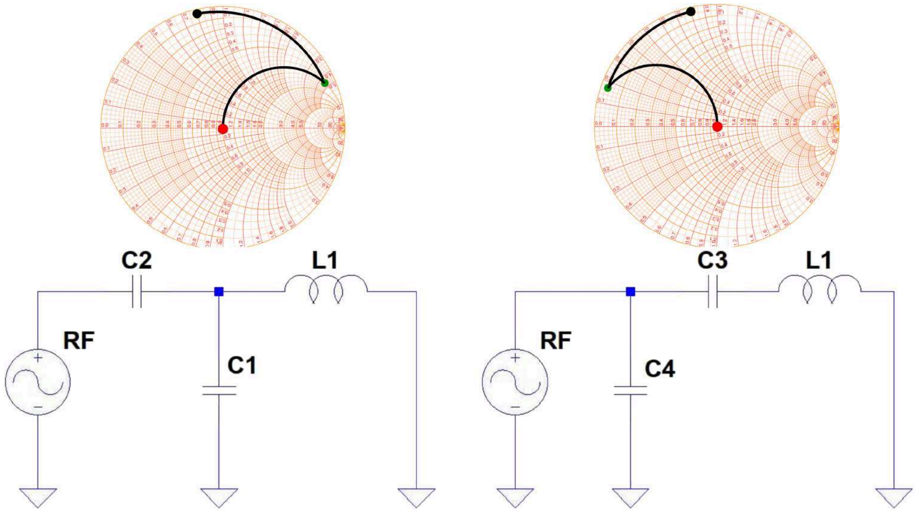

FIGURE 5 | Two capacitor-only matching networks that are suitable for the inductive PR impedance matching system. L1 indicates the plasma antenna, shown in **Figures 1**, **2**. The Smith charts on top of each circuit show the intermediate steps in the impedance transformation.

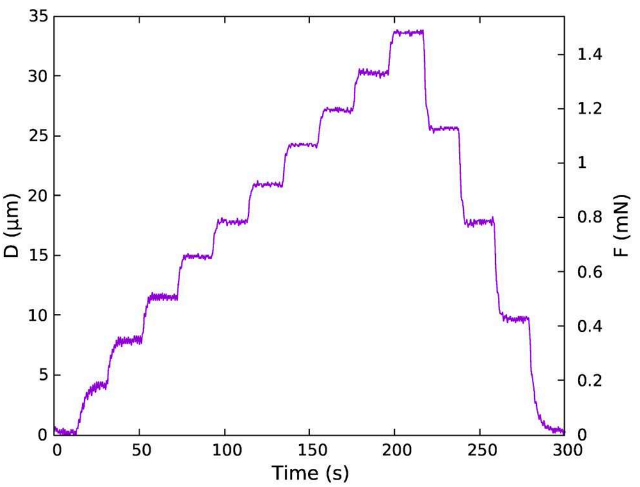

FIGURE 6 | Raw data from thrust balance showing balance displacement *D* over time. In this sequence, the thruster was operated with cold gas starting at 0 to 200 SCCM in steps of 20 SCCM, then return back to 0 SCCM in steps of 50 SCCM. Each step lasts for 20 s. The data sampling rate is 312.5 Hz with a moving average of 256 applied. It is then reduced by taking the mean of every 16 measurements. The basic calibration is measured to be approximately 0.044 mN/μm.

of 256 applied. The data is then reduced by taking the mean every 16 measurements. The result is a measurement clear of frequencies >1 Hz which are most likely noise. This measurement was repeated 4 times to obtain the F_{cg} results shown by green crosses on **Figure 7**. Plenum pressure P_{plenum}, shown in the figure as blue stars, varies from 0 Torr to ∼3 Torr when the argon flow rate increases from 0 SCCM to 200 SCCM.

On first approximation, neglecting the neutral gas pressure term, the axial thrust force generated by the cold gas can be approximated for an isentropic, choked flow (Mach number M = 1) regime by the momentum term [27]

$$F_{cg_{calc}} = \dot{m} c_{cs}$$

where $\dot{m}$ is the argon mass flow rate and $c_s = \sqrt{\frac{\gamma_{Ar} R T_s}{m_{Ar}}}$ is the argon gas sound speed (T_s is the static temperature at the exit of the tube, R = 8.314 J.mol^{-1}K^{-1} is the universal gas constant, γ_{Ar} is the specific heat capacity ratio for argon and m_{Ar} is the argon molar mass). At T_{total} ∼300 K, T_s given by the formula $T_s = \frac{T_{total}}{1+\frac{\gamma-1}{2}M^2}$ is about 225 K, c_s is about 279.2 m/s and the calculated $F_{cg_{calc}}$ shown by the purple solid line on **Figure 7** increases up to ∼1.7 mN for 200 SCCM of argon. Agreement between $F_{cg_{calc}}$ and $F_{cg_{meas}}$ is very good at flow rates below ∼80 SCCM giving confidence that the experimental system is appropriate. The measured thrust $F_{cg_{meas}}$ is somewhat lower than the calculated thrust at high flow which indicates that some unaccounted loss is occurring in the system. This loss is likely due to the boundary layer friction force acting upon the inner wall of the ceramic tube, as described in detail by Ho et al. [25].

3.2. Direct Thrust Measurements: Plasma Thrust

The next step in the characterization of the inductive PR is the measurement of the thrust increase F_p when the plasma is ignited (as shown in **Figure 4**). For the thrust measurement procedure, the prototype is fed with a constant gas flow rate, the balance is allowed a few seconds to settle and the RF is subsequently switched on. A typical measurement of the balance displacement D at plasma ignition and thereafter is shown by a solid purple line in **Figure 8**. D exhibits a rapid increase due to the volumetric gas heating directly by the RF power, followed by a slower increase due to the ceramic wall heating up via ion bombardment and exchanging heat with the propellant as described in detail

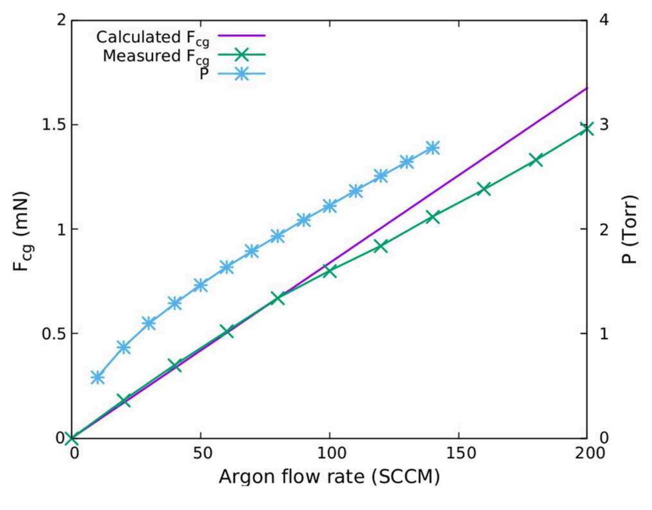

FIGURE 7 | Calculated cold gas thrust F_{cg} (solid purple line), measured thrust F_{cg} (green crosses) in the Wombat chamber (**Figure 3**) and plenum pressure P_{plenum} (blue stars) measured in the Chi-Kung chamber (**Figure 2**), vs. argon flow rate.

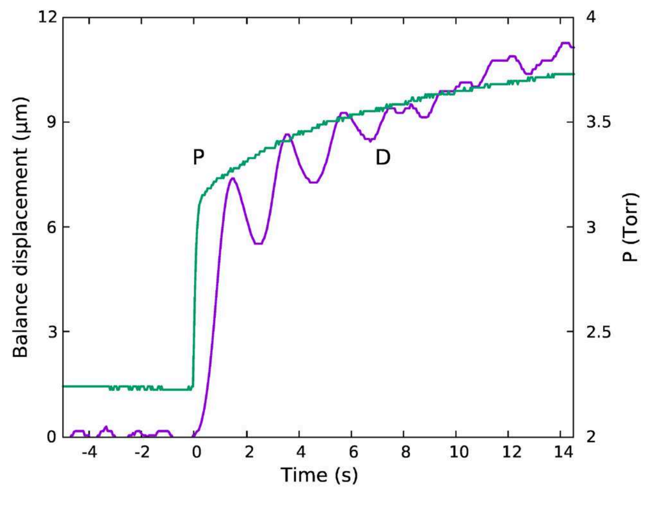

FIGURE 8 | Balance displacement D (solid purple line) measured in the Wombat chamber and plenum pressure P_{plenum} (solid green line) measured in the Chi-Kung chamber, over time for 100 SCCM of argon and 50 W of RF. The RF is switched on at $t = 0$ s. The RF current flowing on the antenna when the RF is switched on is ∼ 4 A RMS.

in Greig et al. [19]. In this figure the effect of the magnetic damper can be observed with the ∼2 s period oscillations in the displacement being damped effectively within a few seconds.

A useful qualitative diagnostic indicative of thrust from the plasma is the plenum pressure P_{plenum}, measured with the inductive PR mounted onto the Chi-Kung chamber (first configuration), and shown by the solid green line on **Figure 8** for the same flow rate and RF power. The displacement and plenum pressure plots are synchronized to the time the RF is switched on ($t = 0\ s$). The variation in P_{plenum} vs. time strongly matches that of the balance displacement. The thrust increase due to the wall heating up can be compared to the resistojet thruster principle and it is found to be a function of gas flow rate, RF power and heat dissipation mechanism. For the present study we only focus on plasma thrust gain at plasma ignition vs. the two main parameters, RF power and argon gas flow rate.

The effect of RF power in the plasma thrust (F_p) for a fixed 100 SCCM argon flow rate is presented in **Figure 9**. The data points in this plot are the average of 8 measurements of 20 s burns for each power and the error bars reflect the distribution of these measurements. The reported thrust does not include the slow resistojet increase which can be seen in **Figure 8**. F_p increases quasi-linearly from 0.14 mN at 20 W to 0.3 mN at 50 W. Since the pressure is of the order of one Torr in the plasma cavity (same as capacitive PR), it is expected that the source of thrust will be mostly from heated neutrals [16]. Previous thrust measurements have been reported for inductive RF sources [21]: for a 6.5 cm-diameter, 9.5 cm-long plasma cavity in which the operating pressure was of the order of a mTorr (∼55 SCCM of argon), the source of thrust was shown to be mostly a result of the maximum electron pressure converted into ion momentum and about 0.5 mN at 100 W.

It is of interest to carry out a basic estimate of the thrust from ions for the mm scale inductive PR inductive RF source for comparison. This was done using a global plasma model (comprising a particle balance and power balance) described in detail by Lieberman and Lichtenberg [14] and previously applied to low pressure RF sources [20] to determine a maximum electron density subsequently used in a plasma thrust model described in detail by Fruchtman [23], and successfully applied to inductively coupled RF thrusters [21, 34]. In the latter, ion-neutral collisions are ignored and the thrust from accelerated ions is given by the maximum electron pressure within the plasma source region, $F_{ion} = qn_eAT_e$, where q is the electron charge magnitude, n_e is the maximum radially averaged density within the source region, A is the cross-sectional area of the source tube, and T_e is the electron temperature. A thrust reduction factor of 0.6 [21] to 0.82 [34] to account for the radial density profile in a cylindrical source is typically used.

A particle balance for a cylindrical plasma which consists in equating the total surface particle loss to the total ionization [14, 20] was initially carried out to determine the electron temperature (assuming a Maxwellian distribution) and the ion Bohm velocity: ignoring the presence of the plenum cavity and using a plasma cavity radius of 2 mm and length of 15 mm, approximately corresponding to the antenna footprint shown in **Figure 4**, and a mid-cavity pressure of 1.1 Torr corresponding to half the plenum pressure measured for 100 SCCM gas flow (**Figure 8**), the electron temperature is about 2.0 eV and the Bohm velocity $u_B = 2.2 \times 10^3$ m/s. The input gas temperature was assumed to be 300 K in first approximation. Such procedure was also carried out in the capacitive PR as described in Charles and Boswell [12] yielding similar output due to the similar geometry and operating gas pressure. These capacitive PR results were later confirmed with direct electrostatic probe measurements and computer simulations [28, 35].

Based on the derivation of the electron temperature with the particle balance, a power balance was subsequently used

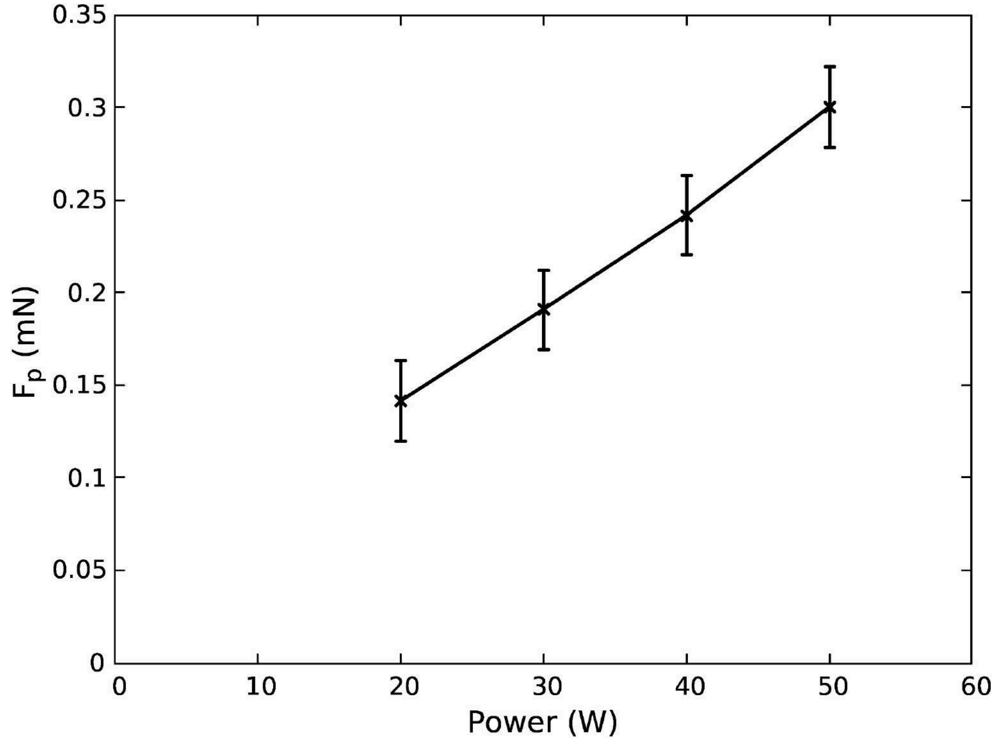

FIGURE 9 | Plasma thrust F_p vs. absorbed RF power for fixed 100 SCCM argon flow (solid line, crosses). The reported thrust in this plot does not include the cold gas F_{cg} component which is ∼0.8 mN, as seen in **Figure 7**.

to estimate the maximum plasma density ($n_{e_{max}} = n_{i_{max}}$) as described in Lafleur et al. [21], Lieberman and Lichtenberg [14], Lieberman et al. [20] and determine the thrust from the electron pressure $n_e T_e$ (which is converted into axial ion momentum as previously detailed in Lafleur et al. [21] and Fruchtman [23], here a thrust factor of 0.6 is used for the radially averaged plasma density, i.e., $n_e = 0.6 n_{e_{max}}$). An approximate power input reduction factor of 0.9 was used to account for the electrical power transfer inefficiency of the matching circuit. This model yields a thrust from ions of 25 μN at 20 W to 63 μN at 50 W of RF power which is between 18% (at 20 W) and 21% (at 50 W) of the measured values. These values are significantly lower than the respective measured thrust values shown in **Figure 9** (140 μN and 300 μN for 20 W and 50 W, respectively). This is an indication that like its predecessor the capacitive PR, the inductive PR behaves like an electrothermal thruster (thrust gain from heated neutrals) rather than a plasma or ion thruster when operated in the configuration of geometry, power and gas flow rate described in this paper.

Having confirmed that the main thrust generation mechanism is neutral gas heating (rapid increase reported in **Figure 9** due to the volumetric gas heating directly by the RF power, followed by a slower increase due to the ceramic wall heating up via ion bombardment and shown in **Figure 8**, the variation in thrust from the ions was investigated as a function of the gas temperature since we initially assumed a gas temperature of 300 K. Assuming a gas temperature increase of a factor of 2 from 300 K to 600 K for 20 W power input gives an electron temperature of 2.3 eV and an ion thrust of 19 μN (down from 25 μN). This simple estimate shows the interplay between thrust generation from heating neutrals and accelerated ions. Global models provide no spatial information and should be complemented by future dedicated computer simulations for further investigation.

Figure 10 shows the effect in F_p of flow rate change when keeping the RF power constant. The knee point seen at about 60 SCCM in this plot is likely to be related to the physical dimensions of the tube and will be the topic of future work. In summary, the comparison between F_p at ignition for this unoptimized thruster and F_{cg} shows a total thrust increase of up to 40%.

The thrust reported so far does not include the slow increase of thrust over time shown in **Figure 8** which is attributed to the ceramic wall heating. This increase however has an important role in the overall performance of the system. To understand better this effect, a simple resistojet experiment was performed using a constant 100 SCCM argon flow into a same dimensions plenum and ceramic tube system. In this experiment, the RF antenna was replaced by a tungsten wire which is placed inside the tube and was heated up to glowing temperatures by 20 W of DC power. The heating element had a diameter of 2 mm and length of ∼1 cm and was placed at a distance L from the exhaust end of the tube. Thrust measurements were made on the Wombat balance by setting up a constant 100 SCCM flow, then turning

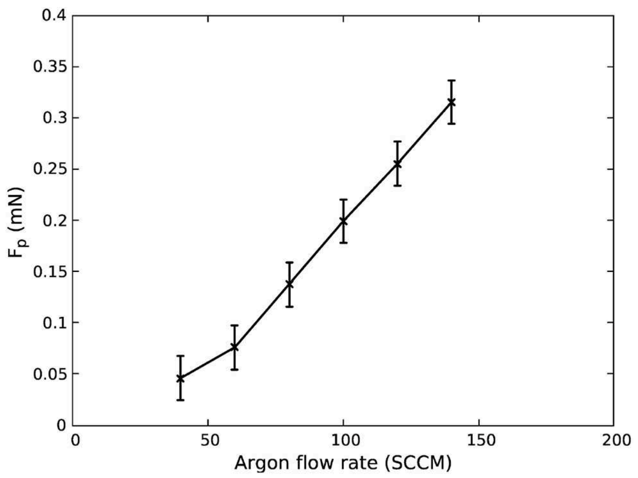

FIGURE 10 | Plasma thrust F_p vs. flow rate for fixed 30 W RF power. The reported thrust in this plot is the thrust due to the RF and does not include the cold gas component.

on the heating element for 20 s and recording the thrust at the end of that period. At $L = 0$ mm, the thrust gain over the cold gas thrust, F_T/F_{cg}, was 1.99 dropping to 1.65 at $L = 5$ mm, 1.34 at $L = 10$ mm, 1.22 at $L = 15$ mm and 1.08 at $L = 20$ mm. The gradual decrease in thrust can be attributed to the cooling down of the gas while traversing the final length of the tube which did not have enough time to reach a temperature equilibrium. An attempt was made to shift the RF antenna of the inductive pocket rocket closer to the exhaust end of the tube to confirm this observation with an RF plasma but the RF matching was affected possibly due to the pressure change near the exhaust end and was impossible to get reliable data. This observation is going to be an important focus on future optimizations of this thruster.

Another point of interest to future optimizations is the formation of parasitic plasma outside the ceramic tube. This is due to the pressure in the vacuum chamber increasing from 10^{-6} Torr to $\sim 10^{-3}$ Torr when the thruster is operational which resulted in enough gas density to ignite and sustain a weak plasma outside the ceramic tube. The effect of this is that a percentage of the injected power is lost to that parasitic plasma which contributes to the lower performance of the thruster. In space, this is less likely to be a problem however it is possible to mitigate it the impact by increasing the vacuum pumping speed and protecting the area directly over the antenna by some inert material to discourage the formation of plasma.

4. CONCLUSIONS

The present study demonstrates total thrust gain at plasma ignition in a small size inductively coupled Pocket Rocket, operating as an electrothermal plasma thruster. A small foot print impedance matching network operating at 40.68 MHz and mounted directly onto the plasma cavity allows reliable measurements of direct thrust with inductive PR immersed in vacuum and attached to an optimized thrust balance. A magnetic damper facilitates the measurement procedure by limiting thermal effects during plasma burns which allows reliable and repeatable measurements. The measured thrust is found to be comparable with the predecessor capacitive PR, however, absolute comparisons are not easy to make due to the unknown losses in the matching network of the capacitive PR vs. and the differently configured thruster balance used for those early experiments. The direct thrust measurements of the inductive PR have confirmed the production of thrust and have highlighted a couple of areas of improvement which will be considered in future work. Based on the reported results, future studies are justified and are expected to incrementally improve the performance of the presented proof of concept.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/supplementary material.

AUTHOR CONTRIBUTIONS

This work presented in this paper is DT's Ph.D. student work and has been supervised by CC and RB.

REFERENCES

1. Goebel DM, Katz I. *Fundamentals of Electric Propulsion: Ion and Hall Thrusters*. Hoboken, NJ: John Wiley & Sons, Inc. (2008). doi: 10.1002/9780470436448
2. California Polytechnic State University. *Cubesat Design Specification rev. 13* (2014). Available online at: http://www.cubesat.org/s/cds_rev13_final2.pdf
3. Kulu E. *Nanosat Database.* (2019). Available online at: https://www.nanosats.eu
4. NASA. *State of the Art of Small Spacecraft Technology*. Hanover, MD: NASA (2018). Available online at: https://sst-soa.arc.nasa.gov/
5. Pascoa JC, Teixeira O, Filipe G. "A review of propulsion systems for cubesats," in: *ASME International Mechanical Engineering Congress and Exposition, Proceedings (IMECE)*. Pittsburgh, PA (2018).
6. Lemmer K. Propulsion for CubeSats. *J Acta Astron.* (2017) **134**:232–43. doi: 10.1016/j.actaastro.2017.01.048
7. Keidar M, Zhuang T, Shashurin A, Teel G, Chiu D, Lukas J, et al. Electric propulsion for small satellites. *Plasma Phys Control Fusion.* (2015) **57**:014005. doi: 10.1088/0741-3335/57/1/014005
8. Mueller J, Hofer R, Ziemer J. Survey of propulsion technologies applicable to cubesats. In: *57th JANNAF Propulsion Meeting.* Pasadena, CA (2010).
9. Tummala AR, Dutta A. *An Overview of Cube-Satellite Propulsion Technologies and Trends*. Aerospace (2017). Available online at: http://www.mdpi.com/2226-4310/4/4/58
10. Leomanni M, Garulli A, Giannitrapani A, Scortecci F. Propulsion options for very low Earth orbit microsatellites. *Acta Astronaut*. (2017) **133**:444–54. doi: 10.1016/j.actaastro.2016.11.001
11. Scharlemann C, Tajmar M, Buldrini N, Krejci D, Seifert B. Propulsion for nanosatellites. In: *32nd International Electric Propulsion Conference*. Wiesbaden (2011).
12. Charles C, Boswell RW. Measurement and modelling of a radiofrequency micro-thruster. *Plasma Sources Sci Technol.* (2012) **21**:022002. doi: 10.1088/0963-0252/21/2/022002

13. Charles C. Plasmas for spacecraft propulsion. *J Phys D Appl Phys*. (2009) **42**:163001. doi: 10.1088/0022-3727/42/16/163001
14. Lieberman MA, Lichtenberg AJ. *Principles of Plasma Discharges and Materials Processing*. 2nd ed. Hoboken, NJ: Wiley (2005). doi: 10.1002/0471724254
15. Charles C, Boswell RW, Bish A. Variable frequency matching to a radiofrequency source immersed in vacuum. *J Phys D Appl Phys*. (2013) **46**:365203. doi: 10.1088/0022-3727/46/36/365203
16. Ho TS, Charles C, Boswell R. Neutral gas heating and ion transport in a constricted plasma flow. *Phys Plasmas*. (2017) **24**:084501. doi: 10.1063/1.4996014
17. Doyle SJ. *Electron, Ion and Neutral Heating in Hollow Cathode Plasma Thrusters*. York: University of York (2019).
18. Ho TS, Charles C, Boswell R. Redefinition of the self-bias voltage in a dielectrically shielded thin sheath RF discharge. *J Appl Phys*. (2018) **123**:193301.
19. Greig A, Charles C, Paulin N, Boswell RW. Volume and surface propellant heating in an electrothermal radio-frequency plasma micro-thruster. *Appl Phys Lett*. (2014) **105**:054102. doi: 10.1063/1.4892656
20. Lieberman MA, Charles C, Boswell RW. A theory for formation of a low pressure, current-free double layer. *J Phys D Appl Phys*. (2006) **39**:3294–304. doi: 10.1103/PhysRevLett.97.045003
21. Lafleur T, Takahashi K, Charles C, Boswell RW. Direct thrust measurements and modelling of a radio-frequency expanding plasma thruster. *Phys Plasmas*. (2011) **18**:080701. doi: 10.1063/1.3610570
22. Takahashi K. Helicon-type radiofrequency plasma thrusters and magnetic plasma nozzles. *Rev Mod Plasma Phys*. (2019) **3**:3. doi: 10.1007/s41614-019-0024-2
23. Fruchtman A. Neutral depletion in a collisionless plasma. In: *IEEE Transactions on Plasma Science*. (2008) **36**:2. doi: 10.1109/TPS.2008.918777
24. Fruchtman A. *Neutral Gas Depletion in Low Temperature Plasma* (2017) **50**:473002. doi: 10.1088/1361-6463/aa87a9
25. Ho TS, Charles C, Boswell RW. A comprehensive cold gas performance study of the pocket rocket radiofrequency electrothermal microthruster. *Front Phys*. (2017) **4**:55. doi: 10.3389/fphy.2016.00055
26. Charles C, Boswell RW. Time development of a current-free double-layer. *Phys Plasmas*. (2004) **11**:3808–12. doi: 10.1063/1.1764829
27. Ho TS. *Supersonic Constricted Plasma Flows*. Canberra, ACT: Research School of Physics and Engineering, The Australian National University (2018).
28. Ho TS, Charles C, Boswell R. Performance modelling of plasma microthruster nozzles in vacuum. *J Appl Phys*. (2018) **123**:173301. doi: 10.1063/1.5012765
29. Charles C, Boswell RW, Bish A, Khayms V, Scholz EF. Direct measurement of axial momentum imparted by an electrothermal radiofrequency plasma micro-thruster. *Front Phys*. (2016) **4**:19. doi: 10.3389/fphy.2016.00019
30. Sokal NO, Sokal AD. Class E-A new class of high-efficiency tuned single-ended switching power amplifiers. *IEEE J Solid-State Circ*. (1975) **10**:168–76. doi: 10.1109/JSSC.1975.1050582
31. Liang W, Charles C, Raymond L, Stuchbery A, Surakitbovorn K, Gu L, et al. An integrated RF power delivery and plasma micro-thruster system for nano-satellites. *Front Phys*. (2018) **6**:115. doi: 10.3389/fphy.2018.00115
32. Chen W, Chinga RA, Yoshida S, Lin J, Chen C, Lo W. A 25.6 W 13.56 MHz wireless power transfer system with a 94% efficiency GaN Class-E power amplifier. In: *IEEE MTT-S International Microwave Symposium Digest*. Montreal, QC (2012). doi: 10.1109/MWSYM.2012.6258349
33. Liang W, Raymond L, Praglin M, Biggs D, Righetti F, Cappelli M, et al. Low-mass RF power inverter for cubesat applications using 3-D printed inductors. *IEEE J Emerg Select Top Power Electr*. (2017) **5**:880–90. doi: 10.1109/JESTPE.2016.2644644
34. Takahashi K, Lafleur T, Charles C, Alexander P, Boswell RW, Perren M, et al. Direct thrust measurement of a permanent magnet helicon double layer thruster. *Appl Phys Lett*. (2011) **98**:141503. doi: 10.1063/1.3577608
35. Greig AD. *Pocket rocket: an electrothermal plasma micro-thruster* (Ph.D. thesis). Canberra, ACT: ANU (2015).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Physics

ORIGINAL RESEARCH
published: 13 March 2020
doi: 10.3389/fphy.2020.00049

The Emergence of Critical Stocks in Market Crash

Shan Lu[1], Jichang Zhao[2,3]* and Huiwen Wang[2,3]

[1] School of Statistics and Mathematics, Central University of Finance and Economics, Beijing, China, [2] School of Economics and Management, Beihang University, Beijing, China, [3] Beijing Advanced Innovation Center for Big Data and Brain Computing, Beijing, China

In complex systems like financial market, risk tolerance of individuals is crucial for system resilience. The single-security price limit, designed as risk tolerance to protect investors by avoiding sharp price fluctuation, is blamed for feeding market panic in times of crash. The relationship between the critical market confidence which stabilizes the whole system and the price limit is therefore an important aspect of system resilience. Using a simplified dynamic model on networks of investors and stocks, an unexpected linear association between price limit and critical market confidence is theoretically derived and empirically verified in this paper. Our results highlight the importance of relatively "small" but critical stocks that drive the system to collapse by passing the failure from periphery to core. These small stocks, largely originating from homogeneous investment strategies across the market, has unintentionally suppressed system resilience with the exclusive increment of individual risk tolerance. Imposing random investment requirements to mitigate herding behavior can thus improve the market resilience.

Keywords: network science, financial system, risk contagion, market resilience, market crash

Edited by:
Matjaž Perc,
University of Maribor, Slovenia

Reviewed by:
Jiangcheng Li,
Yunnan University of Finance and Economics, China
Bikas K. Chakrabarti,
Saha Institute of Nuclear Physics (SINP), India

Correspondence:
Jichang Zhao
jichang@buaa.edu.cn

Specialty section:
This article was submitted to Interdisciplinary Physics, a section of the journal Frontiers in Physics

Received: *29 November 2019*
Accepted: *20 February 2020*
Published: *13 March 2020*

Citation:
Lu S, Zhao J and Wang H (2020) The Emergence of Critical Stocks in Market Crash. Front. Phys. 8:49. doi: 10.3389/fphy.2020.00049

1. INTRODUCTION

Financial markets are characterized by complex systems which give rise to emergent phenomena such as bubbles and crashes occasionally [1]. The price limit for single-security, which usually regards as part of a broader effort to mitigate extreme risk in stock market, has been widely used in China, the US, Japan and Canada, etc. It forbids traders trading stocks at any price above or below a predefined level for the remainder of the day. In Chinese A-share market, for instance, the absolute return permitted is 10% for every regular stock. Though the single-security price limit is designed to be equal for all securities, it functions as a stock-specified tolerance, and helps protect investors by avoiding sharp price declining or jumping. Conversely, price limit critics claim that price limit may be ineffective [2, 3] or even feed panic selling in times of market crash [4]. As the traders are in fear of the potential illiquidity when price limit locks their positions [2], they will collectively sell stocks to seek for liquidity or reduce risk exposure, which in turn smashes market confidence and leads to further downward depression on a wider range of stock prices. For instance, in the 2015–2016 stock market crash in China, such global sell-off has spread so widely that more than 1,000 stocks prices declining to the daily price limit has become nearly normal for investors. From the complex system perspective, if the market confidence is sufficient enough to withstand illiquidity shocks caused by the price limits, a system collapse can be avoided. How the critical market confidence that keeps the market away from collapse reacts to price limits therefore determines the market resilience, which usually described as the ability of a system to adjust its activities to retain stable when shocks arrive [5]. In the context, a clear feature of interest is the presence of single-stock price limits' effects on the critical market confidence and system disruption.

An important depression contagion channel for price limits and market panic is the overlapping portfolios when investors invest in the same equities. As mentioned above, this might be the case if investors sought to hedge their exposure to potential illiquidity of the security that had reached the price limit by trading other stocks in their portfolios. A similar effect might arise if traders who have incurred mark-to-market losses in stocks of price limits face margin calls or are required to reduce their positions to meet the obligation of leverage ratio [6–8]. Additionally, investors may be unwilling to hold securities in fear of that the price limit is driven by information that will affect the value of a wide range of stocks, e.g., the systemic risk [9]. This kind of "loss of market confidence," in particular, plays a non-negligible role in market crisis [10, 11]. Thus, it is worth exploring the exact knowledge of the relationship between the designed price limit that cause liquidity shocks and the more general "loss of confidence" that can provoke throughout the system. It is also of great importance to probe how the investment behavior and market network structure link together as well as how they influence the association between price limit and critical market confidence, from both theoretical and practical perspectives.

We explore these by constructing a bipartite stock network. While previous work heavily depends on numerical simulations of networks that are often assumed to be random [12–14], we use the bipartite network constructed by mutual fund share holding data in the real world. Furthermore, we validate our theoretical analysis with the real world market crash events through dense computing of price information in minute-granularity. By doing so, we are able to understand and explain the market crash in a data-driven way, aiming for better understanding of financial system in practice. It is also worth mentioning that the method of network modeling in our paper provides a feasible way to study stock market in addition to existing studies.

We also introduce a contagion mechanism and its analytical explanation on the network for capturing how the market confidence and price limits may contribute to market collapse in times of crisis. In the contagion model, we regard overlapping portfolio as risk contagion channel and consider price limit, liquidity shocks and loss of confidence as key interactive factors in market crash. While liquidity shocks and loss of confidence are already known as vital ingredients in market crisis as mentioned before, our study extends existing studies by integrating price limit, a factor of individual tolerance, into the contagion process. Thus, the contagion model also distinguishes our work from previous studies.

From the theoretical point of view, our approach recognizes some important differences with other complex systems and inherently challenges existing understandings. While previous models have underlined the importance of "superspreader" in ecosystem stability [11, 15, 16], our results adversely demonstrate the importance of relatively "small," totally "nested" stocks that drives the system to collapse. Those small stocks, largely originating from homogeneous risk-minimizing investment strategy across the market, determine how the critical market confidence reacts to the price limit. In this study, we call these stocks "critical stocks" in the context of stock market crash or "driving nodes" in the context of stock market network.

The word "critical" is borrowed from the terminology "critical points" in system science. As far as we are concerned, critical points of complex dynamical systems are defined as the explosion to infinity of a normally well-behaved quantity [14]. In particular, Sornette et al. [14] suggests that a stock market would crash when the strength of herding behavior increases up to a certain point called the "critical" point. The term "critical" in our study, however, has two meanings. The first meaning involves with "critical market confidence," which shares the same concept with previous studies, indicating that if the market confidence reaches the critical point, the stock market would switch from stable to unstable. The second meaning belongs to "critical stocks," also called "driven nodes" in this study, because they play the decisive roles in the exact values of critical market confidence under the predetermined price limits. While the "critical market confidence" refers to critical points for the system at the macro-level, the "critical stocks" relates to the components of the system at the micro-level that determine the critical points of the whole system.

Unlike other studies in econophysics in which prices or returns are the main focuses to observe market crash [14, 17], we zoom in the determinants of critical points and the procedure of market crash. In particular, although both Sornette et al. [14] and our work are trying to explain the financial market crash from the perspective of complex system and we both agree that herding behavior is the source of market crash, the fundamental ideas of our work are different from theirs. Sornette et al. [14] argues that the herding behavior of the traders who may drastically revise their decision would abruptly produce a sudden unbalance between supply and demand. When the strength of herding behavior reaches to the critical point, a crash happens. In our study, we point out that the herding behavior unintentionally breeds a network with small but critical stocks, who play decisive roles in regulating the relationship between price limit and critical points of market confidence.

We further capture that the order of stocks reaching their limit down prices in the real world is similar to that found in our model: from periphery to core, then from core to the whole system. During this process, the small but critical stocks located on the periphery are essential, as they firstly pass on the failures to the core nodes inside the network, that lead the system reach global failure eventually. In a word, the small stocks are vulnerable to the first-round of failure and are critical to the further rounds of price limits implements and illiquidity propagation, which makes them crucial for system resilience. Note that previous studies have not underlined the importance of small nodes in the systems, our findings inject new and counterintuitive insights into the market supervision and suggest authorities watch the critical small stocks instead of fully attracted by some systemically important ones in practice.

2. MATERIALS AND METHODS

2.1. Model

Here we build a bipartite stock-investor network (see **Figure 1**). While similar networks have been used to study the financial system stability [18–20], the present paper would mainly focus on

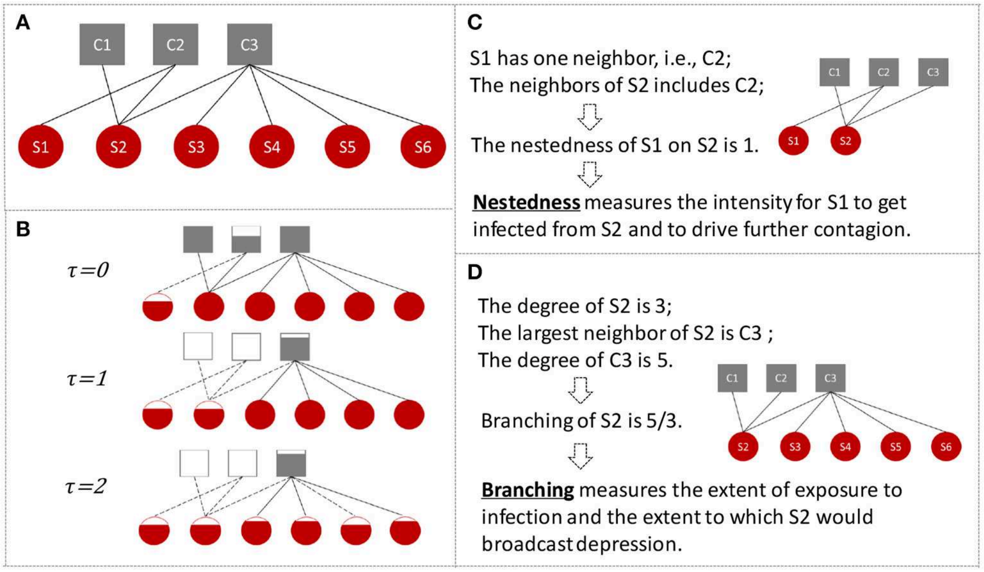
A
C1
C2
C3
S1
S2
S3
S4
S5
S6
B
τ=0
τ=1
τ=2
C
S1 has one neighbor, i.e., C2;
The neighbors of S2 includes C2;
The nestedness of S1 on S2 is 1.
Nestedness measures the intensity for S1 to get infected from S2 and to drive further contagion.
D
The degree of S2 is 3;
The largest neighbor of S2 is C3 ;
The degree of C3 is 5.
Branching of S2 is 5/3.
Branching measures the extent of exposure to infection and the extent to which S2 would broadcast depression.

FIGURE 1 | The stock-investor network illustration. **(A)** Shows the network compositions, where the gray squares are investors and the red circles are stocks. Edges exist only from one kind of nodes to the other and edge weights represent how much the investors invest on the corresponding stocks with respect to market values. It is an undirected weighted network. **(B)** Shows the contagion procedure. The initial shock is stock S1 reaching its price limit and losing c proportions of its market value. The lack of liquidity results from this leads to investor C2 faces a proportion of share holdings being locked. This further makes C2 sell other stocks in hand at $\tau = 1$, i.e., S2, which then conveys the depression to C1, and so on and forth. **(C,D)** Shows the examples of "nestedness" and "branching" that defined in section 2.2.

the stock market crash and how the network structure determines the association between downward price limit and critical market confidence, which features the market resilience.

Define the absolute value of price limit down as c, where $c \in (0, 1)$. The market value of stock i at time τ is $S_{i,\tau}$. For stock i, if

$$\frac{S_{i,\tau} - S_{i,\tau=0}}{S_{i,\tau=0}} \leq -c, \tag{1}$$

stock i reaches its limit down price, what we called it "failed." A high absolute value of limit down allows a stock to withstand larger shocks before it is pushed to suffer the lack of liquidity. We study the consequences of shock initially hitting any single stock at $\tau = 0$, with the shock taking the form of wiping out a fraction c of its initial values. While c is the limit down threshold, this equals to evoking the stock's failure. An initial failure of a stock that reduces the market values of the investors' liquidation ability will elicit the panic selling on other stocks. If the market's demand is less than perfectly elastic, such disposals will result in a short run change in stock price [21, 22]. Subsequently, the externally imposed price limits may dictate additional panic selling which will have a further impact on market prices. See **Figure 1B** for an illustration.

Following the outlined cascading procedure, we integrate the interaction of market confidence into the model as a scaling effect on the liquidity shock. Specifically, the $\tau = 0$ failure by a single stock results in each of its investors' holding portfolio experiencing a $\tau = 1$ shock of magnitude

$$\alpha \frac{A_{m,\tau}}{A_{m,\tau-1}}, \tag{2}$$

where α is the market confidence, $\alpha \in [0, 1]$, and $A_{m,\tau}$ is the total stocks' market value held by investor m at time τ. The term $\frac{A_{m,\tau=1}}{A_{m,\tau=0}}$ defines the degree of illiquidity results from the price limits of failed stocks, by which we assume that a depreciating investor may depress the prices of other stocks in holding portfolio according to the relative illiquidity. The market confidence α adjusts the illiquidity magnitude by multiplying $\frac{A_{m,\tau=1}}{A_{m,\tau=0}}$ and regulates the market resilience accordingly. Market illiquidity is linked directly to confidence effects by formula (2). The assumption is that investors who hold the failed stocks could dispose other stocks in their portfolios at lower prices that related to both their liquidity pressure and market confidence. This process causes the capital position of other investors holding these same stocks to be eroded.

We will now have further, $\tau = 1$, failures of the stocks connected to the initially infected investors if formula (1) is realized. This, in turn, may generate $\tau = 2$ failures of stocks when these $\tau = 1$ failures of stocks convey the price downward pressure to their investors through formula (2). And so on for $\tau = 3$ and further. The details of the contagion model with market confidence $\alpha \in [0, 1]$ and price limit $c \in (0, 1)$ are as follows:

Step 1: $\tau = 0$, we initially shock a single stock i, wiping out c of its market value and it is then considered as failed. We have $S_{m,\tau=1} = (1 - c)S_{m,\tau=0}$, $w_{i,m,\tau=1} = (1 - c)w_{i,m,\tau=0}$.

Step 2: Update the stocks' market value $\forall i$, $S_{i,\tau} = \sum_m w_{i,m,\tau}$.

Step 3: If $\forall i$, $\frac{S_{i,\tau} - S_{i,\tau=0}}{S_{i,\tau=0}} > -c$, no further failures, the algorithm ends. If $\exists i$, $\frac{S_{i,\tau} - S_{i,\tau=0}}{S_{i,\tau=0}} \leq -c$, we call these stocks failed and add them into the stocks list F_τ. The set of neighbors of stocks belong to F_τ is denoted as L_τ.

Step 4: Delete the stocks in F_τ as they reach their down limit prices and are regarded as completely illiquid. Update the investors holding market value, $A_{m,\tau} = \sum_i w_{i,m,\tau}$.

Step 5: $\tau = \tau + 1$, update the investors holding values, i.e., $\forall m \in L_\tau, \forall i, w_{i,m,\tau+1} = \alpha \frac{A_{m,\tau+1}}{A_{m,\tau}}$. The term $\frac{A_{m,\tau+1}}{A_{m,\tau}}$ defines the degree of illiquidity results from the price limits of failed stocks in which we assume that the illiquidity of holding portfolio, a depreciating investor may depress the price of those stocks in the market. The confidence effects on stock prices depression are mild or negligible when $\alpha = 1$, but become more severe as α decrease.

Step 6: Return to Step 2.

The deficiency in the outlined model is that we assume the confidence level remain fixed to be α as the cascade surges through the system. Nevertheless, we believe that it is useful to have a clear understanding of the dynamics of potential system disruption by assuming the confidence level remains universal [23, 24]. Our model captures how the interplay of market confidence and price limits can generate a downward spiral during market crash. More broadly, the critical market confidence that maintains the stability of the system is of primary interest in the presence of risk tolerance of individuals. By unveiling the connection of the two, we obtain the gauging of market resilience in section 2.2 from theoretical perspective and section 3.1 from empirical perspective.

2.2. Theoretical Analysis

By mapping our model onto a generalized process, we show analytically here that there is a region in parameter space where further cascades of failures occur. Denote the market value of investor m holding stock i as $w_{i,m,\tau}$, i.e., the edge weight from node m to node i. The original market value of stock i is $S_{i,\tau=0} = \sum_m w_{i,m,\tau=0}$. The original market value of investor is $A_{m,\tau=0} = \sum_i w_{i,m,\tau=0}$. Denote the stocks fail at τ as F_τ. Denote the investors holding stocks that fail at τ as L_τ. Considering the devaluation in formula (2), the stocks' failure boundary in formula (1) could be written as

$$\frac{\sum_{m\in L_\tau} \alpha\left(1 - \frac{\sum_{f\in F_\tau} w_{f,m,\tau=0}}{A_{m,\tau}}\right) w_{i,m,\tau} + \sum_{m\notin L_\tau} w_{i,m,\tau}}{\sum_m w_{i,m,\tau=0}} \leq 1-c, \tag{3}$$

where m belongs to the investors stock i connects. Following formula (1) and (2), for every initially shocked stock, the market confidence needed to avoid its neighboring stock i's failure at $\tau+1$ could be calculated as

$$\alpha_{c_i} = \frac{(1-c)\sum_m w_{i,m,\tau=0} - \sum_{m\notin L_\tau} w_{i,m,\tau=0}}{\sum_{m\in L_\tau}\left(1 - \frac{\sum_{i\in F_\tau} w_{i,m,\tau=0}}{A_{m,\tau}}\right) w_{i,m,\tau=0}}. \tag{4}$$

Consider $\tau = 0$, we assume that $\frac{\sum_{f\in F_\tau} w_{f,m,\tau=0}}{A_{m,\tau}}$ is rather small because the investors have a wide range of portfolios and one of them would not be comparable with the investors' total holding values. See **Figure S3D** for evidence in the Chinese case. Formula (3) would then be simplified as

$$\frac{\sum_{m\in L_\tau} \alpha w_{i,m,\tau} + \sum_{m\notin L_\tau} w_{i,m,\tau}}{\sum_m w_{i,m,\tau=0}} \leq 1 - c. \tag{5}$$

The intuition behind the numerator is that the market value of stock i consists of two parts: one part held by investors connect to the failed stocks, the other part held by investors do not connect to the failed stocks. Apparently, the ratios of the two are critical in the network cascading. Inspired by this, two indicators for stocks are raised: nestedness and branching.

The nestedness of stock i on stock j is the degree of how stock i would be influenced by stock j's failure, which could be defined as

$$\text{nestedness of stock } i \text{ on stock } j = \frac{\text{the number of common neighbors between } i \text{ and } j}{\text{the degree of } i}. \tag{6}$$

Nestedness basically measures the severity of portfolio overlapping. Note that the nestedness of stock i on stock j is not equal to the nestedness of stock j on stock i. For instance, suppose stock i is held by few investors while these investors hold the other stock j, then the nestedness of stock i on stock j would be one (see **Figure 1C**). But if the stock j has more investors, the nestedness of stock j on stock i would be small. A higher value of nestedness implies a higher value of $\frac{\sum_{m\in L_\tau} w_{i,m,\tau}}{\sum_m w_{i,m,\tau=0}}$ in formula (5) if edge weights are really close to each other, which means the stocks with higher nestedness have greater potential to get infected by other stocks' failures and their failures will drive further rounds of risk contagion (see **Figure 1**).

The branching for stock i is defined as

$$branching = \frac{\text{the highest degree of } i\text{'s neighbors}}{\text{the degree of } i}. \tag{7}$$

Branching takes account of the number of neighbors stock i have by definition and thus reveals the probability of stock i's exposure to other stocks' failures. Remember in each step of contagion, it is only the stocks which share common neighbors with failed stocks that are taken account into the devaluation and have the potential to reach price limits, i.e., formula (3). Therefore, branching unfolds the overall extent of stock i's exposure to random shocks. Additionally, branching also depicts the capacity of spreading risk because it correlates to the number of stocks that one failed stocks could pass the depression to others through common neighbors (see **Figure 1D**). A high level of branching is basically an outcome of investors' highly diversified portfolios. Nestedness and branching altogether govern the potential for the spread of shocks through the network and it is shown that they help provide effective information into the causes of the potential dynamical behavior as well as influence system resilience. The results section will present empirical evidence of the analysis and more insights on the two new indicators.

2.3. Data

In this paper, we consider using mutual funds as proxy of market investors. The dataset of mutual funds stock holdings in China are downloaded from *Wind Information*, containing the market value of shares held by mutual fund for listed stocks on June 30 2015, around the period that the severe Chinese stock market crash happened. It covers 1,512 mutual funds and 2,709 stocks listed on either Shanghai Stock Exchange or Shenzhen Stock Exchange that are the two main stock exchanges in China A-shares market. Existing study has pointed out that though the ownership data is taken at one particular time of the year, it represents noisy yet unbiased estimate of mutual funds investment preferences in that year or at least the days around the reporting date [25]. We group ownership records by mutual fund management companies as we assume that mutual funds under the same management company could have collective actions on an individual stock.

There are 87 mutual fund companies and 2,709 stocks included in the study. To be more specific, the 87 mutual fund companies are nodes of investors, and while the 2,709 stocks listed in the market are nodes of stocks. The edge weights are equal to the sum of market value hold by mutual funds under the same management institutes. Additionally, we deploy the proposed model into the mutual fund and stock network, in which the 1,512 mutual funds are used as in the investor entities in **Figure 1** instead of the 87 mutual fund companies. The network is less denser and the results could be found in **Figures S8, S9**, which are consistent with those from the network of mutual fund companies and stocks.

The timing of stocks reaching their limit down prices are obtained by integrating two sources of information: the statuses of stocks and the stocks' intraday prices. The statuses of stocks acknowledge us whether the stocks reach their limit down prices or not during the trading day. If they did, we pick out the time that the stocks first reached their lowest prices of the day, i.e., their limit down prices, as their moments of failures. The datasets analyzed during the current study are available in the figshare.com repository, https://doi.org/10.6084/m9.figshare.8216582.v2.

3. RESULTS

3.1. Price Limits and Critical Market Confidence: The Undesirable Relationship

The relevant parameters in the model design are the price limit c and market confidence α. The prior interest is the minimum market confidence to guarantee the robustness of the system, or say the critical α, given a fixed price limit. Denote it as α_c. Here we use dataset of Chinese mutual company's holding positions to establish the bipartite network and use numerical simulations on the real world network to illustrate and clarify the intuition underpinning our model.

To make the result independent on the initial shock, we apply a shock to the stocks one at a time and iterate over the stocks set to obtain the averaged outcomes, see **Figure 2A**. The initial shock could, in principle, cause crash of the entire system if $\alpha \leq 1 - c$. The system can switch between stable and unstable, which means that the stock market can either survive and be healthy or completely collapse. More importantly, the phase transition boundary is $\alpha_c = 1 - c$, indicating that the critical market confidence does decrease with the deepening of down price limit. In other words, the system tends to be more resilient to shocks when the individual risk tolerance is higher. However, as the slope of their relationship is not steep enough as expected, the critical market confidence could not be effectively curtailed with respect to the increase of absolute value of downward price limit. Note that the downward price limit is set as -10% in the Chinese stock market, where the critical market confidence is still high according to our model. The micro-level market structure that leads to such association has to be further examined, and it is also necessary to investigate the possibility of proper structures in which the critical market confidence can be more efficiently reduced by rising the absolute value of price limits, or the market resilience is to be enhanced.

3.2. Driving Nodes: The Critical Ones That Prompt System to Collapse

As the market confidence α is felt by every investor holding the initially shocked stock and then deliveries the 'loss of market confidence' depression to portfolios, it would be stocks with the largest critical market confidence that determined the system-wide critical market confidence, i.e., α_c. For every initially shocked stock, denote α_{c_i} as the market confidence needed to avoid its neighboring stock i's failure (neighboring stock i refers to a stock denoted as i that shares at least one common investor with initially shocked stock), whose value at τ could be accordingly derived, as mentioned in section 2.2. Define "driving nodes" as stocks that have common investors with the initially shocked stocks and their α_{c_i} are the largest among others at $\tau = 1$ for the initially shocked stocks. One would expect that these driving nodes play the critical roles in regulating system resilience, i.e., the interconnection between the minimum market confidence that needed to keep system stable and the predetermined price limit.

From the empirical perspective, **Figure 2B** illustrates the proportions of initial shocks on behalf of their neighboring stocks' maximum α_{c_i} at $\tau = 1$. The highest proportions lie on the diagnose of the matrix, demonstrating that it is because the majority of initial shocked stocks connected to at least one of the stocks with $\alpha_{c_i} = 1 - c$ at $\tau = 1$ that drives the system phase transition boundary to be $\alpha_c = 1 - c$ that shown in **Figure 2A**.

Note that the driving nodes are arose from the micro-level network structure, one stock is the driving node of another stock doesn't mean it would be driving node for other stocks. Denote the probability for the stocks to be driving nodes as P_D, calculated as the ratio of the stocks' α_{c_i} equal to $1 - c$ at $\tau = 1$ to all possible initial shocks. **Figure 3A** exhibits that those of high chances being driving nodes are indeed the ones that fail

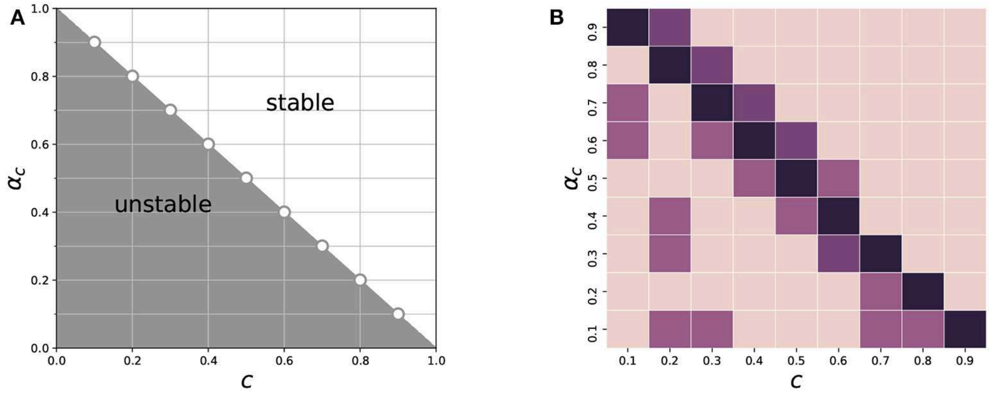

FIGURE 2 | The relationship between critical market confidence and price limit down. **(A)** Shows the phase transition boundary as $\alpha_c = 1 - c$, from which the system switches from stable to unstable. **(B)** Shows the ratios of initially shocked stock whose neighbors' maximum α_{c_i} lie in the relative intervals with respect to c ranging from 0.1 to 0.9. The darker the higher of the ratios. The grids on diagnose are the darkest ones, indicating the majority of initial shocked stocks have at least one neighbor with $\alpha_{c_i} = 1 - c$.

at the early stage. This coincides with the argument that the success of passing the depression at the start-up phase is the crucial component in cascading failures. Additionally, the inset in **Figure 3A** indicates that the probabilities of being driving nodes are low for most stocks while a few have high probabilities of being driving nodes. In spite of this, the initial shocks would cascade and cause the stock network collapses provided that there is at least one driving node for any initial attack. Therefore, the nodes with high likelihoods of being driving nodes are critical in determining the system resilience at macroscopic scale.

On top of the cascading simulation in the bipartite network, the real-cases of market crash also approve the idea that the driving nodes have far-reaching effects on depression contagion. While the contagion model has considered only the case in which there is only one stock failed at the beginning, the initial shock in real market collapse is hard to specify and a more realistic scenario is one in which a network is subjected to simultaneous initial shocks. In fact, when probing the number of newly failed stocks in a minute-granularity manner during market crash, we always find appearance of local peaks (see **Figure S1**). This scenario can be modeled as a sequence of "waves" of newly failed stocks that reaching price limits [26]. And the probabilities of being driving nodes in the bipartite network, i.e., P_D, is referred to as the stocks' capability for driving other stocks reaching price limits down in reality. Besides, the contagion model has implied that market collapse would occur in the presence of at least one driving node. Thus we only use the maximum P_D among the failed stocks in each time slot, denoted as max (P_D), to qualify the overall driving ability of these failed stocks. Interestingly, as shown in **Figure 3B**, the closer to the moment where a wide range of stocks failed, the bigger the max (P_D) is. This suggests that failures of stocks with high probabilities of being driving nodes in the established network are indeed capable of leading the market destruction. On the one hand, the results coincide with the simulation results in **Figure 3A** to certain degree. On the other hand, the results again empahsis the significant influence of these driving nodes to the stability of market, making further examinations of their roles in structure necessary.

3.3. Structural Roles: Small Nodes Take Over and Pass on Risk

In knowledge of the importance of driving nodes, how they emerge from the investing activities is the major concern. From the theoretical perspective, the driving nodes, or say, stocks with $\alpha_{c_i} = 1 - c$ in the present real data case, originates from completely sharing neighbors with initially shocked stocks (recall formula 4). And the probability of being driving nodes are further determined by the exposure to connections with initially shocked stocks. Thus we define nestedness as the ratio of overlapping investors and branching as the degree of a stock relative to its largest investor's degree, as shown in **Figure 1**. In general, the nestedness of one stock on a specific initially shocked stock measures how likely it gets infected by the initial shock and reluctantly becomes driving node to pass on depression contagion. The branching of a stock unfolds the balance between its variety in terms of number of neighbors and the diversity of its largest neighbor' investment portfolio. Therefore, branching measures the potential of a stock for being driving node if other stocks among its neighbors' portfolios got shocked.

Matching the driving nodes' probability P_D with nestedness and branching, we find that the odds of being driving nodes are positively correlated with the other two, see **Figure 4**. On one hand, the stocks which have high nestedness are doomed to have high probability of being driving nodes. On the other hand, the stocks which are of high nestedness tend to possess high level of branching. The P_D of stocks that nestedness equal to 1, in particular, are strictly proportional to their branching. Note that nestedness equal to 1 indicates that all of the stocks' nearest neighboring stocks share completely same neighbor(s) with them but have more neighbors than them. Thus the perfect linearity in the inset of **Figure 4** reveals that the probability for these stocks to be driving nodes depends on their branching, that is, how their investors diversified portfolios. In particular, those of high

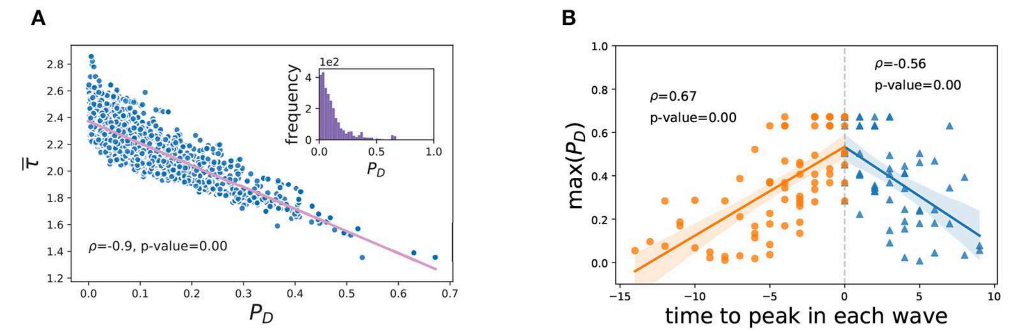

FIGURE 3 | (A) The relationship between $\overline{\tau}$ and P_D. $\overline{\tau}$ is the average cascading steps. P_D is the probability of being driving nodes. The Pearson correlation coefficient and p-value annotated in the plots are for the main axes. The inset in **(A)** describes the distribution of P_D. **(B)** The relationship between the time to the peak moments for stocks reaching price limits and the maximum P_D of failed stocks in every minute. We consider four trading days when market crashes, including June 26, June 29, July 2 and July 3 in 2015, as they considerably speak for the 2015 Chinese market crash and these four days are around the mutual fund ownership data's disclosure date [25]. We first divide each trading day into a couple of non-overlapping time intervals where in each time interval there are stocks reaching down price limits continuously in minute-granularity (see **Figure S1**). These time intervals are called "waves," as they possibly incorporate cascading failure of stocks, respectively. We also detect the peak minute(s) in every wave where the number of failed stocks hits the local maximum, indicating a wide range of stocks' failures is happening. The gray dotted line indicates where the peak moments are and the results before or after the peaks are in different colors. The Pearson correlation coefficients and p-values annotated are for the points separated by the gray dotted line.

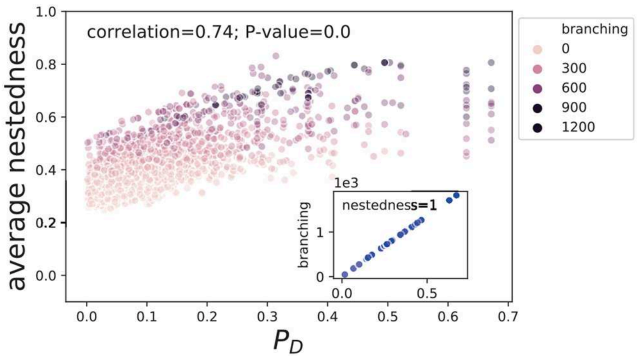

FIGURE 4 | The relationship between average nestedness and P_D. P_D is the probability of being driving nodes. The definitions of nestedness and branching could be found in section 2.2. The correlations annotated in the plots are for the main axes. The inset illustrates the relationship between P_D and branching for stocks with the average nestedness equal to 1.

probabilities of being driving nodes are connected to investors holding a wide range of equities. By being so, the highly nested stocks are the most likely to absorb the depression risk from their influential neighbors at the early stage and broadcast shocks to other branches of the system.

Moreover, we find that those driving nodes are mainly small-cap stocks, with low degrees but high branching and nestedness (see **Figure S2**). The larger-cap stocks, on the contrary, connect to more mutual fund companies, and thus have lower nestedness and branching than the smaller-cap ones. The large gap in average degrees between the large-cap and small-cap stocks implies that while the large-cap stocks are popular among all the mutual fund companies, the small-cap ones could only attract a few mutual fund companies. Additionally, **Figure S3A** shows that a large proportion of stocks are of small degrees whereas a few stocks are held by almost everyone of the mutual fund companies. Unlike the stock degree distribution, the mutual fund companies overall have high degrees, in particular a few mutual fund companies hold nearly two thirds of the listed stocks. Therefore, aiming at minimizing the investment risk, investors like the mutual fund companies who hold large numbers of stocks tend to invest on the popular stocks (large-cap) and the unpopular ones (small-cap). The popular ones become severely overlapped but the unpopular ones do not. This leads to the small-cap stocks become the nodes of high nestedness and branching in the network, making them relatively "small" when compared with their neighboring stocks but become the driving nodes that could turn over the system.

Different from previous studies in which the importance of "super-spreader" in network is emphasized [11, 15], the present fact that driving nodes could largely be recognized by nestedness and branching prompts the idea of nodes with few neighbors but having one important neighbor would play essential part in our story. In other words, the connections to investors that having exposures across a wider set of stocks make the small stocks possess higher risk on taking-over the depression. Stocks that are nested too much on others should be protected first for the sake of the whole system. And the reason behind this is the homogenous investment strategy that seek for not only wide diversity but also preference on some particular stocks for their safety (see **Figure S4**).

Many different mechanisms have been suggested in the literature to account for such a high degree of similarity across portfolios including connections between mutual fund managers and corporate board members, herding behavior and imitation of successful diversification strategies [27, 28]. Another potential reason behinds the similarity of investing pattern is the investing concentration on a range of stocks with high social trust in Chinese stock market as it is believed that stocks with high social trust have smaller crash risks [29]. These prudential investment strategies are designed to enhance the market resilience for shocks. However, they lead to a more densely connected heterogeneous financial market and the emergence of small but critical stocks that take over initial shocks and drive further depression, thus undermine system resilience.

3.4. Risk Contagion: Cascading Patterns Due to Driving Nodes

Considering the particularity of driving nodes in the microstructure and their critical roles in the acceptance and diffusion of depression, they may lead to a formation of stable macroscopic cascading patterns. Understanding such patterns not only helps to recognize the systemic impact of driving nodes, but also offers references for precautionary actions of collapse prevention and even brings about the power of prediction. Given the fact that the stocks of high probabilities of being driving nodes are those of high nestedness, a reasonable path for risk contagion would be from the network periphery to the core and then spread to the entire system. Here we use k-core index as the description of the nodes' locations in network for its effectiveness in detecting cores and peripheries [30]. The left panel in **Figure 5** clearly demonstrates that the initial attack toward the system first hit the periphery, where the driving nodes locate, and then spreads to the inside. By then, a wide range of failures emerges, propagates to the whole network and results in the system collapse. Note that the contagion dynamics are not long-lived, as the simulation always terminates within a few steps, due to the fact that our network is a rather small one (see **Figures S5, S6** for how the cascading proceeds).

We also probe the k-core index distribution using the real-world stocks' failing procedure and find great similarity, see right panel of **Figure 5**. A similar trend of the contagion procedure is found in these four market crashing days, separating by the lunch breaks. To be specific, the order of stocks reaching their limit down prices in the real world is similar: from outside to inside, then from inside to outside (see also for **Figure S7**). Note that the financial crash is extremely rare and the past sample size cannot be large. The statistics based on observations of large population are accordingly restricted. Still, the success in fitting stocks failures within a few representative crashing days validates that our model approximation gives agreement similar to that seen in real market crash.

More importantly, one insight in addition to a previous study, that the root case of the system collapse is the extinction of nodes located in the maximum k-core of the network [16], has emerged. We argue that the driving nodes on the periphery are essential, as they firstly pass on the failures to the core nodes inside the network, that lead the system reach global failure eventually.

Admittedly, the results could only provide slim prediction power of stocks failures due to the fact that the network is built on the mutual funds investors while the individual investors are the majority traders in Chinese stock markets. However, the network is still of valuable representation for the whole market considering that individual investors are easily allured by mutual fund institutions' holding positions and investment trending and that mutual funds occupy a large fraction of overall trading value in mature markets like US, Japan and Hongkong [25]. More importantly, the success on approximating the real-world stocks' failing process at the macro-level endows the contagion model with an early warning capability. And the results also suggest that the small stocks, which play the roles of driving nodes on the network periphery, should be protected or isolated for precautionary purposes. Overall, the contagion model would be good a resemblance to the actual market crash.

4. DISCUSSION

Though the single-security price limit is widely used in stock markets of different counties, the exact knowledge of its influence on market resilience is still unknown. In the proposed bipartite network based on common asset exposure in stock market, we specifically study the relationship between the critical market confidence α_c that could maintain system stability and single-security limit down c in a risk contagion model design. The linear relationship between the two, which signals the verge of the financial system becoming unstable, implies that α_c cannot be reduced significantly when rising the absolute value of c. The fine-tuned price limit, in essence, cannot drastically alter the critical market confidence as expected. From this perspective, the slope of α_c and c would be a new indicator to reflect the system resilience. The results are similar if we use mutual funds as the investors in the network instead of mutual fund companies with respect to network structure and the embedded relationship between critical parameters (see **Figures S8, S9**). This sheds lights on the counterproductive of the accordant single price limit setting in the circumstance of investment behavior pattern like China.

Essentially, the verge of the system stability is awarded by the overall similarity among investments. Even though fund managers are professional investors whose diversification strategies cannot be reduced to random selection of assets, several studies have mentioned that the investing behavior among mutual funds is similar [19, 22, 25]. The herding behavior is more severe in Chinese stock market where we find that similar heterogeneity characterizes the stocks: most stocks are found in the portfolios of a few funds, but some stocks enjoy huge popularity and are held by almost every fund. Stocks with investments from few mutual fund companies are mostly held by those who tend to over-diversify their portfolios. Therefore, they enjoy high nestedness and branching, indicating that they are relatively small in terms of number of investors but have severe portfolio overlapping with other stocks and are able to link the initially shocked stock with other stocks through these investors. On the one hand, they can quickly fail in reaction to the losses of the initially shocked stock's illiquidity even with considerable level of price limit. On the other hand, they increase the chances that other stocks will be exposed to investors who experiencing panic selling in the first round of depression contagion, acting like driving nodes for further system collapse. That is to say, when the failure of such a stock triggers contagious illiquidity because of price limit, a large number of its investors' linkages also increases the potential for contagion to be extremely widespread.

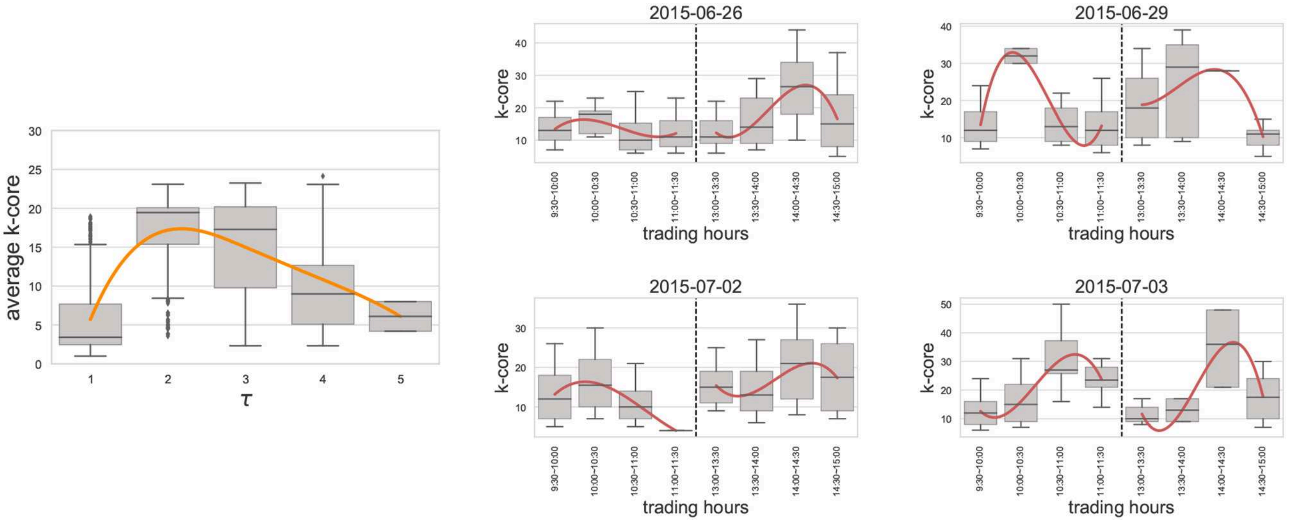

FIGURE 5 | The *k*-core index of stocks reaching the limit down prices. The left panel presents the averaged *k*-core index at each τ in the simulation course from the proposed contagion model. The boxes show the distribution of the averaged *k*-core index at τ. The smooth line (orange) shows the trend of the mean of the averaged *k*-core index at each τ. The right panel which contains four subplots shows the *k*-core indexes of stocks reaching the limit down prices in four market crashing days. The smooth lines (red) depict the trend of the mean of the stocks' *k*-core indexes. The gray dotted lines imply the lunch breaks. For the purpose of clearly visualizing data, outliers have been dropped.

In conclusion, the small stocks are vulnerable to the first-round of failure and are critical to the further rounds of price limits implements and illiquidity propagation. Additionally, knowing that these small stocks are usually blind spots for regulators, our results highlight their critical roles in determining the system resilience in reaction to individual risk tolerance. And it is the herding behavior on diversification strategies that leads to the dominance of small but critical stocks in the system.

One of the possible ways to augment system resilience is to adjust the investing pattern on the whole to avoid the dominance of small but critical stocks. To achieve this, we conduct a series of straightforward random experiments, see **Figure 6**. When we completely randomize the original network, the linearity between α_c and c has become steeper, i.e., $\alpha_c = 1 - 2c$, which is good as the critical market confidence is lower at a certain level of price limit as compared to the original case. In fact, even with randomizing of a part of the original network, the linearity of $\alpha_c = 1 - c$ would be successfully adjusted. This implies that the system resilience could be improved by regulating the whole picture of investment pattern whereas retaining part of the original investment structure. In essence, the randomizations have successfully modify the distribution of both nestedness and branching (see **Figure S10**). The nodes with high nestedness are gradually eliminated with the increase of randomization, indicating the extent of portfolio overlapping has been reduced. As a result, it would be more difficult for the driving nodes to take over failures and the system resilience would then be promoted. Similarly, the number of nodes with high branching have been significantly curtailed, implying the risk broadcasting power and the exposure to risk of driving nodes will be lowered accordingly. The adaption of the exact relationship between α_c and c through randomizing the network in **Figure 6**, in turn, highlights the importance of nestedness and branching in terms of system resilience. The exploratory experiments show that imposing the variated investment strategy on the whole or partly can thus enhance the resilience of the system. More broadly, if market participants could make a compromise between individual profit maximization and system stability enhancement, weakening the herding behavior and rethinking the over-diversified investment strategy simultaneously, the market structure will be reformed as the nestedness and branching are redistributed. As a result, the system resilience would be boosted and the price limit will work better for stabilizing the market.

The unexpected side effect of the principle of profit maximization and risk minimization in the individual level of investors is inherently missed in existing understandings of portfolios. While our results both theoretically and empirically suggest that from the view of system resilience, overlapping portfolios due to herding investments derived from this principle unintentionally forge the emergence of small yet critical stocks that drive the market to collapse. In terms of networking investors and stocks, ideas from system science can help manifest the market crash and inject new insights to the practice of market supervision. And to obtain these insights might be challenging for the classical approaches in finance. Even more importantly, the crash of stock market also offers a new testbed to examine the previous understandings of system science and surprisingly, small nodes, which are conventionally thought to be trivial in risk contagion, emerge to be the most critical parts that reignite the failure cascading and determine the system behavior at the critical phase. By lowering the disassortativeness of the system, the entanglement between small

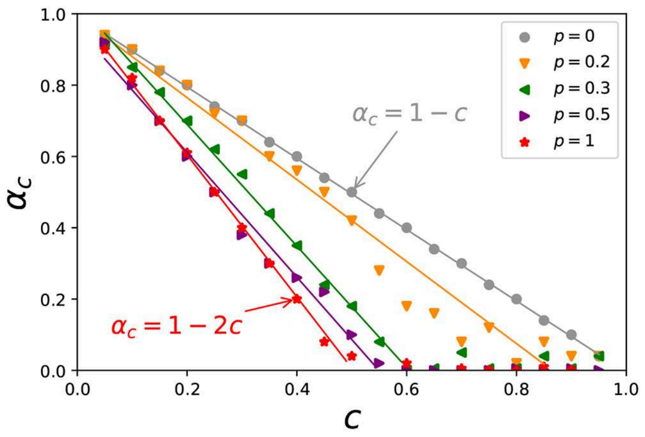

FIGURE 6 | The relationship between α_c and c when randomizing the stock-investor network. p denotes the proportion of which the edges in the original network are randomized. The results for different p are shown in different colors and marker shapes, along with lines of best fit (results with $\alpha_c \approx 0$ are neglected in the lines of best fit). For each p, we delete a proportion of p edges in the original network and generate the same amount of edges apart from the $1-p$ edges that have been kept in the network. Thus, p indicates the extent of randomization. When $p=1$, for example, random bipartite networks that possess the same amounts of nodes and edges with the real network are generated. On the contrary, when $p=0$, there is no randomization and the original network is completely kept, which will give us the results in **Figure 2A**. All the edge weights are set to be equal in the randomization experiments for simplicity. The initial attacked stocks are also randomly selected to initiate the contagion procedures and the procedures are repeated 600 times for each p. Note that the slope of the linearity between α_c and c is -2 in the most random design. The reasons are that, first, under the circumstance of a dense network, the market value of the failed stock is extremely small compared to the market value of all stocks one investors hold, so $\frac{\sum_{f\in F_\tau} w_{f,m,\tau=0}}{A_{m,\tau}}$ in formula (3) is still small (recall that it is the $\tau=0$ that matters). Second, the scheme of randomly linking investor nodes and stock nodes makes the probabilities same for investors holding the failed stocks or not holding the failed stocks, i.e., $\frac{\sum_{m\in L_\tau} \alpha w_{i,m,\tau}}{\sum_m w_{i,m,\tau=0}} = \frac{\sum_{m\notin L_\tau} w_{i,m,\tau}}{\sum_m w_{i,m,\tau=0}} = 0.5$ in formula (5), which gives $\alpha_c = 1-2c$.

but critical parts and instability of the whole system can be effectively weakened, thus leading to enhancement of system resilience. Our results may be of interest to policy markers tasked with developing regulation to promote market-wide stability and venue operators interested in designing effective trading strategies.

5. CONCLUSION

From the perspective of system science, this paper both theoretically and empirically shows the small stocks, which are conventionally thought to be trivial in risk contagion, surprisingly emerge to be the most critical parts that reignite the failure cascading from periphery to core. These stocks also result in the inefficiency of enhancing system resilience with the exclusive increment of price limit. As the emergence of these small but critical stocks stems from herding behavior of investment, imposing random investment strategy in portfolio diversification can lead to improvement of system resilience.

The paper inevitably has limits. While the contagion model provides insightful findings and well-explains market crashes in real world, how to empirically quantify the confidence factor in contagion model is a key challenge for future work. In addition, the current study has only focused on the Chinese stock market. Stock markets in other countries would also be brought into consideration in the future.

DATA AVAILABILITY STATEMENT

The datasets analyzed during the current study are available in the figshare.com repository, https://doi.org/10.6084/m9.figshare.8216582.v2.

AUTHOR CONTRIBUTIONS

JZ, SL, and HW designed the research and wrote the paper. SL and JZ performed the research and prepared the figures.

ACKNOWLEDGMENTS

This research was financially supported by National Natural Science Foundation of China (Grant Nos. 71420107025 and 71871006). This manuscript has been released as a Pre-Print at https://arxiv.org/abs/1908.07244 [31].

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphy.2020.00049/full#supplementary-material

REFERENCES

1. Stavroglou SK, Pantelous AA, Stanley HE, Zuev KM. Hidden interactions in financial markets. *Proc Natl Acad Sci USA*. (2019) **116**:10646–51. doi: 10.1073/pnas.1819449116
2. Bildik R, Gülay G. Are price limits Effective? Evidence from the istanbul stock exchange. *J Finan Res*. (2006) **29**:383–403. doi: 10.1111/j.1475-6803.2006.00185.x
3. Kim KA, Rhee SG. Price limit performance: evidence from the Tokyo stock exchange. *J Finan*. (1997) **52**:885–901. doi: 10.1111/j.1540-6261.1997.tb04827.x
4. Shen S, Goh B. *China Stock Market Freezing up as Sell-off Gathers Pace*. (2015). Reuters. Available online at: https://www.reuters.com/article/us-china-stocks-idUSKCN0PI04Q20150708
5. Gao J, Barzel B, Barabási AL. Universal resilience patterns in complex networks. *Nature*. (2016) **530**:307–12. doi: 10.1038/nature16948
6. Anderson N, Webber L, Noss J, Beale D, Crowley-Reidy L. The resilience of financial market liquidity. *Bank of England Financial Stability Paper No. 34*. (2015). Available online at: https://ssrn.com/abstract=3204599
7. Bardoscia M, Battiston S, Caccioli F, Caldarelli G. Pathways towards instability in financial networks. *Nat Commun*. (2017) **8**:14416. doi: 10.1038/ncomms14416
8. Gao YC, Tang HL, Cai SM, Gao JJ, Stanley HE. The impact of margin trading on share price evolution: a cascading failure model investigation. *Physica A*. (2018) **505**:69–76. doi: 10.1016/j.physa.2018.03.032
9. Brugler J, Linton OB, Noss J, Pedace L. The cross-sectional spillovers of single stock circuit breakers. *Bank England Working Paper No. 759*. (2018). doi: 10.2139/ssrn.3297499

10. May RM, Arinaminpathy N. Systemic risk: the dynamics of model banking systems. *J R Soci Interf*. (2010) **7**:823–38. doi: 10.1098/rsif.2009.0359
11. Arinaminpathy N, Kapadia S, May RM. Size and complexity in model financial systems. *Proc Natl. Acad Sci USA*. (2012) **109**:18338–343. doi: 10.1073/pnas.1213767109
12. Caccioli F, Shrestha M, Moore C, Farmer JD. Stability analysis of financial contagion due to overlapping portfolios. *J Bank Finan*. (2014) **46**:233–45. doi: 10.1016/j.jbankfin.2014.05.021
13. Caccioli F, Farmer JD, Foti N, Rockmore D. Overlapping portfolios, contagion, and financial stability. *J Econ Dyn Cont*. (2015) **51**:50–63. doi: 10.1016/j.jedc.2014.09.041
14. Sornette D. *Why Stock Markets Crash: Critical Events in Complex Financial Systems*. New Jersey, NJ: Princeton University Press (2017).
15. Haldane AG, May RM. Systemic risk in banking ecosystems. *Nature*. (2011) **469**:351. doi: 10.1038/nature09659
16. Morone F, Del Ferraro G, Makse HA. The K-core as a predictor of structural collapse in mutualistic ecosystems. *Nat Phys*. (2019) **15**:95. doi: 10.1038/s41567-018-0304-8
17. Stanley H, Mantegna R. *An Introduction to Econophysics*. Cambridge: Cambridge University Press (2000). doi: 10.1017/CBO9780511755767
18. Huang X, Vodenska I, Havlin S, Stanley HE. Cascading failures in bi-partite graphs: model for systemic risk propagation. *Sci Rep*. (2013) **3**:1219. doi: 10.1038/srep01591
19. Delpini D, Battiston S, Caldarelli G, Riccaboni M. Systemic risk from investment similarities. *PLoS ONE*. (2019) **14**:e0217141. doi: 10.1371/journal.pone.0217141
20. Poledna S, Martínez-Jaramillo S, Caccioli F, Thurner S. Quantification of systemic risk from overlapping portfolios in the financial system. *arXiv:180200311*. (2018).
21. Cifuentes R, Ferrucci G, Shin HS. Liquidity risk and contagion. *J Eur Econ Assoc*. (2005) **3**:556–66. doi: 10.1162/1542476054472946
22. Coval J, Stafford E. Asset fire sales (and purchases) in equity markets. *J Finan Econ*. (2007) **86**:479–512. doi: 10.1016/j.jfineco.2006.09.007
23. Motter AE, Lai YC. Cascade-based attacks on complex networks. *Phys Rev E*. (2002) **66**:065102. doi: 10.1103/PhysRevE.66.065102
24. Motter AE. Cascade control and defense in complex networks. *Phys Rev Lett*. (2004) 93:098701. doi: 10.1103/PhysRevLett.93.098701
25. Lu S, Zhao J, Wang H, Ren R. Herding boosts too-connected-to-fail risk in stock market of China. *Physica A*. (2018) **505**:945–64. doi: 10.1016/j.physa.2018.04.020
26. Tanizawa T, Paul G, Cohen R, Havlin S, Stanley HE. Optimization of network robustness to waves of targeted and random attacks. *Phys Rev E*. (2005) **71**:047101. doi: 10.1103/PhysRevE.71.047101
27. Cohen-Cole E, Kirilenko A, Patacchini E. Trading networks and liquidity provision. *J Finan Econ*. (2014) **113**:235–51. doi: 10.1016/j.jfineco.2014.04.007
28. Corsi F, Marmi S, Lillo F. When micro prudence increases macro risk: the destabilizing effects of financial innovation, leverage, and diversification. *Operat Res*. (2016) **64**:1073–88. doi: 10.1287/opre.2015.1464
29. Li X, Wang SS, Wang X. Trust and stock price crash risk: evidence from China. *J Bank Finan*. (2017) **76**:74–91. doi: 10.1016/j.jbankfin.2016.12.003
30. Kitsak M, Gallos LK, Havlin S, Liljeros F, Muchnik L, Stanley HE, et al. Identification of influential spreaders in complex networks. *Nat Phys*. (2010) **6**:888–93. doi: 10.1038/nphys1746
31. Lu S, Zhao J, Wang H. The emergence of critical stocks in market crash. *arXiv:190807244*. (2019).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Physics

ORIGINAL RESEARCH
published: 31 March 2020
doi: 10.3389/fphy.2020.00083

Diffusion Entropy and the Path Dimension of Frictional Finger Patterns

Kristian Stølevik Olsen[1]* and *James Matthew Campbell*[1,2]

[1] Department of Physics, University of Oslo, Oslo, Norway, [2] College of Engineering, Swansea University, Swansea, United Kingdom

The authors investigate, using both analytical and numerical methods, the entropy associated with a diffusion process inside frictional finger patterns. The entropy obtained from simulations of diffusion inside the pattern is compared to analytical predictions based on an effective continuum description. The analytical result predicts that the entropy depends in a particular way on the path dimension of the system, which governs the scaling of simple paths in the system. The findings indicates that there is a close analogy between the frictional fingers in the continuum and minimum spaning trees on the lattice, as the path dimension is found, through studies of the entropy, to be close to the defining value for the minimum spanning tree universality class.

Keywords: entropy, anomalous diffusion, universality, fractal geometry, frictional finger pattern, pattern analysis

OPEN ACCESS

Edited by:
Giancarlo Ruocco,
Center for Life NanoScience (IIT), Italy

Reviewed by:
Matteo Colangeli,
University of L'Aquila, Italy
Constantinos Simserides,
National and Kapodistrian University of Athens, Greece

****Correspondence:***
Kristian Stølevik Olsen
k.s.olsen@fys.uio.no

Specialty section:
This article was submitted to Condensed Matter Physics, a section of the journal Frontiers in Physics

Received: *09 January 2020*
Accepted: *09 March 2020*
Published: *31 March 2020*

Citation:
Olsen KS and Campbell JM (2020) Diffusion Entropy and the Path Dimension of Frictional Finger Patterns. Front. Phys. 8:83. doi: 10.3389/fphy.2020.00083

1. INTRODUCTION

Patterns with complex geometry and topology are ubiquitous in Nature. When transport processes take place inside such patterns, their dynamical properties are typically anomalous due to the non-trivial geometry. The case of anomalous diffusive transport, where the mean-square displacement scales non-linearly with time, has been studied in some detail since the 1980s and remains a popular topic to this day [1–9]. While many of the systems studied in this context are idealized and synthetic, like that of hierarchical fractals, real systems are noisy and often have a geometry that is too complex to be exactly captured by the simplified models. In order to utilize the simplified models, one has to identify the right set of relevant geometric properties when a scaled-up effective continuum description is used. These coarser geometric properties then determine the long-time dynamical properties of diffusing particles in the geometry, like the mean square displacement and the entropy. The entropy associated with anomalous diffusion processes has been studied in some detail in the framework of fractional diffusion equations [10–13]. We here instead consider diffusion with spatially dependent diffusivity, where the analytical form of the diffusivity is linked to the systems geometry. It is the aim of this paper to investigate the entropy for such an effective continuum description of diffusion in frictional finger patterns, and the associated insight it brings into the systems coarser geometric properties.

Fricitonal patterns are space-filling bifurcating two-dimensional geometries that arise due to instabilities in frictional fluids. Experimentally, the frictional finger patterns are produced by preparing a mixture of glass beads and liquid in a Hele-Shaw cell before pumping out of liquid from the center of the cell. When the cell has open ends, this forces air into the glass bead / liquid mixture resulting in a deformation of the boundary of the mixture [14, 15]. The deformation takes place where the energy needed to move the boundary is the smallest. A simulated version of the pattern is shown in **Figure 1A**, based on the algorithms discussed in Olsen et al. [16].

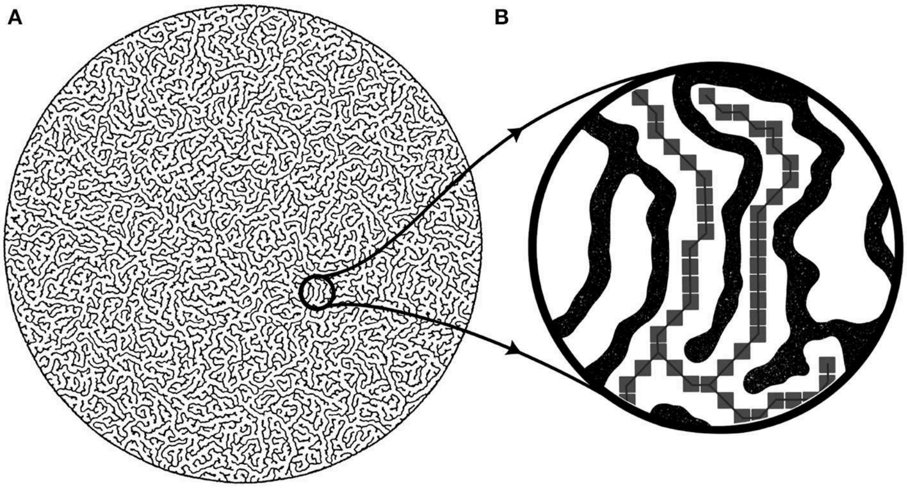

FIGURE 1 | Figure showing a frictional finger pattern **(A)** together with the one-dimensional skeletonized tree-structure obtained by contracting the finger widths to zero **(B)**. The skeletonized pattern is pixelized before random walkers are released as discussed in section 3.

It was recently hypothesized by the authors that since the frictional finger patterns are formed through a optimal path-finding process it may belong to the geometric universality class of minimum spanning trees (MST)[16]. These are tree-structures constructed on a lattice by assigning a random energy to every lattice link and finding the loopless configuration of links that globally minimizes the total system energy [17]. Universality in this context refers to the exponents reflecting geometric properties of a pattern, like the fractal dimensions d_f or minimum path dimension d_m, the latter being the scaling exponent connecting Euclidean and intrinsic distance measures. Since on loopless structures there is an unique path connecting any two points we will simply refer to d_m as the path dimension. The 2D MST class has $d_f = 2$ and $d_m = 1.22 \pm 0.01$ [17]. Direct measurements of the path dimension in the frictional fingers have proven to be difficult, due to the noisy and complex nature of the patterns. In particular, the path dimension can be estimated locally by fixing a sample point in the system and considering the average length $\overline{\ell}$ of paths out to a radius of Euclidean length r. To obtain a global "coarse-grained" estimate $\overline{d_m}$ for the path dimension of the system this local path dimension should be averaged over many sample points. This gives a path dimension of $\overline{d_m} = 1.25 \pm 0.03$ [16]. The path dimension can also be estimated by treating the pattern as a tree structure and using branching statistics, similar to the study of river networks. This method in stead gives $\overline{d_m} = 1.20 \pm 0.03$ [16]. While both of these measurements are consistent with the MST class, they are inconsistent with each other. Rather that directly measuring the path dimension we will here use the diffusing particles as a probe of the system geometry. In particular, since the entropy is a measure of how fast the diffusion process is relaxing toward equilibrium, the entropy will be a function of the systems geometry and will therefore give us some insight into the value of $\overline{d_m}$.

The rest of this paper is outlined as follows. Section 2 discusses the diffusion entropy associated with the effective continuum description, which is based on a simple power-law scaling of the diffusivity. The entropy associated with the corresponding Fokker-Planck equation is calculated analytically and compared to results obtained using fractional diffusion equations. Section 3 discusses a new numerical implementation of random walkers in frictional finger patterns that is expected to increase both efficiency and accuracy. The entropy is calculated, and compared to analytical predictions. Finding the best fit of the analytical prediction as system parameters are varied gives a value for the path dimension. Concluding remarks are offered in section 4.

2. ENTROPY OF THE EFFECTIVE CONTINUUM DESCRIPTION

To study the diffusion process in the frictional finger patterns on a large length scale we use state-dependent diffusion equations where the diffusivity can depend of the particle position. Microscopically, Brownian particles move throughout the pattern with a constant diffusivity. However, the collisions with the walls of the confining geometry affects the macroscopic transport properties. We imagine that after a sort of coarse-graining or homogenization procedure the diffusion process can be described by an overdamped Langevin equation of the form

$$\dot{x}_a = \sqrt{2D(x)}\eta_a(t) \tag{1}$$

where a is the spatial component of the vector and η is a delta-correlated white noise with $\langle \eta_a(t)\eta_b(t')\rangle = 2\delta_{ab}\delta(t-t)$. The diffusivity is here assumed to be isotropic but inhomogeneous. When going from a Langevin description on the microscopic scale to an evolution equation for the particle density on the macroscopic scale one has to decide on which stochastic calculus to be used, as discussed in the classical books by Risken [18] and Van Kampen [19]. This problem, known as the Itô-Stratonovich dilemma, results in different forms of the macroscopic equations that differ in the presence of an additional drift term proportional to the gradient of the diffusivity. Recently it has also been showed that these different form of the density evolution equation can be obtained as scaling limits of a random walk on a lattice where inhomogeneities are associated with bonds and/or vertices of the lattice [20].

In the case of diffusion in frictional finger patterns we chose the diffusion law associated with the Hänggi-Klimontovich convention, where no drift term associated with diffusivity gradients are present. This choice of diffusion law together with the Einstein relation was recently used to identify the form of the spatially dependent diffusivity for transport in the frictional fingers, where under an isotropy assumption one has $D(r) = D_0 r^{-\xi}$ [16]. As is the case for perfectly hierarchical fractals, the exponent ξ is related to the fractal dimensions of the pattern as $\xi = d_f - 2 + d_m = d_m$, where we used the space-filling property of the finger pattern [1, 16].

The corresponding density evolution equation in the Hänggi-Klimontovich interpretation takes the following form

$$\partial_t \rho(r,t) = \partial_r [rD(r)\partial_r \rho(r,t)] \tag{2}$$

This generalized Ficks equation has a well-known solution for radial power-law diffusivity, taking the form [6]

$$\rho(r,t) = \frac{2+\xi}{2\pi\Gamma\left(\frac{d_s}{2}\right)} \left[\frac{1}{D_0(2+\xi)^2 t}\right]^{\frac{d_s}{2}} \exp\left(-\frac{r^{2+\xi}}{D_0(2+\xi)^2 t}\right) \tag{3}$$

where $d_S = 4/(2+\xi)$ is the spectral dimension and the normalization used is $\int dr(2\pi r)\rho = 1$. This solution is typically thought of as a smoothened out envelope of the discrete set of probabilities associated with the vertices of a fractal, as discussed in the original paper [6]. There are several properties of this solution that are only approximately shared with the actual frictional finger system. For example, the solution is completely isotropic $\partial_\theta \rho(r,t) = 0$. Since diffusing particles will be forces to move along fingers in our pattern, we know that locally the system is very anisotropic. However, on large time and length scales, the different anisotropies are expected to cancel to produce an approximately isotropic behavior. Furthermore, the solution assumed a single globally well-defined path dimension d_m, while it is known that in noisy real systems this dimension can vary locally. Again we expect that on large space and time scales the inhomogeneities average out, producing a single global path dimension $\overline{d_m}$ as discussed in the introduction. The predicted second moment of the solution Equation (3) was tested against the mean-square displacement of random walk simulations in the pattern with reflecting boundary conditions in previous work and was seen to agree well with the simulations, adding to its validity as an effective model [16].

Given the above solution Equation (3) the entropy of the diffusion process can be calculated analytically. What type of entropy we consider is not of great importance here, as long as it is the same entropy that is calculated later in section 3 in the numerical methods. This is because at the end of the day, we are interested in using the numerical measurements of the entropy as an indirect measurement of the path dimension for the frictional fingers. From an information theoretic perspective there are dozens of entropies that could be considered, most of which can be thought of as an analytical continuation of the Shannon-Gibbs entropy which is recovered as some entropic parameter is tuned correctly [21]. We here consider the Shannon-Gibbs formula as it is not only readily calculated but also closer connected to the entropy familiar in extensive thermodynamics.

The Shannon-Gibbs entropy for the particle density takes the form [22]

$$H[\rho] = -\int dV \rho(x) \log \rho(x). \tag{4}$$

We will write Equation (3) in the form $\rho(r,t) = A(t,\xi)\exp(-r^{2+\xi}/a(t,\xi))$ for notational simplicity. According to Equation (4) we then have the entropy in terms of a non-integer moment:

$$H[\rho(t),\xi] = \frac{\langle r^{2+\xi}\rangle}{a(t,\xi)} - \log A(t,\xi). \tag{5}$$

Since our distribution is a simple shifted Gaussian a change of variables easily allows us to find this moment. Using the integral

$$\int_0^\infty dx x^\mu e^{-x^\nu/a} = \frac{a^{\frac{\mu+1}{\nu}}}{\nu}\Gamma\left(\frac{\mu+1}{\nu}\right). \tag{6}$$

With $\mu = 3+\xi$ and $\nu = 2+\xi$, the entropy can be calculated as

$$H[\rho(t),\overline{d_m}] = \frac{2}{2+\overline{d_m}}\left[1+\log\left(D_0(2+\overline{d_m})^2\right)\right] + \log\left[\pi\Gamma\left(\frac{4}{2+\overline{d_m}}\right)\right] + \frac{2}{2+\overline{d_m}}\log t, \tag{7}$$

where we used $\xi \approx \overline{d_m}$ based on the above discussion. Interestingly the associated entropy production $\dot{H} = 2/(t(2+\overline{d_m}))$ has also been obtained by using a two-dimensional diffusion equation with Caputo or Weyl fractional-time derivative [13].

In Equation (7) we see that the global path dimension $\overline{d_m}$ determines the temporal evolution of the entropy. As expected, a higher path dimension, meaning a more disordered system geometry, will give a lower entropy production since the diffusion process is more hindered. Using the same integral formula as above it is also easily shown from the solution Equation (3) that the mean-square displacement takes the form

$$\langle r^2\rangle = \frac{\Gamma(2\alpha)\,[D_0(2+\overline{d_m})^2]^\alpha}{\Gamma(\alpha)} t^\alpha \tag{8}$$

where the diffusion exponent is given by $\alpha = 2/(2+\overline{d_m})$. Note that the diffusion exponent also governs the temporal scaling of the entropy $H \sim \alpha \log t$.

3. RESULTS FROM NUMERICAL SIMULATION

To calculate the numerical entropy we construct a simplified discrete random walk-based model for the diffusion process. To

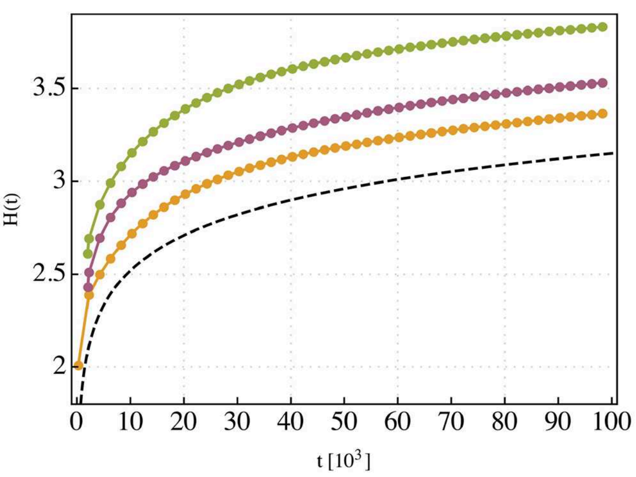

FIGURE 2 | The entropy associated with three randomly chosen initial positions inside the finger pattern. The point were chosen inside a circle of $\sim 50\%$ pattern radius in order to avoid the outer boundaries of the pattern. The dotted line is a reference line $0.62 \log_{10}(t)$.

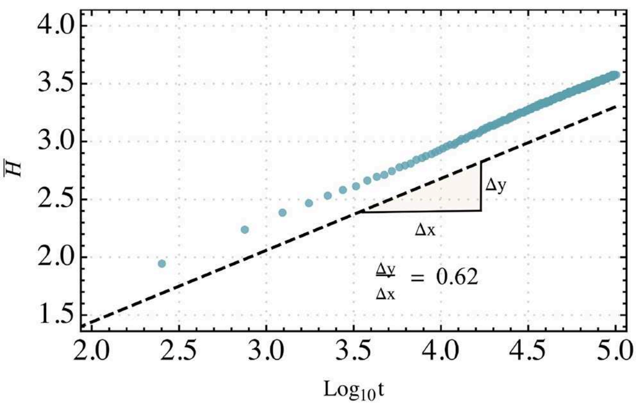

FIGURE 3 | Averaged entropy $\overline{H}$ shown in semi-log scale. The entropy grows with a linear slope of $\alpha = 0.62$. Since the entropy of the three different initial positions may vary at short times we have ignored the first 20% of the datapoints when performing the linear fit. The slope $\alpha = 0.62$ is in good agreement with the expected minimum spanning tree value.

make these simulations more efficient, we make some simplifying assumptions. The biggest simplification comes from applying a topological contraction on the pattern so that the finger widths are set zero, effectively turning the problem into a one dimensional one. The resulting skeletonized version of the pattern, showed in **Figure 1B**, is what we will release random walkers on. This topological simplification will not change the main geometric features of the pattern, since the folding and connectedness of every branch is conserved.

When performing the numerical simulations the one-dimensional skeletonized pattern is discretized before a discrete random walk process is released. In the resulting discrete "morphological graph" of the pattern there are no additional inhomogeneities associated with transition probabilities over links as all the inhomogeneity we are interested in stems from the pattern itself. In practice, the discretization is obtained by pixelating the skeletonized pattern and treating the pixels as sites for the random walker. A cartoon of the pixels are shown in **Figure 1B**. A random walker jumps to one of its neighboring pixels, including diagonal neighbors, with equal probability. Since the code is ran with a very large number of particles, we estimate the number of particles that move to a given neighboring pixel according to a binomial distribution. Hence, at every time step we only need as many random numbers as there are neighbors for a given pixel rather than one number for every particle as in traditional random walk methods.

To calculate the entropy numerically we use the Gibbs-Shannon formula for the discrete random walk

$$H_{\text{num}}(t) = -\sum_{\text{pixels } i} \rho_i(t) \log \rho_i(t) \qquad (9)$$

where $\rho_i(t)$ is the probability of finding a particle at pixel i at time t. This probability is straightforwardly calculated as the ratio of the number of particles at pixel i at time t to the total number of particles in the system

$$\rho_i(t) = \frac{n_i(t)}{N}. \qquad (10)$$

The system is initialized with all particles released at the same position, as the analytical solution assumes a Dirac delta-like initial condition. **Figure 2** shows the entropy of the simulation for three different randomly chosen initial positions close to the center of the pattern. We see that while the temporal scaling agrees, they have different zero-point entropies. By inspection of Equation (7) we see that it is possible to have the same temporal scaling but a different zero-point entropy is the diffusion constant D_0 is allowed to vary throughout the system. This may indicate that a more realistic diffusivity is $D(x, y) = D_0(x, y)r^{-\xi}$ where D_0 is a slowly varying function taking into account small inhomogeneities in the pattern not captured by the simplified power-law model.

Figure 3 shows the average entropy $\overline{H}(t) = \sum_{i=1}^{3} H_i(t)$ where i runs over the three different initial positions. The entropy shows a very convincing growth proportional to log t over several decades. The best fit for the slopes of the entropies in **Figure 2** is given by $\alpha = 0.62$. This is consistent with the global estimate of the path dimension $\overline{d_m} \approx 1.22$, which is exactly the path dimension of minimum spanning trees [17]. This value for the path dimension is consistent with the ones obtained in earlier work, although the value obtained through the entropy is much closer to the MST value [16]. This significantly strengthens our belief that the frictional finger pattern lies in the MST universality class and can be seen as a continuum analog of the lattice MST.

4. CONCLUSION

In this paper we have studied the entropy of diffusion in frictional finger patterns. In addition to being a measure of how fast the non-equilibrium process is evolving, the entropy is also considered as a tool for studying the systems coarser geometry as the diffusing particles explore the large-scale structure at late times. Our results show that the (coarse) path dimension takes a value close to $\overline{d_m} = 1.22$ which is the defining value of the minimum spanning tree universality class. This strengthens the current hypothesis that the frictional fingers belong to this class.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

KO performed analytical calculations and wrote the paper. JC developed numerical code crucial for the paper, analyzed the pattern, and aided in the writing process.

FUNDING

This work was supported by the Research Council of Norway through the Center of Excellence funding scheme, Project No. 262644(PoreLab).

ACKNOWLEDGMENTS

The authors thank Eirik G. Flekkøy, Luiza Angheluta, and Signe Kjelstrup for insightful input and discussions during this work.

REFERENCES

1. ben Avraham D, Havlin S. *Diffusion and Reactions In Fractals And Disordered Systems*. Cambridge: Cambridge University Press (2000). doi: 10.1017/CBO9780511605826
2. Bouchaud J-P, Georges A. Anomalous diffusion in disordered media: statistical mechanisms, models and physical applications. *Phys Rep*. (1990) **195**:127–293. doi: 10.1016/0370-1573(90)90099-N
3. de Gennes PG. La percolation: un concept unificateur (Percolation a unifying concept). *La Recherche*. (1976) **919**:72–82.
4. Havlin S, Djordjevic ZV, Majid I, Stanley HE, Weiss GH. *Phys Rev Lett*. (1984) **53**:178. doi: 10.1103/PhysRevLett.53.178
5. Havlin S, Kiefer JE, Weiss GH. Anomalous diffusion on a random comblike structure. *Phys Rev A*. (1987) **36**:1403. doi: 10.1103/PhysRevA.36.1403
6. O'Shaughnessy B, Procaccia I. Analytical solutions for diffusion on fractal objects. *Phys Rev Lett*. (1985) **54**:455. doi: 10.1103/PhysRevLett.54.455
7. B'nichou O, Illien P, Oshanin G, Sarracino A, Voituriez R. Diffusion and subdiffusion of interacting particles on comblike structures. *Phys Rev Lett*. (2015) **115**:220601. doi: 10.1103/PhysRevLett.115.220601
8. Tamm MV, Nazarov LI, Gavrilov AA, Chertovich AV. Anomalous diffusion in fractal globules. *Phys Rev Lett*. (2015) **114**:178102. doi: 10.1103/PhysRevLett.114.178102
9. Tan P, Liang Y, Xu Q, Mamontov E, Li J, Xing X, et al. Gradual crossover from subdiffusion to normal diffusion: a many-body effect in protein surface water. *Phys Rev Lett*. (2018) **120**:248101. doi: 10.1103/PhysRevLett.120.248101
10. Hoffmann KH, Essex C, Schulzky C. Fractional diffusion and entropy production. *J Non Equilib Thermodyn*. (1998) **23**:166–75. doi: 10.1515/jnet.1998.23.2.166
11. Essex C, Schulzky C, Franz A, Hoffmann KH. Tsallis and R'nyi entropies in fractional diffusion and entropy production. *Physica A Stat. Mech. Appl*. (2000) **284**:299–308. doi: 10.1016/S0378-4371(00)00174-6
12. Li X, Essex C, Davison M, Hoffmann KH, Schulzky C. Fractional diffusion, irreversibility and entropy. *J Non Equilib Thermodyn*. (2003) **28**:279–91. doi: 10.1515/JNETDY.2003.017
13. Aghamohammadi A, Fatollahi AH, Khorrami M, Shariati A. Entropy as a measure of diffusion. *Phys Lett A*. (2013) **377**:1677–81. doi: 10.1016/j.physleta.2013.05.015
14. Sandnes B, Knudsen HA, Måløy, KJ, Flekkøy, EG. Labyrinth patterns in confined granular-fluid systems. *Phys Rev Lett*. (2007) **99**:038001. doi: 10.1103/PhysRevLett.99.038001
15. Knudsen HA, Sandnes B, Flekkøy EG, Maåløy KJ. Granular labyrinth structures in confined geometries. *Phys Rev E*. (2008) **77**:021301. doi: 10.1103/PhysRevE.77.021301
16. Olsen KS, Flekkøy EG, Angelutha L, Campbell JM, Maåløy KJ, Sandnes B. Geometric universality and anomalous diffusion in frictional fingers. *N J Phys*. (2019) **21**:063020.
17. Dobrin R, Duxbury PM. Minimum spanning trees on random networks. *Phys Rev Lett*. (2001) **86**:5076. doi: 10.1103/PhysRevLett.86.5076
18. Risken H. *The Fokker-Planck Equation*. Berlin; Heidelberg: Springer-Verlag (1984). doi: 10.1007/978-3-642-96807-5
19. Van Kampen NG. *Stochastic Processes in Physics and Chemistry*, Vol. 1. Amsterdam: Elsevier (1992).
20. Andreucci D, Cirillo ENM, Colangeli M, Gabrielli D. Fick and fokker-planck diffusion law in inhomogeneous media. *J Stat Phys*. (2019) **174**:469. doi: 10.1007/s10955-018-2187-6
21. Rapisarda A, Thurner S, Tsallis C. Nonadditive entropies and complex systems. *Entropy*. (2019) **21**:538. doi: 10.3390/e21050538
22. Wehrl A. The many facets of entropy. *Rep Math Phys*. (1991) **30**:119–29. doi: 10.1016/0034-4877(91)90045-O

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 03 April 2020
doi: 10.3389/fphy.2020.00096

Non-linear Thermo-Optical Properties of MoS_2 Nanoflakes by Means of the Z-Scan Technique

Soghra Mirershadi[1,2†], Farhad Sattari[3†], Afshin Alipour[3†] and Seyedeh Zahra Mortazavi[4†]*

[1] Department of Engineering Sciences, Faculty of Advanced Technologies, University of Mohaghegh Ardabili, Namin, Iran, [2] Department of Engineering Sciences, Faculty of Advanced Technologies, Sabalan University of Advanced Technologies (SUAT), Namin, Iran, [3] Department of Physics, Faculty of Sciences, University of Mohaghegh Ardabili, Ardabil, Iran, [4] Physics Department, Faculty of Science, Imam Khomeini International University, Qazvin, Iran

The non-linear thermo-optical response of MoS_2 nanoflakes was investigated using the Z-scan technique, employing TM00-mode with a CW-laser diode operating at a wavelength of 532 nm. The systems were found to display a strong non-linear response, dominated by non-linear refraction. The effect of the thickness of the MoS_2 layer, deposited on a glass substrate, on the non-linear susceptibility was studied. Furthermore, in this study, the effects of modifying the thickness of the MoS_2 nanoflakes on the non-linear optical phenomena, such as self-focusing and self-defocusing was investigated for the first time. In all cases, the non-linear absorption and refraction were determined. The corresponding third-order susceptibilities and second-order hyperpolarizability were calculated to be as large as 10–7 (esu) and 10–32 (esu), under laser excitation, respectively. Showing large third-order optical non-linearity suggests the potential of the MoS_2 nanoflakes in photonics applications.

Keywords: Z-scan technique, non-linearity, refractive index, MoS_2, non-linear susceptibility, hyperpolarizability

OPEN ACCESS

Edited by:
Samit K. Ray,
Indian Institute of Technology
Kharagpur, India

Reviewed by:
Jianming Wen,
Kennesaw State University,
United States
Yilin Sun,
Tsinghua University, China

Correspondence:
Soghra Mirershadi
s.mirershadi@uma.ac.ir

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Optics and Photonics, a section of the journal Frontiers in Physics

Received: *15 January 2020*
Accepted: *13 March 2020*
Published: *03 April 2020*

Citation:
Mirershadi S, Sattari F, Alipour A and Mortazavi SZ (2020) Non-linear Thermo-Optical Properties of MoS_2 Nanoflakes by Means of the Z-Scan Technique. Front. Phys. 8:96. doi: 10.3389/fphy.2020.00096

INTRODUCTION

For the past few decades, the subjects that have stolen the spotlight of the theoretical and experimental interests are the linear and non-linear optical response of semiconductors [1, 2]. Recently, among the novel two-dimensional (2D) materials, transition metal dichalcogenides (TMDCs) with the general formula MX_2, where M refers to a transition metal and X refers to a chalcogen (S, Se, or Te), has shown particularly promising electronic and optoelectronic properties [3–5]. Molybdenum disulfide (MoS_2) is one of the most typical TMDCs. A monolayer of MoS_2 is a semiconductor with a direct bandgap of 1.8 eV [6]. This property of MoS_2 can largely compensate for the weakness of the gapless graphene, which is an essential factor in making it possible to be used in the next-generation switching, optoelectronic and photonic devices, such as optical limiters, mode-lockers, and *Q*-switchers [7, 8]. A wide range of research has been conducted on the non-linear properties of few-layer MoS_2 structures including saturable absorption, non-linear absorption and non-linear refraction [9–11].

However, realizing the photophysical properties and discovering the potential for usage of few-layer MoS_2 compounds in optical applications come hand in hand with studying the non-linear optical properties of these structures. There are various techniques for measuring the non-linear refractive index. The Z-scan is a sensitive and standard technique that offers simplicity and high sensitivity and was the technique employed for measuring the sign and the magnitude of the non-linear refractive index, as well as the non-linear absorption coefficient [12–15].

A methodical first-principles study of the second-order non-linear optical properties of the MX_2 (M =Mo, Wand X = S, Se) was carried out by Chung-Yu Wang et al. [16]. They demonstrated that the second-order non-linear optical susceptibility [$\chi^{(2)}$] of the MX_2 monolayer, over the whole range of the optical photon energy, is large and comparable to that of GaAs. Zhang et al. [17] used the Z-scan technique at various wavelengths, with femtosecond pulses. They examined the layer number and the excitation wavelength effects on the optical non-linearity of mono- and few-layers of WS_2 and MoS_2, and discovered that the monolayer WS_2 exhibited high optical non-linearities, having a two-photon absorption coefficient of $\sim 1.0 \times 10^4$ cm/GW. Moreover, Nitesh Dhasmana et al. demonstrated a dual absorption characteristic of the exfoliated MoS_2 dispersed in 1-Methyl-2-pyrrolidinone using an open aperture Z-scan technique [18]. They showed that, due to non-linear optical scattering, a saturable absorption in MoS_2, at low fluences, and a deviation from this saturable absorption, at higher fluences, can be observed. Zhang et al. experimentally verified that MoS_2 has a broadband saturable absorption response from visible to the near-infrared band [19]. They also demonstrated the first ultrafast photonics application of MoS_2 saturable absorber for passive laser mode-locking operation, and experimentally, generated nanosecond pulses. They showed that, contrary to other saturable absorbers similar to graphene, the MoS_2 saturable absorber has excellent mode-locking ability. Non-linear optical properties of WS_2 and MoS_2, obtained from both open and closed aperture Z-scan techniques using a picosecond mode-locked Nd: YAG laser operating at a wavelength of 1,064 nm, were investigated by Bikorimana et al. [20]. Both WS_2 and MoS_2 showed non-linear saturable absorption, whereas WSe_2 and $Mo_{0.5}W_{0.5}S_2$ exhibited non-linear two-photon absorption. A large two-photon absorption coefficient, β, as high as $+1.91 \times 10^{-8}$ cm/W was attained for $Mo_{0.5}W_{0.5}S_2$, and a non-linear refractive index of $n_2 = -2.47 \times 10^{-9}$ cm^2/W was determined for the WSe_2 sample.

In the previous works, the structural and optical properties of the MoS_2 nanoflakes, grown on different substrates, such as silicon and quartz, by one-step thermal chemical vapor deposition were studied [21, 22]. In addition, the Z-scan technique was employed to study the non-linear optical properties of the obtained nanoflakes [22].

Here, the non-linear thermo-optical properties of MoS_2 nanoflakes, deposited on a glass substrate with various layer thickness, using the Z-scan technique, employing CW-laser diodes operating at 532 nm wavelength, were investigated. From the open-aperture Z-scan data it was realized that these two-dimensional structures exhibit two-photon absorption. The sign and the magnitude of the third-order non-linearity, by the means of the closed aperture Z-scan, was estimated.

EXPERIMENTAL

Preparation of Materials

Thin films containing mono and few-layer MoS_2 were synthesized by one-step thermal chemical vapor deposition on Si/SiO_2 substrate. Details of the experimental procedures were described elsewhere [22]. First, in order to remove the native oxides from the silicon substrates, these substrates were soaked in hydrofluoric acid (HF) (10%). Then, by means of RF magnetron sputtering, 200 nm SiO_2 was sputtered on the substrates. Afterward, the MoO_3 powder (MERCK 99.5%, 50 mg) was collected in a ceramic boat and put in the highest-temperature thermal zone of the tube. The sulfur powder (TITRACHEM 99,0%, 500 mg) in another ceramic boat, near gas flow direction, was placed at the lower-temperature entrance of the tube (150–200°C). The substrate was placed face down on top of the boat containing MoO_3 in the hot zone. The furnace was heated from the room temperature up to 900°C within 20 min (45°C/min), in Ar gas flow (400 sccm) for 30 min. In order to start the reaction of the precursors under the Ar gas flow of 200 sccm as the carrier gas, the final temperature was kept constant for 1 h, and finally, the furnace was cooled to room temperature.

For the Z-scan measurement and the subsequent investigation of the effect of the MoS_2 layer's thickness on the non-linear optical properties, the following procedures were carried out. First, the prepared thin film was immersed in the solution of ethanol (64%) and DI water (36%) and treated by an ultrasonic homogenizer, with the power of 80 W for 15 min. The treated solution was drop cast on a glass substrate and dried in atmospheric condition. In principle, film thickness depends on the volume of the solution dispersed. In this work, films with thicknesses of 150, 230, and 420 nm (sample 1, sample 2, and sample 3, respectively) were achieved using different volumes of the solution.

CHARACTERIZATION TECHNIQUES

Structure and the Bragg reflections of the MoS_2 nanoflake were characterized by X-ray diffraction (XRD) (Cu Kα X-ray radiation source, Ital structure model MPD 3000), with 2θ within the range of 2–70°. The scanning speed and step intervals were 1°/min and 0.02°, respectively. UV–visible diffuse reflectance spectra (UV–Vis DRS) were recorded on a Sinco S4100 spectrophotometer. The thickness of the thin films of the MoS_2 structure was measured using Dektak XT Profilometer (Bruker). Scanning electron microscopy (SEM) (LEO 1430VP) and atomic force microscopy (AFM) (Nanosurf) were used to determine the morphology and roughness of the thin films of the MoS_2 nanoflake, respectively.

Non-linear Optical Measurement

The third-order non-linear optical properties of the two-dimensional MoS_2 structure were measured using the single-beam Z-scan technique. Briefly, the Z-scan technique depends on the evaluation of the transmission of a sample when it is irradiated by a focused Gaussian laser beam and its position is translated through the beam waist along the propagation direction. As the sample experiences different laser intensities at different positions, the recordings of the transmission, as a function of the Z coordinate, provide information about the third-order non-linear effects [12]. The two measurable quantities are non-linear absorption and non-linear refraction. The non-linear absorption is determined by the "open-aperture"

Z-scan, while the non-linear refraction is measured using the “closed-aperture” Z-scan. Furthermore, this technique gives the real and the imaginary parts of the third-order susceptibility, $\chi^{(3)}$, second-order hyperpolarizability, γ, and also provides important information regarding the non-linear optical properties of the materials [12]. A diode laser operating at a wavelength of 532 nm was used as the light source. The laser beam was focused using a 100 mm focal length lens onto the sample's surface. These experimental conditions produced an intensity of 1.236×10^3 Wcm^{-2} at the focal point. The laser beam radii at the focal point were measured using a CCD camera and were found to be $\omega_0 = 39.3\,\mu\text{m}$. The Raleigh length, $Z_0 = \frac{(\pi\omega_0^2)}{\lambda}$, was determined to be $Z_0 = 9.11$ mm, i.e., a value much larger than the film's thickness, which is an essential prerequisite for the Z-scan experiments. The sample was moved along the Z-axis using a translation stage. An aperture with a diameter of 2 mm was located before the photodetector placed away from the beam's focus. The transmittance of the aperture, S, was determined to be ~0.54 by $S = 1 - \exp(\frac{-2r_a^2}{\omega_a^2})$. In this equation, r_a is the radius of the aperture, and ω_a is the beam waist on the aperture. Closed aperture data were obtained by measuring the transmitted beam intensity from the sample as a function of the sample's position along the propagation direction. The measurements were repeated after removing the aperture in order to obtain the open aperture data [12]. In this work, the third-order non-linear optical parameters of the MoS_2 nanoflakes with different thicknesses were studied. In order to evaluate the effect of glass substrate to the Z-scan results, measuring the non-linearity of substrate without MoS_2 nanoflakes is performed and results show that glass substrate couldn't show measurable effect on the Z-scan results.

RESULTS AND DISCUSSION

Structural Studies

The structural properties of the nanoflakes were investigated by XRD analysis. The X-ray diffraction pattern of the MoS_2 nanoflakes is shown in **Figure 1**. The main peak at $2\theta = 14.4°$ is attributed to the (003) plane of 3R-MoS_2. The other corresponded peaks observed at $2\theta = 44.2°$ and 58.2° are related to the (009) and (110) planes, respectively (RefCode: 01–077-0341). Furthermore, the peak at $2\theta = 37.3°$ corresponds to the (200) plane of MoO_2, while another peak at $2\theta = 18.4°$ is attributed to the (100) plane of MoO_2 (RefCode: 01–086-0135).

SEM micrographs of the MoS_2 nanoflakes, with different magnifications, are shown in **Figures 2A,B**. These images show the nanoflakes are grown with sharp edges and are in great abundance. On the other hand, AFM is applied in order to characterize the topography and surface quality of the prepared samples. **Figures 2C,D** represent the phase profile of the MoS_2 film and the corresponding height profile as typical. It indicates that the MoS_2 film is continuous and the average roughness is about 80 nm. This sample's surface roughness or optical path is appropriate for z-scan measurement.”

Furthermore, the optical diffuse reflectance spectra were used to calculate the values of the optical band gap energies for the MoS_2 nanoflakes. **Figure 3A** presents the diffuse reflectance spectra of the MoS_2 nanoflakes as a function of the wavelength. The results confirmed that the valence to conduction band transport of electrons, as a result of the absorption of the incident photon, leads to the reduction of the intensity of light at the relevant wavelength. Consequently, the relative percentage of the transmission to reflectance is diminished [23]. As it can be seen in **Figure 3A**, for the MoS_2 nanoflakes, by reducing the wavelength of the incident photons from 493 to 354 nm, the reflectance is decreased.

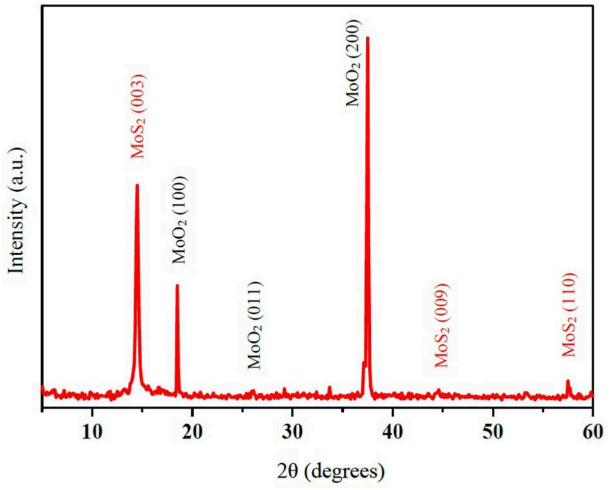

FIGURE 1 | X-ray diffraction pattern of the MoS_2 nanoflakes.

The band gap energy was calculated using a reflectance technique by the Mott and Davis theory [24]. Experimental values for the bandgap were attained by extrapolating the linear region of the curves to the zero absorption at $(\alpha h\nu)^2 = 0$, for the direct allowed transitions [12, 23]. **Figure 3B** displays the plot of $(\alpha h\nu)^2$ as a function of the incident photon energy ($h\nu$), for the direct allowed transitions of the MoS_2 nanoflakes. The optical band gap for the MoS_2 nanoflakes is calculated to be 1.91 eV.

Non-linear Optical Studies

In **Figure 4**, the closed aperture Z-scan curves of the MoS_2 nanoflakes are presented. In order to investigate the effect of samples' thickness on the non-linear optical properties, films with different thicknesses of 150, 230, and 420 nm (sample 1, sample 2, and sample 3, respectively) were used. As can be seen, both films of the MoS_2 nanoflakes with thicknesses of 150 and 230 nm exhibit a pre-focal peak followed by a post-focal valley. Thus, inferring a negative value of the non-linear refractive index, indicating a self-defocusing behavior, for these samples [25, 26]. As the thickness of the film of the MoS_2 nanoflakes reaches to 420 nm, a pre-focal valley and a post-focal peak represent a positive value for the non-linear refractive index that signifies a self-focusing behavior in this sample.

Using the theoretical fit, via the experimental data of the closed aperture Z-scan curve, $\Delta T_{P-V} = T_P - T_V$ can be obtained. After the ΔT_{P-V} is derived from the difference of the transmittance

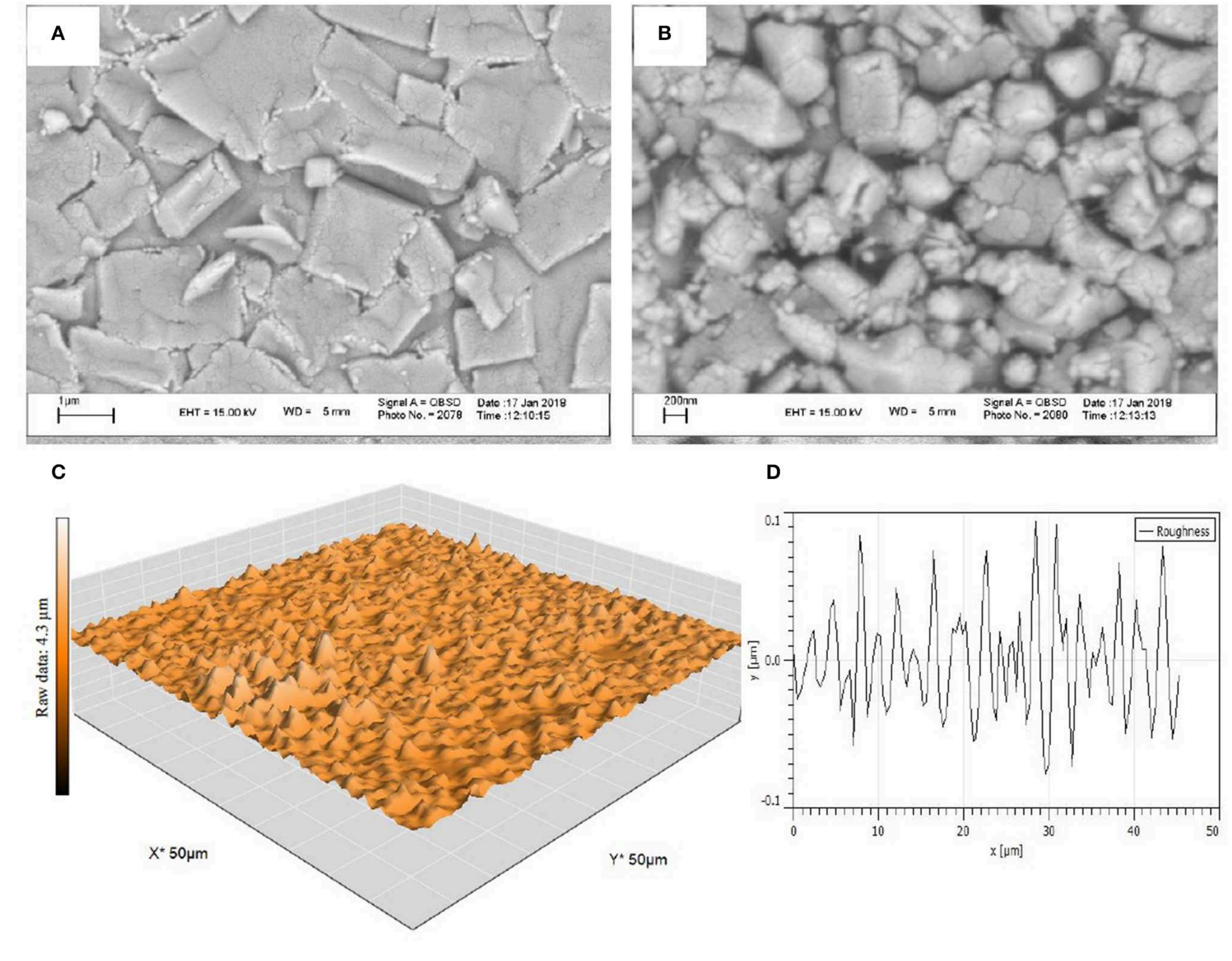

FIGURE 2 | (A,B) SEM micrographs of the MoS_2 nanoflakes with different magnificence. **(C)** Phase profile of the MoS_2 film and **(D)** the corresponding height profile as typical.

between the peak and valley, the magnitude of the phase shift $|\Delta\varphi_0|$ is given by:

$$|\Delta\varphi_0| = \frac{\Delta T_{P-V}}{0.406\,(1-S)^{0.27}} \tag{1}$$

Here, S is the fraction of the beam transmitted through the aperture [25]. The non-linear refractive index, n_2, can be calculated as:

$$n_2 = \frac{|\Delta\varphi_0|}{\left(\frac{2\pi}{\lambda}\right) I_0 L_{eff}} \tag{2}$$

Where I_0 is the peak on-axis irradiance at the focal point, and the effective thickness can be expressed by the following form:

$$L_{eff} = \frac{\left(1 - e^{-\boldsymbol{\alpha_0 l}}\right)}{\boldsymbol{\alpha_0}} \tag{3}$$

Where the linear absorption coefficient is $\alpha_0 = \frac{-1}{l}\ln\left(\frac{I}{I_0}\right)$, in which l is the thickness of the thin film, I and I_0 are the transmitted and the incident intensities, respectively [12, 25]. The results of the calculations are tabulated in **Table 1**. The linear absorption coefficient was measured through the conventional method, in the linear regime of the experiment. By changing the thickness of the film from 150 to 230 and 420 nm (sample 1, sample 2, and sample 3, respectively), the linear absorption coefficient varies from 9.51×10^4 cm^{-1} to 6.91×10^4 cm^{-1} and 4.21×10^4 cm^{-1}, respectively. The closed aperture Z-scan results showed that the non-linear refractive index decreases substantially in the MoS_2 nanoflakes as the thickness of the films is increased from 150 to 230 nm. By further increasing the thickness of the film up to 420 nm, the sign of the non-linear refractive index changes from negative to positive, and also the magnitude of the non-linear refractive index increases to 3.92×10^{-3} (cm^2/W). It is worth mentioning that a CW laser was used and so a thermal effect is also accompanied by the non-linear

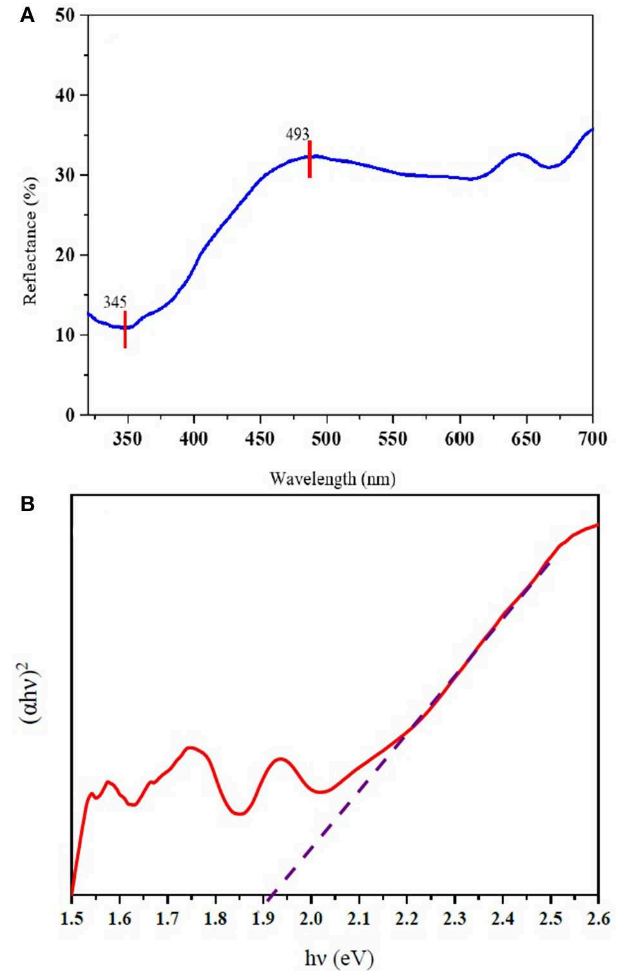

FIGURE 3 | (A) Diffuse reflectance spectra of the MoS_2 nanoflakes. **(B)** Tauc's plot of $(\alpha h\nu)^2$ as a function of photon energy (hν) for the MoS_2 nanoflakes.

optical properties. This fact induces an enhancement in the obtained non-linear refractive index. So obviously, the thermo-optical non-linear response of the nanoflakes was investigated.

Recently, Jiang et al. investigated the role of the dipole moment in characterizing the non-linear optical behavior of materials [27]. To better understand the relevance between the dipole moments and the NLO responses, a flexible dipole model was suggested. This model clearly confirmed that the induced dipole oscillations via the external field, rather than the intrinsic dipole moments, determine the NLO responses. Therefore, in order to attain a large SHG effect, the focus should be on the non-centrosymmetric crystals containing flexible chemical bonds and not merely on the crystals with large polarity, since for the non-centrosymmetric crystals, the microscopic second-order susceptibility of the relevant active groups is additively superposed. So, the large third-order non-linear optical susceptibility in the MoS_2 nanoflakes could depend on the intrinsic dipole moments in the centrosymmetric crystal. On the other hand, the open-aperture Z-scan curves of the

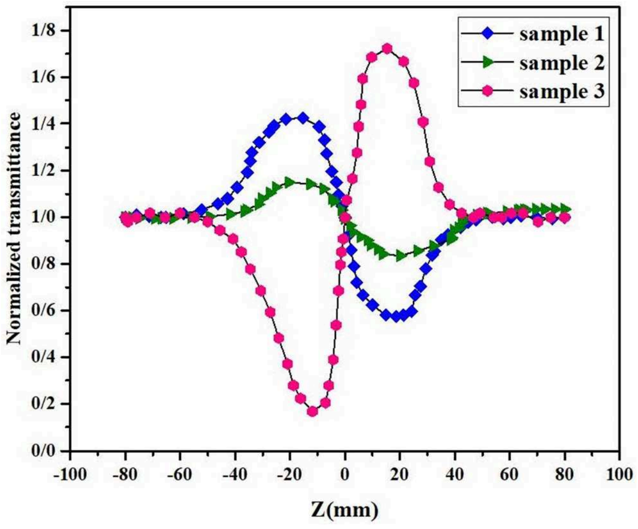

FIGURE 4 | Closed aperture Z-scan traces of MoS_2 nanoflakes with various thickness: 150 nm (sample 1), 230 nm (sample 2), and 420 nm (sample 3).

MoS_2 nanoflakes, with three different thicknesses (150, 230, and 420 nm) in **Figure 5**, show a transmission minimum, when the samples reach the focal point. A normalized transmittance valley, indicating the existence of a two-photon absorption behavior, corresponding to positive sign non-linear absorption ($\beta > 0$).

The saturable absorption coefficient, β, can be determined by:

$$\beta = \frac{q_0}{I_0 L_{eff}} \tag{4}$$

q_0 can be obtained by fitting the experimental plots with the theoretical plots, where, the normalized change in the transmitted intensity ($\Delta T(Z) = T(Z)$-1) can be approximated by the following equation:

$$\Delta T\,(Z) \approx \frac{q_0}{2\sqrt{2}}\left[\frac{1}{1+\frac{Z^2}{Z_0^2}}\right] \tag{5}$$

Van Stryland et al. [25] the obtained values of the non-linear absorption coefficient, β, for the MoS_2 layers with different thicknesses are shown in **Table 1**.

Interestingly, the non-linear absorption coefficient increases substantially by augmenting the thickness of the MoS_2 layers.

The real and the imaginary parts of the third-order susceptibility, $\chi^{(3)}$, can be given in the following forms [28]:

$$Re\,\chi^3\,(esu) = \left(\frac{10^{-4}\varepsilon_0 c^2 n_0^2}{\pi}\right) n_2\left(\frac{cm^2}{W}\right), \tag{6}$$

and

$$Im\chi^3\,(esu) = \left(\frac{10^{-2}\varepsilon_0 c^2 n_0^2 \lambda}{4\pi^2}\right)\beta\left(\frac{cm}{W}\right). \tag{7}$$

Where, ε_0 and c describe the vacuum permittivity and the speed of light in vacuum, respectively. Based on this theory, the

TABLE 1 | Non-linear optical parameters of the MoS_2 layers with different thicknesses.

MoS_2 layer thickness	n_2 (cm^2/W)	β (cm^2/W)	$\chi^{(3)}$ (m^2/v^2)	$Im\chi^{(3)}$ (esu)	$Re\chi^{(3)}$ (esu)	γ (esu)
150 nm	-2.3×10^{-3}	63.07	0.18×10^{-7}	15.44×10^{-2}	1.32	1.50×10^{-32}
230 nm	-0.9×10^{-3}	72.86	0.07×10^{-7}	17.80×10^{-2}	0.51	0.61×10^{-32}
420 nm	3.9×10^{-3}	80.53	0.31×10^{-7}	19.68×10^{-2}	2.25	2.5×10^{-32}

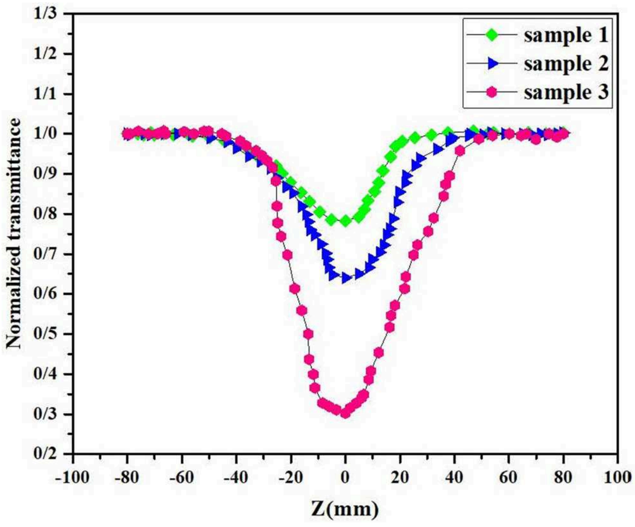

FIGURE 5 | Open aperture Z-scan traces of MoS_2 nanoflakes with various thickness: 150 nm (sample 1), 230 nm (sample 2), and 420 nm (sample 3).

results of the experiments are used to calculate the third-order susceptibility of the MoS_2 layers with different thicknesses. The results of these calculations are also presented in **Table 1**.

Finally, the corresponding second-order hyperpolarizability, γ, (i.e., the polarizability per molecule) is defined as:

$$\gamma = \frac{\chi^{(3)}}{NL^4} \tag{8}$$

Where N is the number of molecules per unit volume, and the local field correction factor, L, can be given by [29].

$$L = \frac{n_0^2+2}{3} \tag{9}$$

The values of γ, for the MoS_2 nanoflakes, are given in **Table 1**. The obtained values of $\chi^{(3)}$ and γ show that these two-dimensional structures exhibit large second-order hyperpolarizability and signify their potential applications as NLO materials. Our study concluded that these two-dimensional materials are potential candidates for non-linear optical applications.

CONCLUSION

In conclusion, in this study, the non-linear thermo-optical response of the MoS_2 nanoflakes was investigated by means of the Z-scan technique. It was found that by increasing the thickness of the MoS_2 layers, the non-linear refractive index is increased, and the TPA process can occur with high probability, leading to the focusing of the Gaussian beam. Also, the MoS_2 nanoflakes were found to exhibit very large values of third-order non-linear susceptibility. Furthermore, the MoS_2 nanoflakes presented a self-defocusing behavior (i.e., negative non-linear refraction) in low thicknesses and a self-focusing behavior (positive non-linear refraction) in high thicknesses. Also, these nanoflakes showed two-photon absorption in all samples. Thus, the magnitude and the sign of the non-linear refractive index can be tuned by modifying the thickness of the MoS_2 nanoflakes. Also, the strong non-linear optical properties observed in these samples can be attributed to the intrinsic dipole moments in the centrosymmetric crystal. It has been demonstrated that these two-dimensional MoS_2 structures provide a suitable medium to observe and quantify two-photon absorption. It is concluded that the MoS_2 nanoflakes development, as non-linear optical materials and devices, is a new and exciting opportunity.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/supplementary material.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

REFERENCES

1. Irimpan L, Krishnan B, Nampoori VPN, Radhakrishnan PJ. Luminescence tuning and enhanced nonlinear optical properties of nanocomposites of ZnO–TiO_2. *Colloid Interface Sci.* (2008) **324**:99–104. doi: 10.1016/j.jcis.2008.04.056

2. El Ouazzani H, Dabos-Seignon S, Gindre D, Iliopoulos K, Todorova M, Bakalska R, et al. Novel styrylquinolinium dye thin films deposited by pulsed laser deposition for nonlinear optical applications. *J Phys Chem.* (2012) **116**:7144–52. doi: 10.1021/jp2118218

3. Lee HS, Min SW, Chang YG, Park MK, Nam T, Kim H, et al. MoS_2 nanosheet phototransistors with thickness-modulated optical energy gap. *Nano Lett.* (2012) **12**:3695–700. doi: 10.1021/nl301485q
4. Yin Z, Li H, Jiang L, Shi Y, Sun Y, Lu G, et al. Single-layer MoS2 phototransistors. *ACS Nano.* (2011) **6**:74–80. doi: 10.1021/nn2024557
5. Dashora A, Ahuja U, Venugopalan K. Electronic and optical properties of MoS_2 thin films: feasibility for solar cells. *Comput Mater Sci.* (2013) **69**:216–21. doi: 10.1016/j.commatsci.2012.11.062
6. Ganatra R, Zhang Q. Few-layer MoS_2: a promising layered semiconductor. *ACS Nano.* (2014) **8**:4074–99. doi: 10.1021/nn405938z
7. Bao QL, Zhang H, Wang Y, Ni Z, Yan Y, Shen ZX, et al. Atomic-layer graphene as a saturable absorber for ultrafast pulsed lasers. *Adv Funct Mater.* (2009) **19**:3077–83. doi: 10.1002/adfm.200901007
8. Zhang H, Tang DY, Zhao LM, Bao QL, Loh KP. Large energy mode locking of an erbium-doped fiber laser with atomic layer graphene. *Opt Exp.* (2009) **17**:17630–5. doi: 10.1364/OE.17.017630
9. Li P, Liang B, Su M, Zhang Y, Zhao Y, Zhang M, et al. 980-nm Q-switched photonic crystal fiber laser by MoS_2 saturable absorber. *Appl Phys*. (2016) **122**:150. doi: 10.1007/s00340-016-6433-9
10. Khudyakov DV, Borodkin AA, Lobach AS, Mazin DD, Vartapetov SK. Optical nonlinear absorption of a few-layer MoS_2 under green femtosecond excitation. *Appl Phys B*. (2019) **125**:73. doi: 10.1007/s00340-019-7167-2
11. Li Y, Dong N, Zhang S, Zhang X, Feng Y, Wang K, et al. Two photon absorption in monolayer MoS_2. *Laser Photonics Rev.* (2015) **9**:427–34. doi: 10.1002/lpor.201500052
12. Mirershadi S, Ahmadi-Kandjani S, Zawadzka A, Rouhbakhsh H, Sahraoui B. Third order nonlinear optical properties of organometal halide perovskite by means of the Z-scan technique. *Chem Phys Lett.* (2016) **674**:7–13. doi: 10.1016/j.cplett.2016.01.04
13. Abdallah B, Zidan MD, Allahham A. Deposition of ZnS films by RF magnetron sputtering: structural and optical properties using Z-scan technique. *Int J Mod Phys.* (2019) **33**:1950348. doi: 10.1142/S021797921950348X
14. Majlesara MH, Dehghani Z. (2012). Measurement of nonlinear responses and optical limiting behavior of Tio_2/Ps nano-composite by single beam technique with different incident intensities. *Int J Mod Phys B*. **5**:277–83. doi: 10.1142/S2010194512002139
15. Arun Gaur DK, Sharma KS, Singh N. Photoexcited carrier lifetime in direct and indirect band gap crystals on the Z-Scan technique at 532 nm. *Int J Mod Phys.* (2007) **21**:3029–34. doi: 10.1142/S021797920703751X
16. Chung-Yu W, Guang-Yu G. Nonlinear optical properties of transition-metal dichalcogenide MX2 (M = Mo, W; X = S. Se) monolayers and trilayers from first-principles calculations. *J Phys Chem C*. (2015) **119**:13268–76. doi: 10.1021/acs.jpcc.5b01866
17. Zhang S, Dong N, McEvoy N, O'Brien M, Winters S, Berner NC, et al. Direct observation of degenerate two-photon absorption and its saturation in WS2 and MoS_2 monolayer and few-layer films. *ACS Nano.* (2015) **9**:7142–50. doi: 10.1021/acsnano.5b03480
18. Dhasmana N, Fadil D, Kaul AB, Thomas J. Investigation of nonlinear optical properties of exfoliated MoS2 using Photoacoustic Zscan. *MRS Adv.* (2016) **1**:1–7. doi: 10.1557/adv.2016.456
19. Zhang H, Lu SB, Zheng J, Du J, Wen SC, Tang DY, et al. Molybdenum disulfide (MoS2) as abroadband saturable absorber for ultra-fast photonics. *Opt Exp.* (2014) **22**:7249–60. doi: 10.1364/OE.22.007249
20. Bikorimana S, Lama P, Walser A, Dorsinville R, Anghel S, Mitioglu A, et al. Nonlinear optical responses in two-dimensional transition metal dichalcogenide multilayer: WS2, WSe2, MoS2 and Mo0.5W0.5S2. *Opt Exp.* (2016) **24**:20685–95. doi: 10.1364/OE.24.020685
21. Bayesteh S, Mortazavi SZ, Reyhani A. Investigation on nonlinear optical properties of MoS2 nanoflakes grown on silicon and quartz substrates. *J Phys D.* (2018) **51**:195302–10. doi: 10.1088/1361-6463/aab808
22. Bayesteh S, Mortazavi SZ, Reyhani A. Role of precursors' ratio for growth of two-dimensional MoS2 structure and investigation on its nonlinear optical properties. *Thin Solid Films.* (2018) **663**:37–43. doi: 10.1016/j.tsf.2018.08.013
23. Mirershadi S, Sattari F, Golghasemi Sorkhabi S, Shokri AM. Pressure-induced optical band gap transition in methylammonium lead halide perovskites. *J Phys Chem*. (2019) **123**:12423–8. doi: 10.1021/acs.jpcc.9b02744
24. Mott NF, Davis EA. *Electronic Processes in Non-Crystalline Materials*. Oxford: Clarendon (1979).
25. Van Stryland EW, Sheik-Bahae M, Kuzyk MG, Dirk CW. *Z-Scan Measurements of Optical Nonlinearities Characterization Techniques and Tabulations for Organic Nonlinear Materials*. Marcel Dekker, Inc (1998). p. 655–692.
26. Joshi JH, Kalainathan S, Kanchan DK, Joshi MJ, Parikh KD. Effect of l-threonine on growth and properties of ammonium dihydrogen phosphate crystal. *Arab J Chem.* (2020) **13**:1532–50. doi: 10.1016/j.arabjc.2017.12.005
27. Jiang X, Zhao S, Lin Z, Luo J, Bristwe PD, Guan X, et al. The role of dipole moment in determining the nonlinear optical behavior of materials: ab initio studies on quaternary molybdenum tellurite crystals. *J Mater Chem*. (2014) **2**:530–7. doi: 10.1039/C3TC31872A
28. Golian Y, Dorranian D. Effect of thickness on the optical nonlinearity of gold colloidal nanoparticles prepared by laser ablation. *Opt Quant Electron.* (2014) **46**:809–19. doi: 10.1007/s11082-013-9792-z
29. Couris S, Koudoumas E, Ruth AA, Leach SJ. Concentration and wavelength dependence of the effective third order susceptibility and optical limiting of C60 in Toluene. *Phys At Mol Opt Phys.* (1995) **2**:4537–54. doi: 10.1088/0953-4075/28/20/015

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 03 April 2020
doi: 10.3389/fphy.2020.00060

Onsager-Symmetry Obeyed in Athermal Mesoscopic Systems: Two-Phase Flow in Porous Media

Mathias Winkler[1]*, *Magnus Aa. Gjennestad*[1], *Dick Bedeaux*[2], *Signe Kjelstrup*[2], *Raffaela Cabriolu*[3] and *Alex Hansen*[1]

[1] *PoreLab, Department of Physics, Norwegian University of Science and Technology, Trondheim, Norway,* [2] *PoreLab, Department of Chemistry, Norwegian University of Science and Technology, Trondheim, Norway,* [3] *Department of Physics, Universita' degli Studi di Cagliari, Monserrato, Italy*

We compute the fluid flow time correlation functions of incompressible, immiscible two-phase flow in porous media using a 2D network model. Given a properly chosen representative elementary volume, the flow rate distributions are Gaussian, and the integrals of time correlation functions of the flows are found to converge to a finite value. The integrated cross-correlations become symmetric, obeying Onsager's reciprocal relations. These findings support the proposal of a non-equilibrium thermodynamic description for two-phase flow in porous media.

Keywords: non-equilibrium thermodynamics, immiscible two-phase-flow, porous media, network model, fluctuations

OPEN ACCESS

Edited by:
Antonio F. Miguel,
University of Evora, Portugal

Reviewed by:
Steffen Berg,
Shell, Netherlands
Aydin Murat,
Ontario Tech University, Canada

****Correspondence:***
Mathias Winkler
mathias.winkler@ntnu.no

Specialty section:
This article was submitted to Interdisciplinary Physics, a section of the journal Frontiers in Physics

Received: *13 January 2020*
Accepted: *26 February 2020*
Published: *03 April 2020*

Citation:
Winkler M, Gjennestad MA, Bedeaux D, Kjelstrup S, Cabriolu R and Hansen A (2020) Onsager-Symmetry Obeyed in Athermal Mesoscopic Systems: Two-Phase Flow in Porous Media. Front. Phys. 8:60. doi: 10.3389/fphy.2020.00060

1. INTRODUCTION

Athermal fluctuations occur in a number of phenomena in nature and are important to biology, chemistry, and physics [1–3]. Currently, an active effort is being made to better understand the statistical physics of such systems, and its use is realized in a growing number of research areas [1, 3–8]. A particular example is granular materials, the constituents of which are macroscopic. In the absence of an external driving force, the material will stay in its current configuration, sharing some properties with non-ergodic systems [3]. However, when a granular material is exposed to an external force, a great number of states may be visited, resulting in solid- or fluid-like behavior as a response to that force.

One area that seems not to have been analyzed in these terms is flow driven through porous media. Such flows are important for numerous geological and technical processes, for example, in oil production, CO_2 sequestration, water transport in aquifers, or heterogeneous catalysis. An important class of porous media flows is the simultaneous flow of two immiscible fluids. In such a system, clusters of the two fluid phases, traveling through the porous medium, are constantly forced to split and recombine. Thus, the fluid configuration in the pore space changes, leading to fluctuations in the flow rate of each phase (fractional flow rate), as well as in the total flow rate. These fluctuations are of a mechanical nature and are different from but analogous to thermal fluctuations on the molecular level. The fluctuations appear on a mesoscopic scale much larger than the molecular scale of statistical thermodynamics, yet the mesoscopic scale that is defined by the pore sizes of the medium is very small compared to the overall system. In the most extreme cases, the pores can be in the nanosize regime, while the system of interest, for instance, in chalk oil reservoirs [9], has geological dimensions.

Our long-term aim is to find a non-equilibrium thermodynamic description for such flow systems. The challenge is then to define a suitable representative elementary volume (REV) where the essential assumption of local equilibrium, as expressed by the ergodic hypothesis, and microscopic reversibility hold. The statistical foundation of the theory was spelled out a long time ago [10]. Experimental [11] and computational [12, 13] evidence now exist that the ergodic hypothesis can be expected to hold for immiscible two-phase flow in two-dimensional porous media of a minimum number of links.

Here, the aim is to move one step forward and examine the idea of time-reversal invariance or the microscopic reversibility of fluctuations [10, 14]. Thus, our interest is in the time correlation functions of the flows. On the molecular scale, thermal fluctuations have correlation functions that are connected to transport coefficients. This is formulated in the Green–Kubo relations, which are frequently employed in molecular dynamic simulations. The method goes back to Onsager's regression hypothesis [15, 16], which says that the decay of molecular fluctuations is governed by the same laws as the relaxation of macroscopic non-equilibrium disturbances. For the Onsager reciprocal relations [15, 16] to apply, microscopic reversibility must hold. The idea of the present work is to apply Green–Kubo-like relations to the fluctuations in an REV of a network model. The Green–Kubo relations for the molecular level apply to global equilibrium. In the present approach, we will use similar expressions but for fluctuations on the mesoscale, thereby extending or expanding the Green–Kubo-scheme. We shall see that the system models a time reversal-invariant process and that the integrated flux correlations satisfy Onsager symmetry. A similar approach to fluctuations in hydrodynamic dispersion was taken by Flekkøy et al. [17].

In the present study, we use a dynamic pore network model to simulate immiscible two-phase flow in porous media. This model, introduced in the late nineties [18], was developed for and calibrated against the bead-filled Hele-Shaw cell used for experimentally studying such flow. For example, the pressure fluctuations in the dynamic network model were shown to match very well with those observed in the experimental system [19]. A recent comparison between model and experiment may be found in Zhao et al. [20]. In the present implementation of the model, we use a network based on a hexagonal lattice, where the links represent pore throats and have a spatially uncorrelated distribution of link radii. The steady flow of two incompressible and immiscible fluids is driven by a constant pressure difference across the network, leading to a steady state with fluctuating fluid flow. The flow properties fluctuate around well-defined values, and the system is in a non-equilibrium steady state on the network level of description. A correlation of the two flows is thus unavoidable. But what is the nature of this correlation? The answer will have an impact on how we may proceed with a theoretical description of the flows.

A crucial question that needs to be answered is whether the dynamic network model is ergodic or not. Flekkøy and Pride [21] argue that immiscible two-phase flow at the pore scale is only ergodic if the interfaces do not move or move very little. However, at the molecular scale, there is ergodicity. Hence, ergodicity may be lost through the way the system is described. When the pores are assembled into a porous medium so that the motion of the fluids is a collective phenomenon, Savani et al. argue and demonstrate, based on the dynamic pore network model, that ergodicity is restored [12, 13]. This was done by rewriting time averages for the dynamic network model as configurational averages and thereby constructing the ensemble probability function explicitly.

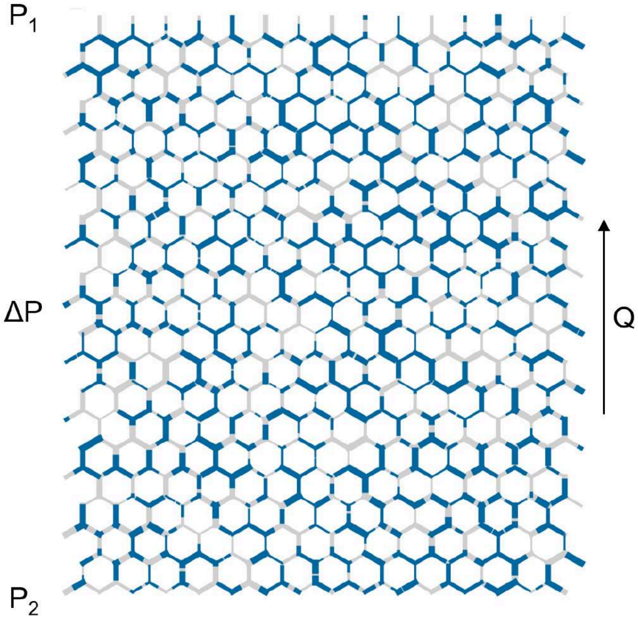

FIGURE 1 | Illustration of the network model. The network is occupied by two immiscible fluids (blue and gray colors). The equally long links (pores) have an hourglass-shaped structure and a random distribution of diameters.

If one considers a steady state of immiscible two-phase flow, as we will in our model, we shall see that the fluctuations become Gaussian around a steady mean.

Hence, the concept of the REV at steady state is highly relevant and important for how we build a theory that can help us understand transport in porous media.

2. MODEL

The transport of the two immiscible fluids through a two-dimensional porous medium is represented by a dynamical pore network model [18, 22]. This model has been in development for over two decades and has a record of explaining experimental and theoretical results in steady and transient two-phase flow in porous media [11–13, 18, 20, 22–25]. In this model, the two fluids are separated by interfaces and are flowing in a network of links that are connected at nodes. The network has a honeycomb structure, as illustrated in **Figure 1**, with equally long links and a distribution of link radii. The radii are drawn from a uniform distribution in the interval 0.1 to 0.4 L, where L is the length of the links. The flow rate q_{ij} inside a link connecting nodes j and i is given by:

$$q_{ij} = -g_{ij}[p_j - p_i - c_{ij}(\mathbf{z_{ij}})]. \tag{1}$$

Here, p_j and p_i are the pressures at nodes j and i, c_{ij} is the capillary pressure, and g_{ij} is the conductivity of the link. The links have an hourglass shape, and thus the capillary pressure is a function of the interface positions, $\mathbf{z_{ij}}$:

$$c_{ij}(\mathbf{z_{ij}}) = \frac{2\gamma}{r_{ij}} \sum_{z \in \mathbf{z_{ij}}} (\pm 1) \left\{1 - \cos(2\pi\chi(z))\right\}. \quad (2)$$

Here, γ is the surface tension, r_{ij} is the radius of link ij, and

$$\chi(z) = \begin{cases} 0, & \text{if } z < \beta r_{ij}, \\ \frac{z-\beta r_{ij}}{L-2\beta r_{ij}}, & \text{if } \beta r_{ij} < z < L - \beta r_{ij}, \\ 1, & \text{if } z > L - \beta r_{ij}. \end{cases} \quad (3)$$

The effect of the χ-function is to introduce zones of length βr_{ij} at each end of the links where the pressure discontinuity of any interface is zero. The conductivity of the link, g_{ij}, contains a geometrical factor and the effective viscosity of the link:

$$g_{ij} = \frac{\pi r_{ij}^4}{8L\mu(S_{w,ij})}. \quad (4)$$

Here, r_{ij} is the radius of the link, and the viscosity is defined as

$$\mu(S_{w,ij}) = S_{w,ij}\mu_w + (1 - S_{w,ij})\mu_n, \quad (5)$$

with $S_{w,ij}$ being the saturation (i.e., the volume fraction) of the wetting phase in link ij. Simulations were carried out, applying a constant global pressure drop ΔP across the network. Periodic boundary conditions are used in all directions. The local pressures p_i are determined by solving the Kirchhoff equations. Further details of the model and solution methods can be found in Sinha et al. [22] and Gjennestad et al. [24]. For each link, the flow rate q_{ij} is calculated using Equation (1), and the positions of the interfaces are advanced with an appropriately small, constant time step of 10^{-5} s. A constant time step is used to facilitate a convenient calculation of the time-correlation functions. The simulations were started with a random distribution of the two liquid phases and were propagated at least 300,000 time steps to allow the system to reach a steady state. Statistics for the time correlation functions were collected for 9.7 million time steps. The length of a link in the network was set to 1 mm. We report results for each set of parameters as averages of at least 30 runs using different starting configurations of the two fluids. Volume flow rates and velocities refer to network averages. The properties of the steady-state flow are determined by the pressure drop across the system, ΔP, the total wetting saturation of the network, the surface tension, γ, and the viscosity of the two fluids. We have ensured that the same steady-state flow averages are obtained from different initial distributions of the two fluids, including an initial configuration where the two phases are completely separated (i.e., each phase is in a single connected cluster). Furthermore, it has been tested that a different link radii configuration drawn from the same uniform distribution does not change the steady flow averages nor the appearance of the time correlation functions computed in this study.

We investigated time correlation functions and their long-time convergence for two choices of the parameter set μ_w, μ_n, and γ, which are viscosities of the wetting and the non-wetting phase and the surface tension, respectively.

Case (A) had viscosities $\mu_w = \mu_n = 10^{-3}$ Pa·s and different choices for the pressure gradient (100–200 kPa/m) and the surface tension γ = 3–30 mN/m. At steady state, the capillary number was Ca $\approx 10^{-3}$ – 10^{-2}. Here, the capillary number is defined as Ca $= v\mu_n/\gamma$, with v being the seepage velocity. The case was chosen to represent flow where the two fluids are interchangeable with respect to their viscous dissipation.

Case (B) had $\mu_w = 5\mu_n$ ($\mu_w = 10^{-3}$ Pa·s) and $\gamma = 0$. This case is typical at high flow rates where the contribution from γ essentially can be neglected, and the capillary number goes to infinity. It can be considered as a limiting case, chosen to elucidate the behavior when viscous forces dominate and the surface tension is negligible. Data were collected for wetting phase saturation S_w = 0.25, 0.5, and 0.75. Here, S_w is the volume fraction of the wetting phase of the total pore volume in the network.

3. RESULTS AND DISCUSSION

We report first that the fluctuations in flow velocities are Gaussian when a suitable representative volume (REV) is chosen. We proceed to give the structure of the time correlation functions for the REV. The results for what we will call from now the Green–Kubo coefficients for the network follow from this.

3.1. Fluctuations

In case (A), the resistance is determined by the positions of the interfaces in the links only, as the two phases have the same viscosity. In case (B), the resistance to flow in link i is inversely proportional to the effective viscosity. The positions of the interfaces are then irrelevant, as there is no surface tension and hence, no capillary pressure.

A typical example of fluctuations in the total volume flow Q for case (A), with a pressure gradient $\Delta P/\Delta x$ of 100 kPa/m and surface tension 30 mN/m, is shown in **Figure 2**. By plotting the statistical frequency of the flow rate or the seepage velocity, we obtain a Gaussian distribution. This is shown in **Figure 3**, where we plot the statistical frequency of the fluid velocity (counts) on a logarithmic scale vs. $\text{sgn}(v-v_0)(v-v_0)^2$. In such a plot, a Gaussian distribution appears as a triangle, and this behavior is very well followed by the data. There is only a very small asymmetry in the distribution, which is to be expected as the fluid velocity cannot be less than zero. A regular plot of the distributions is presented for the seepage velocity and the velocities of the wetting and non-wetting phases in **Figure 4**. The distributions are shown for two network sizes of case (A), one with 30×20 links and one twice the size, with 60×40 links. The shown distributions are normalized to the area, and the variance of the larger network is half the width of that of the smaller network. So, in spite of the apparent noise seen in **Figure 2**, one obtains the distributions shown in **Figure 4**, which has its analog in molecular statistical thermodynamics, basic to thermodynamic equilibrium properties. This allows us

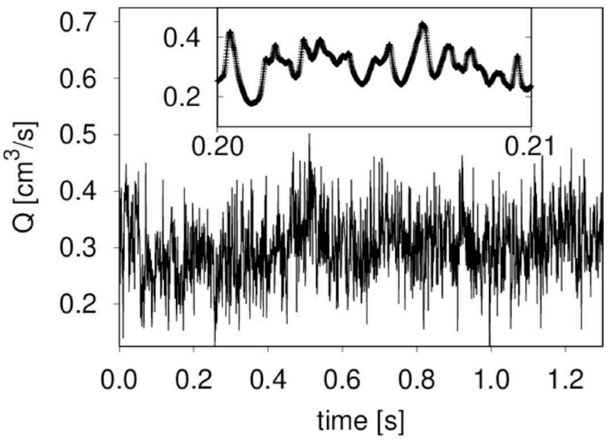

FIGURE 2 | Fluctuations of the volumetric flow Q for case A (see text). The pressure gradient was 100 kPa/m. During the time span shown, 1.3 s, a volume twice the total volume of the network has passed. The inset shows the same quantity over a short time interval.

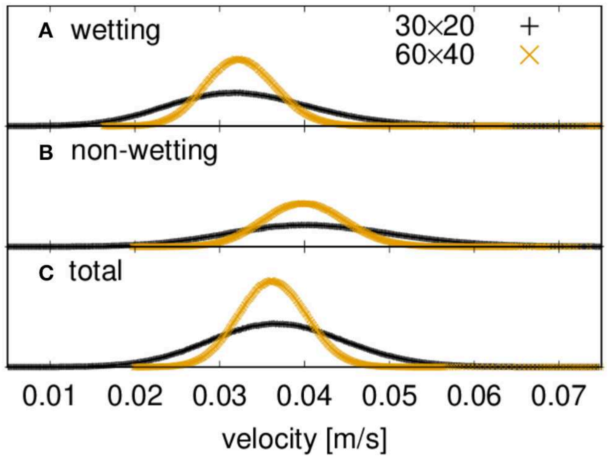

FIGURE 4 | Distributions of network-averaged instantaneous fluid velocities for two network sizes (30×20, 60×40).

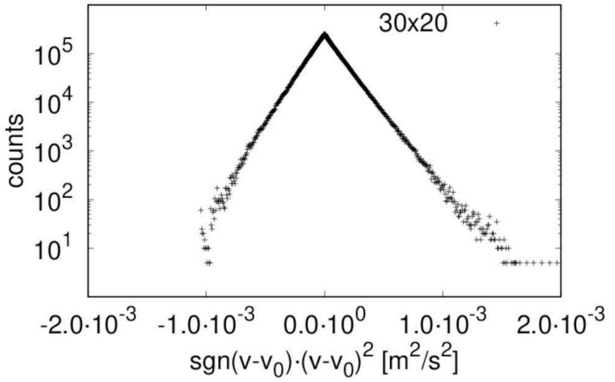

FIGURE 3 | Distribution of network-averaged instantaneous total (wetting and non-wetting) fluid velocity for network size 30 x 20 (case A). v_0 is the total average velocity and sgn is the sign function.

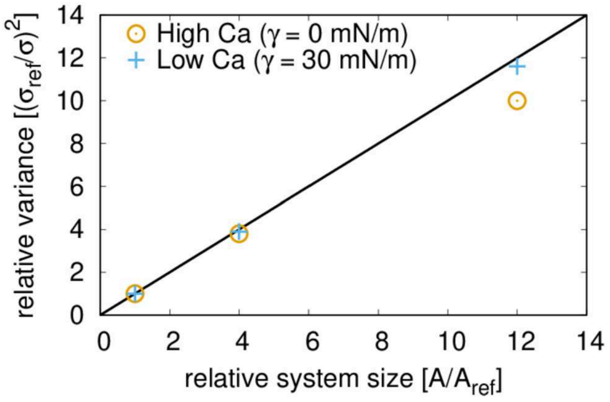

FIGURE 5 | Scaling of the variance σ^2 of the flow velocity size distributions with system size.

to proceed with the next step and construct time correlation functions for the meso-level.

3.2. Network Size and Representative Volume

Ideally, the system size of the simulation is sufficiently large and representative of the statistical ensemble. In this, the entropy and other thermodynamic properties are proportional to the system size (i.e., they are extensive [26]). With the Gaussian nature, one may expect that the inverse variance $1/\sigma^2$ of the fluctuations is proportional to the system size, i.e., the area A. This relation is plotted in **Figure 5** for network models of dimension 30×20, 60×40, and 120×60 links. It shows that this requirement is well met by a system with low Ca, but systems with higher Ca may be more susceptible to possible size effects. The size of the REV will be system-dependent; see Savani et al. [13]. But in the present cases, (A) and (B), an REV can be defined for a range of Ca. The results for the REV comply with the meso-level analog we are seeking.

3.3. Time Correlation Functions

With a well-defined REV, and with Gaussian fluctuations established, we can proceed to define the time correlation functions C_{RS} for the fluctuating quantities R and S at the meso-level:

$$C_{RS}(\tau) = \langle \delta R(0)\delta S(\tau)\rangle = \langle R(0)S(\tau)\rangle - \langle R\rangle\langle S\rangle, \quad (6)$$

where the brackets $\langle\cdots\rangle$ indicate ensemble averages.

The fluctuation from the mean, δR, is defined as

$$\delta R(t) = R(t) - \langle R\rangle, \quad (7)$$

and

$$\langle \delta R\rangle = 0. \quad (8)$$

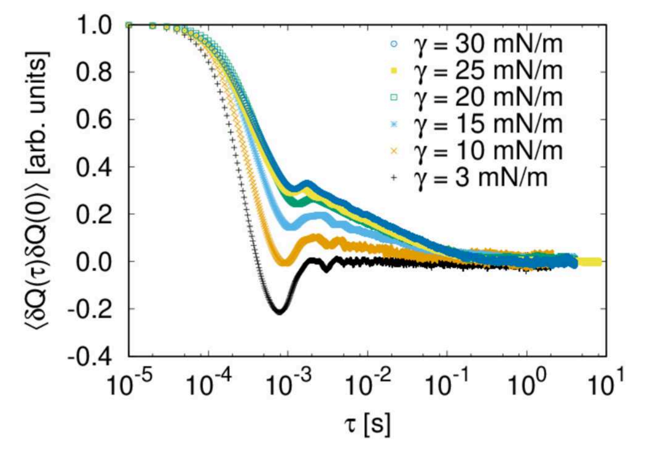

FIGURE 6 | Dependence of (scaled) time correlation function ⟨$Q(\tau)Q(0)$⟩ on the surface tension γ (case (A) with $\Delta P/\Delta x$ = 100 kPa/m and network size of 30×20 links).

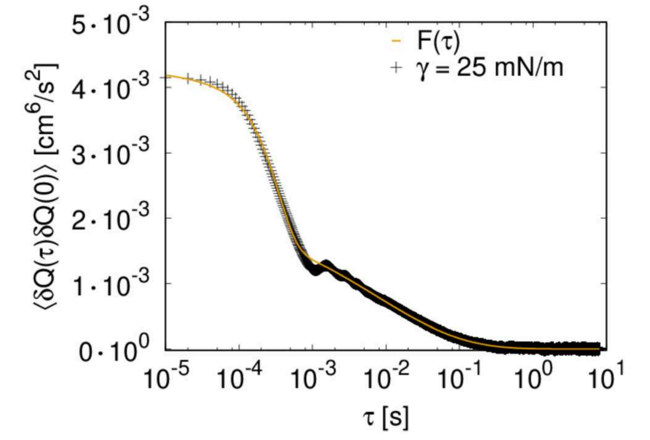

FIGURE 7 | Fit of function F(t) (Equation 9) to the autocorrelation function of the total volume flow for case (A) with γ = 25 mN/m ($\Delta P/\Delta x$ = 100 kPa/m and 30×20 links).

Figure 6 shows the time correlation functions of the total flow rate Q for different choices of the surface tension. After a rapid decay on a short timescale (below 10^{-3} s) there is a slower, logarithmic decay which is more pronounced for larger values of γ (between 1 and 100 ms). These two regimes are followed by a slow long-timescale decay. The rapid decay appears on the timescale that corresponds to the time necessary to evolve the flow by one average link volume. As shown in **Figure 6**, the decay is somewhat faster for smaller surface tensions as the total flow velocity is higher.

The regime of the logarithmic decay is within the time of evolving the flow by the total volume of the network and is more pronounced for higher surface tensions and thus higher capillary pressures in the pores. Hence, the decay corresponds to parts of the flow that are slow-moving or frustrated. These are the regimes of interest here. They contain the relative movements of the two flows in terms of their mutual displacement.

TABLE 1 | Fitting parameters for $F(t)$ (see Equation 9) to the autocorrelation functions of the total flow for different values of γ.

γ [mN/m]	a	b	τ_1	τ_2	α	β
30	0.28	0.17	0.39	22	1.40	0.52
25	0.21	0.23	0.34	6.6	1.75	0.36
20	0.13	0.50	0.49	0.059	2.69	0.16
15	0.15	0.22	0.39	0.096	2.28	0.16
10	0.11	0.043	0.31	0.26	2.09	0.15

The units of the fitting parameters are [cm^6/s^2]·10^{-2} (a and b), and [ms] (τ_1, τ_2).

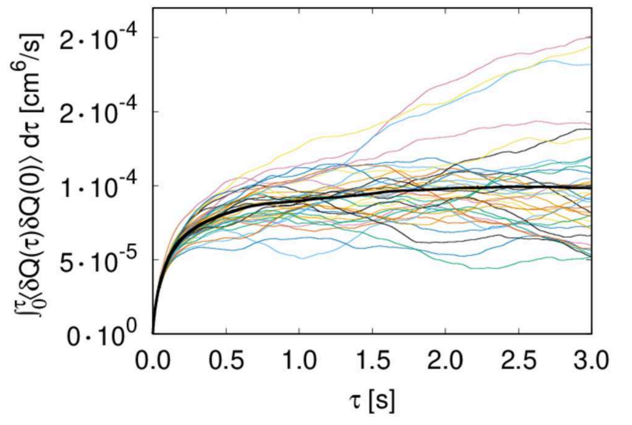

FIGURE 8 | Convergence of the integrated time correlation function [case (A), γ = 30 mN/m and 30×20 links]. The colored lines represent individual trajectories; the thicker black line is the average of all trajectories.

It is interesting to note some similarities with time correlation functions of glass [27] or yield-stress fluids [28]. In these cases, the autocorrelation functions, like the self-scattering function, can be fitted to a function of the form:

$$F(t) = a\exp[-(t/\tau_1)^\alpha] + b\exp[-(t/\tau_2)^\beta]. \tag{9}$$

We attempted a fit of $F(t)$ to the autocorrelation functions of the total flow; see **Figure 7**. Satisfying fits could be obtained with the exception that the local minimum and maximum at around 10^{-3} s and in some cases the flat top (at times $< 10^{-4}$ s) are not well-described. Fit parameters for the different choices of γ are summarized in **Table 1**.

3.4. Convergence and Symmetry

The Green–Kubo method employs integrals of suitable time correlation functions C_{RS} (as defined in Equation 6) to compute coefficients, L_{RS}:

$$L_{RS} = \int_0^\infty C_{RS}(\tau)d\tau. \tag{10}$$

The convergence of the integral over the time correlation function of the total flow is shown in **Figure 8**. As in molecular dynamics, where the Green–Kubo method is normally used, the convergence is slow, and statistics have to be collected over long timescales and/or multiple trajectories to achieve convergence of the integral when τ is approaching infinity.

TABLE 2 | Integrated time correlation functions for case (A) and three different settings of the pressure drop across the network.

$\Delta P/\Delta x$ [kPa/m]	100	150	200
Λ_{ww} [cm^6/s]·10^{-4}	0.46	0.72	1.59
Λ_{nn} [cm^6/s]·10^{-4}	0.69	1.51	1.88
Λ_{wn} [cm^6/s]·10^{-4}	−0.11	−0.35	−0.73
Λ_{nw} [cm^6/s]·10^{-4}	−0.10	−0.32	−0.70

Λ_{ij} are obtained from Equation (11). The uncertainties for $\Lambda_{i,j}$ are estimated to be less than 21%.

TABLE 3 | Volumetric flow rates Q and fluid velocities v for case (A) and three different settings of the pressure drop across the network.

$\Delta P/\Delta x$ [kPa/m]	100	150	200
Q [cm^3/s]	0.308	0.869	1.565
Q$_w$ [cm^3/s]	0.139	0.401	0.723
Q$_n$ [cm^3/s]	0.169	0.467	0.841
v [m/s]	0.037	0.104	0.188
v$_w$ [m/s]	0.033	0.096	0.174
v$_n$ [m/s]	0.041	0.113	0.203
v$_n$-v$_w$ [m/s]	0.007	0.017	0.028

TABLE 4 | Integrated time correlation functions for case (B) and three different choices for the saturation.

S_w	0.25	0.5	0.75
Λ_{ww} [cm^6/s]·10^{-4}	0.011	0.006	0.001
Λ_{nn} [cm^6/s]·10^{-4}	0.212	0.113	0.020
Λ_{wn} [cm^6/s]·10^{-4}	−0.046	−0.024	−0.005
Λ_{nw} [cm^6/s]·10^{-4}	−0.048	−0.027	−0.005

Λ_{ij} are obtained from Equation (11). Uncertainties for $\Lambda_{i,j}$ are estimated to be less than 21%.

TABLE 5 | Volumetric flow rates Q and fluid velocities v for case (B) and three different choices for the saturation.

S_w	0.25	0.5	0.75
Q [m^3/s]	0.763	0.503	0.359
Q$_w$ [cm^3/s]	0.137	0.208	0.248
Q$_n$ [cm^3/s]	0.626	0.295	0.110
v [m/s]	0.92	0.61	0.44
v$_w$ [m/s]	0.066	0.050	0.040
v$_n$ [m/s]	0.10	0.072	0.055
v$_n$-v$_w$ [m/s]	0.034	0.022	0.015

We computed the integrals for the autocorrelation and cross-correlation functions of the wetting and non-wetting phases,

$$\Lambda_{ij} = \int_0^\infty [\langle Q_i(\tau)Q_j(0)\rangle - \langle Q_i\rangle\langle Q_j\rangle]d\tau, \qquad (11)$$

with the indexes $i, j = n, w$ referring to the non-wetting and wetting phase, respectively. The results are listed in **Table 2** for case (A), where the surface tension is 30 mN/m and the fluids have the same viscosity, for three different choices of ΔP, the pressure difference across the network. **Table 3** lists the corresponding flow properties for case (A). **Table 4** summarizes results for infinite capillary numbers (case B), where the surface tension is zero but the fluids have different viscosities, for different choices of the saturation. Corresponding flow properties for case (B) are tabulated in **Table 5**.

In both cases (A) and (B), we find that the integral cross-correlations obey the Onsager reciprocal relation, $\Lambda_{ij} = \Lambda_{ji}$, within the statistical error in the simulations. This result is new for a meso-level description like the one used here and is encouraging for the overall aim; to create a non-equilibrium thermodynamic description for the macroscopic level. The finding applies to a well-defined REV, for which we have a Gaussian distribution of fluctuations, analogous to the corresponding distribution on the molecular level.

It is interesting that the cross coefficients are all negative. This makes sense for network flow, where one component cannot advance faster (on average) than the mean flow unless the other component advances slower (on average).

The Λ_{ij}s for case (B), where the surface tension is zero, show an extreme limit property, because the cross-correlation functions obey $\Lambda_{ww}\Lambda_{nn} - \Lambda_{wn}\Lambda_{nw} \approx 0$. A singular matrix of coefficients is the essence of complete coupling of the two fluids' flows; they are linearly dependent. For all choices of saturation, $\Lambda_{ww} = \zeta^2\Lambda_{nw}$, where ζ is a constant [14]. On the other hand, for case (A), where the surface tension differs from zero, a deviation of this dependency is observed. The linear dependence of the fluxes in case (B) can thus be associated with a lack of capillary forces. This can be understood in the following way: in case (B), the variation of mobility in a given link is a function of the saturation in the link only. However, if the mobility in one link is increased, it has to decrease elsewhere. In contrast, for case (A), the variations in link mobility depend on the interface position, and a change in the link mobility can take place without affecting the mobilities of other links.

The value of ζ for case (B) can be deduced by looking at the coefficients in **Table 4**. Within the accuracy, we find $\zeta^2 \approx 20$ or $\zeta = 4.5 \pm 0.5$ for all S_w. The value is close to the ratio of fluid viscosities, which will describe the dissipation.

All coefficients in **Table 4** show a dependence on the saturation, decreasing in value as the saturation increases. Here, the wetting fluid is the more viscous fluid, and the increase in saturation reduces the effective permeability. The coefficients also show a dependence on the pressure drops across the network (see **Table 2**), increasing with higher pressure drops. This is consistent with a higher effective permeability at higher pressures. In fact, the dependence of the volumetric flow is non-linear, as can be seen in **Table 3**. A non-linear dependence of a flow rate on the pressure difference is a well-known phenomenon in immiscible two-phase flow[29].

To investigate the origin of the Onsager symmetry in more detail, we examined the cross-correlations in **Figure 9**. For molecular systems, one can find transport coefficients using the Green–Kubo method (see [30]), and Onsager reciprocal relations apply given time-reversal invariance. We have found that time-reversal invariance also applies on the mesoscopic level, as formulated by: $C_{AB}(\tau) = C_{BA}(\tau)$. This equality is illustrated in **Figure 9**. There is agreement except for values at very short timescales, where the contribution to the integral is negligible.

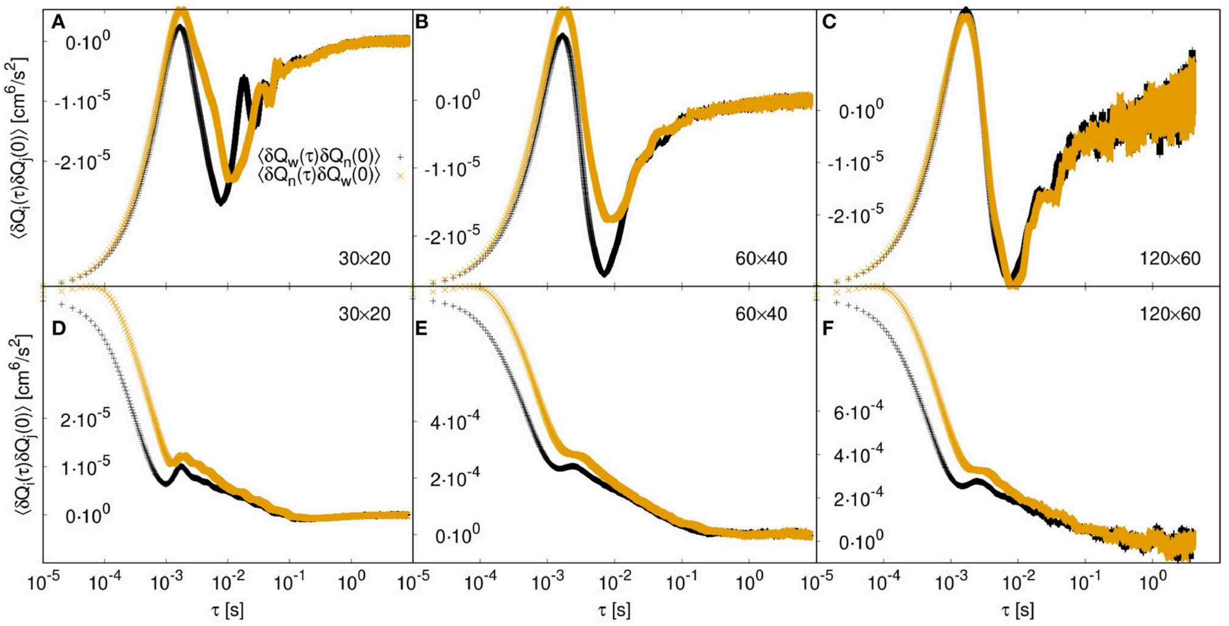

FIGURE 9 | Cross-correlation functions of wetting (Q_w) and non-wetting (Q_n) flows. The upper panels **(A–C)** show the cross-correlations for case (B) with zero surface tension and increasing network size from **(A)** to **(C)**. The lower panels **(D–F)** show the cross-correlations for case (A) with γ = 30 mN/m. The network size is increasing from left to right from 30×20 to 60×40 to 120×60 links.

Moreover, for case (B), the deviation from symmetry at small time values is attributable to the finite system size. It vanishes for larger network sizes. This is shown in the upper three panels of **Figure 9**.

4. CONCLUSIONS

Our investigation of time correlation functions has revealed interesting parallels between the time correlation functions of two immiscible fluids in a porous media, those observed for glass and stress-yield fluids, and those for molecular fluctuations. A network with incompressible fluids has been used as a model for the porous medium, but the findings should not be restricted to this. We have been able for the first time to find Onsager symmetry in athermal fluctuations on the meso level. The symmetry of the coefficients implies time-reversal invariance or microscopic reversibility of fluctuations also on the meso level, in agreement with recent experimental [11] and computational evidence [12, 13]. Time-reversal invariance is here understood as $C_{AB}(\tau) = C_{BA}(\tau)$, holding for all timescales except very short timescales.

We found that the structure of the time correlation functions depends on the surface tension. Integrals over auto- and cross-correlation functions of an REV were found to converge, and the integrals of the cross-correlation functions essentially obeyed Onsager's symmetry. The coefficients obtained in this manner may have a relation to porous medium permeabilities. Further research on time correlation functions to compute the transport properties of immiscible-two phase flow is therefore encouraged.

One may ask whether the dynamic pore network model we have used really reflects the dynamics of real porous media or whether ergodicity has been built into it somehow, ergodicity that is not there in real porous media. Apart from doing experiments that will answer this question, one may repeat the measurements here but based on other models that differ substantially from ours, such as the Lattice Boltzmann Method.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

MW conceptualized the study, carried out the simulation, and wrote the first draft. MG assisted with the computational work. All authors contributed to data analysis, and in developing the theory as well as shaping the manuscript to its final form.

FUNDING

This work was supported by the Research Council of Norway through its Centres of Excellence funding scheme, project number 262644. MW is grateful for a postdoc scholarship from the Department of Physics, the Norwegian University of Science and Technology, NTNU, Trondheim.

ACKNOWLEDGMENTS

The authors would like to thank Prof. Daan Frenkel for helpful discussions of this work. We thank the reviewers for providing us with exceptionally good and challenging reviews.

REFERENCES

1. Ben-Isaac E, Park Y, Popescu G, Brown FLH, Gov NS, Shokef Y. Effective temperature of red-blood-cell membrane fluctuations. *Phys Rev Lett*. (2011) **106**:238103. doi: 10.1103/PhysRevLett.106.238103
2. Gnoli A, Petri A, Dalton F, Pontuale G, Gradenigo G, Sarracino A, et al. Brownian ratchet in a thermal bath driven by coulomb friction. *Phys Rev Lett*. (2013) **110**:120601. doi: 10.1103/PhysRevLett.110.120601
3. Bi D, Henkes S, Daniels KE, Chakraborty B. The statistical physics of athermal materials. *Annu Rev Condens Matter Phys*. (2015) **6**:63–83. doi: 10.1146/annurev-conmatphys-031214-014336
4. Kanazawa K, Sano TG, Sagawa T, Hayakawa H. Minimal Model of stochastic athermal systems: origin of non-gaussian noise. *Phys Rev Lett*. (2015) **114**:090601. doi: 10.1103/PhysRevLett.114.090601
5. Clewett JPD, Wade J, Bowley RM, Herminghaus S, Swift MR, Mazza MG. The minimization of mechanical work in vibrated granular matter. *Sci Rep*. (2016) **6**:28726. doi: 10.1038/srep28726
6. Weber SC, Spakowitz AJ, Theriot JA. Nonthermal ATP-dependent fluctuations contribute to the *in vivo* motion of chromosomal loci. *Proc Natl Acad Sci USA*. (2012) **109**:7338–43. doi: 10.1073/pnas.1119505109
7. Dabelow L, Bo S, Eichhorn R. Irreversibility in active matter systems: fluctuation theorem and mutual information. *Phys Rev X*. (2019) **9**:021009. doi: 10.1103/PhysRevX.9.021009
8. Kumar N, Soni H, Ramaswamy S, Sood AK. Flocking at a distance in active granular matter [Journal Article]. *Nat Commun*. (2014) **5**:4688. doi: 10.1038/ncomms5688
9. Tang H, Yan B, Chai Z, Zuo L, Killough J, Sun Z. Analyzing the well-interference phenomenon in the eagle ford shale/austin chalk production system with a comprehensive compositional reservoir model. *SPE Reserv Eval Eng*. (2019) **22**:827–41. doi: 10.2118/191381-PA
10. de Groot SR, Mazur P. *Non-equilibrium Thermodynamics*. New York, NY: Dover (1984).
11. Erpelding M, Sinha S, Tallakstad KT, Hansen A, Flekkøy EG, Måløy KJ. History independence of steady state in simultaneous two-phase flow through two-dimensional porous media. *Phys Rev E*. (2013) **88**:053004. doi: 10.1103/PhysRevE.88.053004
12. Savani I, Sinha S, Hansen A, Bedeaux D, Kjelstrup S, Vassvik M. A Monte Carlo algorithm for immiscible two-phase flow in porous media. *Transport Porous Med*. (2017) **116**:869–88. doi: 10.1007/s11242-016-0804-x
13. Savani I, Bedeaux D, Kjelstrup S, Vassvik M, Sinha S, Hansen A. Ensemble distribution for immiscible two-phase flow in porous media. *Phys Rev E*. (2017) **95**:023116. doi: 10.1103/PhysRevE.95.023116
14. Kjelstrup S, Bedeaux D. *Non-Equilibrium Thermodynamics of Heterogeneous Systems*. World Scientific (2008) Available online at: https://www.worldscientific.com
15. Onsager L. Reciprocal relations in irreversible processes. I. *Phys Rev*. (1931) **37**:405–26. doi: 10.1103/PhysRev.37.405
16. Onsager L. Reciprocal relations in irreversible processes. II. *Phys Rev*. (1931) **38**:2265–79. doi: 10.1103/PhysRev.38.2265
17. Flekkøy EG, Pride SR, Toussaint R. Onsager symmetry from mesoscopic time reversibility and the hydrodynamic dispersion tensor for coarse-grained systems. *Phys Rev E*. (2017) **95**:022136. doi: 10.1103/PhysRevE.95.022136
18. Aker E, Jørgen Måløy K, Hansen A, Batrouni GG. A two-dimensional network simulator for two-phase flow in porous media. *Transport Porous Med*. (1998) **32**:163–86. doi: 10.1023/A:1006510106194
19. Aker E, Måløy KJ, Hansen A, Basak S. Burst dynamics during drainage displacements in porous media: Simulations and experiments. *Eur Phys Lett*. (2000) **51**:55–61. doi: 10.1209/epl/i2000-00331-2
20. Zhao B, MacMinn CW, Primkulov BK, Chen Y, Valocchi AJ, Zhao J, et al. Comprehensive comparison of pore-scale models for multiphase flow in porous media. *Proc Natl Acad Sci USA*. (2019) **116**:13799–806.
21. Flekkøy EG, Pride SR. Reciprocity and cross coupling of two-phase flow in porous media from Onsager theory. *Phys Rev E*. (1999) **60**:4130–7. doi: 10.1103/PhysRevE.60.4130
22. Sinha S, Gjennestad MA, Vassvik M, Hansen A. A dynamic network simulator for immiscible two-phase flow in porous media. *arXiv e-prints* arXiv:1907.12842 (2019).
23. Sinha S, Bender AT, Danczyk M, Keepseagle K, Prather CA, Bray JM, et al. Effective rheology of two-phase flow in three-dimensional porous media: experiment and simulation. *Transport Porous Med*. (2017) **119**:77–94. doi: 10.1007/s11242-017-0874-4
24. Gjennestad MA, Vassvik M, Kjelstrup S, Hansen A. Stable and efficient time integration of a dynamic pore network model for two-phase flow in porous media. *Front Phys*. (2018) **6**:56. doi: 10.3389/fphy.2018.00056
25. Sinha S, Gjennestad MA, Vassvik M, Winkler M, Hansen A, Flekkøy EG. Rheology of high-capillary number two-phase flow in porous media. *Front Phys*. (2019) **7**:65. doi: 10.3389/fphy.2019.00065
26. Kjelstrup S, Bedeaux D, Hansen A, Hafskjold B, Galteland O. Non-isothermal transport of multi-phase fluids in porous media. Constitutive equations. *Front Phys*. (2019) **6**:150. doi: 10.3389/fphy.2018.00150
27. Reichman DR, Charbonneau P. Mode-coupling theory. *J Stat Mech Theory Exp*. (2005) **2005**:P05013. doi: 10.1088/1742-5468/2005/05/p05013
28. Levashov VA. Contribution to viscosity from the structural relaxation via the atomic scale Green-Kubo stress correlation function. *J Chem Phys*. (2017) **147**:184502. doi: 10.1063/1.4991310
29. Longeron DG. Influence of very low interfacial tensions on relative permeability. *Soc Petrol Eng J*. (1980) **20**:391–401. doi: 10.2118/7609-PA
30. Liu X, Schnell SK, Simon JM, Krüger P, Bedeaux D, Kjelstrup S, et al. Diffusion coefficients from molecular dynamics simulations in binary and ternary mixtures. *Int J Thermophys*. (2013) **34**:1169–96. doi: 10.1007/s10765-013-1482-3

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Physics

REVIEW
published: 24 April 2020
doi: 10.3389/fphy.2020.00121

Metal Filled Nanostructured Silicon With Respect to Magnetic and Optical Properties

Petra Granitzer and Klemens Rumpf*

Institute of Physics, University of Graz, Graz, Austria

Within this work the utilization of nanostructured silicon as host material for filling with various magnetic nanostructures is reviewed whereas the magnetic and optical properties of the gained composite systems are elucidated. The metal filling of the pores is mainly performed by electroless deposition or by electrodeposition which is discussed by means of some examples. Furthermore, two different types of porous silicon (PSi) morphology are used for the deposition procedure. On the one hand microporous silicon offering luminescence in the visible range is utilized as template material. It offers a branched morphology with a structure size between 2 and 5 nm. In this case not only the magnetic response is investigated but also the influence of the metal filling on the optical properties. On the other hand mesoporous and macroporous silicon in it's low pore regime is employed which offers straight pores with diameters up to 90 nm. In this case the magnetic response strongly depends on the size, the geometry and the spatial distribution of the metal deposits within the pores. A crucial role plays also the morphology of the porous silicon, especially the distance between adjacent pores which is an important parameter regarding magnetic interactions.

Keywords: porous silicon, electrodeposition, magnetic nanostructures, luminescence, magnetic properties

OPEN ACCESS

Edited by:
Simona Binetti,
University of Milano-Bicocca, Italy

Reviewed by:
Tatiana S. Perova,
Trinity College Dublin, Ireland
Shinji Yae,
University of Hyogo, Japan

****Correspondence:***
Petra Granitzer
petra.granitzer@uni-graz.at

Specialty section:
This article was submitted to Optics and Photonics, a section of the journal Frontiers in Physics

Received: *18 February 2020*
Accepted: *26 March 2020*
Published: *24 April 2020*

Citation:
Granitzer P and Rumpf K (2020) Metal Filled Nanostructured Silicon With Respect to Magnetic and Optical Properties. Front. Phys. 8:121. doi: 10.3389/fphy.2020.00121

INTRODUCTION

Porous silicon can be fabricated by various methods such as anodization [1], stain etching [2], metal assisted etching [3], galvanic etching [4], reactive ion etching [5], or laser ablation [6]. The obtained morphology reaches from microporous silicon offering a structure size between 2 and 5 nm to mesoporous silicon with pore diameters up to 50 nm and further to macroporous silicon with pores up to a few micrometers [7]. A popular fabrication technique which allows the tuning of the porous morphology is anodization of the silicon wafer. In this case beside the doping concentration of the silicon wafer the resulting porous silicon morphology mainly depends on the applied current density and the HF concentration. An increase of the current density generally leads to an increase of the pore diameter and a concomitant decrease of the distance between the pores. A reduction of the HF concentration results in higher porosity which is related to the decrease of the pore diameter with increasing HF concentration [8].

Microporous silicon offers a branched morphology with interconnected channels leading to a huge surface area up to 1,000 m^2/cm^3 [9], depending on the porosity which is defined by the ratio of the pore diameter and the wall thickness but do not give information about the actual dimensions. Due to the small structure size entailing quantum confinement effects, light emission in the visible range is observed [10]. Quantum size effects are responsible not only for the photoluminescence but also for the porous silicon formation [11]. Light emission has been observed in various spectral ranges. The most intense and the most investigated luminescence band of PSi is in the red regime. In this case the photoluminescence peak offers a broad range from about 590 nm to about 950 nm [7] which can be influenced by e.g., etching conditions, oxidation of the PSi, temperature treatment, or hydrostatic pressure [12]. But also further photoluminescence bands in the infrared and in the green-blue region could be observed [13]. Considering the life time dependence of the different bands one can say that the red band, also named the slow band, offers decay times of a few microseconds, also does the infrared band, whereas the green-blue one shows a fast decay in the nanosecond regime [14–16]. The slow decay of the photoluminescence is attributed to carrier recombination through localized states offering an energy distribution and size disorder [17]. The lifetime of the fast band is explained by quasi direct recombination in the silicon crystallites or by oxide related effects [18, 19]. Generally of high interest is the increase of the quantum yield of the luminescence which recently has been reported to extend 32% at room temperature due to supercritical drying. The porous structure and the silicon grains are better retained by this drying process and the formation of non-radiative defects occurring during usual drying in air is reduced [20]. An increase of the quantum yield to 53–61% can be obtained by Si/SiO_2 core/shell nanoparticles which emit at 1.5 eV (∼826 nm). The oxidation of the silicon nanocrystals was carried out by high pressure water vapor annealing. In the case of silicon powder, obtained by the same method, which emits at 1.9 eV (∼652 nm), a quantum yield of about 30% could be reached [21].

Beside photoluminescence also electroluminescence of porous silicon has been intensely investigated. The setup is arranged as semitransparent metal/PSi/c-silicon/Al-electrode and it shows a rectifying junction behavior. This behavior is explained by radiative transition due to electron and hole injection in quantized states in porous silicon [22]. The quantum efficiency of the luminescence of such arrangements is quite low. Since porous silicon offers a high resistivity caused by the pore formation process [23] efforts have been made to improve the conductivity e.g., by metal filling of the pores [24]. A further approach is to use the metal filling of microporous silicon to enhance the intensity of the photoluminescence. Metals such as iron, nickel or cobalt are employed and they lead to an enhancement of the photoluminescence of porous silicon due to the passivation of the silicon dangling bonds formed after the anodization process [25–27]. Gold and silver nanoparticles placed on porous silicon feature the localized surface plasmon resonance whereas the porous silicon as dielectric spacing layer enhances the plasmon resonance [28]. The metal filling procedure can be carried out either electroless or by electrodeposition.

Mesoporous silicon fabricated by anodization offers dendritic straight pores which are separated from each other. Using highly doped n-silicon the pore formation is due to electrical breakdown dominated by tunneling [29]. The pore diameter as well as the porosity increase with increasing current density and it decreases with increasing HF concentration [7]. In a pore diameter regime between 25 and 100 nm the morphology can be tuned quite accurately by varying the applied current density obtaining a quasi-regular pore arrangement [30]. Filling of such pores with magnetic metals results in a magnetic response dependent on the size, geometry, and arrangement of the deposits. The coercivity decreases with increasing elongation of the deposits, the magnetic anisotropy between easy and hard axis magnetization increases with increasing structure length. The morphology of the template, especially the degree of dendritic growth also plays a crucial role which influences the magnetic crosstalk between neighboring pores [31].

MICROPOROUS SILICON

Porous silicon has been produced in the 1950's by Uhlir [32] coincidentally as byproduct of electropolishing experiments. After a long silence, in 1990 Canham and Lehmann discovered the light emitting properties of this material and ascribed it to quantum confinement effects [1]. Thereafter microporous silicon has been under intense investigation. Beside photo- and electroluminescence also electrical- [33] and thermal [34] conductivity as well as optical [35, 36] and mechanical [37] properties have been examined. In the middle of the 1990's V. Lehmann showed that not only microporous but also meso- and macroporous silicon can be produced by anodization [29, 38, 39]. The resulting morphology depends on the electrochemical parameters and the doping concentration of the silicon wafer. To obtain microporous morphology low and medium doped silicon (10^{14}-10^{18} cm^{-3}) is used, mesoporous structures are achieved in using highly doped silicon ($>10^{17}$ cm^{-3}) and macropores are formed in low/medium doped silicon (10^{14}-10^{17} cm^{-3}) [29].

Fabrication and Metal Filling of Microporous Silicon

The most common fabrication process is anodization of c-silicon in aqueous hydrofluoric acid solution. As wetting agent ethanol is used which also influences the porosity, higher ethanol content leading to higher porosity [40]. The anodization process allows the tuning of the porous morphology quite easily. A further porosification method is stain etching, which means electroless etching in hydrofluoric acid solution containing an oxidizing agent [41]. Using this electroless procedure the tunability of the porous structure can be reached less easily, mainly by varying the type and content of the oxidizing agent [42]. The most common method to determine the porosity and the size of the porous morphology is gas adsorption but also an indirect method, using the peak position of the photoluminescence spectra, can be used to estimate the structure size [43]. Microporous silicon with a structure size below 5 nm, offering light emitting properties, poses a high challenge for metal pore filling.

P-type silicon (100) with 10–20 Ωcm has been used to produce porous silicon with two morphologies. The porous silicon offers macropores which are covered by a microporous layer of about 2 μm thickness with pore diameters of about 3 nm. These small pores are filled with zinc from an aqueous $ZnSO_4$ solution by electrodeposition. The results show that in the case of such confined nanopores electrodeposition is not diffusion limited [44]. Electrodeposition of platinum within

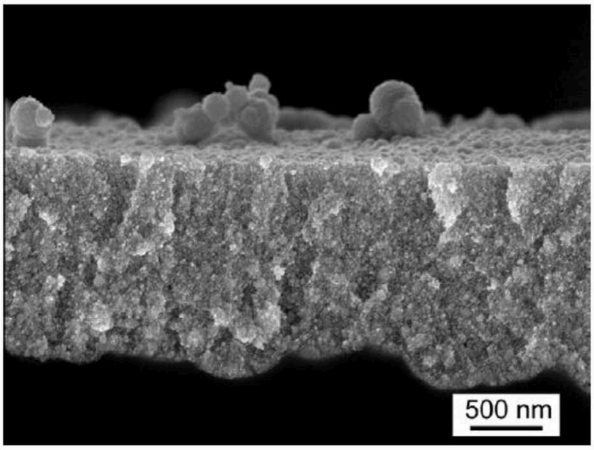

FIGURE 1 | Cross-sectional SEM image of hydrophobic porous silicon filled with Pt. Fukami et al. [46], Elsevier by permission.

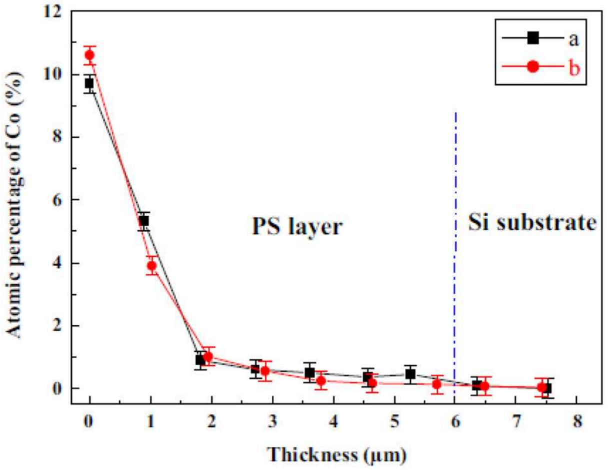

FIGURE 2 | Depth dependence of the atomic percentage of Co obtained from EDX measurements showing (a) sample before temperature treatment and (b) after annealing at 400°C. Bouzouraa et al. [27], Elsevier by permission.

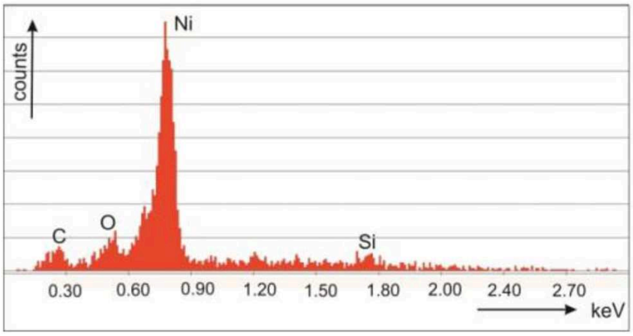

FIGURE 3 | EDX-spectrum taken at the pore bottom of a cross-section of a Ni filled microporous silicon sample. Granitzer et al. [47], with permission.

hydrophobic and hydrophilic porous silicon was carried out using a solution consisting of 0.1 M K_2PtCl_4 and 0.5 M NaCl, Ag was deposited using a 0.1 M $AgNO_3$ and 0.5 M KNO_3 solution. It has been shown that Pt is deposited within the pores in using hydrophobic structures, whereas it is only deposited on the top surface in the case of hydrophilic porous silicon. Ag deposits for both chemically modified structures equally [45]. In **Figure 1** Pt deposited within a porous silicon layer is depicted. Using hydrophobic porous structures the deposition reaction is enhanced and the diffusion limited condition, occurring in nanopores, is suppressed [46].

A further approach to fill porous structures is immersion plating. Immersion of as etched PSi into an 0.5 M $CoCl_2$ solution for 120 min shows Co inside the pores with a concentration decrease toward the pore bottom [27]. **Figure 2** shows the atomic percentage of Co in dependence on the porous layer thickness, obtained from EDX investigations.

Ni deposition within microporous silicon has been performed in utilizing an aqueous 0.1 M $NiCl_2$ solution. First a current density of 0.6 $mAcm^{-2}$ has ben applied before the electroless deposition. The duration of the electroless reaction process was varied between 0 and 11 min [26]. Pulsed electrodeposition is employed to deposit Ni from the so-called Watts electrolyte consisting of $NiCl_2$ (45 g/l) and $NiSO_4$ (300 g/l) within porous silicon. A current density of 10 $mAcm^{-2}$ and a frequency of 2 Hz has been applied. The deposition time has been enhanced from 5 till 15 min leading to a filling of the porous layer down to the bottom, which was evidenced by EDX spectra [47]. **Figure 3** depicts an EDX spectrum taken at the pore tips [47].

Optical Properties of Microporous Silicon

Considering the photoluminescence of microporous silicon, the emission can be tuned due to the structure size in the region from the near-infrared to the ultraviolet in the case of a hydrogen passivated surface. Oxidation of the surface leads to a red shift of about 1 eV showing that quantum confinement and surface passivation, both determine the electronic states. This red-shift can be explained by recombination attendant to a trapped electron or exciton [48]. Since the quantum confinement model suggested by Canham [1] offers some deficiencies other models have been taken into account such as that the light emission comes from siloxane molecules formed on the surface [49] which could be disproved by thermal oxidation experiments [50]. A further approach to describe the luminescence is the surface states model, which involves the trapping and localization of photoexcited carriers in silicon related boundary states of the crystallites [51]. In the case of a structure size larger than 3 nm, offering a red/orange luminescence, the quantum confinement model is sufficient and surface passivation plays a minor role on the radiative recombination mechanism [48]. Radiative relaxation processes are influenced by quantum confinement and not affected by oxidation, whereas non-radiative relaxation processes are affected by the surface chemistry [52]. Continuous wavelength and time resolved photoluminescence of silicon nanocrystals embedded in a SiO_2 matrix can be used to

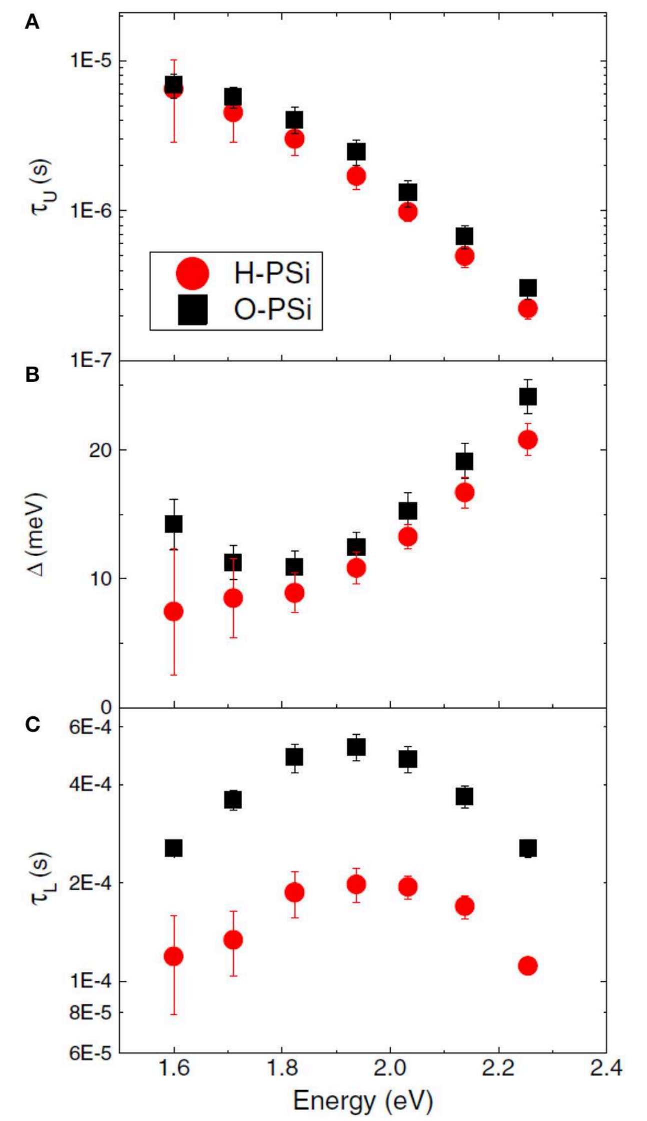

FIGURE 4 | Comparison of hydrogen terminated and oxidized porous silicon, showing **(A)** upper singlet lifetimes, **(B)** excitonic energy splitting, and **(C)** lower triplet lifetimes. Arad-Vosk et al. [52], Springer with permission.

distinguish between microscopic and macroscopic characteristics of the decay [53]. Microscopic characteristics of the decay are attributed to quantum confinement effects and the macroscopic ones are affected by the environment of the silicon nanocrystal [53]. The decay results from two neighboring levels splitted in their energy and it has been found that the upper level decay is shorter (microseconds) than the lower level decay (milliseconds) [53]. **Figure 4** depicts the slower triplet lifetimes which are due to the oxidation of the freshly prepared silicon and the faster singlet lifetimes which are not influenced by the surface chemistry as well as the energy splitting [52].

Metal filling of luminescent porous silicon can influence the optical properties of the material. The deposition of Au nanoparticles within the pores results in the excitation of surface plasmon polaritons and an increase of the photoluminescence intensity is observed [54]. Considering the interaction between porous silicon and Au nanoparticles on its surface, not only plasmon effects influence the optical properties but also the porosity and surface chemistry of the nanocomposite. The plasmonic behavior depends on the particle size and shape as well as on the refractive index of the surrounding medium. The latter one explains the dependence of the optical response on the porosity of the porous silicon [55]. The interaction of the plasmons of the Au nanoparticles with the excitons generated in the semiconductor result in a modification of the emission and absorption properties [56]. Considering the photoluminescence lifetime of an Au/PSi system, generally a decrease due to the plasmonic effects compared to plain PSi is expected. In [55] an increase of the decay times, which are determined by the fit of a stretched exponential, has been observed which could be explained by the surface chemistry of the samples. In the case of Au particles deposited on porous silicon nanowires an enhancement as well as a quenching of the photoluminescence intensity, depending on the deposition time and the solution concentration, has been reported [57]. An increase of the photoluminescence intensity and a blue shift of the spectrum has been reported in the case of silver nanoparticles positioned in the vicinity of silicon nanocrystals with a SiO_2 spacer of a few nanometers in-between [58]. **Figure 5** shows the comparison of the photoluminescence obtained from plain samples and samples containing Ag particles. The intensity increase can be explained by the coupling of the silicon nanocrystals to the surface plasmons of the Ag particles.

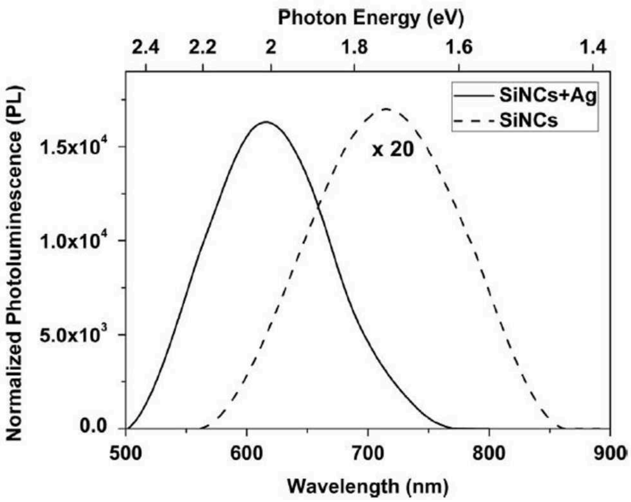

FIGURE 5 | Photoluminescence spectra of only silicon nanocrystals (dashed line) and silicon nanocrystals with Ag particles deposited on the silicon surface (solid line). The latter case shows an increase of the luminescence intensity as well as a blue shift of the spectrum of about 100 nm. Gardelis et al. [58], Elsevier by permission.

An increase of the photoluminescence intensity could also be observed by electrodepositing Ag particles from an $AgNO_3$ solution by varying the deposition time and the concentration

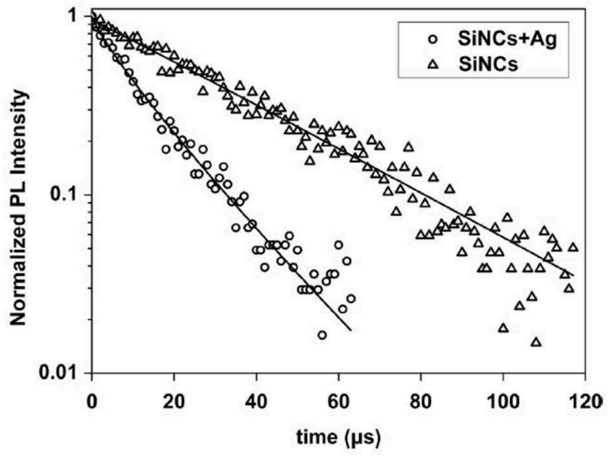

FIGURE 6 | Comparison of the decay times of microporous samples containing Ag particles and samples without Ag. Gardelis et al. [58], Elsevier by permission.

of the electrolyte [59]. This behavior of the luminescence is furthermore shown after LiCl treatment of porous silicon as well as an enhancement of the minority carrier lifetime due to the passivation of the dangling bonds at the silicon surface [60].

The decay times obtained from time resolved measurements and fitting with a stretched exponential function show an increase of the recombination rate in the Ag particle loaded samples (**Figure 6**). Together with the enhanced photoluminescence, this is a hint that the Ag particles cause an increase of the emission rate of the silicon nanocrystals [58].

The photoluminescence of porous silicon with incorporated Ni has been investigated with respect to the immersion time of the samples in an Ni-salt solution. An increase of the luminescence intensity has been observed with increasing immersion time from 1 to 7 min and a subsequent decrease of the intensity with further increasing immersion time. Furthermore, a blue shift of the spectrum has been measured [26]. **Figure 7** depicts these findings. The photoluminescence intensity enhancement is explained by the reaction between the Ni ions and the porous silicon surface.

Stain etched porous silicon powder filled with Ni by electroless plating in a $NiCl_2$ solution has been investigated concerning the photoluminescence of the composite system. The samples offer a dependence on the concentration of the $NiCl_2$ solution (25–500 g/l). With increasing concentration the intensity decreases and increases again. The decreasing intensity behavior is explained by the increase of surface defects by the desorption of H-atoms forming a low quality oxide layer, the following increase of the intensity with higher concentration is due to the formation of a better quality oxide. The peak position of the photoluminescence spectrum also depends on the $NiCl_2$ concentration. First a blue-shift is observed which could be caused by a reduction of the emitter size due to oxidation and with further increase of the concentration the wavelengths are red shifted [61].

In the case of using pulsed electrodeposition for the Ni filling of luminescent porous silicon the optical properties have been investigated in dependence on the deposition time, ranging from 5 to 15 min [47]. A small enhancement of the photoluminescence intensity and a blue-shift of the spectrum is observed with increasing deposition time. The decay times show a decrease from about 300 μs for bare porous silicon to about 100 μs for a metal filled sample (deposition time 15 min). This result can be interpreted by a decrease of the radiative lifetime. Generally radiative processes prevail the non-radiative processes at room temperature. The decay behavior of porous silicon can be explained by an excitonic two-level model with the upper excitonic singlet-triplet state and the ground state [52]. Coupling of the plasmonic modes of the Ni with the emitter (silicon) can occur due to the direct vicinity of the emitter and the deposited Ni, only separated by a native oxide layer, and lead to the increase of the photoluminescence intensity [47].

In contrast to the findings above, quenching of the photoluminescence has been reported after Co electrodeposition within porous silicon due to oxidation of the emitting centers. An additional blue emission band around 490 nm occurs due to silanol formation at the surface [62].

MAGNETIC RESPONSE OF METAL FILLED MICROPOROUS SILICON

Filling of porous silicon with magnetic materials allows to tune the magnetic properties of the composite by the electrochemical deposition parameters as well as by the morphology of the porous silicon matrices. Considering luminescent porous silicon, due to the branched morphology also the metal deposits offer an interconnnected structure. Due to the pore-size (a few nanometers) the metal structures are in the superparamagnetic range but because of the interconnection between them ferromagnetic behavior is observed [47]. Microporous silicon filled with a magnetic material renders a system offering both, light emitting and magnetic properties and thus the nanocomposite is promising for magneto-optical on-chip devices. Temperature dependent magnetization measurement do not give any hint of superparamagnetism. In the case of isolated nanoparticles a clear peak showing the transition temperature between superparamagnetic behavior and blocked state would be observed. The investigated magnetic response strongly depends on the metal deposition time, which means the amount of metal inside the pores. An increase of the coercivity H_C from 150 to 450 Oe with increasing deposition time from 5 to 15 min is observed (**Figure 8A**). In this case the magnetic field was applied parallel to the surface. Furthermore, the magnetic anisotropy between easy axis and hard axis magnetization shows a dependence on the filling time (**Figure 8B**). This anisotropic behavior is film-like and becomes more pronounced with increasing metal filling and thus more interconnections between the metal deposits. The magnetic easy axis corresponds to an external magnetic field applied parallel to the surface, which is in contrast to mesoporous silicon (see in section Optical Properties of Metal Filled Mesoporous Silicon). Two hysteresis curves, one measured

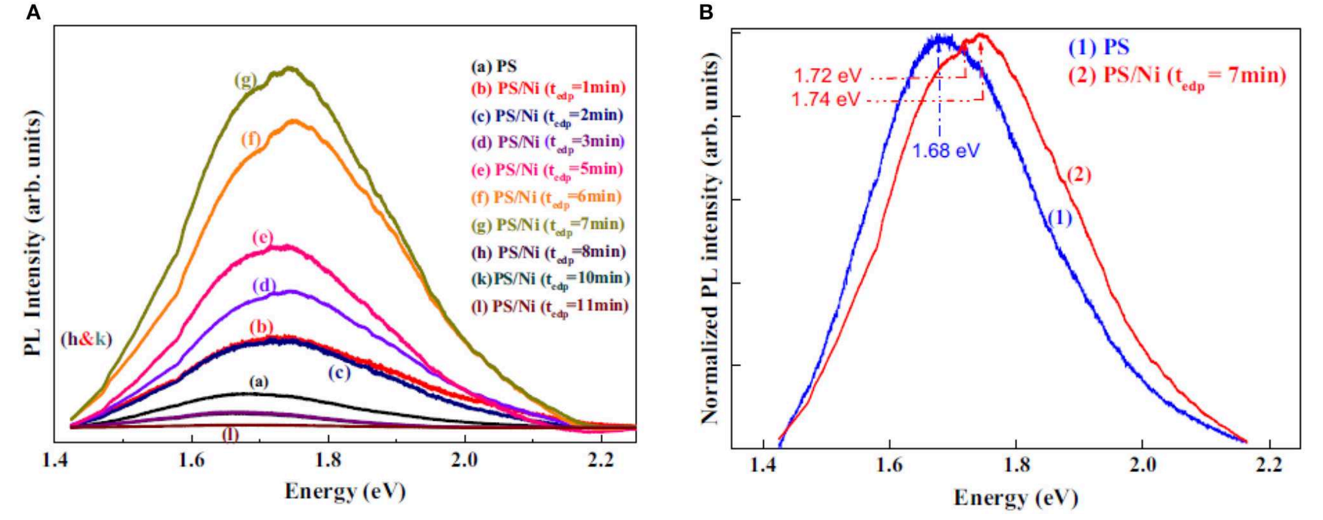

FIGURE 7 | Photoluminescence spectra of Ni filled porous silicon in dependence on the immersion time. **(A)** Photoluminescence spectra for porous silicon and Ni/PSi for an immersion time from 1 to 11 min. **(B)** Normalized spectra showing the blue shift of Ni filled PSi (immersion time 7 min). Amdouni et al. [26], Elsevier by permission.

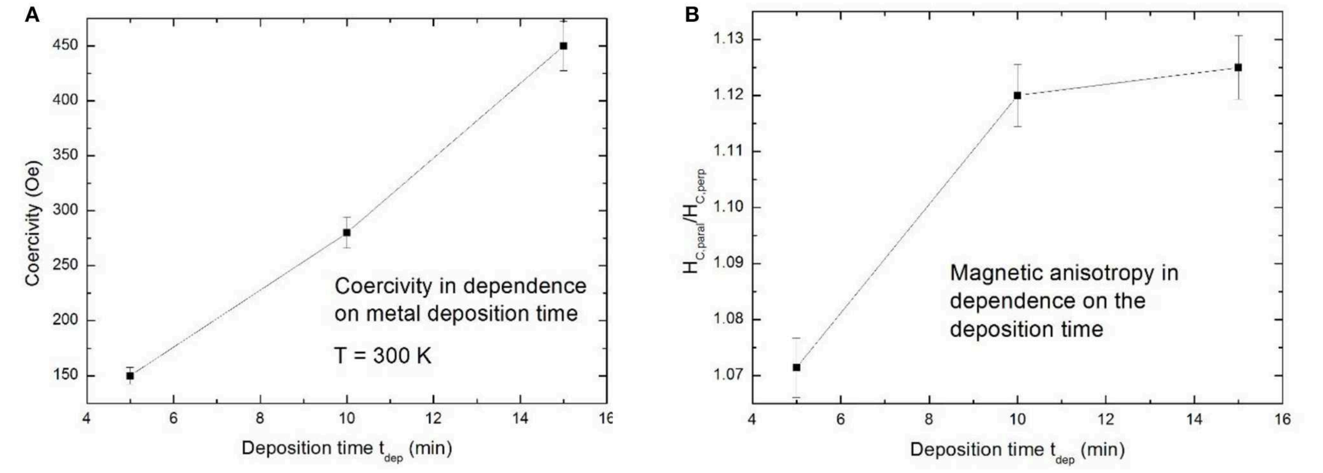

FIGURE 8 | (A) Coercivity of microPSi/Ni in dependence on the deposition time. **(B)** Magnetic anisotropy ($H_{c,paral}/H_{c,perp}$) vs. filling time. Granitzer et al. [47], with permission.

with an applied field parallel and the other one with an applied field perpendicular to the sample surface are depicted in **Figure 9**.

The magnetic response of Ni deposited within porous silicon powder has been investigated with respect to the plating time and the concentration of the solution. The results show a superparamagnetic behavior, except for a plating time of 15 min. The measured magnetization increases with the plating time till a value of 360 min and afterwards decreases again [61]. **Figure 10** shows the hysteresis for various plating times (a) and the normalized saturated magnetization in dependence on the deposition time (b) [61].

Also a non-superparamagnetic behavior has been observed for Co deposited within porous silicon particles, but all samples show a low coercivity [62] which indicates a low dispersion of Co within the porous structure offering only weak magnetic crosstalk.

MESOPOROUS SILICON

Mesoporous silicon offers in contrast to microporous silicon straight pores with more or less dendritic growth. The dendritic grows mainly depends on the electrochemical etching parameters such as applied current density, bath temperature and electrolyte concentration. Keeping the temperature and electrolyte concentration constant the regularity of the pore arrangement increases with increasing current density [63]. The dendrite-size and especially their length is determined by the distance between the pores. If this inter-pore distance exceeds

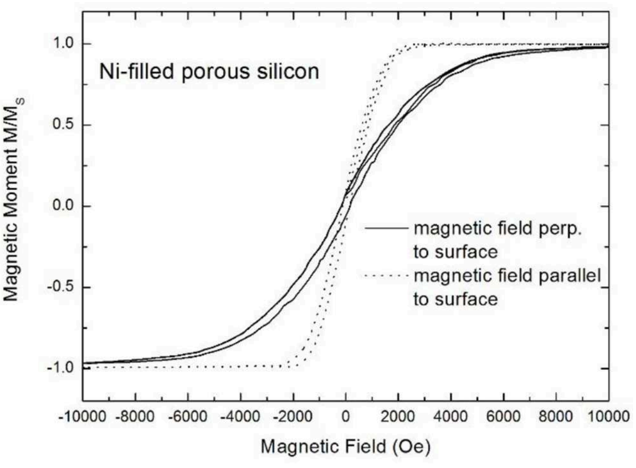

FIGURE 9 | Field dependent magnetization curves measured with an applied field parallel (dotted line) and perpendicular (solid line) to the sample surface showing that the magnetic easy axis is parallel to the surface. Granitzer et al. [47], with permission.

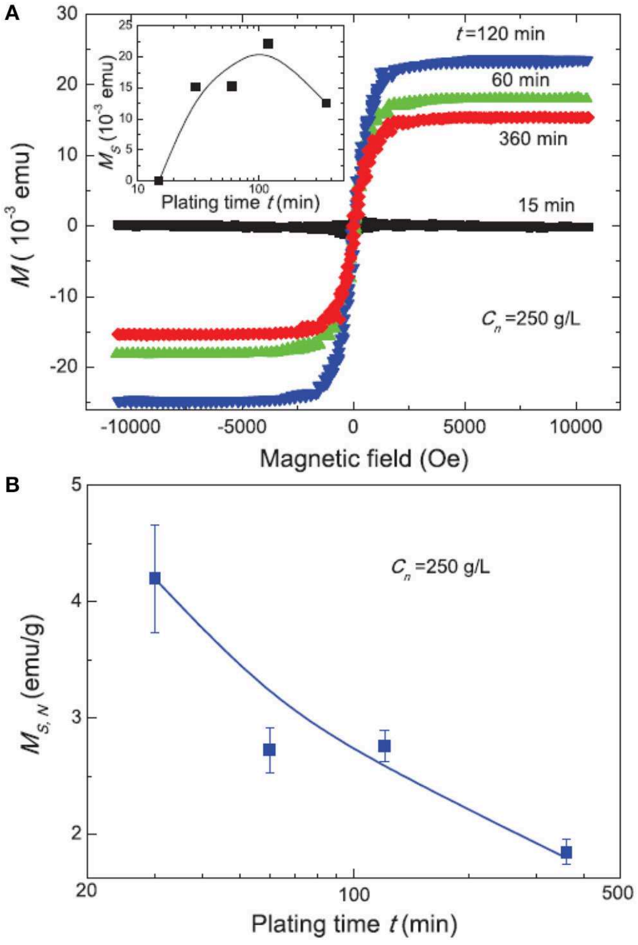

FIGURE 10 | Field dependent magnetization curves of Ni/Psi in dependence on the deposition time **(A)** and normalized saturation magnetization in dependence on the deposition time **(B)**. The inset in **(A)** shows the saturated magnetization vs. time. Reproduced from Nakamura et al. [61], with permission of AIP Publishing.

twice the thickness of the space charge region the silicon skeleton is not any longer free of charge carriers which favors the formation of side pores [7]. The growth direction depends on the crystal orientation and this feature increases when the applied current density comes close to the critical current density at the pore tips. The fastest growth rate is along the (100) direction [7].

Fabrication and Metal Filling of Mesoporous Silicon

Mesoporous and macroporous silicon with pore diameters in the lower regime is generally fabricated in using a highly n- or p- doped silicon wafer. A standard formation procedure is anodic dissolution of bulk crystalline silicon in an aqueous hydrofluoric acid solution. In applying a high current density the breakthrough regime is enabled. In this case illumination of n-type silicon to generate holes for the silicon dissolution is not necessary. In the mesoporous regime straight self-organized nanoholes are formed. The pore diameter and the concomitant pore distance can be adjusted quite accurately by the electrochemical parameters.

A further common technique is stain etching, which means electroless etching of silicon in the presence of a solution generally containing acidic fluoride and an oxidant which injects holes into the valence band [64]. Using this electroless formation process it is more difficult to adjust the porous silicon morphology and the obtained porous layers are quite thin (few micrometers). In adding Fe^{3+}, VO_2^+, or Ce^{4+} as oxidizing agent, thicker (>10 μm), more homogeneously and reproducible porous layers can be formed than with standard solutions [65].

Beside its large surface area, low electrical and thermal conductivity, mesoporous silicon can also offer photoluminescence in the visible range for structure sizes small enough for quantum confinement effects. Mesoporous silicon powder has been produced by anodization technique, subsequent detaching of the porous layer by applying a high current to exceed the critical current density, and finally hand milling of the detached porous layer. The samples were then treated by high pressure water vapor annealing for stabilization and to enhance the luminescence of the material [66].

In the following discussion porous silicon samples produced by anodization are considered. Since the morphology of the material can be tuned by the formation process, porous silicon is an adequate host material for pore filling. Beside attaching molecules to the surface for biomedical [67–69] or sensor [70–72] applications, the loading of the pores with metals has been investigated intensely. Requirements for this purpose are straight and separated pores to guarantee an arrangement of metal nanostructure arrays. The filling mechanism of different metals has been examined with the goal to control the metal deposition to produce tailored metal nanorods within the pores.

It has been reported that Fe deposition in low doped n-type silicon starts at the pore tips and grows along the walls. Mass

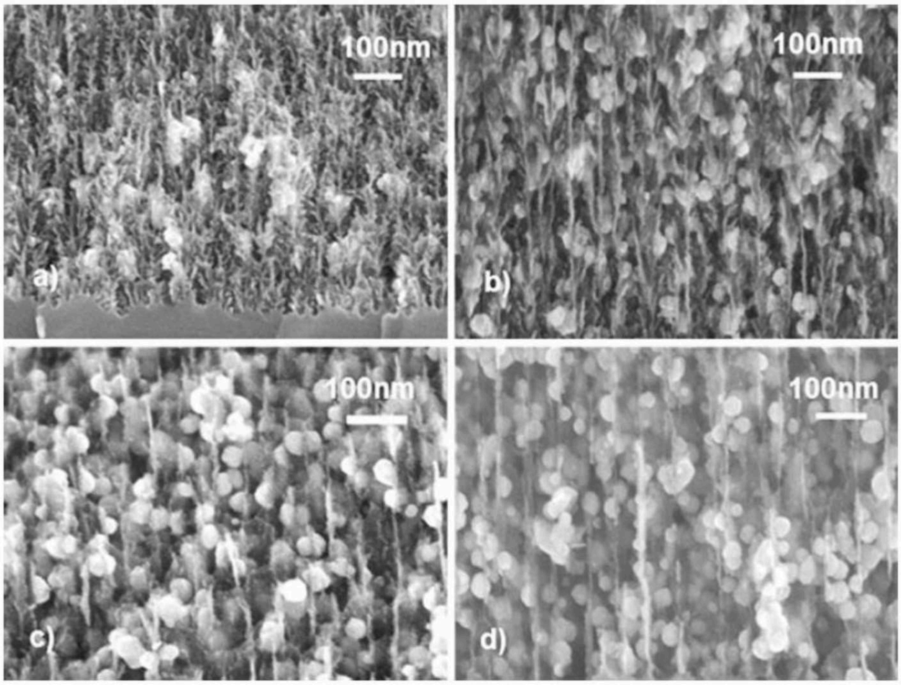

FIGURE 11 | Cross-sectional scanning electron micrographs of porous silicon offering different morphologies with infiltrated Ni structures. The Ni deposition depends on the porous silicon porosity and it can be seen that for lower porosity the Ni particles are smaller than in the case of higher porosity. This behavior is due to faster exhaustion of the electrolyte and thus a longer growth time. Michelakaki et al. [80], Springer Nature with permission. In **(a,b)**, the samples offer a porosity of about 70% and the Ni nanoparticle size is ∼23 nm. With increasing porosity to 86% and 88% shown in **(c,d)**, the Ni particle size increases to about 33 nm.

transport limited conduction which accumulates electrons at the sample surface has been avoided in applying a low current density and in using a porous silicon morphology with wide pore opening [73]. Fe electrodeposition on p-type porous silicon is performed under cathodic conditions and under illumination, using n-type porous silicon, illumination can be omitted. Electron transfer reaction via the conduction band occurs in n-type porous silicon but in the case of p-type porous silicon electrons are generated by illumination [74]. Fe incorporated by electrodeposition within mesoporous silicon nucleates on the walls of the pores and on the surface. Generally the metal formation depends on the applied current density, pH of the electrolyte and the surface chemistry of the pore walls. Performing the deposition on an H-terminated porous silicon surface, Fe forms clusters at the surface. In using porous silicon with a native oxide layer Fe is deposited as randomly distributed particles within the pores [75]. Beside electrodeposition the Fe incorporation can also be performed by electroless deposition from the Fe salt solution [76].

Ni deposition can also be performed either electroless or electrochemically. In the case of immersion plating an aqueous ammonium fluoride solution containing $NiSO_4$ is employed. Performing the deposition under room temperature metallic Ni is observed at the surface, whereas SiO_2 is not detected [77]. Displacement deposition of Ni from an $NiSO_4$, NH_4F solution carried out at 60°C starts to grow at the pore walls rather than on the surface due to the Ni atoms mass transport toward the pore tips [78]. Displacement deposition of Ni produces a non-uniform deposition in high aspect ratio pores (∼200) at the beginning of the reaction process due to a dominance of the mass transport. With increasing deposition time the Ni distribution along the pores becomes more uniform because the deposition rate decreases and the process is dominated by interfacial reaction [79].

In using electrodeposition for the metal filling of the pores, either dc or pulsed deposition can be employed. Pulsed electrodeposition under constant current density and applying rectangular pulses with a pulse duration of 1 s and a pulse delay time of 25 s is shown for different porous silicon porosities in **Figure 11** [80]. The pulsed current deposition leads to nucleation of metal nanoparticles within the pores. A crucial parameter is the pulse delay time which can be estimated from the dynamics of the voltage vs. time graphs [81].

The shape of the metal deposits can be adjusted by the pulse-frequency and the applied current density. A variation of the frequency between 0.025 and 0.2 Hz results in a shape modification from spherical particles to elongated structures with an aspect ratio of about 100 [82]. In **Figure 12** the dependence of the elongation of Ni structures on the pulse duration is depicted [83]. The spatial distribution of the deposits within the pores can be modified by the current density. Applying a current density between 5 and 25 mA/cm^2, the packing density of the metal deposits within the pores can be modified by the pulse frequency. In the case of densely packed particles the time between the pulses is 5 s and for loosely packed particles it is 20 s [82].

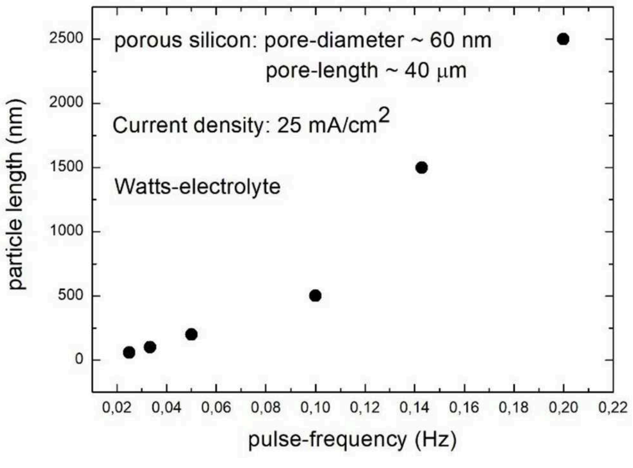

FIGURE 12 | Pulse frequency vs. elongation of Ni deposits. With increasing pulse frequency the length of the Ni structures is increased. Granitzer et al. [83], ECS with permission.

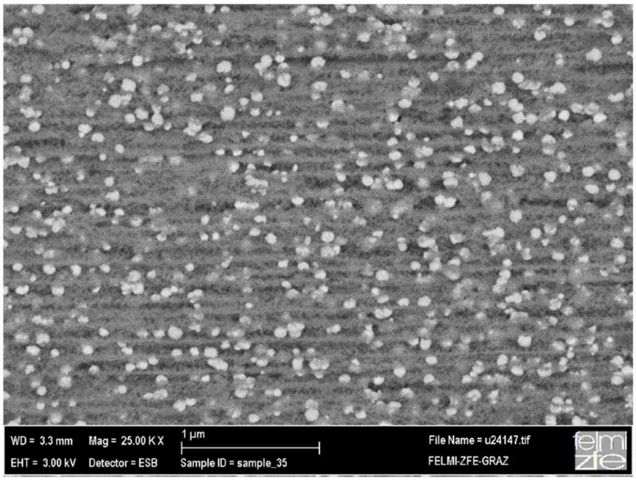

FIGURE 13 | Cross-sectional back scattered electron micrograph (BSE) showing the distribution of sphere-like Ni deposits within porous silicon. Rumpf et al. [82], IOP with permission.

In applying a current density of 25 mAcm^{-2} and a pulse frequency of 0.025 Hz more or less homogeneous deposition of spherical Ni particles with a moderate packing density is achieved and can be seen in **Figure 13** [82].

Optical Properties of Metal Filled Mesoporous Silicon

Not only from microporous but also from mesoporous silicon in the lowermost size regime luminescence can be observed. Furthermore, it can be utilized as template for metal particle deposition on the surface and act as SERS sensitive material. Mesoporous silicon produced by a standard chemical route, offering pore diameters of about 15 nm, and subsequent Ag particle deposition on its surface is shown to be Surface Enhanced Raman Scattering (SERS) active for rhodamine 6G and crystal violet. In using an excitation wavelength of 514.5 nm the rhodamine 6G and the silver plasmons are in resonance with the excitation light resulting in a large surface enhancement [84]. Ag nanoparticles deposited on porous silicon by immersing the sample into an $AgNO_3$ solution are used for single molecule detection by SERS. The Raman response is investigated in employing Cyanine and Rhodamine 6G. If the particle plasmon resonance coincides with the molecule electronic resonance strong Raman enhancement is observed [85]. The utilization of mesoporous silicon as template material in comparison to microporous silicon for Ag particle deposition shows a broader particle size distribution as well as a localized surface plasmon resonance closer to the excitation wavelength. Due to the high SERS sensitivity ultralow concentrations of dye molecules can be detected [86]. Furthermore, Ag nanostructures in SiO_2/p-silicon is found as SERS active substrate in a broad spectral range [87]. Au nanoparticles modified porous silicon shows a strong fluorescence enhancement due to plasmon resonance which is used for highly sensitive DNA detection [88]. Porous silicon with a Au nanoparticle layer on top offers light absorption by the Au particles leading to localized surface plasmon resonance. The wavelength of the plasmon resonance depends not only on the Au nanostructures but also on the refractive index of the surrounding medium. Plasmon resonance is sensitive to changes in the refractive index only close to the dielectric/metal interface [89]. Metal decorated porous silicon structures are feasible as sensors, especially to detect molecules.

TABLE 1 | Dependence of the coercivity H_C on the particle elongation.

Particle size [nm]	H_C [Oe]	M_R/M_S
60	500	0.54
500	350	0.45
1000	270	0.28

Magnetic Response of Metal Filled Mesoporous Silicon

Mesoporous silicon with its tunable morphology by the formation parameters, its straight pores and its quasi regular pore arrangement is an adequate system to embedded magnetic nanostructures and to adjust the resulting magnetic properties. On the one hand the magnetic response strongly depends on the size, shape and distribution of the metal deposits and on the other hand the morphology of the porous silicon plays a crucial role, especially with respect to magnetic interactions between deposits of adjacent pores. The magnetic response of the nanocomposite samples allows to draw conclusions from the shape of the deposits. The coercivity H_C decreases with the elongation of the embedded metal structures whereas the reduced remanence (magnetic remanence M_R/saturation magnetization M_S) increases with increasing elongation approaching a wire-like behavior [90], summarized in **Table 1**.

TABLE 2 | Magnetic features measured at T = 4, 100 K and T = 250 K summarized for conventional etched samples and magnetic field assisted etched ones.

	Coercivity [Oe] mag. field perp. to surface	Coercivity [Oe] mag. field paral. to surface	Remanence M/M_S [emu] mag. field perp. to surface	Remanence M/M_S [emu] mag. field paral. to surface
T = 4 K (conv.)	270	180	0.42	0.36
T = 100 K (conv.)	200	110	0.40	0.28
T = 250 K (conv.)	160	75	0.38	0.22
T = 4 K (mag.)	660	190	0.85	0.38
T = 100 K (mag.)	570	125	0.81	0.34
T = 250 K (mag.)	540	100	0.78	0.28

Granitzer et al. [91], Springer with permission.

Furthermore, the packing density of the metal deposits within the pores influences the magnetic characteristics. Densely packed metal particles magnetically interact within the pores leading to a wire like behavior and thus to a smaller coercivity compared to isolated particles with comparable size. Field dependent magnetization measurements performed at various temperatures ranging from 4 to 300 K shows that the coercivity decreases with increasing temperature [91]. Some values for T = 4,100, and 250 K are depicted in **Table 2**.

The magnetic response is also strongly influenced by the morphology of the porous silicon template, especially regarding the dendritic pore growth [92]. A reduction of the side pore growth could be achieved by magnetic field assisted etching, in applying a magnetic field of 8 T perpendicular to the sample surface during the anodization process [93]. With increasing side pore length the effective mean distance between the pores decreases resulting in stronger magnetic cross talk between metal deposits within adjacent pores. A further important parameter is beside the dendritic pore growth, the roughness of the pores because the embedded metal deposits grow with concomitant morphology. In the case of occurrence of side pores the metal deposits offer strong roughness which results in magnetic stray fields, influencing the magnetic response and reducing the coercivity of the nanocomposite [92]. Performing magnetization measurements with an applied field parallel and perpendicular to the sample surface one sees that the magnetic easy axis corresponds to the pore direction. The magnetic anisotropy between these two magnetization directions increases with increasing length of the metal deposits and with decreasing magnetic interactions between structures of neighboring pores. **Figure 14** shows the magnetic anisotropy for magnetic field assisted etched porous silicon with elongated Ni structures deposited within the pores.

Considering Fe deposits within porous silicon, in and out of plane magnetization measurements show a magnetic anisotropy originating from the elongated shape of the deposits [75]. The easy axis corresponds to the pore direction with a reduced remanence (M_R/M_S) of 0.6 [75] which is a typical value for such nanocomposites including magnetic cross talk.

These investigations show that control of the porous silicon formation resulting in more or less ordered pores with a minimum on roughness are a precondition to suppress magnetic coupling between the deposited metal structures of adjacent pores. By adjusting the metal deposition parameters accurately desired size and shape of the deposited nanostructures can be obtained.

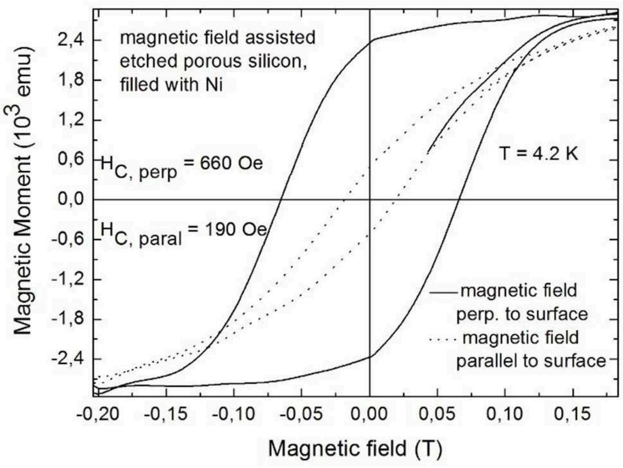

FIGURE 14 | Magnetic field dependent measurements performed with an external field applied perpendicular (solid line) and parallel (dotted line) to the sample surface. Granitzer et al. [94], AIP with permission.

CONCLUSIONS

In this work the utilization of porous silicon in the microporous and mesoporous regime as template material for metal deposition has been discussed. The optical properties such as the luminescence and the associated decay times of microporous silicon have been addressed also with respect to metal filling of these templates which modifies and improves the light emission. The plasmon resonance of the deposited metal structures can be exploited to enhance the luminescence and in combination with dye molecules the system can act as active SERS material. Due to the employment of magnetic metals nanocomposites with adjusted magnetic features strongly correlated to the size, shape, and distribution of the deposits can be achieved. In the case of microporous silicon optical and magnetic properties are merged on one material level which makes the system interesting for magneto-optical applications. A further key parameter concerning the magnetic features is the morphology of the porous silicon, especially the distance between the pores which is strongly influenced by dendritic pore growth. A reduction of the side pore length results in a decrease of the magnetic crosstalk and thus approaches the magnetic characteristics of individual magnetic nanostructures. Since these nanocomposites offer a silicon substrate which is applicable in today's microtechnology the discussed systems are promising to act as component in on-chip devices.

AUTHOR CONTRIBUTIONS

PG and KR wrote this review article and they are responsible for the content of the work.

REFERENCES

1. Canham L. Silicon quantum wire array fabrication by electrochemical and chemical dissolution of wafers. *Appl Phys Lett*. (1990) **57**:1046. doi: 10.1063/1.103561
2. Kolasinski KW. New approaches to the production of porous silicon by stain etching. In: Granitzer P, Rumpf K, editors. *Nanostructured Semiconductors - From Basic Research to Applications*. Singapore: Pan Stanford Publishing (2014). p. 45–72.
3. Chartier C, Bastide S, Levy-Clement C. Metal-assisted chemical etching of silicon in HF-H_2O_2. *Electrochim Acta*. (2008) **53**:5509–16. doi: 10.1016/j.electacta.2008.03.009
4. Kolasinski KW. The mechanism of galvanic/metal-assisted etching of silicon. *Nanoscale Res Lett*. (2014) **9**:432. doi: 10.1186/1556-276X-9-432
5. Gianneta V, Olziersky A, Nassiopoulou AG. Si nanopatterning by reactive ion etching through an on-chip self-assembled porous anodic alumina mask. *Nanoscale Res Lett*. (2013) **8**:71. doi: 10.1186/1556-276X-8-71
6. Zhang J, Zhao L, Rosenkranz A, Song C, Yan Y, Sun T. Nanosecond pulsed laser ablation of silicon-finite element simulation and experimental validation. *J Micromech Microeng*. (2019) **29**:075009. doi: 10.1088/1361-6439/ab208b
7. Lehmann V. *Electrochemistry of Silicon - Instrumentation, Science, Materials and Applications*. Weinheim: Wiley-VCH (2002).
8. Herino R, Bomchil G, Barla K, Bertrand C, Ginoux JL. Porosity and pore size distributions of porous silicon layers. *J Electrochem Soc*. (1987) **134**:1994. doi: 10.1149/1.2100805
9. Herino R. Pore size distribution in porous silicon. In: Canham LT, editor. *Properties of Porous Silicon*. London: INSPEC (1997). p. 89–97.
10. Cullis AG, Canham LT. Visible light emission due to quantum size effects in highly porous crystalline silicon. *Nature*. (1991) **353**:335. doi: 10.1038/353335a0
11. Lehmann V, Jobst B, Muschik T, Kux A, Petrova-Koch V. Correlation between optical properties and crystallite size in porous silicon. *Jpn J Appl Phys*. (1993) **32**:2095. doi: 10.1143/JJAP.32.2095
12. Zeman J, Zigone M, Rikken GLJA, Martinez G. Hydrostatic pressure effects on the porous silicon luminescence. *J Phys Chem Solids*. (1995) **56**:655–61. doi: 10.1016/0022-3697(94)00259-2
13. Fauchet PM, Tsybeskov L, Peng C, Duttagupta SP, von Behren J, Kostoulas Y, et al. Light-emitting porous silicon: materials science, properties, and device applications. *IEEE J Select Top Quant Electron*. (1995) **1**:1126. doi: 10.1109/2944.488691
14. Canham LT. Luminescence bands and their proposed origins in highly porous silicon. *Phys Stat Sol*. (1995) **190**:9. doi: 10.1002/pssb.2221900102
15. Mauckner G, Hamann J, Rebitzer W, Baier T, Thonke K, Sauer R. Origin of the infrared band from porous silicon. *Mater Res Soc Symp Proc*. (1995) **358**:489. doi: 10.1557/PROC-358-489
16. Li P, Wang G, Ma Y, Fang R. Origin of the blue and red photoluminescence from aged porous silicon. *Phys Rev B*. (1998) **58**:4057. doi: 10.1103/PhysRevB.58.4057
17. Grivickas V, Linnros J. Free-carrier absorption and luminescence decay of porous silicon. *Thin Solid Films*. (1995) **255**:70–3. doi: 10.1016/0040-6090(94)05606-E
18. Trojanek F, Maly P, Pelant I, Hospodkova A, Kohlova V, Valenta J. Picosecond dynamics of photoexcited carriers in free-standing porous silicon. *Thin Solid Films*. (1995) **255**:77–9. doi: 10.1016/0040-6090(94)05610-P
19. M'Ghaleth R, Maaref H, Mihalcescu I, Vial JC. Porous silicon: photoluminescence decay in the nanosecond range. *Microelectron J*. (1999) **30**:695–8. doi: 10.1016/S0026-2692(99)00013-0
20. Joo J, Defforge T, Loni A, Kim D, Li ZY, Sailor MJ, et al. Enhanced quantum yield of photoluminescent porous silicon prepared by supercritical drying. *Appl Phys Lett*. (2016) **108**:153111. doi: 10.1063/1.4947084
21. Gelloz B, Juangsa FB, Nozaki T, Asaka K, Koshida N, Jin L. Si/SiO_2 core/shell luminescent silicon nanocrystals and porous silicon powders with high quantum yield, long lifetime, and good stability. *Front Phys*. (2019) **7**:47. doi: 10.3389/fphy.2019.00047
22. Koshida N, Koyama H. Visible electroluminescence from porous silicon. *Appl Phys Lett*. (1992) **60**:347. doi: 10.1063/1.106652
23. Lehmann V, Hofmann F, Möller F, Grüning U. Resistivity of porous silicon: a surface effect, *Thin Solid Films*. (1995) **255**:20. doi: 10.1016/0040-6090(94)05624-M
24. Herino R. Impregnation of porous silicon. In: Canham L, editor. *Properties of Porous Silicon*. London: INSPEC (1997). p. 66–77.
25. Rahmani M, Moadhen A, Zaibi MA, Elhouichet H, Oueslati M. Photoluminescence enhancement and stabilisation of porous silicon passivated by iron. *J Lumin*. (2008) **128**:1763–6. doi: 10.1016/j.jlumin.2008.04.003
26. Amdouni S, Rahmani M, Zaibi MA, Oueslati M. Enhancement of porous silicon photoluminescence by electroless deposition of nickel. *J Lumin*. (2015) **157**:93–7. doi: 10.1016/j.jlumin.2014.08.041
27. Bouzouraa MB, Rahmani M, Zaibi MA, Lorrain N, Hajji L, Oueslati M. Optical study of annealed cobalt-porous silicon nanocomposites. *J Lumin*. (2013) **143**:521–5. doi: 10.1016/j.jlumin.2013.05.050
28. Lublow M, Kubala S, Veyan JF, Chabal YJ. Colored porous silicon as support for plasmonic nanoparticles. *J. Appl. Phys*. (2012) **111**:084302. doi: 10.1063/1.3703469
29. Lehmann V, Stengl R, Luigart A. On the morphology and the electrochemical formation mechanism of mesoporous silicon. *Mater Sci Eng B*. (2000) **69–70**:11–22. doi: 10.1016/S0921-5107(99)00286-X
30. Granitzer P, Rumpf K, Pölt P, Albu M, Chernev B. The interior interfaces of a semiconductor/metal nanocomposite and their influence on its physical properties. *Phys Stat Sol*. (2009) **6**:2222–7. doi: 10.1002/pssc.200881730
31. Granitzer P, Rumpf K, Ohta T, Koshida N, Poelt P, Reissner M. Magnetic field assisted etching of porous silicon as a tool to enhance magnetic characteristics. *ECS Trans*. (2013) **50**:55. doi: 10.1149/05037.0055ecst
32. Uhlir A Jr. Electrolytic shaping of germanium and silicon. *Bell System Tech J*. (1956) **35**:333–47. doi: 10.1002/j.1538-7305.1956.tb02385.x
33. Ram SK. Electrical transport in porous silicon. In: Canham LT, editor. *Handbook of Porous Silicon*. Cham: Springer (2018). p. 263–79.
34. Zhao Y, Yang L, Kong L, Nai MH, Liu D, Wu J, et al. Ultralow thermal conductivity of single-crystalline porous silicon nanowires. *Adv Func Mater*. (2017) **27**:1702824. doi: 10.1002/adfm.201702824
35. Sohn H. Refractive index of porous silicon. In: Canham LT, editor. *Handbook of Porous Silicon*. Cham: Springer (2018). p. 231–43.
36. Fujii M, Diener J. Optical birefringence of porous silicon. In: Canham LT, editor. *Handbook of Porous Silicon*. Cham: Springer (2018). p. 245–53.
37. Canham L. Mechanical properties of porous silicon. In: Canham LT, editor. *Handbook of Porous Silicon*. Cham: Springer (2018). p. 213–20.
38. Lehmann V, Grüning U. The limits of macropore array fabrication. *Thin Solid Films*. (1997) **297**:13. doi: 10.1016/S0040-6090(96)09478-3
39. Lehmann V, Rönnebeck S. The physics of macropore formation in low-doped p-type silicon. *J Electrochem Soc*. (1999) **146**:2968. doi: 10.1149/1.1392037
40. Halimaoui A. Determination of the specific surface area of porous silicon from its etch rate in HF solutions. *Surf Sci Lett*. **306**:L550–4 (1994). doi: 10.1016/0039-6028(94)91176-2
41. Kolasinski K. Porous silicon formation by stain etching. In: Canham LT, editor. *Handbook of Porous Silicon*. Cham: Springer (2018). p. 35–48.
42. Kolasinski K, Yadlovskiy J. Stain etching of silicon with V_2O_5. *Phys Stat Sol*. (2011) **8**:1749–53. doi: 10.1002/pssc.201000063
43. Ossicini S, Pavesi L, Priolo F. *Light Emitting Silicon for Microphotonics*. Berlin: Springer (2003).
44. Koda R, Fukami K, Sakka T, Ogata YH. A physical mechanism for suppression of zinc dendritescaused by high efficiency of the electrodeposition within confined nanopores. *ECS Electrochem Lett*. **2**:D9 (2013). doi: 10.1149/2.010302eel
45. Koda R, Fukami K, Sakka T, Ogata YH. Electrodeposition of platinum and silver into chemically modified microporous silicon electrodes. *Nanoscale Res Lett*. (2012) **7**:330. doi: 10.1186/1556-276X-7-330
46. Fukami K, Koda R, Sakka T, Urata T, Amano K, Takaya H, et al. Platinum electrodeposition in porous silicon: the influence of surface solvation effects on a chemical reaction in a nanospace. *Chem Phys Lett*. (2012) **542**:99–105. doi: 10.1016/j.cplett.2012.05.078
47. Granitzer P, Rumpf K, Poelt P, Reissner M. Magnetic characteristics of Ni-filled luminescent porous silicon. *Front Chem*. (2019) **7**:41. doi: 10.3389/fchem.2019.00041

48. Wolkin MV, Jorne J, Fauchet PM, Allan G, Delerue C. Electronic states and luminescence in porous silicon quantum dots: the role of oxygen. *Phys Rev Lett*. (1999) **82**:197. doi: 10.1103/PhysRevLett.82.197
49. Brandt MS, Fuchs HD, Stutzmann M, Weber J, Cardona M. The origin of visible luminescencefrom "porous silicon": a new interpretation. *Solid State Commun*. (1992) **81**:307–12. doi: 10.1016/0038-1098(92)90815-Q
50. Petrova-Koch V, Muschik T, Kux A, Meyer BK, Koch F, Lehmann V. Rapid-thermal-oxidized porous Si–the superior photoluminescent Si. *Appl Phys Lett*. (1992) **61**:943. doi: 10.1063/1.107736
51. Koch F, Petrova-Koch V, Muschik T. The luminescence of porous Si: the case for the surface state mechanism. *J Lumin*. (1993) **57**:271–81. doi: 10.1016/0022-2313(93)90145-D
52. Arad-Vosk N, Sa'ar A. Radiative and nonradiative relaxation phenomena in hydrogen- and oxygen-terminated porous silicon. *Nanoscale Res Lett.* (2014) **9**:47. doi: 10.1186/1556-276X-9-47
53. Dovrat M, Goshen Y, Jedrzejewski J, Balberg I, Sa'ar A. Radiative versus nonradiative decay processes in silicon nanocrystals probed by time-resolved photoluminescence spectroscopy. *Phys Rev B*. (2004) **69**:155311. doi: 10.1103/PhysRevB.69.155311
54. Vainshtein JS, Goryachev DN, Ken OS, Sreseli OM. Surface plasmon polaritons in a composite system of porous silicon and gold. *Semiconductors.* (2015) **49**:442–7. doi: 10.1134/S1063782615040260
55. de la Mora MB, Bornacelli J, Nava R, Zanella R, Reyes-Esqueda JA. Porous silicon photoluminescence modification by colloidal gold nanoparticles: plasmonic, surface and porosity roles. *J Lumin*. (2014) **146**:247–55. doi: 10.1016/j.jlumin.2013.09.053
56. Achermann M. Exciton–plasmon interactions in metal–semiconductor nanostructures. *J Phys Chem Lett.* (2010) **1**:2837. doi: 10.1021/jz101102e
57. Tang H, Liu C, He H. Surface plasmon enhanced photoluminescence from porous silicon nanowires decorated with gold nanoparticles. *RCS Adv*. (2016) **6**:59395. doi: 10.1039/C6RA06019F
58. Gardelis S, Gianneta V, Nassiopoulou AG. Twenty-fold plasmon-induced enhancement of radiative emission rate in silicon nanocrystals embedded in silicon dioxide. *J Lumin*. (2016) **170**:282–7. doi: 10.1016/j.jlumin.2015.10.029
59. Cetinel A, Artunc N, Tarhan E. The growth of silver nanostructures on porous silicon for enhanced photoluminescence: the role of $AgNO_3$ concentration and deposition time. *Eur Phys J Appl Phys*. (2019) **86**:11301. doi: 10.1051/epjap/2019190013
60. Azaiez K, Zaghouani RB, Khamlich S, Meddeb H, Dimassi W. Enhancement of porous silicon photoluminescence property by lithium chloride treatment. *Appl Surf Sci*. (2018) **441**:272–6. doi: 10.1016/j.apsusc.2018.02.006
61. Nakamura T, Adachi S. Properties of magnetic nickel/porous-silicon composite powders. *AIP Adv.* (2012) **2**:032167. doi: 10.1063/1.4754152
62. Munoz-Noval A, Sanchez-Vaquero V, Torres-Costa V, Gallach D, Ferro-Llanos V, Serrano JJ, et al. Hybrid luminescent/magnetic nanostructured porous silicon particles for biomedical applications. *J Biomed Optics.* (2011) **16**:025002. doi: 10.1117/1.3533321
63. Granitzer P, Rumpf K. Mesoporous silicon utilized as matrix for 3-dimensional arrays of ferromagnetic nanostructures. In: Burness LT, editor. *Mesoporous Materials: Properties, Preparation and Applications.* New York, NY: Nova Science Publishing (2009). p. 99–120.
64. Nahidi M, Kolasinski KW. Effects of stain etchant composition on the photoluminescence and morphology of porous silicon. *J Electrochem Soc.* (2006) **153**:C19 (2006). doi: 10.1149/1.2129558
65. Kolasinski KW. Charge transfer and nanostructure formation during electroless etching of silicon. *J Phys Chem C*. (2010) **114**:22098. doi: 10.1021/jp108169b
66. Gelloz B, Loni A, Canham L, Koshida N. Luminescence of mesoporous silicon powders treated by high-pressure water vapor annealing. *Nanoscale Res Lett*. (2012) **7**:382. doi: 10.1186/1556-276X-7-382
67. Rodriguez GA, Lawrie JL, Weiss SM. Nanoporous silicon biosensors for DNA sensing. In: Santos HA, editor. *Porous Silicon for Biomedical Applications*. Cambridge: Woodhead Publishing (2014). p. 304–33.
68. Kim B, Sun S, Varner JA, Howell SB, Ruoslahti E, Sailor MJ. Securing the payload, finding the cell, and avoiding the endosome: peptide-targeted, fusogenic porous silicon nanoparticles for delivery of siRNA. *Adv Mater*. (2019) **31**:1902952. doi: 10.1002/adma.201902952
69. Tieu T, Alba M, Elnathan R, Cifuetes-Rius A, Voelcker NH. Advances in porous silicon-based nanomaterials for diagnostic and therapeutic applications. *Adv Therap*. (2019) **2**:1800095. doi: 10.1002/adtp.201800095
70. De Stefano L. Porous silicon optical biosensors: still a promise or a failure? *Sensors*. (2019) **19**:4776. doi: 10.3390/s19214776
71. Barillaro G. Porous silicon gas sensing. In: Canham LT, editor. *Handbook of Porous Silicon*. Cham: Springer (2018). p. 845–56.
72. Ozdemir S, Gole JL. The potential of porous silicon gas sensors. *Curr Opin Solid State Mater Sci*. (2007) **11**:92. doi: 10.1016/j.cossms.2008.06.003
73. Renaux C, Scheuren V, Flandre D. New experiments on the electrodeposition of iron in porous silicon. *Microelectron Reliabil.* (2000) **40**:877–9. doi: 10.1016/S0026-2714(99)00331-5
74. Harraz FA, Sakka T, Ogata YH. A comparative electrochemical study of iron deposition onto n- and p-type porous silicon prepared from lightly doped substrates. *Electrochim Acta*. (2005) **50**:5340–8. doi: 10.1016/j.electacta.2005.03.013
75. Bardet B, Defforge T, Negulescu B, Valente D, Billoue J, Poveda P, et al. Shape-controlled electrochemical synthesis of mesoporous Si/Fe nanocomposites with tailored ferromagnetic properties. *Mater Chem Front*. (2017) **1**:190. doi: 10.1039/C6QM00040A
76. Miu M, Kleps I, Ignat T, Simion M, Bragaru A. Study of nanocomposite iron/porous silicon material. *J Alloys Compounds*. (2010) **496**:265–8. doi: 10.1016/j.jallcom.2010.01.058
77. Harraz FA, Sasano J, Sakka T, Ogata YH. Different behavior in immersion plating of nickel on porous silicon from acidic and alkaline fluoride media. *J Electrochem Soc*. (2003) **150**:C277. doi: 10.1149/1.1562595
78. Xu C, Zhang X, Tu K-N, Xie Y. Nickel displacement deposition of porous silicon with ultrahigh aspect ratio. *J Electrochem Soc.* (2007) **154**:D170. doi: 10.1149/1.2430690
79. Xu C, Li M, Zhang X, Tu K-N, Xie Y. Theoretical studies of displacement deposition of nickel into porous silicon with ultrahigh aspect ratio. *Electrochim Acta.* (2007) **52**:3901–9. doi: 10.1016/j.electacta.2006.11.007
80. Michelakaki E, Valalaki K, Nassiopoulou AG. Mesoscopic Ni particles and nanowires by pulsed electrodeposition into porous Si. *J Nanopart Res*. (2013) **15**:1499. doi: 10.1007/s11051-013-1499-3
81. Munoz-Noval A, Gallach D, Garcia MA, Ferro-Llanos V, Herrero P, Fukami K, et al. Characterization of hybrid cobalt-porous silicon systems: protective effect of the matrix in the metal oxidation. *Nanoscale Res Lett.* (2012) **7**:495. doi: 10.1186/1556-276X-7-495
82. Rumpf K, Granitzer P, Hilscher G, Pölt P. Interacting low dimensional nanostructures within a porous silicon template. *J Phys Conf Ser*. (2011) **303**:012048. doi: 10.1088/1742-6596/303/1/012048
83. Granitzer P, Rumpf K, Koshida N, Pölt P, Michor H. Electrodeposited metal nanotube/nanowire arrays in mesoporous silicon and their morphology dependent magnetic properties. *ECS Trans*. (2014) **58**:139. doi: 10.1149/05832.0139ecst
84. Zeiri L, Rechav K, Porat Z, Zeiri Y. Silver nanoparticles deposited on porous silicon as a surface-enhanced raman scattering (SERS) active substrate. *Appl Spectr*. (2012) **66**:294–9. doi: 10.1366/11-06476
85. Virga A, Rivolo P, Frascella F, Angelini A, Descrovi E, Geobaldo F, et al. Silver nanoparticles on porous silicon: approaching single molecule detection in resonant SERS regime. *J Phys Chem C*. (2013) **117**:20139. doi: 10.1021/jp405117p
86. Kosovic M, Balarin M, Ivanda M, Derek V, Marcius M, Ristic M, et al. Porous silicon covered with silver nanoparticles as surface-enhanced raman scattering (SERS) substrate for ultra-low concentration detection. *Appl Spectr*. (2015) **69**:1417. doi: 10.1366/14-07729
87. Yakimchuk D, Kaniukov E, Bundyukova V, Osminkina L, Teichert S, Demyanov S, et al. Silver nanostructures evolution in porous SiO_2/p-Si matrices for wide wavelength surface-enhanced Raman scattering applications. *MRS Commun.* (2018) **8**:95. doi: 10.1557/mrc.2018.22
88. Wang J, Jia Z. Metal nanoparticles/porous silicon microcavity enhanced surface plasmon resonance fluorescence for the detection of DNA. *Sensors*. (2018) **18**:661. doi: 10.3390/s18020661
89. Balderas-Valadez RF, Schürmann R, Pacholski C. One spot-two sensors: porous silicon interferometers in combination with gold nanostructures

showing localized surface plasmon resonance. *Front Chem*. (2019) **7**:593. doi: 10.3389/fchem.2019.00593

90. Rumpf K, Granitzer P, Koshida N, Poelt P, Reissner M. Magnetic interactions between metal nanostructures within porous silicon. *Nanoscale Res Lett*. (2014) **9**:412. doi: 10.1186/1556-276X-9-412
91. Granitzer P, Rumpf K, Ohta T, Koshida N, Poelt P, Reissner M. Porous silicon/Ni composites of high coercivity due to magnetic field-assisted etching. *Nanoscale Res Lett*. (2012) **7**:384. doi: 10.1186/1556-276X-7-384
92. Rumpf K, Granitzer P, Koshida N, Pölt P, Michor H. Morphology controlled magnetic interactions in metal embedded porous silicon nanostructures. *ECS J Solid State Sci Technol*. (2015) **4**:N41. doi: 10.1149/2.0221505jss
93. Hippo D, Nakamine Y, Urakawa K, Tsuchiya Y, Mizuta H, Koshida N, et al. Formation mechanism of 100-nm-scale periodic structures in silicon using magnetic-field-assisted anodization. *Jpn J Appl Phys*. (2008) **47**:7398. doi: 10.1143/JJAP.47.7398
94. Granitzer P, Rumpf K, Ohta T, Koshida N, Reissner M, Poelt P. Enhanced magnetic anisotropy of Ni nanowire arrays fabricated on nano-structured silicon templates. *Appl Phys Lett*. (2012) **101**:033110. doi: 10.1063/1.4738780

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

REVIEW
published: 14 May 2020
doi: 10.3389/fphy.2020.00128

Multiphoton Microscopy of Oral Tissues: Review

Rosa M. Martínez-Ojeda [1], María D. Pérez-Cárceles [2], Lavinia C. Ardelean [3], Stefan G. Stanciu [4] and Juan M. Bueno [1]**

[1] Laboratorio de Óptica, Instituto Universitario de Investigación en Óptica y Nanofísica, Universidad de Murcia, Murcia, Spain, [2] Departmento de Medicina Legal y Forense, IMIB-Arrixaca, Facultad de Medicina, Universidad de Murcia, Murcia, Spain, [3] Department of Technology of Materials and Devices in Dental Medicine, "Victor Babes" University of Medicine and Pharmacy Timisoara, Timisoara, Romania, [4] Center for Microscopy-Microanalysis and Information Processing, University Politehnica of Bucharest, Bucharest, Romania

Multiphoton microscopy (MPM) is currently acknowledged as a very powerful method for the visualization and analysis of tissues in biomedicine. It allows high resolution, deep optical sectioning and reduced photodamage. MPM does not require labeling and is deployable both *in-vivo* and *ex-vivo*, which simplifies the diagnostic procedure compared to traditional histology approaches based on excisional biopsy, tissue fixation and staining. Among the important applications of MPM in medicine, differentiation of healthy from pathological tissues has gained massive interest over the past years, but MPM is also very useful for acquiring new insights on how various pathologies originate and progress. In this work we review the use of MPM in imaging assays focused on investigating unlabeled oral tissues (teeth and oral mucosa) and discuss a series of important results which hold potential for enabling a next generation of oral tissue characterization/diagnostic frameworks. The surveyed literature shows that non-linear optical imaging tools can significantly contribute to achieve a better understanding of oral cavity tissues, by allowing the accurate analysis of morphological structures and relevant biochemical processes.

Keywords: multiphoton microcopy, oral tissues, caries, enamel, dentine, squamous cell carcinoma

OPEN ACCESS

Edited by:
*Udo Jochen Birk,
University of Applied Sciences HTW
Chur, Switzerland*

Reviewed by:
*Susanne Fenz,
Julius Maximilian University of
Würzburg, Germany
Bernd H. Zinselmeyer,
Washington University in St. Louis,
United States*

****Correspondence:***
*Stefan G. Stanciu
stefan.stanciu@cmmip-upb.org
Juan M. Bueno
bueno@um.es*

Specialty section:
*This article was submitted to
Medical Physics and Imaging,
a section of the journal
Frontiers in Physics*

Received: *22 October 2019*
Accepted: *03 April 2020*
Published: *14 May 2020*

Citation:
*Martínez-Ojeda RM,
Pérez-Cárceles MD, Ardelean LC,
Stanciu SG and Bueno JM (2020)
Multiphoton Microscopy of Oral
Tissues: Review. Front. Phys. 8:128.
doi: 10.3389/fphy.2020.00128*

INTRODUCTION

The oral cavity represents the first part of the digestive tract, and the main entry for nutrients and different environmental components in the human body. Its main function is to begin the process of digestion by mastication, but speech or breathing are also important functions. Pathologies of the oral cavity can result from a wide range of causes, one of the main sources consisting in harmful components in ingested liquids or food. These may lead either to the genesis of oral diseases or, after being dissolved by saliva and ingested, may cause other minor to major health problems with respect to the gastrointestinal system, and not only. Besides toxic food, tobacco and alcohol, other environmental factors may also be responsible for the occurrence of various oral diseases, ranging from dental caries [1, 2] to oral squamous cell carcinoma [3–5]. In general, all the structures in the oral cavity, including teeth and soft tissues will be at some point affected during one's life by various types of pathologies, or by age related modifications.

Although teeth have a damage resistant structure, this might fail due to environmental factors, diseases or habits. The appearance of cavities may cause pain and other consequent oral problems. However, these are not the most severed pathologies the oral cavity is exposed to. According to current data, half of all cancers in humans appear in the squamous epithelium, which is also lining the soft surfaces in the oral cavity [6]. More specifically, oral squamous cell carcinoma (SCC) represents about 90% of the malignant lesions developed in the oral cavity [7].

At present, various complementary imaging techniques are being used for characterizing hard and soft tissues in the oral cavity, each exhibiting its own strengths and limitations. Since Rontgen discovered radiographs in the late 1800's, radiation was selectively used for the diagnostics and therapy of oral tissues [8]. Periapical and cephalometric radiographs have been used for detecting caries, analyzing bone structures and for planning implantological interventions. Their main disadvantages relate to the interposition of anatomic structures, and the potential harmful effects of the ionizing radiation [9]. Even in latest generation panoramic radiographs, direct ionizing damage and indirect damage from the free radicals created during the ionization of water molecules within cells is associated with a risk of cancer. Computed Tomography (CT), first reported experimentally in the 1970's, combines the concepts of x-ray imaging (performed under different angles) with the advantages of computer technology to provide cross-sectional images of the scanned tissue region, allowing its tomographic inspection [10]. A subsequent technology, Cone-Beam Computed Tomography (CBCT) [11] allows faster acquisition speed and lower radiation exposure, down to 10 times less compared to conventional CT. This diagnostic imaging technology is now widely popular for oral medicine being capable of providing three-dimensional representations of teeth and jaws. Besides limitations in resolution, its main shortcomings are exposure to ionizing radiation (which is reduced, but still exists), and inability to simultaneously image calcified and non-calcified dental tissues, particularly important in regenerative endodontics [12, 13]. Another medical-imaging technique that has been shown to be very valuable for imaging oral tissues is Magnetic Resonance Imaging (MRI), which uses non-ionizing radiation from the radiofrequency band of the electromagnetic spectrum, thus reducing the irradiation hazards that the patient needs to face. MRI can be used for imaging pulp attached to the periodontal membrane, and obviously other soft tissues in the oral cavity. However, it cannot easily visualize teeth because of their high mineral content. Sweep Imaging with Fourier Transformation (SWIFT) [14], an update to conventional MRI, overcomes these limitations, enabling the three-dimensional visualization of both soft and hard (enamel, denting, cortical bone) oral tissues, while also reducing acquisition time [15]. SWIFT-based MRI has the potential to precisely determine the extent of carious lesions and simultaneously assess pulpal tissue [15]. The usefulness of these aforementioned techniques is biased by their limited resolution (lying in the millimeter range). To address this, several safe and non-invasive optical techniques have been introduced to the field of oral tissue imaging over the past couple of decades. Among these, Optical Coherence Tomography (OCT) has successfully been used for the detection and microscale characterization of tooth and periodontal disorders. Applications such as root canal imaging, diagnosis of vertical root fractures, dental microstructure assessment, detection of recurrent caries and loss of marginal integrity of fixed restorations demonstrate its huge potential in dentistry [16–18]. Compared to conventional OCT, Polarization-Sensitive OCT (PS-OCT) [19] provides better resolution and can therefore image also early enamel lesions and secondary caries. It can also be used for the assessment of dentin and cement demineralization and remineralization, representing a useful diagnostic instrument for prevention and early intervention [20]. The resolution achievable with OCT is positioned between the resolution available with ultrasound-based techniques, and with point-scanning optical techniques, e.g., Confocal Laser Scanning Microscopy (CLSM). This latter technique, CLSM, is also well-suitable for *in vivo* clinical studies (in implementations for endomicroscopy), being able to non-invasively provide optical sections of both hard and soft tissues in the oral cavity [21–23]. Importantly, the current gold standard for the diagnostics of soft oral tissues remains the histopathological exam, consisting on brightfield microscopy of tissues that are excised, fixed and stained. However, this approach presents important disadvantages such as long diagnosis time, invasiveness, artifacts, sampling error, time consumption, high costs, and interpretive variability [24–26]. In **Figure 1**, we present a series of images that are representative for these techniques above discussed.

During the past couple of decades, Multiphoton Microscopy (MPM) has emerged as a powerful tool to explore the structure and function of biological samples, and especially of tissues. This is mainly because MPM techniques can non-invasively acquire optical sections (virtual biopsies) in unlabeled tissues, containing information that is very relevant for diagnostic purposes. These non-linear techniques are based on the theory of quantum transition through photons proposed by Nobel Laureate Maria Göppert-Mayer [32], and the first MPM experimental implementation was demonstrated by Denk, Strickler and Webb in Cornell University in 1990 [33]. During the non-linear processes that take place, the sample absorbs two or three infrared photons and emits a unique photon of shorter wavelength. This can occur via different physical processes, that may take place quasi-simultaneously, e.g., fluorescence or harmonic generation, which have been thoroughly discussed in previous reviews [34–36]. The MPM imaging modes are Two-Photon Excitation Fluorescence (2PEF), Three-Photon Excitation Fluorescence (3PEF), Second Harmonic Generation (SHG) and Third Harmonic Generation (THG), all of these being capable to image tissues in a label-free manner, based on their endogenous contrast. The use of infrared light sources allows deeper penetration into the tissues and reduced scattering [37, 38], with reduced photodamage compared to other optical sectioning techniques working in the visible range, such as CLSM [39].

The aforementioned MPM modalities are complementary and provide biological information that is relevant from different perspectives, including morphology, structural organization and cell metabolism. In brief, the energy (i.e., light) released by fluorophores during 2PEF allows the visualization of different biological components, such as elastin, keratin, melanin, nicotinamide adenine dinucleotide (NAD^+/NADH) or flavin adenine dinucleotide (FAD). Selective probing of these autofluorescent tissue components by 2PEF followed by arithmetic operations for distinct signals enable the non-invasive assessment of important information such as cell morphology, size variation of cell nuclei, blood vessel hyperplasia, or inflammatory reaction related aspects [40–42]. 3PEF relies on the

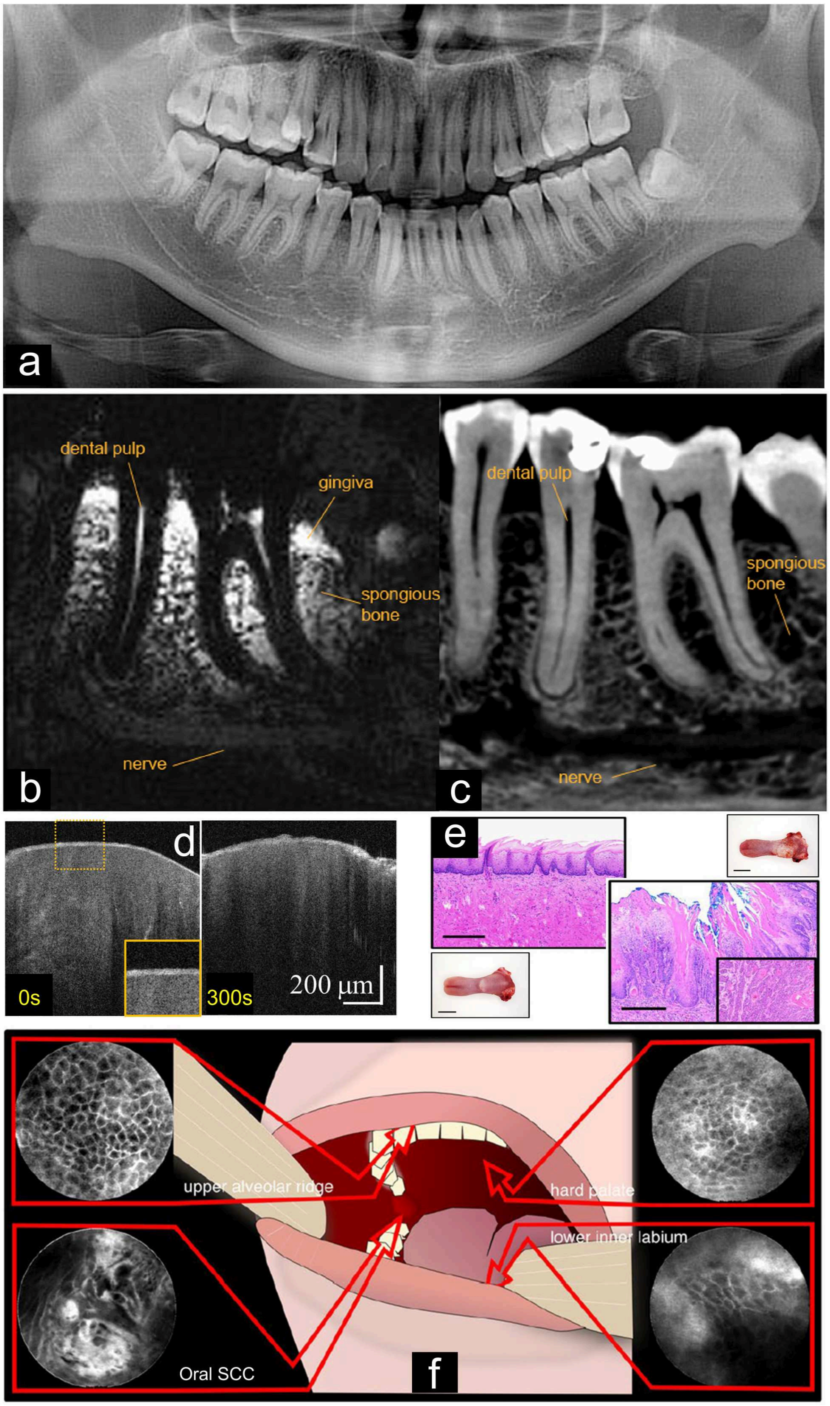

FIGURE 1 | Hard and soft tissues of the oral cavity imaged with routinely used investigation techniques. **(a)** Panoramic dental radiograph. **(b)** Sagital section through MRI and **(c)** CBCT images of the lower jaw. **(d)** Time-series OCT images of a healthy tooth obtained before and after demineralization **(e)** gross observation, and brightfield microscopy images on hematoxylin and eosin (H&E) stained healthy and malignant tongue tissue. **(f)** Confocal laser endomicroscopy images collected on various oral tissues. Artwork reuse permissions: **(a)** adapted from Kim et al. [27], **(b,c)** adapted from Flügge et al. [28], **(d)** adapted from Tsai et al. [29], **(e)** adapted from Kosugi et al. [30], **(f)** adapted from Aubreville et al. [31], all under the Creative Commons CC BY license.

same principles as 2PEF, but usually uses longer laser wavelength to excite the fluorescent molecules, which translates to reduced out of focus light, less tissue scattering, and hence higher penetration [37, 43]. However, 3PEF on tissues is more difficult to achieve and collect, hence related studies are much scarcer compared to the studies that deal with 2PEF imaging of tissues.

We find important to mention here that the potential of 2PEF and 3PEF for probing unlabeled *in-vivo, ex-vivo* or fixed tissues can be further augmented by equipping such systems with time-correlated single photon counting options to enable Fluorescence Life-Time Imaging Microscopy (FLIM) measurements. In addition to the information gained from the intensity of a fluorescent signal, its lifetime provides information on the biophysical environment (e.g., ion and oxygen concentrations, temperature, or pH) of the respective fluorophore [44]. Furthermore, 2PEF/3PEF-FLIM can provide information on a fluorophore's conformational or molecular binding state, whose assessment is also relevant in the context of tissue characterization based on endogenous fluorescence. In a very insightful recent review [45], the authors discuss the importance of FLIM for evaluating cell metabolism.

SHG signals are exclusively originated by non-centrosymmetric structures (e.g., fibrillar collagen, microtubules and skeletal muscle) [35] and THG signal arises from interfaces within the specimen exhibiting a refractive index mismatch [46, 47]. The ability of SHG for probing at high spatial resolution the collagen distribution in tissues facilitates a precise and non-invasive assessment of extracellular matrix modifications specific to various pathologies, enabling consistent diagnostic possibilities [35]. As nicely demonstrated by Kuzmin et al. [48] THG complements SHG, as a result of its ability to image interfaces hosting lipid-rich molecules, allowing thus the visualization of cells and nuclei, or the investigation of vascularization related aspects.

Most MPM imaging studies performed to date on *ex-vivo* and fixed tissues have been performed with tabletop systems, either custom built/modified or commercialized. Such systems can operate in both upright or inverted configurations and require coupling with an appropriate laser source [49]. *In-vivo* imaging on animal models can be done with MPM systems adapted for intravital assays, which require that enough space is available under the objective for positioning the subject of investigation [50].

Most importantly, MPM imaging is also available at present in the form of clinically validated multiphoton tomographs, enabling *in-vivo* assessment of human skin [51]. In addition, a compact non-contact clinically-adapted MPM system has been recently used to image in-vivo the human eye [52]. This implementation offers real-time tissue visualization and the possibility to perform objective analyses of pathological or surgically modified ocular tissues [53–55]. Although, to the best of our knowledge, *in-vivo* imaging has not been yet demonstrated for organs positioned inside the human body (except for exposed brain [56]), recent progress in multiphoton endomicroscopy [57] suggests that such applications are within reach.

Given the advantages above discussed, MPM imaging is likely to become soon one of the default tools for tissue characterization. It can both augment conventional histopathology (e.g., by enabling lower sampling errors), or even replace it entirely in some scenarios. However, there are still challenges on the road to achieving this. For example, interpretation of MPM data can pose problems to histopathologists (who are trained on conventional modalities, e.g., brightfield microscopy of stained tissues). Collecting MPM datasets inside the human body is difficult due to intrinsic tissue movement that cannot be controlled. Non-invasive imaging of deep tissues with MPM is difficult due to light scattering and attenuation. However, despite of all these, there are still many applications where MPM imaging was demonstrated to be very useful. Therefore, MPM is currently regarded as a highly efficient tool for the study of pathological tissues, ranging from epithelial tissues [42], and internal organs [58] to ocular structures [54, 59] or brain tissue [43, 46, 60]. In this article, we review MPM applications focused on the study of oral tissues, addressing previous efforts dealing with visualization and analysis of different structures in the oral cavity, and objective diagnostic methods that were reported to detect and discriminate various oral diseases.

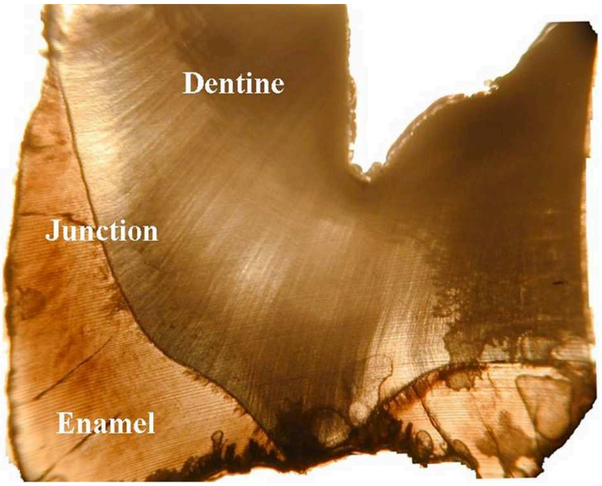

FIGURE 2 | Photo of a longitudinal dental section. Courtesy of Dr. Stephen Chin-Ying, Department of Preventive Dentistry, National University of Singapore. Image size ~4 × 4 mm^2.

MULTIPHOTON MICROSCOPY OF THE TOOTH

Tooth Structure

Enamel, dentine, cementum and pulp represent the structural components of the tooth [61]. An example of a dental section is presented in **Figure 2**.

Enamel is the hard and highly mineralized cover of the tooth crown. It is the most mineralized tissue in the human body, mostly made up of apatite crystals. When the tooth emerges into the oral cavity, ameloblasts (cells that produce the enamel) disappear; for this reason, enamel can't be regenerated. The dentin-enamel junction separates enamel and dentin.

Dentine is also a mineralized tissue, but elastic, avascular, and composed of apatite and collagen. The dentine structure is formed by packed tubules along its entire thickness. Tubules are responsible for tooth hydration and transmission of the physical signals and they are found along the entire structure, running

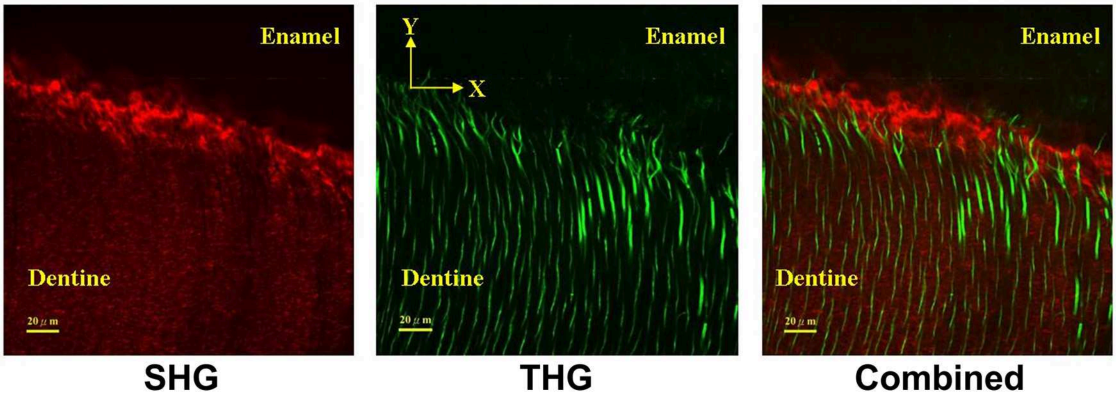

FIGURE 3 | SHG image (in the forward direction) from a location near the dentine-enamel junction (left). THG image from the same location clearly showing the dentinal tubules (center). Superposition of SHG and THG signals (right). It should be noted that the collected THG signals are much weaker compared to the SHG signals and consequently higher incident laser power or higher detector gain was required for image acquisition. Artwork: courtesy of Prof. Fu-Yen Kao (Institute of Biophotonics, National Yang Ming University, Taiwan), images were acquired during the experiment reported in [66].

from the pulp to the enamel and cementum. The diameter and density of the tubules increases toward the pulp [61].

The innermost part of the tooth is the pulp, formed by connective tissue that feeds and regenerates the dentinal collagen through the cells called odontoblasts. Moreover, it is richly innervated with sensory afferents, mostly involved in pain mediation.

The root is the part of the tooth covered by gingival tissue. At this location, dentine is covered by a connective tissue, cementum [62], made of a hard bone-like connective tissue that grows in concentric bands, which increases in thickness throughout life. The cementum also joins the periodontal ligament to the tooth, fixing it to the alveolar bone.

The most common tooth disease is caries, which is characterized by demineralization and degeneration of the organic matrix (i.e., collagen denaturation) [1]. Microorganisms in dental plaque are mainly responsible for caries initiation, but caries have also been related to systemic diseases [63]. Caries affect both enamel (coronal caries) and cementum (root caries) and, in later stages, the dentine or even the pulp.

MPM of Tooth Components

Investigating at high resolution the highly mineralized composition of the tooth typically requires processes of decalcification and staining that may alter its natural structure. This can be avoided by employing MPM to characterize the structure of dental pieces, which has been demonstrated as an important alternative tool.

To our knowledge, Kao et al. in 2000 were the first to report that the tooth generates SHG signal in the back-scattered direction (or epi-illuminated mode), but the details provided for the employed samples were limited. Moreover, the quality of the SHG images presented in that study was low compared to current standards, and probably due to this, the SHG signal was (erroneously) attributed to the highly organized structures of the enamel that encapsulates the dentine [64]. Later, the same group showed SHG and THG images of the dentine in transmission (or forward-scattered) mode [65]. In this second effort the images were acquired in areas near the dentine-enamel junction of a dental section, which allowed observing that the enamel does not generate any of the two considered non-linear signals. Instead, THG images revealed the tubule structure of the dentine, since this kind of MPM signal is sensitive to interfaces and boundaries, while SHG images were found to simply exhibit dentinal collagen content. The absence of SHG in enamel was also confirmed by a wavelength-dependent study [66]. An example of SHG and THG images collected near the dentine-enamel junction are depicted in **Figure 3**.

Non-linear signals from dental structures have also been explored in additional efforts dealing with other MPM imaging approaches. For example, in Chen et al. [67, 68] the authors focused on collecting submicron epi-illuminated MPM images of the enamel, dentine and periodontal ligaments. It was found that enamel exhibits a strong 2PEF signal, revealing the structures of the enamel rods; the dentine presents not only 2PEF signal, but also SHG. It is important to highlight here that the contrast provided by SHG imaging was not only useful to analyze the peritubular dentine structure, but also to distinguish the less mineralized circumpulpal dentine areas that were found to generate only SHG signal, and no 2PEF. That is, the more mineralized the dentinal structure becomes, the higher the 2PEF emission. In addition, due to their dominant collagen-based composition, clear observation of periodontal ligaments was also possible with SHG. The complementarity of SHG and 2PEF signals with respect to imaging hard oral tissues is shown also in **Figure 4** where we present a pair of SHG-2PEF images collected on the dentinal area of a tooth. For a direct comparison they are presented with the same color scale [69]. Noteworthy, the inherent confocality of MPM allows three-dimensional (3D) projections to be built, which facilitates the visualization of interesting features and their placement into a relevant topographic context. In **Figure 5** a 3D reconstruction of the dentine from SHG imaging is shown, and tubules can clearly be observed [69].

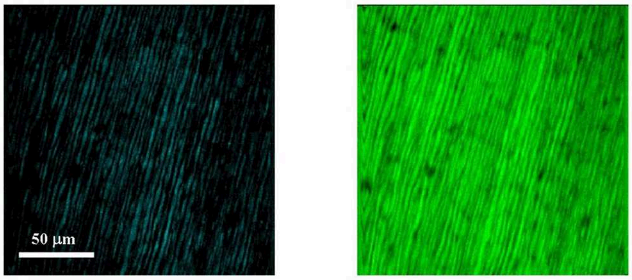

FIGURE 4 | SHG (left) and 2PEF (right) images recorded at the dentine. Since images share the color bar, it can directly be observed how the former is much dimmer compared to the latter. Adapted from Bueno et al. [69].

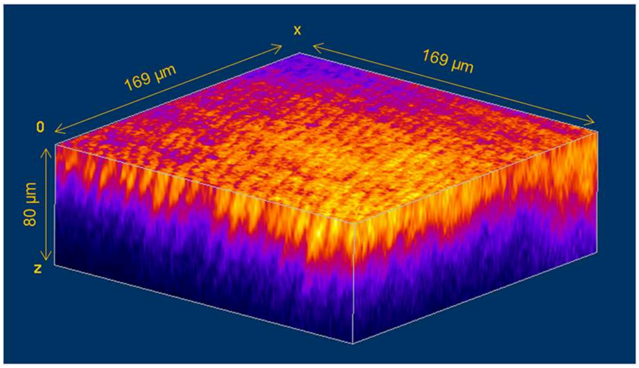

FIGURE 5 | 3D projection of the dentinal area. Adapted from Bueno et al. [69].

Elbaum and colleagues also reported high-resolution 3D images of tooth dentine, but in their experiment this was achieved based on SHG and THG images [70] in the forward-scattered mode. In their quest to explore the architecture of the dentine tubules and the surrounding collagen distribution, they were able to image depth locations up to 200 microns into the sample. The processed 3D reconstructions enabled the visualization of individual tubules and the collagen fibrils mesh around them with an optical resolution of about 1 micron. An important conclusion of their work is that collagen fibrils are organized perpendicularly to the tubules, however close to the dentin-enamel junction they lie also along the long axis of the tubules.

Another experiment, this time dealing with back-scattered MPM, showed that no significant SHG signals can be collected in this configuration on enamel, but THG images successfully revealed its prism structure and distribution [71]. The employed imaging configuration showed stronger THG signal compared to those reported in previous studies performed in transmission [65, 66], allowing thus a penetration depth exceeding 300 μm below the natural tooth surface. It was useful to observe that at the superficial enamel layer, the prisms were found to be organized in a honeycomb structure perpendicular to the tooth surface, and that this direction becomes parallel to the surface as deeper enamel layers are imaged.

In another effort focused on studying the dentin-enamel junction with MPM [72], 2PEF and SHG signals were simultaneously acquired in back and forward directions, respectively. The superposition of image pairs allowed clearly visualizing the junction. It should be noted here that 2PEF images revealed the transitional zone of the enamel as a non-fluorescent irregular line corresponding to the interface between dentin and enamel. This line of dim intensity was associated with very low concentration of protein and hence the lack of endogenous fluorophores. Similar to previous studies, SHG signal was not present in enamel, suggesting an absence of non-centrosymmetric proteins. While enamel prisms were found to exhibit high 2PEF signals, the inner aprismatic enamel (located close to the junction) showed a homogeneous low 2PEF signal. In the same experiment, dentine was observed to provide both SHG and 2PEF signals. The latter was hypothesized to be related to the odontoblast process which involves fluorescent proteins. SHG signals were much weaker compared to 2PEF (see **Figure 4**), what might indicate a low amount of highly non-centrosymmetric molecular assemblies as collagen and microtubules. The dentinal SHG intensity was found however to increase from the junction toward the inner dentine.

The work by Pan et al. took MPM imaging of teeth one step further by addressing tooth morphogenesis [73]. MPM images were used to investigate the development of the tooth in neonate mice from birth to the seventh day after. Results showed that predentina emits solely SHG signal, but dentine structures were observed to provide both SHG and 2PEF. Enamel, odontoblast and ameloblast were also found to exhibit strong 2PEF signal.

Another component of the tooth, dental cementum, has also been analyzed through SHG microscopy by H. Aboulfadl et al. [74]. The work showed that collagen fibers are distributed along two directions: radial (i.e., pointing more or less perpendicularly to the root surface) and circumferential (perpendicular to the radial and oriented parallel to the surface), which we regard as an important finding.

Non-linear multimodal microscopy [75, 76] combining Coherent Anti-Stokes Raman Scattering (CARS) [77], SHG, THG, and 2PEF was able to provide information not only on the tooth structure, but also on biochemical and biomolecular aspects [75]. Multimodal imaging revealed the microtubule structure nearby the dentin-enamel junction, and although CARS did not add extra information to that showed by 2PEF in dentine, this experiment demonstrated that its enhanced optical sectioning capability makes it a useful alternative tool for tooth analysis.

In preparation of future dental tissue engineering, Traphagen et al. [78] demonstrated the ability of MPM microscopy to characterize decellularized and demineralized teeth (while preserving the natural extracellular matrix). 2PEF appeared to be distributed throughout both natural decellularized tissue samples, although it was more prominent in the former. In addition, compared to the decellularized tooth tissue, SHG showed (as objectively measured by means of parameters such orientation index, entropy, and collagen density) higher collagen fiber density, and lower degree of organization in the natural one. Work related to demineralized teeth was also reported

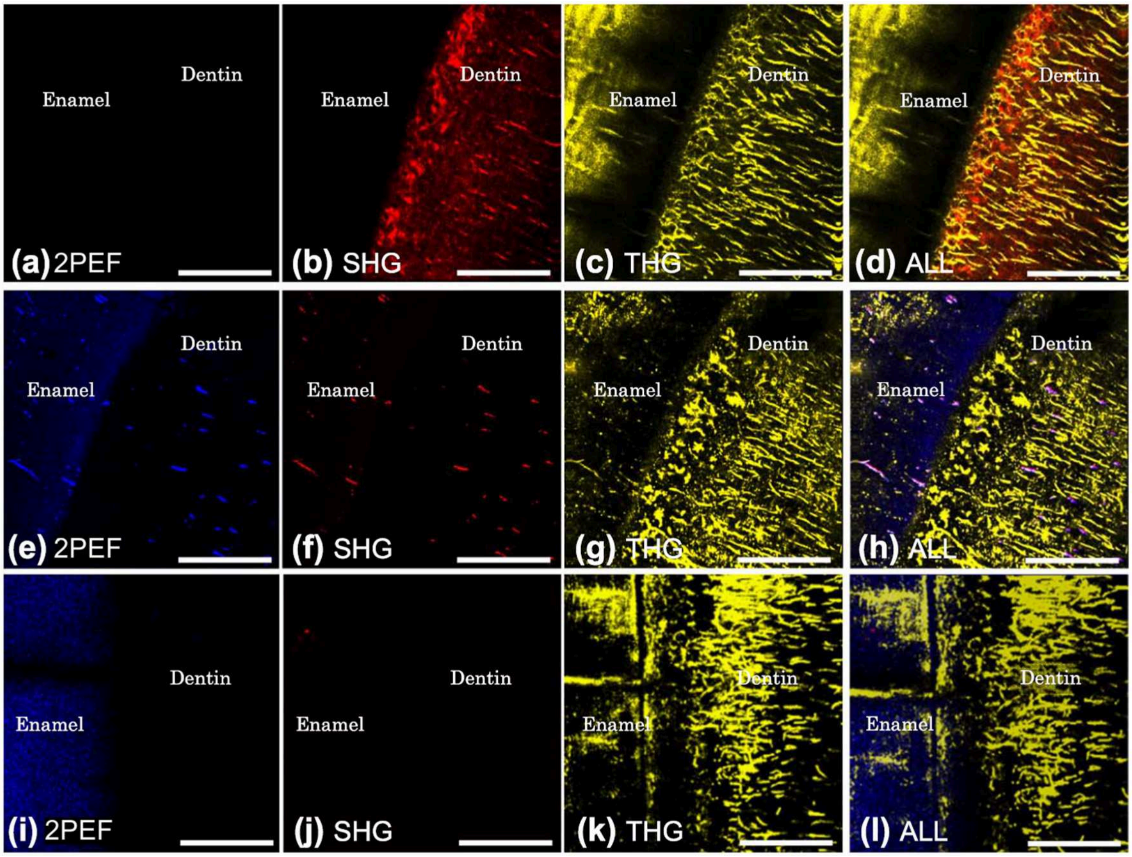

FIGURE 6 | 2PEF, SHG, and THG imaging of extant and fossil teeth of crocodilians. **(a–d)** Images of an extant Alligator tooth acquired under **(a)** 2PEF, **(b)** SHG, **(c)** THG microscopies, and **(d)** overlay of three channels. **(e–h)** Images of a fossil Alligator (1.5 Ma) tooth under **(e)** 2PEF, **(f)** SHG, **(g)** THG microscopies, and **(h)** overlay of three channels. **(i–l)** Images of a fossil Kem Kem crocodilian (93 Ma) tooth under **(i)** 2PEF, **(j)** SHG), **(k)** THG microscopies, and **(l)** 2PEF, SHG and THG overlay. All scale bars, 40 μm. Ma, millions of years. Adapted with permission from Chen et al. [80] © The Optical Society.

by Atmeh et al. [79], where the authors employed 2PEF microscopy to investigate the remineralising potential of certain materials on totally demineralised dentine. Although the signal intensity depended on the sample, these showed microscopic features of matrix remineralisation (including front, intra and intertubular mineralisation).

Most of the previously discussed works propose MPM microscopy as a very valuable tool to characterize various aspects of extant teeth (i.e., extracted teeth from living humans or animals). However, this technique has also been reported to be very useful to explore the anatomy of dentinal tubules in ancient fossil teeth [80]. In particular, THG imaging yielded (unexpectedly) strong signals when dealing with submicrometer level anatomy. When compared to extant teeth, the visualization of fossilized dentine tubules revealed a strong morphological correlation, confirming that the dentinal tubule structures have remained relatively constant through time. This is illustrated in **Figure 6**, which also very nicely demonstrates the complementarity of the 2PEF, SHG, and THG signals for teeth imaging, in general.

MPM for the Analysis of Dental Diseases and Abnormalities

Enamel covers the tooth crown and protects the inner tissues from bacterial infection, as well as from mechanical, thermal, and chemical attacks. Any disorder of this most external dental structure may allow acids and bacteria to penetrate into the inner tissues, which can lead to dental diseases. In the previous section of the manuscript we reviewed published works that aimed to explore sources of non-linear optical contrast in the tooth and the way they can be used to document various properties of this structure. In this part, we review additional work that focused on analyzing dental abnormalities and diseases, such as caries.

Girkin et al. were among the first to report an MPM image showing early dental caries in an intact tooth and proposed MPM techniques as a diagnostic tool in dentistry [81, 82]. In a different effort, MPM microscopy allowed imaging the tooth from the outer part up to a depth of ∼500 microns [83]. The imaged lesion depths compared well with the depths measured by physically sectioning the teeth. While healthy tooth tissue exhibits strong 2PEF signal, as discussed also in the previous section, this is available to a lesser extent in carious tooth tissue [84]. Caries

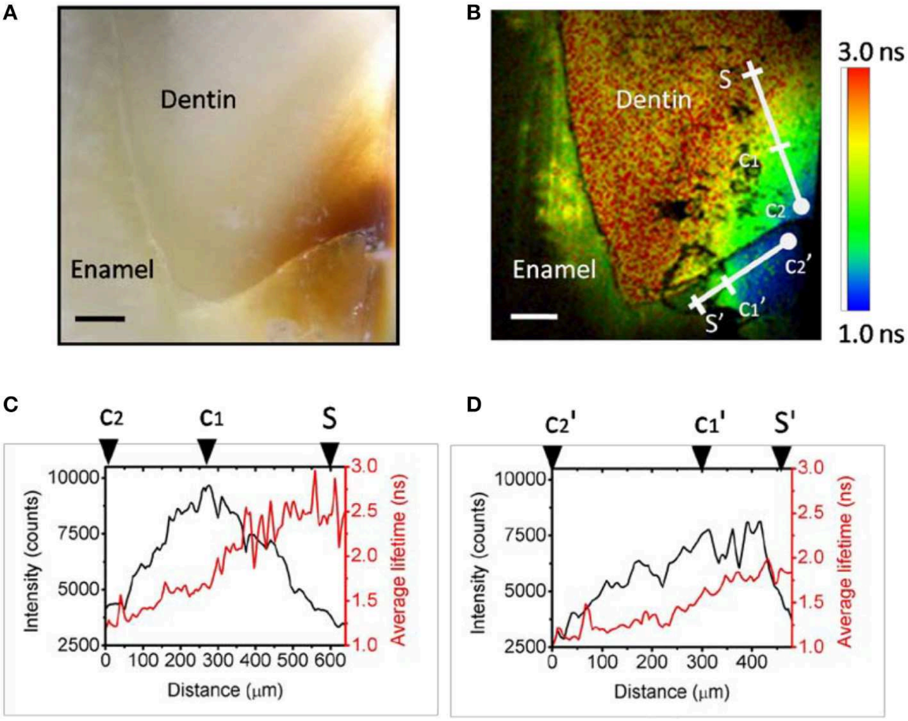

FIGURE 7 | 2PEF-FLIM of dental tissues. **(A)** An epi-illuminated image of the investigated carious dental sample. **(B)** The corresponding 2PEF-FLIM image showing carious regions with greatly reduced lifetime (blue) compared to healthy regions. The scale bar is 200 μm. The intensity (black) and average lifetime (red) line profiles are obtained from the 2PEF-FLIM image depicted in **(B)**, for **(C)** dentin and **(D)** enamel, respectively. Reprinted with permission from Lin et al. [85] © The Optical Society.

appear as a dark spot within a brightly fluorescent tooth, so in a 2PEF inverted image these will appear bright within a dark background, hence the decayed tissue can be well-highlighted with this imaging modality.

In another interesting work, the surfaces of teeth with abnormal enamel were studied and compared to the surfaces of intact human teeth as a basis for future clinical applications [71]. The investigated samples included white spot lesions, cracks, and the artificially-lased (irradiated) enamel. Since prisms within dental enamel present a homogeneous structure of hydroxyapatite crystals, the detected THG signal was thought to originate from the organic-matrix-filled interprismatic space rather than from inorganic crystalline regions inside the prisms. In diseased enamel crystal inhomogeneity appears, and this abnormality reflects in THG signals generated inside the prism. It was hypothesized that white spot lesions are mainly caused by mineral loss (with presence of crystal inhomogeneity), and the THG images collected in this study were coherent with this assumption, depicting different degrees of mineral loss. Moreover, unlike in sound enamel, SHG signal was also found in teeth regions harboring white spot lesions, which is believed to occur due to a symmetry breakage taking place under certain strain. In general, natural cracks may be a result from mechanical damages, thermal stress, and the stress-strain around the cracks; the MPM images were in agreement with this hypothesis. As expected, SHG signals depicting strains originated at the sites of the cracks. In addition, THG images revealed both the cracks and the enamel prisms beside them. Finally, in irradiated human tooth enamel, THG images collected on superficial layers revealed heat-induced cracks. The THG signals originating from the interprismatic space were observed to drastically decrease due to the melting of this superficial layer that takes place during irradiation. The strain at the cracks was also present in SHG images, and both THG and SHG signals were observed around the heat-induced cracks. At deeper locations (with reduced energy absorption and lower prism melting effects) strain sensitive SHG signals were found to become weaker but THG generated from the interprismatic spaces recovered.

In other works, focused on investigating how MPM signals can be used to indicate the tooth's health state, Lin and colleagues confirmed previous results on dental MPM sources and used fluorescence lifetime analysis to differentiate normal dental tissues from caries [85]. The latter were found to present a noticeable decrease in the lifetime of present endogenous fluorophores in both enamel and dentin (**Figure 7**). These results suggest that 2PEF-FLIM's usefulness for identifying carious tissues based on their specific fluorescence lifetime signatures represents an important asset that can potentially augment the current ways of identifying and assessing early sub-surface lesions that can be linked to teeth degradation during carries development. Terrer et al. tackled a similar problem as well and demonstrated that SHG and 2PEF intensities of human dentine were strongly modified during the tooth caries process, as a result of the degradation of the dentinal organic matrix [86].

In addition, they proposed the SHG/2PEF ratio as a reliable parameter to follow dental caries. The usefulness of combining these two complementary MPM signals for detection and classification of carious stages has been recently demonstrated by Slimani et al. [87], who correlate the SHG/2PEF ratio (which they regard as an indicator of the organic matrix denaturation) with the International Caries Detection and Assessment System (ICDAS).

Other important MPM efforts were focused on addressing endodontic infections. The elimination of microbes from the infected root canal system is currently completed by specific medication, but the effectiveness of the disinfection may be limited by microorganisms present in the dentinal tubules. In this sense, ZnO and TiO_2 nanoparticles represent relevant therapeutic agents, given their photoactive and bacterial inhibiting properties. Trunina et al. proposed to use MPM microscopy to visualize the penetration of these nanomaterials in the human tooth tissue [88, 89], in the frame of an *in-vitro* study. While ZnO nanoparticles produced SHG signal, TiO_2 generated 2PEF. Using these two imaging modalities it was observed that ZnO particles penetrate up to 45 microns into the enamel and dentine, respectively, while TiO_2 nanoparticles only penetrated 5 microns. This study suggested thus ZnO as a more promising material for penetration imaging, indicating also that dentinal permeability is one order of magnitude higher than that of enamel (for these types of particles).

Another aspect investigated with MPM was mineral density in the dentine collagen and enamel, which was shown to increase with age [61]. In light of this, a number of methods are emerging to estimate the age of an individual based on the tooth's composition [90]. Bueno et al. have recently reported results on this idea, by using MPM signals as an aging indicator [69]. The procedure relies on exploiting collagen denaturation with age and combines SHG and 2PEF images collected on dentine. The usefulness of such approaches is especially important in forensics, as they can characterize and help identify corpses that have been submerged in water or exposed to high temperatures as a consequence of natural disasters.

MULTIPHOTON MICROSCOPY OF THE ORAL MUCOSA

Oral Mucosa

The oral mucosa is formed by two layers, epithelium and connective tissue [7], and divided into three categories: masticatory (keratinized epithelium), lining (non-keratinized epithelium) and specialized mucosa. The mucosa surrounding the teeth (gingiva), due to its permeability, may be easily passed through by antigens.

One of the major diseases of the oral mucosa is cancer. The golden standard for the diagnostics of this pathology consist in biopsy under general anesthesia, combined with histopathological analysis of the excised tissue following fixation and staining. This method is very useful to evaluate histochemical and morphological changes in the tissue. However, it is subjected to the well-known disadvantages of traditional histopathology, which we discussed in the Introduction. These issues can be overcome or alleviated with the help of fluorescence spectroscopy [91] and MPM [6, 92–94], which have been demonstrated over the past years as very useful non-invasive tools to explore the oral mucosa.

Multiphoton Microscopy of the Oral Mucosa

One of the most important capabilities of MPM consists in its efficiency to characterize the structure and composition of tissues based on the fluorescence of endogenous fluorophores, which has also been demonstrated in the case of the oral mucosa. For example, Wu and collaborators found that 2PEF signals from NADH and FAD in the epithelium can be achieved with an excitation wavelength of 810 nm [95]. These autofluorescence signals are closely linked to cellular metabolism, and their ratio (known as the metabolic redox ratio) has been demonstrated in several landmark studies to reveal metabolism aspects unavailable with other techniques [41, 96]. The authors reported as well-significant SHG signals in the stroma, which were proposed as a sensitive indicator to separate the epithelial layer from underlying stroma. SHG intensity depends on the content and organization of the collagen fibers within the tissue [35, 40, 97], and in the case of the oral mucosa stroma these were highly organized, leading thus to consistent SHG signals. In a different study, Zhuo et al. used MPM to obtain images of the elastin fibers in the oral mucosa [98], and similar to the previous publication, they reported NADH and FAD autofluorescence in the epithelium under 810 nm excitation (however, the 2PEF signal corresponding to NADH was found to be higher at 730 nm). They also found significant SHG signals arising from the stroma and took SHG image analysis one step forward by quantitatively studying the distribution of the collagen (based on both number of fibers and inter-space measurements). In addition, they were able to acquire images and study the morphology of salivary glands, which are surrounded by collagen. MPM signals achieved from NADH, FAD, and collagen were also found to be important in light of their potential utility as biomarkers in the precancerous development of the epithelium [95].

In vivo MPM (yielding virtual biopsies) has also been demonstrated in the human oral mucosa [99]. Images of the epithelium and the lamina propria (connective tissue lying beneath the epithelium) were recorded without producing any damage. The performed measurements lasted about 30 min and the deepest imaged plane was located at 280 μm, which proposes MPM as an important tool for diagnosing pathologies of the mucosa *in vivo*.

Other studies that we find important to mention were focused on characterizing the vocal folds, which are located inside of the vocal tract and vibrate due to the air exhaled by the lungs to originate one's voice. Any problem in the vocal folds might limit the phonation mechanism and hence the treatment of

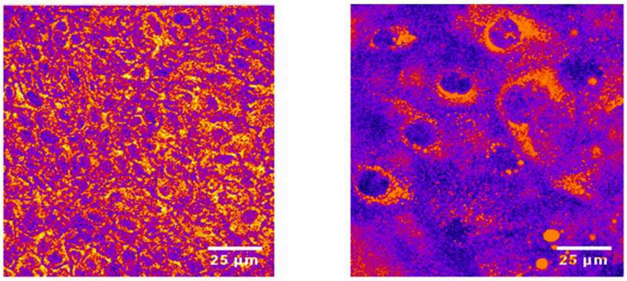

FIGURE 8 | 2PEF images of normal (left) and cancerous (right) tissues. These MPM images are courtesy of Prof. Nirmala Ramanujam (Duke Cancer Institute, Duke University) and Dr. Melissa C. Skala (Dpt. Biomedical Engineering, University of Wisconsin-Madison). They take part of the image set acquired and classified during the experiment reported in Skala et al. [6].

specific diseases, such as nodules and polyps, and requires thus immediate attention. Diagnosing such pathologies requires efficient detection, accurate location and, after surgery, reliable monitoring. In 2012, MPM was proposed as a tool for diagnosis and monitoring of diseases in the vocal folds [100]. Both the extracellular matrix and their overall morphological structure were studied by complementary MPM modalities. SHG and 2PEF images showed the geometry of the collagen and the elastin within the lamina propria, respectively. Two other MPM studies were focused on analyzing the healing process of the vocal folds [101, 102], and very recently, a system combining MPM and nanotomography has been successfully used for their 3D restoration [103].

In a different type of contribution compared to the rest discussed in this section, Sriram et al. [104] have recently introduced a step-by-step MPM imaging and sample preparation protocol for the non-invasive and label-free imaging of monolayer and three-dimensional organotypic cultures of the skin and oral mucosa.

Multiphoton Microscopy of Oral Mucosa Pathologies

Oral SCC originates from a mutation and an uncontrolled proliferation of keratocytes (located in the epithelium of the oral mucosa) [105]. Epithelial dysplasia precedes SCC and involves microscopic changes in the stratified squamous epithelium [106]. According to Vargas et al. [92], the stages of epithelial dysplasia are classified into: mid (dysplastic cells are found in the basal layer of the epithelium), moderate (dysplastic cells are extended through epithelium) and high dysplasia (where cell invasion occupies a large fraction of the epithelium). In carcinoma *in situ* the entire epithelium is covered with dysplasia. In oral SCC there is an alteration of the basement membrane due to invasive cells entering the collagen-based stroma. As an example, **Figure 8** depicts representative MPM images of tissues diagnosed as normal and cancerous.

In 2004, Wilder-Smith et al. explored the feasibility of MPM to act as an efficient tool for oral SCC early detection, and to represent a non-invasive (and faster) alternative to traditional diagnostic approaches that involve biopsy [93]. A variety of relevant structures were clearly noticeable in the images they collected in an *in vivo* hamster model. Also, an important finding of their study was the gradual reduction of the collagen fibers and a loss of the collagen's normal structure. The achieved results were in agreement with traditional histopathological assays that require biopsy, tissue fixation and staining. Later, the same authors combined MPM and OCT to increase the effectiveness of non-invasive oral SCC diagnostic (also in a hamster model) [94]. In this study, blood vessels, epithelial and subepithelial layers, the basement membrane and even the epithelial invasion were imaged by means of OCT, while the structure of collagen and elastin fibers, as well as the vessels, were visualized with MPM. This approach was demonstrated as being very useful with respect to the aimed purpose, and MPM signals were found to be very valuable for following (and understanding) the carcinogenesis process based on changes in the collagen matrix, organization loss, and reduction in length and number of collagen fibers. In a more recent experiment, Elagin et al. [107] showed in an animal model (7,12-dimethylbenz[a]anthracen (DMBA)-induced hamster oral carcinoma) that the complementarity of MPM and OCT can also be exploited for efficiently distinguishing *in vivo* between benign papilloma and papilloma that are either dysplastic or affected by SCC. While the MPM images presented in this study allowed extracting important cellular features such nuclear-cytoplasmic ratio or nuclear density, OCT was demonstrated to provide microvascular maps which were also useful with respect to assessing the pathological state of the characterized tissues.

In a different effort, this time dealing with a mouse model, a protocol combining an established carcinoma model with MPM imaging was used to quantify the invasion of tumor cells; a multi-vectorial visualization of lingual tumor spread was reported [108]. The incorporation of a spectroscopic system into an MPM device also provided interesting results concerning the detection of oral SCC at different precancerous stages. Autofluorescence signals were found to decrease in dysplastic hamsters and furthermore, the *in vivo* emission spectrum showed differences as a function of depth and excitation wavelength, in mid, moderate/high grade dysplasia animals compared to normal ones [109]. The accurate evaluation of the 2PEF spectral characteristics also enabled the identification of unique signatures that can be used to delineate normal oral mucosa from neoplasia [110]. In another study addressing oral carcinogenesis in hamsters, 2PEF and SHG signals showed a significant increase in the epithelium thickness and changes in the morphology and distribution of the keratocytes during different stages of the addressed pathology [92]. Skala et al. also found a similar increase occurring in the epithelium thickness, as well as in the keratin layer thickness [6].

In order to increase the performance of MPM imaging of oral mucosa tissues, the use of gold nanorods for targeting cancerous cells was proposed and demonstrated in an animal model with induced oral SCC [111, 112]. Gold is an inertial biocompatible material and the nanorods require low beam power to be excited. Injected gold nanorods reduced the background of 2PEF images and enhanced contrast. In addition, these particles were also shown to allow improved 3D reconstruction of the vessels [113].

Besides studies on animal models, MPM imaging of oral cancerous tissues was also carried out in humans. Tsai et al. performed an *ex vivo* MPM study on the cancerous mucosa of patients with oral SCC [114]. MPM revealed additional histopathological features compared to traditional histology (irregular epithelial stratification, cytological abnormalities and basement membrane interruption, among others), demonstrating thus the added value of such diagnostics assays. The authors reported changes in the patterns of collagen fibers and in the size and nuclei morphology of the parenchymal cells from SCC tissues. In addition, compared to normal mucosa, actin filamentous structures were more abundant in tumor cells, and an increased damage in the squamous epithelium was found. Also in a human model, Cheng et al. used in 2013 an intravital multi-harmonic microscope to image oral SCC [115]. SHG and THG signals of *in vivo* human tissues were acquired, and SHG microscopy was found useful to image the distribution of the collagen fibers in the lamina propria, whereas THG provided information about the keratocytes present in the epithelium and on the red blood cells in the capillaries.

In other studies, NADH and FAD fluorescence lifetime changes have also been reported in oral SCC tissues relative to healthy tissues, in both living and *ex vivo* experimental conditions [96, 116–118]. As discussed in the introductory part, FLIM relies on the measurement of the time it takes for an excited fluorophore to return to its ground state. This technique was found to be a valuable tool to extract information about the glycolysis and the oxidative phosphorylation in the cellular metabolism [117]. A decrease in the NADH fluorescence lifetime (averaged over the entire epithelium) was found in precancerous compared with normal tissues. On the opposite, a FAD lifetime (averaged) increase was observed in precancerous tissues [96]. However, intracellular variability of NADH and FAD fluorescence lifetimes was found to increase when comparing precancerous and normal cells. These results can be used in identifying cues that precede oral SCC. Furthermore, time-resolved 2PEF images were minimally affected by tissue morphology, endogenous absorbers, and illumination [117], which can significantly increase the robustness of MPM diagnostic assays. In the work of Teh et al. [118], a reduction in the lifetime of NADH fluorescence was found to occur in precancerous tissues compared with healthy tissues. While 2PEF intensity (excitation wavelength, 745 nm) of collagen increased in precancerous tissues, SHG signal decreased. The 2PEF/SHG ratio was found to be higher in precancerous tissues compared to healthy ones, a finding that is also useful for diagnostic purposes. Changes in 2PEF signals corresponding to tryptophan were also shown in different layers of dysplastic epithelium, which can contribute to a better understanding on how cancers of the oral mucosa originate.

Noteworthy, recent efforts have showed that MPM imaging can be significantly augmented by emerging techniques in artificial intelligence. For example, Huttunen et al. have recently demonstrated that MPM images collected on transversal fragments of epithelial tissues can be classified with high precision as being healthy or dysplastic by using a deep learning method that does not require extensively large training datasets [119]. Such methods can play an important role in facilitating the spread of MPM modalities in clinical settings to enable novel non-invasive *in vivo* diagnostic applications, including those potentially addressing oral tissues. This will be further facilitated by deep learning methods developed for virtual staining. Such methods enable the representation of images collected based on endogenous contrast (and not only) to the color schemes most familiar to histopathologists [120], which is also likely to bring significant added value to forthcoming clinical MPM imaging applications addressing oral tissue assessment and diagnostics.

CONCLUSIONS

The aim of this work has been to review previous studies focused on applying MPM to better understand oral tissues and to develop related applications, placing main attention on MPM imaging of dental pieces and oral mucosa, in both healthy and pathological conditions.

Although teeth are made of different types of tissues, the complementarity of available MPM techniques allows observing in detail these distinct structural regions. 2PEF signals originating from cellular constituents provide morphological and functional information. Both enamel and dentin emit strong 2PEF signal, helping to reveal various relevant features in detail (such as enamel prisms and dentinal tubules). Even though the spectral profiles of these signals are similar, they can be differentiated by lifetime analysis. 2PEF emission in the dentine tubules is thought to be associated with the odontoblast process containing fluorescent proteins. 2PEF intensity will vary due to protein content variation.

THG signals arise from interfaces with an abrupt change in the index of refraction. In the case of tooth imaging, this type of signal revealed the fine prism structure of the enamel (interprismatic spaces) but could only be detected in the back-scattered imaging direction. Unlike in enamel, both forward and backward THG images can be acquired on the tooth dentine. THG images showed the dentinal microtubule structure as a result of the sensitivity of THG signals to structural boundaries. While THG provides contrast for the observation of the tubules, SHG can be observed throughout the dentine, due to highly crystallized (and non-centrosymmetric) non-mineral organic matrix based on collagen fibrils.

It is also important to highlight that previous work showed that MPM imaging can be used to explore dental caries and abnormal enamel. Morphological and biochemical modifications suffered by teeth can be characterized by combining different non-linear signals, proposing MPM as a tool capable of significant clinical applications oriented toward the evaluation of dental diseases.

MPM's utility for the characterization and diagnostics of oral tissues has also been demonstrated in a series of important studies focused on imaging soft tissues, both healthy and pathological (mainly oral SCC), at high optical resolution and at different depth locations. In soft oral tissues, 2PEF, SHG, and THG allow the visualization of morphological structures, such as keratocytes, collagen, erythrocytes, and monitoring tissue modifications that occur due to various pathologies. We also find important to mention that 2PEF lifetime imaging of the NADH and FAD endogenous chromophores provides inestimable information about the metabolism state of the tissue, contributing thus to the potential of MPM for monitoring changes produced in living oral tissues during the progression of various pathologies, including cancer.

In summary, the findings reviewed herein suggest that MPM represents a valuable tool for the *ex-vivo* and *in-vivo* characterization of hard and soft tissues of the oral cavity. With respect to soft tissues, the diagnosis of cancers is the most widely addressed problem so far, but other pathologies specific to the oral cavity could equally benefit of the advantages offered by MPM modalities. Furthermore, the field of dentistry can also take profit of MPM's valuable contrast mechanisms, exploiting these for the diagnostics of early caries and for the accurate visualization of enamel abnormalities.

AUTHOR CONTRIBUTIONS

JB and SS defined the structure of this review and identified relevant works. RM-O and JB wrote the initial version of this manuscript. SS, MP-C, and LA revised the manuscript and incremented on top of the initial version. All authors read, revised, and approved the final version of the manuscript.

FUNDING

JB acknowledges the support of the Fundación Séneca - Agencia de Ciencia y Tecnología de la Región de Murcia (grant 19897/GERM/15). SS is thankful for the financial support provided by the Executive Agency for Higher Education, Research, Development and Innovation Funding (UEFISCDI) via grant PN-III-P1-1.1-TE-2016-2147 (CORIMAG).

REFERENCES

1. Selwitz RH, Ismail AI, Pitts NB. Dental caries. *Lancet*. (2007) **369**:51–9. doi: 10.1016/S0140-6736(07)60031-2
2. Pitts NB, Zero DT, Marsh PD, Ekstrand K, Weintraub JA, Ramos-Gomez F, et al. Dental caries. *Nat Rev Dis Prim*. (2017) **3**:17030. doi: 10.1038/nrdp.2017.30
3. de Camargo Cancela M, Voti L, Guerra-Yi M, Chapuis F, Mazuir M, Curado MP. Oral cavity cancer in developed and in developing countries: population-based incidence. *Head Neck*. (2010) **32**:357–67. doi: 10.1002/hed.21193
4. Mello FW, Melo G, Pasetto JJ, Silva CAB, Warnakulasuriya S, Rivero ERC. The synergistic effect of tobacco and alcohol consumption on oral squamous cell carcinoma: a systematic review and meta-analysis. *Clin Oral Investig*. (2019) **23**:2849–59. doi: 10.1007/s00784-019-02958-1
5. Jiang X, Wu J, Wang J, Huang R. Tobacco and oral squamous cell carcinoma: a review of carcinogenic pathways. *Tob Induc Dis*. (2019) **17**:1–9. doi: 10.18332/tid/111652
6. Skala MC, Squirrell JM, Vrotsos KM, Eickhoff JC, Gendron-Fitzpatrick A, Eliceiri KW, et al. Multiphoton microscopy of endogenous fluorescence differentiates normal, precancerous, and cancerous squamous epithelial tissues. *Cancer Res*. (2005) **65**:1180–6. doi: 10.1158/0008-5472.CAN-04-3031
7. Massano J, Regateiro FS, Januário G, Ferreira A. Oral squamous cell carcinoma: review of prognostic and predictive factors. *Oral Surg Oral Med Oral Pathol Oral Radiol Endod*. (2006) **102**:67–76. doi: 10.1016/j.tripleo.2005.07.038
8. Mupparapu M. Diagnostic imaging in dentistry. *Dent Clin North Am*. (2016) **60**:xi–xiii. doi: 10.1016/j.cden.2015.10.001
9. Brenner DJ, Doll R, Goodhead DT, Hall EJ, Land CE, Little JB, et al. Cancer risks attributable to low doses of ionizing radiation: assessing what we really know. *Proc Natl Acad Sci USA*. (2003) **100**:13761–6. doi: 10.1073/pnas.2235592100
10. Whitehouse RW. Computed tomography. In: Cassar-Pullicino VN, Mark Davies A, Editors. *Measurements in Musculoskeletal Radiology*. Berlin: Springer Berlin Heidelberg (2008). p. 15–29.
11. Sukovic P. Cone beam computed tomography in craniofacial imaging. *Orthod Craniofacial Res*. (2003) **6**:31–6. doi: 10.1034/j.1600-0544.2003.259.x
12. Blattner TC, George N, Lee CC, Kumar V, Yelton CDJ. Efficacy of cone-beam computed tomography as a modality to accurately identify the presence of second mesiobuccal canals in maxillary first and second molars: a pilot study. *J Endod*. (2010) **36**:867–70. doi: 10.1016/j.joen.2009.12.023
13. Diogenes A, Simpn S, Law AS. Regenerative endodontics. In: Berman L, Hargreaves K, editors. *Cohen's Pathways of the Pulp*. St Louis: Elsevier (2015). p. 447–73.
14. Idiyatullin D, Corum C, Park JY, Garwood M. Fast and quiet MRI using a swept radiofrequency. *J Magn Reson*. (2006) **181**:342–9. doi: 10.1016/j.jmr.2006.05.014
15. Idiyatullin D, Corum C, Moeller S, Prasad HS, Garwood M, Nixdorf DR. Dental magnetic resonance imaging: making the invisible visible. *J Endod*. (2011) **37**:745–52. doi: 10.1016/j.joen.2011.02.022
16. Feldchtein FI, Gelikonov GV, Gelikonov VM, Iksanov RR, Kuranov R V, Sergeev AM, et al. *In vivo* OCT imaging of hard and soft tissue of the oral cavity. *Opt Express*. (1998) **3**:239–50. doi: 10.1364/OE.3.000239
17. Wilder-Smith P, Otis L, Zhang J, Chen Z. Dental OCT. In: Drexler W, Fujimoto JG, editors. *Optical Coherence Tomography: Technology and Applications*. Berlin: Springer Berlin Heidelberg (2008). p. 1151–82. doi: 10.1007/978-3-540-77550-8_37
18. Machoy M, Seeliger J, Szyszka-Sommerfeld L, Koprowski R, Gedrange T, Wozniak K. The use of optical coherence tomography in dental diagnostics: a state-of-the-art review. *J Healthc Eng*. (2017) **2017**:7560645. doi: 10.1155/2017/7560645
19. de Boer JF, Hitzenberger CK, Yasuno Y. Polarization sensitive optical coherence tomography – a review. *Biomed Opt Express*. (2017) **8**:1838–73. doi: 10.1364/BOE.8.001838
20. de Paula Eduardo C, Aranha ACC, Muller Ramalho K, Bello-Silva MS, de Freitas PM. Laser dentistry research. In: Convissa RA, editor. *Principles and Practice of Laser Dentistry*. Saint Louis: Mosby (2015). p. 303–314. doi: 10.1016/B978-0-323-06206-0.00017-5
21. Contaldo M, Serpico R, Lucchese A. *In vivo* imaging of enamel by reflectance confocal microscopy (RCM): non-invasive analysis of dental surface. *Odontology*. (2014) **102**:325–29. doi: 10.1007/s10266-013-0110-9
22. Contaldo M, Di Stasio D, Santoro R, Laino L, Perillo L, Petruzzi M, et al. Non-invasive *in vivo* visualization of enamel defects by reflectance confocal microscopy (RCM). *Odontology*. (2015) **103**:177–84. doi: 10.1007/s10266-014-0155-4
23. Clark AL, Gillenwater AM, Collier TG, Alizadeh-Naderi R, El-Naggar AK, Richards-Kortum RR. Confocal microscopy for real-time detection of oral cavity neoplasia. *Clin Cancer Res*. (2003) **9**:4714–21. doi: 10.1117/1.1805558
24. Tezuka F. Diagnostic validity and variability of histopathology. *Rinsho Byori*. (1994) **42**:902–6.
25. Chatterjee S. Artefacts in histopathology. *J Oral Maxillofac Pathol*. (2014) **18**:111–6. doi: 10.4103/0973-029X.141346
26. Logan RM, Goss AN. Biopsy of the oral mucosa and use of histopathology services. *Aust Dent J*. (2010) **55**:9–13. doi: 10.1111/j.1834-7819.2010.01194.x
27. Kim J, Lee HS, Song IS, Jung KH. DeNTNet: deep neural transfer network for the detection of periodontal bone loss using panoramic dental radiographs. *Sci Rep*. (2019) **9**:17615. doi: 10.1038/s41598-019-53758-2
28. Flügge T, Hövener JB, Ludwig U, Eisenbeiss AK, Spittau B, Hennig J, et al. Magnetic resonance imaging of intraoral hard and soft tissues using

an intraoral coil and FLASH sequences. *Eur Radiol*. (2016) **26**:4616–23. doi: 10.1007/s00330-016-4254-1

29. Tsai MT, Wang YL, Yeh TW, Lee HC, Chen WJ, Ke JL, et al. Early detection of enamel demineralization by optical coherence tomography. *Sci Rep*. (2019) **9**:17154. doi: 10.1038/s41598-019-53567-7
30. Kosugi A, Kasahara M, Yang L, Nakamura-Takahashi A, Shibahara T, Mori T. Method for diagnosing neoplastic lesions by quantitative fluorescence value. *Sci Rep*. (2019) **9**:7833. doi: 10.1038/s41598-019-44287-z
31. Aubreville M, Knipfer C, Oetter N, Jaremenko C, Rodner E, Denzler J, et al. Automatic classification of cancerous tissue in laserendomicroscopy images of the oral cavity using deep learning. *Sci Rep*. (2017) **7**:11979. doi: 10.1038/s41598-017-12320-8
32. Göppert-Mayer M. Elementary processes with two quantum transitions. *Ann der Phys*. (2009) **18**:466–79. doi: 10.1002/andp.200910358
33. Denk W, Strickler JH, Webb WW. Two-photon laser scanning fluorescence microscopy. *Science*. (1990) **248**:73–6. doi: 10.1126/science.2321027
34. Zipfel WR, Williams RM, Webb WW. Nonlinear magic: multiphoton microscopy in the biosciences. *Nat Biotechnol*. (2003) **21**:1369–77. doi: 10.1038/nbt899
35. Campagnola PJ, Dong CY. Second harmonic generation microscopy: principles and applications to disease diagnosis. *Laser Photonics Rev*. (2011) **5**:13–26. doi: 10.1002/lpor.200910024
36. Mazumder N, Balla NK, Zhuo GY, Kistenev Y V, Kumar R, Kao FJ, et al. Label-free non-linear multimodal optical microscopy—basics, development, and applications. *Front Phys*. (2019) **7**:170. doi: 10.3389/fphy.2019.00170
37. Helmchen F, Denk W. Deep tissue two-photon microscopy. *Nat Methods*. (2005) **2**:932–40. doi: 10.1038/nmeth818
38. Bueno JM, Ávila FJ, Artal P. Comparison of second harmonic microscopy images of collagen-based ocular tissues with 800 and 1045 nm. *Biomed Opt Express*. (2017) **8**:5065–74. doi: 10.1364/BOE.8.005065
39. Masters BR, So PTC. Confocal microscopy and multi-photon excitation microscopy of human skin *in vivo*. *Opt Express*. (2001) **8**:2–10. doi: 10.1364/OE.8.000002
40. Williams RM, Zipfel WR, Webb WW. Multiphoton microscopy in biological research. *Curr Opin Chem Biol*. (2001) **5**:603–8. doi: 10.1016/S1367-5931(00)00241-6
41. Georgakoudi I, Quinn KP. Optical imaging using endogenous contrast to assess metabolic state. *Annu Rev Biomed Eng*. (2012) **14**:351–67. doi: 10.1146/annurev-bioeng-071811-150108
42. Balu M, Zachary CB, Harris RM, Krasieva TB, König K, Tromberg BJ, et al. *In vivo* multiphoton microscopy of basal cell carcinoma. *JAMA Dermatol*. (2015) **151**:1068–74. doi: 10.1001/jamadermatol.2015.0453
43. Horton NG, Wang K, Kobat D, Clark CG, Wise FW, Schaffer CB, et al. *In vivo* three-photon microscopy of subcortical structures within an intact mouse brain. *Nat Photonics*. (2013) **7**:205–9. doi: 10.1038/nphoton.2012.336
44. Suhling K, Hirvonen LM, Levitt JA, Chung PH, Tregidgo C, Le Marois A, et al. Fluorescence lifetime imaging (FLIM): basic concepts and some recent developments. *Med Photonics*. (2015) **27**:3–40. doi: 10.1016/j.medpho.2014.12.001
45. Kolenc OI, Quinn KP. Evaluating cell metabolism through autofluorescence imaging of NAD(P)H and FAD. *Antioxid Redox Signal*. (2019) **30**:875–89. doi: 10.1089/ars.2017.7451
46. Witte S, Negrean A, Lodder JC, De Kock CPJ, Silva GT, Mansvelder HD, et al. Label-free live brain imaging and targeted patching with third-harmonic generation microscopy. *Proc Natl Acad Sci USA*. (2011) **108**:5970–5. doi: 10.1073/pnas.1018743108
47. Zhang Z, de Munck JC, Verburg N, Rozemuller AJ, Vreuls W, Cakmak P, et al. Quantitative third harmonic generation microscopy for assessment of glioma in human brain tissue. *Adv Sci*. (2019) **6**:1900163. doi: 10.1002/advs.201900163
48. Kuzmin NV, Wesseling P, Hamer PC de W, Noske DP, Galgano GD, Mansvelder HD, et al. Third harmonic generation imaging for fast, label-free pathology of human brain tumors. *Biomed Opt Express*. (2016) **7**:1889. doi: 10.1364/BOE.7.001889
49. Lefort C. A review of biomedical multiphoton microscopy and its laser sources. *J Phys D Appl Phys*. (2017) **50**:423001. doi: 10.1088/1361-6463/aa8050
50. Perrin L, Bayarmagnai B, Gligorijevic B. Frontiers in intravital multiphoton microscopy of cancer. *Cancer Rep*. (2020) **3**:e1192. doi: 10.1002/cnr2.1192
51. Klemp M, Meinke MC, Weinigel M, Röwert-Huber HJ, König K, Ulrich M, et al. Comparison of morphologic criteria for actinic keratosis and squamous cell carcinoma using *in vivo* multiphoton tomography. *Exp Dermatol*. (2016) **25**:218–22. doi: 10.1111/exd.12912
52. Ávila FJ, Gambín A, Artal P, Bueno JM. *In vivo* two-photon microscopy of the human eye. *Sci Rep*. (2019) **9**:10121. doi: 10.1038/s41598-019-46568-z
53. Bueno JM, Palacios R, Pennos A, Artal P. Second-harmonic generation microscopy of photocurable polymer intrastromal implants in *ex-vivo* corneas. *Biomed Opt Express*. (2015) **6**:2211–19. doi: 10.1364/BOE.6.002211
54. Ávila FJ, Artal P, Bueno JM. Quantitative discrimination of healthy and diseased corneas with second harmonic generation microscopy. *Transl Vis Sci Technol*. (2019) **8**:51. doi: 10.1167/tvst.8.3.51
55. Bueno JM, Ávila FJ, Martínez-Garciá MC. Quantitative analysis of the corneal collagen distribution after *in vivo* cross-linking with second harmonic microscopy. *Biomed Res Int*. (2019) **2019**:3860498. doi: 10.1155/2019/3860498
56. Kantelhardt SR, Kalasauskas D, König K, Kim E, Weinigel M, Uchugonova A, et al. *In vivo* multiphoton tomography and fluorescence lifetime imaging of human brain tumor tissue. *J Neurooncol*. (2016) **127**:473–82. doi: 10.1007/s11060-016-2062-8
57. Dilipkumar A, Al-Shemmary A, Kreiß L, Cvecek K, Carlé B, Knieling F, et al. Label-free multiphoton endomicroscopy for minimally invasive *in vivo* imaging. *Adv Sci*. (2019) **6**:1801735. doi: 10.1002/advs.201801735
58. Stanciu SG, Xu S, Peng Q, Yan J, Stanciu GA, Welsch RE, et al. Experimenting liver fibrosis diagnostic by two photon excitation microscopy and bag-of-features image classification. *Sci Rep*. (2014) **4**:4636. doi: 10.1038/srep04636
59. Stanciu SG, Ávila FJ, Hristu R, Bueno JM. A study on image quality in polarization-resolved second harmonic generation microscopy. *Sci Rep*. (2017) **7**:15476. doi: 10.1038/s41598-017-15257-0
60. Perry SW, Burke RM, Brown EB. Two-photon and second harmonic microscopy in clinical and translational cancer research. *Ann Biomed Eng*. (2012) **40**:277–91. doi: 10.1007/s10439-012-0512-9
61. Arola DD, Gao S, Zhang H, Masri R. The tooth: its structure and properties. *Dent Clin*. (2017) **61**:651–68. doi: 10.1016/j.cden.2017.05.001
62. Shahmoradi M, Bertassoni LE, Elfallah HM, Swain M. Fundamental structure and properties of enamel, dentin and cementum. In: Ben-Nissan B, editor. *Advances in Calcium Phosphate Biomaterials*. Berlin: Springer Berlin Heidelberg (2014). p. 511–47. doi: 10.1007/978-3-642-53980-0_17
63. Matsumoto-Nakano M. Dental caries. In: *Reference Module in Biomedical Sciences*. Elsevier (2014). Available online at: https://www.sciencedirect.com/science/article/pii/B9780128012383000015?via%3Dihub
64. Kao F-J, Wang Y-S, Huang M-K, Huang S-L, Cheng PC. Second-harmonic generation microscopy of tooth. *Opt Sensing Imaging Manip Biol Biomed Appl*. (2000) **4082**:119. doi: 10.1117/12.390534
65. Fu-Jen K, Chin-Ying SH. "Harmonic generation microscopy of dental sections," *Proc. SPIE 5630, Optics in Health Care and Biomedical Optics: Diagnostics and Treatment II* (2005). doi: 10.1117/12.577285
66. Kao F-J. The use of optical parametric oscillator for harmonic generation and two-photon uv fluorescence microscopy. *Microsc Res Tech*. (2004) **63**:175–81. doi: 10.1002/jemt.20026
67. Chen M-H, Chen W-L, Sun Y, Fwu PT, Lin M-G, Dong C-Y. Three-dimensional tooth imaging using multiphoton and second harmonic generation microscopy. *Lasers Dent XIII*. (2007) **6425**:642503. doi: 10.1117/12.698840
68. Chen M-H, Chen W-L, Sun Y, Fwu PT, Dong C-Y. Multiphoton autofluorescence and second-harmonic generation imaging of the tooth. *J Biomed Opt*. (2007) **12**:064018. doi: 10.1117/1.2812710
69. Bueno JM, Martínez-Ojeda RM, Ávila FJ, Fernández-Escudero AC, López-Nicolás M, Pérez-Carceles MD. Multiphoton imaging microscopy of dental pieces as a tool in forensic sciences. In: *Program and Abstract Book Focus on Microscopy 2019*. London (2009). p. 266.
70. Elbaum R, Tal E, Perets AI, Oron D, Ziskind D, Silberberg Y, et al. Dentin micro-architecture using harmonic generation microscopy. *J Dent*. (2007) **35**:150–5. doi: 10.1016/j.jdent.2006.07.007

71. Chen S-Y, Hsu C-YS, Sun C-K. Epi-third and second harmonic generation microscopic imaging of abnormal enamel. *Opt Express*. (2008) **16**:11670–9. doi: 10.1364/OE.16.011670

72. Cloitre T, Panayotov IV, Tassery H, Gergely C, Levallois B, Cuisinier FJG. Multiphoton imaging of the dentine-enamel junction. *J Biophotonics*. (2013) **6**:330–7. doi: 10.1002/jbio.201200065

73. Pan PY, Chen RS, Ting CL, Chen WL, Dong CY, Chen MH. Multiphoton microscopy imaging of developing tooth germs. *J Formos Med Assoc*. (2014) **113**:42–9. doi: 10.1016/j.jfma.2012.03.016

74. Aboulfadl H, Hulliger J. Absolute polarity determination of teeth cementum by phase sensitive second harmonic generation microscopy. *J Struct Biol*. (2015) **192**:67–75. doi: 10.1016/j.jsb.2015.08.011

75. Wang Z, Zheng W, Stephen Hsu CY, Huang Z. Epi-detected quadruple-modal nonlinear optical microscopy for label-free imaging of the tooth. *Appl Phys Lett*. (2015) **106**:033701. doi: 10.1063/1.4906447

76. Wang Z, Zheng W, Lin J, Hsu C-Y, Huang Z. Integrated coherent raman scattering and multiphoton microscopy for label-free imaging of the dentin in the tooth. *Multiphot Microsc Biomed Sci XIV*. (2014) **8948**:89482N. doi: 10.1117/12.2039683

77. Pezacki JP, Blake JA, Danielson DC, Kennedy DC, Lyn RK, Singaravelu R. Chemical contrast for imaging living systems: molecular vibrations drive CARS microscopy. *Nat Chem Biol*. (2011) **7**:137–45. doi: 10.1038/nchembio.525

78. Traphagen SB, Fourligas N, Xylas JF, Sengupta S, Kaplan DL, Georgakoudi I, et al. Characterization of natural, decellularized and reseeded porcine tooth bud matrices. *Biomaterials*. (2012) **33**:5287–96. doi: 10.1016/j.biomaterials.2012.04.010

79. Atmeh AR, Chong EZ, Richard G, Boyde A, Festy F, Watson TF. Calcium silicate cement-induced remineralisation of totally demineralised dentine in comparison with glass ionomer cement: tetracycline labelling and two-photon fluorescence microscopy. *J Microsci*. (2015) **257**:151–60. doi: 10.1111/jmi.12197

80. Chen Y-C, Lee S-Y, Wu Y, Brink K, Shieh D-B, Huang TD, et al. Third-harmonic generation microscopy reveals dental anatomy in ancient fossils. *Opt Lett*. (2015) **40**:1354–7. doi: 10.1364/OL.40.001354

81. Girkin JM, Hall AF, Creanor SL. Multi-photon imaging of intact dental tissue. In: *Proceedings of the 4th Annual Indiana Conference*. Indianapolis (1999). p. 155–168.

82. Girkin JM, Hall AF, Creanor SL. Two-photon imaging of intact dental tissue. *Dent Caries*. (2000) **2**:317–25.

83. Girkin JM. Optical physics enables advances in multiphoton imaging. *J Phys D Appl Phys*. (2003) **36**:R250–8. doi: 10.1088/0022-3727/36/14/204

84. Hall A, Girkin JM. A review of potential new diagnostic modalities for caries lesions. *J Dent Res*. (2004) **83**:89–94. doi: 10.1177/154405910408 301s18

85. Lin P-Y, Lyu H-C, Hsu C-YS, Chang C-S, Kao F-J. Imaging carious dental tissues with multiphoton fluorescence lifetime imaging microscopy. *Biomed Opt Express*. (2011) **2**:149–58. doi: 10.1364/BOE.2.000149

86. Terrer E, Panayotov IV, Slimani A, Tardivo D, Gillet D, Levallois B, et al. Laboratory studies of nonlinear optical signals for caries detection. *J Dent Res*. (2016) **95**:574–9. doi: 10.1177/0022034516629400

87. Slimani A, Tardivo D, Panayotov IV, Levallois B, Gergely C, Cuisinier F, et al. Multiphoton microscopy for caries detection with ICDAS classification. *Caries Res*. (2018) **52**:359–66. doi: 10.1159/000486428

88. Trunina NA, Popov AP, Lademann J, Tuchin VV, Myllylä R, Darvin ME. Two-photon-excited autofluorescence and second-harmonic generation microscopy for the visualization of penetration of TiO 2 and ZnO nanoparticles into human tooth tissue *ex vivo*. *Biophoton Photon Solut Better Heal Care III*. (2012) **8427**:84270Y. doi: 10.1117/12.924073

89. Trunina NA, Darvin ME, Kordas K, Sarkar A, Mikkola JP, Lademann J, et al. Monitoring of TiO 2 and ZnO nanoparticle penetration into enamel and dentine of human tooth *in vitro* and assessment of their photocatalytic ability. *IEEE J Sel Top Quantum Electron*. (2014) **20**:133–40. doi: 10.1109/JSTQE.2013.2276082

90. Maber M, Liversidge HM, Hector MP. Accuracy of age estimation of radiographic methods using developing teeth. *Forensic Sci Int*. (2006) **159**:S68–73. doi: 10.1016/j.forsciint.2006.02.019

91. Ingrams DR, Dhingra JK, Roy K, Perrault DF, Bottrill ID, Kabani S, et al. Autofluorescence characteristics of oral mucosa. *Head Neck*. (1997) **19**:27–32.

92. Vargas G, Shilagard T, Ho KH, McCammon S. Multiphoton autofluorescence microscopy and second harmonic generation microscopy of oral epithelial neoplasms. In: *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society: Engineering the Future of Biomedicine, EMBC 2009*, St. Paul: Minneapolis (2009). p. 6311–3.

93. Wilder-Smith P, Osann K, Hanna N, El Abbadi N, Brenner M, Messadi D, et al. *In vivo* multiphoton fluorescence imaging: a novel approach to oral malignancy. *Lasers Surg Med*. (2004) **35**:96–103. doi: 10.1002/lsm.20079

94. Wilder-Smith P, Krasieva T, Jung W-G, Zhang J, Chen Z, Osann K, et al. Noninvasive imaging of oral premalignancy and malignancy. *J Biomed Opt*. (2005) **10**:051601. doi: 10.1117/1.2098930

95. Wu Y, Qu JY. Two-photon autofluorescence spectroscopy and second-harmonic generation of epithelial tissue. *Opt Lett*. (2005) **30**:3045. doi: 10.1364/OL.30.003045

96. Skala MC, Riching KM, Gendron-Fitzpatrick A, Eickhoff J, Eliceiri KW, White JG, et al. *In vivo* multiphoton microscopy of NADH and FAD redox states, fluorescence lifetimes, and cellular morphology in precancerous epithelia. *Proc Natl Acad Sci USA*. (2007) **104**:19494–9. doi: 10.1073/pnas.0708425104

97. Hristu R, Stanciu SG, Tranca DE, Stanciu GA. Improved quantification of collagen anisotropy with polarization-resolved second harmonic generation microscopy. *J Biophotonics*. (2017) **10**:1171–9. doi: 10.1002/jbio.201600197

98. Zhuo S, Chen J, Jiang X, Xie S, Chen R, Cao N, et al. The layered-resolved microstructure and spectroscopy of mouse oral mucosa using multiphoton microscopy. *Phys Med Biol*. (2007) **52**:4967–80. doi: 10.1088/0031-9155/52/16/017

99. Tsai M-R, Chen S-Y, Shieh D-B, Lou P-J, Sun C-K. *In vivo* optical virtual biopsy of human oral mucosa with harmonic generation microscopy. *Biomed Opt Express*. (2011) **2**:2317–28. doi: 10.1364/BOE.2.002317

100. Miri AK, Tripathy U, Mongeau L, Wiseman PW. Nonlinear laser scanning microscopy of human vocal folds. *Laryngoscope*. (2012) **122**:356–63. doi: 10.1002/lary.22460

101. Heris HK, Miri AK, Ghattamaneni NR, Li NYK, Thibeault SL, Wiseman PW, et al. Microstructural and mechanical characterization of scarred vocal folds. *J Biomech*. (2015) **48**:708–11. doi: 10.1016/j.jbiomech.2015.01.014

102. Yildirim M, Quinn KP, Kobler JB, Zeitels SM, Georgakoudi I, Ben-Yakar A. Quantitative differentiation of normal and scarred tissues using second-harmonic generation microscopy. *Scanning*. (2016) **38**:684–93. doi: 10.1002/sca.21316

103. Kazarine A, Kolosova K, Gopal AA, Wang H, Tahara R, Rammal A, et al. Multimodal virtual histology of rabbit vocal folds by nonlinear microscopy and nano computed tomography. *Biomed Opt Express*. (2019) **10**:1151. doi: 10.1364/BOE.10.001151

104. Sriram G, Sudhaharan T, Wright GD. Multiphoton microscopy for noninvasive and label-free imaging of human skin and oral mucosa equivalents. In: *Methods in Molecular Biology*. Totowa, NJ: Humana Press (2019). p. 1–18.

105. Scully C, Bagan J. Oral squamous cell carcinoma overview. *Oral Oncol*. (2009) **45**:301–8. doi: 10.1016/j.oraloncology.2009.01.004

106. Lumerman H, Freedman P, Kerpel S. Oral epithelial dysplasia and the development of invasive squamous cell carcinoma. *Oral Surg Oral Med Oral Pathol Oral Radiol Endod*. (1995) **79**:321–9. doi: 10.1016/S1079-2104(05)80226-4

107. Karabut M, Kiseleva EB, Sirotkina MA, Kuznetsov SS, Matveev LA, Moiseev AA, et al. Multiphoton tomography and multimodal OCT for *in vivo* visualization of oral malignancy in the hamster cheek pouch. *Photonic Solut Better Heal Care VI*. (2018) **10685**:1–8. doi: 10.1117/12.2306210

108. Gatesman Ammer A, Hayes KE, Martin KH, Zhang L, Spirou GA, Weed SA. Multi-photon imaging of tumor cell invasion in an orthotopic mouse model of oral squamous cell carcinoma. *J Visualized Experiments: JoVE* (2011) 2941. doi: 10.3791/2941

109. Edward K, Shilagard T, Qiu S, Vargas G. Two-photon autofluorescence spectroscopy of oral mucosa tissue. *Multiphot Microsc Biomed Sci XI*. (2011) **7903**:79031Q. doi: 10.1117/12.875049

110. Pal R, Edward K, Ma L, Qiu S, Vargas G. Spectroscopic characterization of oral epithelial dysplasia and squamous cell carcinoma using multiphoton autofluorescence micro-spectroscopy. *Lasers Surg Med*. (2017) **49**:866–73. doi: 10.1002/lsm.22697

111. Motamedi S, Shilagard T, Koong L, Vargas G. Feasibility of using gold nanorods for optical contrast in two photon microscopy of oral carcinogenesis. *Proc SPIE 7576, Reporters, Markers, Dyes, Nanoparticles, and Molecular Probes for Biomedical Applications II* (2010) 75760Z. doi: 10.1117/12.843036

112. Motamedi S, Shilagard T, Edward K, Koong L, Qui S, Vargas G. Gold nanorods for intravital vascular imaging of preneoplastic oral mucosa. *Biomed Opt Express*. (2011) **2**:1194. doi: 10.1364/BOE.2.001194

113. Olesiak-Banska J, Waszkielewicz M, Obstarczyk P, Samoc M. Two-photon absorption and photoluminescence of colloidal gold nanoparticles and nanoclusters. *Chem Soc Rev*. (2019) **48**:4087–117. doi: 10.1039/C8CS00849C

114. Tsai MR, Shieh D Bin, Lou PJ, Lin CF, Sun CK. Characterization of oral squamous cell carcinoma based on higher-harmonic generation microscopy. *J Biophotonics*. (2012) **5**:415–24. doi: 10.1002/jbio.201100144

115. Cheng Y-H, Lin C-F, Shih T-F, Sun C-K. A novel intravital multi-harmonic generation microscope for early diagnosis of oral cancer. *Opt Biopsy XI*. (2013) **8577**:85770R. doi: 10.1117/12.2001593

116. Sun Y, Phipps J, Elson DS, Stoy H, Tinling S, Meier J, et al. Fluorescence lifetime imaging microscopy: *in vivo* application to diagnosis of oral carcinoma. *Opt Lett*. (2009) **34**:2081. doi: 10.1364/OL.34.002081

117. Shah AT, Skala MC. *Ex vivo* label-free microscopy of head and neck cancer patient tissues. *Multiphot Microsc Biomed Sci XV*. (2015) **9329**:93292B. doi: 10.1117/12.2075583

118. Teh SK, Zheng W, Li S, Li D, Zeng Y, Yang Y, et al. Multimodal nonlinear optical microscopy improves the accuracy of early diagnosis of squamous intraepithelial neoplasia. *J Biomed Opt*. (2013) **18**:036001. doi: 10.1117/1.JBO.18.3.036001

119. Huttunen MJ, Hristu R, Dumitru A, Floroiu I, Costache M, Stanciu SG. Multiphoton microscopy of the dermoepidermal junction and automated identification of dysplastic tissues with deep learning. *Biomed Opt Express*. (2020) **11**:186. doi: 10.1364/BOE.11.000186

120. Rivenson Y, Wang H, Wei Z, de Haan K, Zhang Y, Wu Y, et al. Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning. *Nat Biomed Eng*. (2019) **3**:466. doi: 10.1038/s41551-019-0362-y

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

BRIEF RESEARCH REPORT
published: 17 June 2020
doi: 10.3389/fphy.2020.00212

A Logistic Model for Counting Crowds and Flowing Particles

Byung Mook Weon[1,2,3]*

[1] *Soft Matter Physics Laboratory, School of Advanced Materials Science and Engineering, SKKU Advanced Institute of Nanotechnology (SAINT), Sungkyunkwan University, Suwon, South Korea,* [2] *Research Center for Advanced Materials Technology, Sungkyunkwan University, Suwon, South Korea,* [3] *Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, United States*

Counting how many people or particles pass through a specific space within a specific time is an interesting question in applied physics and social science. Here a logistic model is developed to estimate the total number of moving crowds or flowing particles. This model sheds light on a collective contribution of crowd or particle growth rate and transient probability within a specific space. This model may offer a basic concept to understand transport dynamics of moving crowds and flowing particles.

Keywords: flowing particles, logistic model, crowds, growth dynamics, particle mobility

Edited by:
Wenxiang Xu,
Hohai University, China

Reviewed by:
Zhigang Zhu,
Hohai University, China
Hao Hu,
Anhui University, China

Correspondence:
Byung Mook Weon
bmweon@skku.edu

Specialty section:
This article was submitted to Soft Matter Physics, a section of the journal Frontiers in Physics

Received: *06 April 2020*
Accepted: *18 May 2020*
Published: *17 June 2020*

Citation:
Weon BM (2020) A Logistic Model for Counting Crowds and Flowing Particles. Front. Phys. 8:212. doi: 10.3389/fphy.2020.00212

How many people or particles have passed there? This question is simple but significant in many physical, biological, and social situations [1–3]. Counting the total number of moving crowds or flowing particles is often a difficult task because of complexity in mobility and transport dynamics. Conceptually, this question is similar to a population dynamics that is controlled by birth and death rates or immigration and emigration [3]. In mathematical biology, the simplest population growth model is the Malthusian exponential model where the total population increases exponentially with time [4]. The logistic model is widely established in many fields for modeling and forecasting populations [5]. The logistic growth dynamics describes that the total population grows exponentially at early times and saturates to an upper limit at late times, producing a typical *S*-shaped curve. The upper limit represents a capacity limit in the system. In a confined space, there may be a capacity limit and thus the logistic model would be appropriate in crowd or particle counting.

In this article, the logistic model is developed to understand moving crowds or flowing particles for the total number estimation. This model sheds light on a collective contribution of crowd or particle mobility and growth rate to the total number. This model is applicable for both of static and mobile crowds and particles, probably offering a new framework for understanding transport dynamics of static or mobile crowds and particles.

First, consider a physical situation for flowing particles (conceptually, identical for moving crowds), where a fixed number of flowing particles occupy a limited number of positions in a space, as illustrated in **Figure 1**. As flowing particles move through a space together like flowing crowds [6–9], the total number of particles initially increases with time, reach a peak for a while, and eventually diminishes with time. In this situation, the number of particles can be modeled by a combination of particle growth and decay dynamics. This physical situation can be modeled with the factors of the first (or final) particle contribution a, the rate of growth (or decay) b, and the maximum capacity in the place c (physically, c is set by a multiple of the occupation space α and the population density β as $c = \alpha\beta$).

Next, to quantify the hydrodynamic aspects of flowing particles [10, 11], the average transient time d is considered as follows. The transient time is the spent duration for particles to stay by

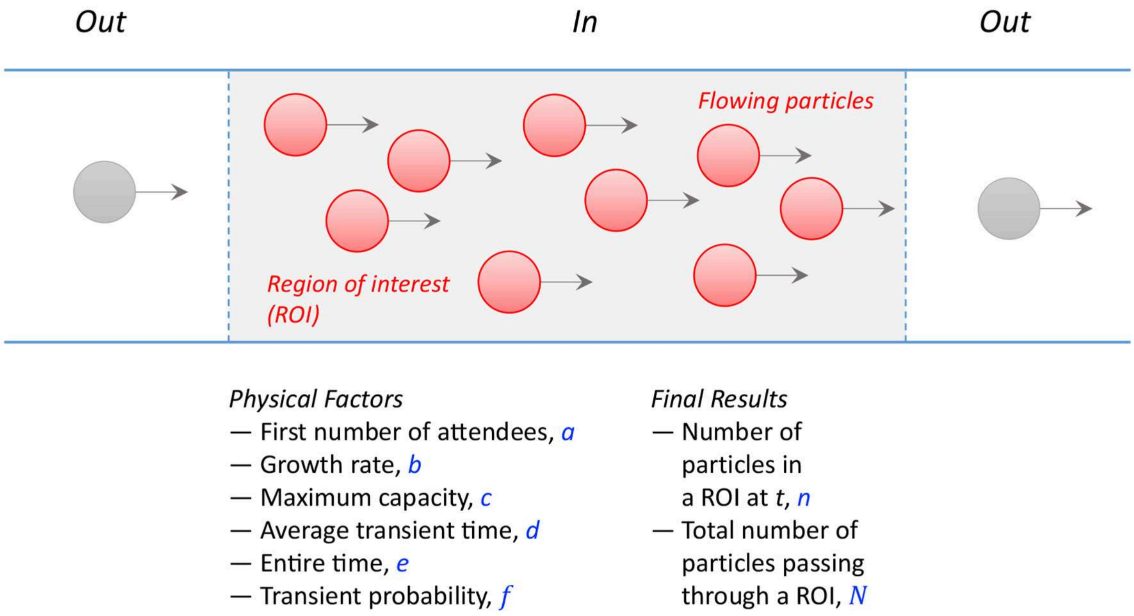
Out
In
Out
Flowing particles
Region of interest (ROI)
Physical Factors
— First number of attendees, a
— Growth rate, b
— Maximum capacity, c
— Average transient time, d
— Entire time, e
— Transient probability, f
Final Results
— Number of particles in a ROI at t, n
— Total number of particles passing through a ROI, N

FIGURE 1 | Illustration of a situation: when particles are flowing in a region of interest (ROI, gray) and their physical factors are given ($a \sim f$), an important question is how many particles (n or N) have passed through the ROI during a period.

occupying the limited positions and is responsible for the particle mobility. Assuming the entire time e for growth and decay, the transient probability f is calculated as $f = d/e$. By taking the transient probability, the particle mobility can be quantified.

The transient probability is useful to characterize the nature of static or mobile particles. For instance, let's think about the following two situations. In the first case, most particles may stay to pass through for a while (e.g., for 30 min) during the entire time (e.g., for 2 h), suggesting the transient probability to be $f = \frac{30}{120} = 0.25$ on average. In the second case, most particles may stay for a while (e.g., for 110 min) during the entire time (e.g., for 2 h), indicating $f = \frac{110}{120} = 0.92$. The first case corresponds to *mobile* particles ($f \ll 1.0$), while the second case to *static* particles ($f \approx 1.0$).

To describe static or mobile particles with the logistic model, the logistic growth dynamics is applied prior to a peak as [2, 4, 5]:

$$n(t) = \frac{ac}{a + (c - a)e^{-bt}}, \tag{1}$$

and after passing a peak, the logistic decay dynamics is applied as:

$$n(t) = \frac{ac}{a + (c - a)e^{-b(e-t)}}. \tag{2}$$

Here $n(t)$ is the number of particles at a moment and is determined by the first (or final) number of particles a, the growth (or decay) rate b, the maximum capacity c, the average transient time d, the entire time e, the transient probability $f = d/e$, and the peak time $g = \frac{1}{2}e$. By integrating $n(t)$ with respect to t and dividing it by the average transient time, the total number N can be estimated as:

$$N = \frac{\int n(t)\,\mathrm{d}t}{d}. \tag{3}$$

As demonstrated in **Figure 2**, the logistic model is appropriate to evaluate how the particle number changes with time by the physical factors in the logistic model. In **Figure 2A**, for the physically feasible conditions, $a = 1.0$, $b = 0.2$, $c = 100$, $d = 30$, and $e = 120$ are assumed (here, time is normalized). Controlling the factors, the particle number for static or mobile particles is counted during particle growth (Equation 1) and decay dynamics (Equation 2). For simplicity, the growth dynamics is assumed to be symmetric with the decay dynamics. In **Figure 2B**, the contribution of the growth rate b is tested by fixing the other conditions in **Figure 2A** except for the variable b [$a = 1.0$, $c = 100$, $d = 30$, and $e = 120$]. Interestingly, the total number significantly increases with the growth rate b. In **Figure 2C**, the contribution of the transient probability f is tested by fixing the other conditions in **Figure 2A** except for the variable f [$a = 1.0$, $b = 0.2$, $c = 100$, and $e = 120$]. Interestingly, the total number is inversely proportional to the transient probability f (this is evident from $N \propto d^{-1}$ and $d = ef$ in Equation 3). The collective contribution of the growth rate and the transient probability is illustrated in **Figure 2D** [by fixing $a = 1.0$, $c = 100$, and $e = 120$], showing that the total number is significantly affected by the transient probability for most b values ($b \gtrsim 0.2$); that is, the particle mobility is crucial to determine the total number of flowing particles.

The logistic model is appropriate to characterize the nature of static or mobile particles. The total number of particles is

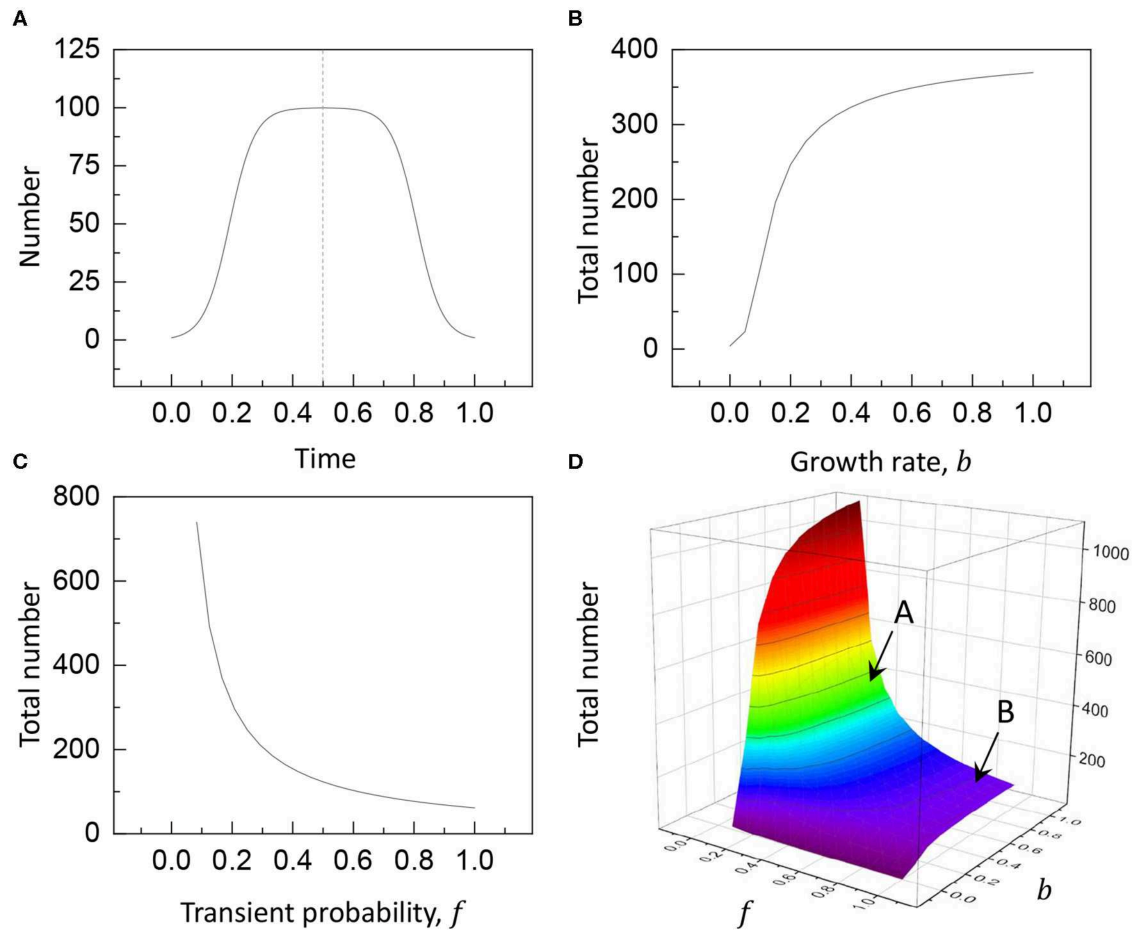

FIGURE 2 | The logistic model: **(A)** the particle number $n(t)$ changes with time [$a = 1.0$, $b = 0.2$, $c = 100$, $d = 30$, and $e = 120$], **(B)** by the contribution of the growth rate b [by fixing $a = 1.0$, $c = 100$, $d = 30$, and $e = 120$], **(C)** by the contribution of the transient probability f [by fixing $a = 1.0$, $b = 0.2$, $c = 100$, and $e = 120$], and **(D)** by the collective contribution of the growth rate b and the transient probability f [by fixing $a = 1.0$, $c = 100$, and $e = 120$]. Here, the total number increases by 3.7 times when the transient probability decreases to $f = 0.25$ (mobile particles, marked A) from $f = 0.95$ (static particles, marked B) for the same growth rate $b = 1.0$.

illustrated in **Figure 2D** as a function of the transient probability and the growth rate [by fixing $a = 1.0$, $c = 100$, and $e = 120$]. Most interestingly, the total number is significantly affected by the transient probability, rather than the growth rate. In particular, the total number significantly increases by 3.7 times when the transient probability decreases to $f = 0.25$ ($N = 3.7c$ as marked A) from $f = 0.95$ ($N = 0.97c$ as marked B) for the same growth rate $b = 1.0$. This result clearly shows why mobile particles are more than static particles. It is noteworthy that the logistic model is applicable for both static and mobile particles by simply adjusting the physical factors. To generalize the result, the total number of particles becomes more than the maximum capacity for mobile particles ($N \gg c$) and becomes less than or equal to the maximum capacity for static particles ($N \leqslant c$).

To demonstrate the validity of the logistic model, a simulation of falling balls through a triangle grid of pegs was tested with help of the Physics Education Technology (PhET) interactive simulations (https://phet.colorado.edu) [12, 13]. In **Figure 3** (see **Movie S1**), the number of red balls in the triangle grid increases with time at $t < g$ and decreases with time at $t > g$. The measured ball number is compared with the logistic model with $a = 1.0$, $b = 0.135$, $c = 38$, $d = 42.2$, and $e = 165$ ($N/c = 2.6$), providing a good agreement between simulation and model.

The logistic growth or decay dynamics is applicable to describe the number of moving crowds or flowing particles in a region of interest, based on which the total number of particles passing through the region can be estimated with *a posteriori* fit for the data, as demonstrated in **Figure 3**. Finding the parameters of the logistic dynamics is crucial for the model to work. For *a priori* or real-time estimation of the parameters, the early data can be analyzed and used to predict the late data. From Equation (1), the logistic differential equation is given as $\frac{dn(t)}{dt} = bn(t)(1 - \frac{n(t)}{c})$ where the growth rate b and the carrying capacity c can be estimated from *a priori* or real-time data (the first number a can be set to be 1.0). To determine the total number N in

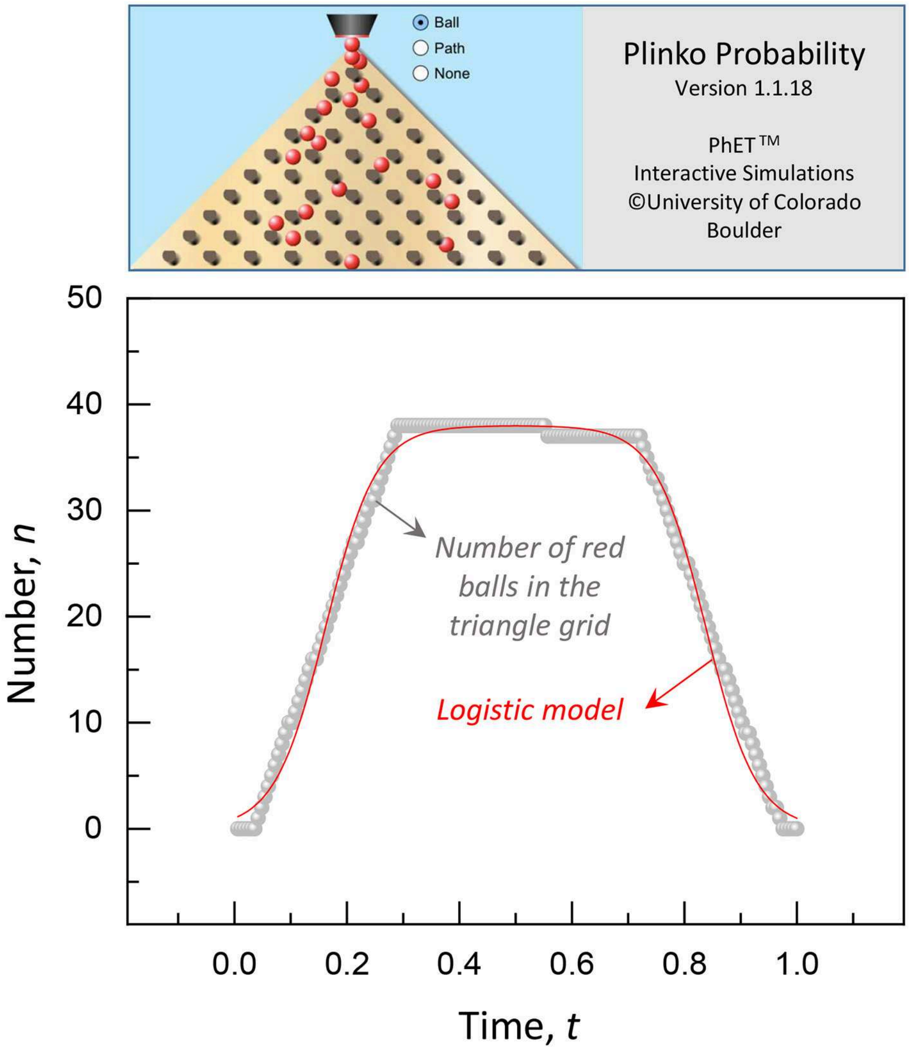

FIGURE 3 | Simulation of falling balls through a triangle grid of pegs for the Plinko Probability by the PhET interactive simulations (**Movie S1**). The number of red balls in the triangle grid increases with time at $t < g$ and decreases with time at $t > g$, showing a good agreement with the logistic model with $a = 1.0$, $b = 0.135$, $c = 38$, $d = 42.2$, and $e = 165$ ($N/c = 2.6$).

Equation (3), the average transient time d can be obtained from *a priori* or real-time information and the entire time e (about twice the peak time $g = \frac{1}{2}e$) can be given or determined in real situations.

Counting the total number of particles, both *a priori* and *a posteriori*, can be applied for human crowds and planning crowd safety in places of public assembly [14–16]. In principle, human crowds are likely to stay or move in a place like flowing particles [11]. Conceptually, flowing particles are identical with moving crowds. Direct counting methods would be available with many modern technologies such as artificial intelligence, drone, and visual analysis [14, 15] to count the total number in many situations. However, direct counting would be expensive and time consuming. The approximate counting of the total number with the logistic model may be useful and applicable to estimate the particle transport through porous media in applied physics, the total number of clients visiting a store in economics [15], the crowd size of a protest in sociology [16], and the growth dynamics of bacteria in a specific colony in biology [17]. Further studies are required to verify the applicability of the logistic model in a variety of systems.

In conclusion, this study shows a theoretical frame of the logistic growth or decay dynamics that would be appropriate to estimate the total number of moving crowds or flowing particles. As demonstrated here, the model is available for both static and mobile crowds and particles. The numerical demonstration of the logistic model clearly shows how the instantaneous particle number changes with time according to the particle mobility and the growth dynamics. Practically, in physical, social, or ecological situations, the logistic model is applicable by identifying the transient probability and the growth rate to count or estimate the total number.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

BW organized the research, conducted the research, analyzed the data, and wrote the manuscript.

ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at [18] (BW). This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2016R1D1A1B01007133, 2019R1A6A1A03033215) and also supported by the Korea Evaluation Institute of Industrial Technology funded by the Ministry of Trade, Industry and Energy (20000423, Developing core technology of materials and processes for control of rheological properties of nanoink for printed electronics).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphy.2020.00212/full#supplementary-material

Movie S1 | This movie shows a simulation with falling balls.

REFERENCES

1. Watson R, Yip P. How many were there when it mattered? *Significance.* (2011) **8**:104–7. doi: 10.1111/j.1740-9713.2011.00502.x
2. Jin W, McCue SW, Simpson MJ. Extended logistic growth model for heterogeneous populations. *J Theor Biol.* (2018) **445**:51–61. doi: 10.1016/j.jtbi.2018.02.027
3. Marchetti C, Meyer PS, Ausubel JH. Human population dynamics revisited with the logistic model: how much can be modeled and predicted? *Technol Forecast Soc Change.* (1996) **52**:1–30. doi: 10.1016/0040-1625(96)00001-7
4. Stokes M. Population ecology at work: managing game populations. *Nat Educ Knowl.* (2012) **3**:5. Available online at: https://www.nature.com/scitable/knowledge/library/population-ecology-at-work-managing-game-populations-50937864/
5. Verhulst PF. Notice sur la loi que la population poursuit dans son accroissement. *Corres Math Phys.* (1838) **10**:113–21.
6. Low DJ. Following the crowd. *Nature.* (2000) **407**:465–6. doi: 10.1038/35035192
7. Ouellette NT. Flowing crowds. *Science.* (2019) **363**:27–8. doi: 10.1126/science.aav9869
8. Helbing D, Molnar P. Social force model for pedestrian dynamics. *Phys Rev E.* (1995) **51**:4282–6. doi: 10.1103/PhysRevE.51.4282
9. Karamouzas I, Skinner B, Guy SJ. Universal power law governing pedestrian interactions. *Phys Rev Lett.* (2014) **113**:238701. doi: 10.1103/PhysRevLett.113.238701
10. Bain N, Bartolo D. Dynamic response and hydrodynamics of polarized crowds. *Science.* (2019) **363**:46–9. doi: 10.1126/science.aat9891
11. Hughes RL. The flow of human crowds. *Annu Rev Fluid Mech.* (2003) **35**:169–82. doi: 10.1146/annurev.fluid.35.101101.161136
12. Perkins K, Adams W, Dubson M, Finkelstein N, Reid S, Wieman C, et al. PhET: interactive simulations for teaching and learning physics. *Phys Teach.* (2006) **44**:18–23. doi: 10.1119/1.2150754
13. Wieman CE, Adams WK, Perkins KK. PhET: simulations that enhance learning. *Science.* (2008) **322**:682–3. doi: 10.1126/science.1161948
14. Botta F, Moat HS, Preis T. Quantifying crowd size with mobile phone and Twitter data. *R Soc Open Sci.* (2015) **2**:150162. doi: 10.1098/rsos.150162
15. Henke LL. Estimating crowd size: a multidisciplinary review and framework for analysis. *Bus Stud J.* (2016) **8**:27–38. Available online at: https://www.abacademies.org/articles/bsjvol8number1.pdf#page=31
16. Still GK. Crowd science and crowd counting. *Impact.* (2019) **1**:19–23. doi: 10.1080/2058802X.2019.1594138
17. Sibilo R, Perez JM, Hurth C, Pruneri V. Surface cytometer for fluorescent detection and growth monitoring of bacteria over a large field-of-view. *Biomed Opt Exp.* (2019) **10**:2101–16. doi: 10.1364/BOE.10.002101
18. Weon BM. A logistic model for flowing particles. *arXiv:1910.11995v2.* (2019) Available online at: https://arxiv.org/abs/1910.11995

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

BRIEF RESEARCH REPORT
published: 14 July 2020
doi: 10.3389/fphy.2020.00249

The Capon Method for Mercury's Magnetic Field Analysis

Simon Toepfer[1], Yasuhito Narita[2,3], Daniel Heyner[3] and Uwe Motschmann[1,4]*

[1] Institut für Theoretische Physik, Technische Universität Braunschweig, Braunschweig, Germany, [2] Space Research Institute, Austrian Academy of Sciences, Graz, Austria, [3] Institut für Geophysik und Extraterrestrische Physik, Technische Universität Braunschweig, Braunschweig, Germany, [4] DLR Institute of Planetary Research, Berlin, Germany

Characterization of Mercury's internal and external magnetic field is one of the primary goals of the magnetometer experiment on board the BepiColombo MPO (Mercury Planetary Orbiter) spacecraft. A novel data analysis tool is developed to determine the Gauss coefficients in the multipole expansion using Capon's minimum variance projection method. The construction of the estimator is presented along with a test against the numerical simulation data of Mercury's magnetosphere and a comparison with the least square fitting method shows, that Capon's estimator is in better agreement with the coefficients, implemented in the simulation, than the least square fit estimator.

Keywords: diagonal loading, Gauss coefficients, least–squares method, magnetic field analysis, Capon's method

OPEN ACCESS

Edited by:
Luca Sorriso-Valvo,
National Research Council, Italy

Reviewed by:
Tommaso Alberti,
Institute for Space Astrophysics and Planetology (INAF), Italy
Igor Ivanovich Alexeev,
Lomonosov Moscow State University, Russia

Correspondence:
Simon Toepfer
s.toepfer@tu-braunschweig.de

Specialty section:
This article was submitted to Space Physics, a section of the journal Frontiers in Physics

Received: *31 March 2020*
Accepted: *05 June 2020*
Published: *14 July 2020*

Citation:
Toepfer S, Narita Y, Heyner D and Motschmann U (2020) The Capon Method for Mercury's Magnetic Field Analysis. Front. Phys. 8:249. doi: 10.3389/fphy.2020.00249

1. INTRODUCTION

The reconstruction of planetary magnetic fields is one of the most important goals of a magnetometer experiment on board an orbiting spacecraft. Various inversion methods have successfully been applied to the data of former missions that visited different planets in our solar system. For example, generalized inversion [1] and elastic net regression [2] have been applied to the reconstruction of Jupiter's internal magnetic field. The weighted least square fit [3] and robust regression [4] appeared as useful methods for the analysis of Saturn's magnetic field. The Earth's magnetic field has been analyzed among other methods by using the maximum entropy method [5]. All these methods will be useful tools for Mercury's magnetic field analysis, which is one of the primary goals of the magnetometer experiment on board the BepiColombo mission. In this work we present an alternative method, namely Capon's method, for the analysis of Mercury's internal magnetic field.

Capon's method [6], also known as minimum variance distortionless response estimator (MVDR) [7], was introduced for reconstructing the velocities and wave vectors of seismic waves measured on an array of sensors on the Earth's surface. In space plasma physics, the method has first been successfully applied to the analysis of plasma waves in the terrestrial magnetosphere [8]. Later on, the method was extended for the mode decomposition of magnetic fields [9]. This establishes a basis to separate the planetary magnetic field from the total measured field in Mercury's magnetosphere.

The separation of the internal magnetic field from the external parts of the field, which are generated by currents flowing in the magnetosphere is important for the reconstruction of the internal field. There exists a paraboloid model of Mercury's magnetosphere [10] which has successfully been applied to the analysis of Mercury's internal magnetic field [11, 12]. Since Capon's method is applied to the analysis of Mercury's internal magnetic field for the first time, here only the internal parts of the field are considered in the parametrization as a proof of concept.

Concerning to the BepiColombo mission, in this work magnetic field data resulting from the plasma interaction of Mercury with the solar wind are simulated and Capon's method is applied to the magnetic field data to analyze Mercury's internal magnetic field.

2. PARAMETRIZATION AND INVERSION METHODS

2.1. Parametrization of Mercury's Magnetic Field

The parametrization of planetary magnetic fields is based on the Gauss representation [13]. If only data in curl-free regions are analyzed, Ampère's law $\partial_{\underline{x}} \times \underline{B} = 0$, where $\underline{B}$ is the magnetic field vector and $\partial_{\underline{x}}$ is the spatial derivative, yields the existence of a scalar potential Φ, so that $\underline{B} = -\partial_{\underline{x}}\Phi$. In general, Φ is composed of internal and external parts. In the following only the internal parts Φ_i will be considered. For the parametrization of the internal dipole and quadrupole fields the scalar potential is expanded into spherical harmonics

$$\Phi_i = R_M \sum_{l=1}^{2} \left(\frac{R_M}{r}\right)^{l+1} \sum_{m=0}^{l} \left[g_l^m \cos(m\lambda) + h_l^m \sin(m\lambda)\right] P_l^m\big(\cos(\theta)\big), \tag{1}$$

where planetary centered coordinates with radius r, azimuth angle $\lambda \in [0, 2\pi]$, and polar angle $\theta \in [0, \pi]$ are chosen. R_M indicates the radius of Mercury and P_l^m are the Schmidt-normalized associated Legendre polynomials of degree l and order m. The expansion coefficients g_l^m and h_l^m are the internal Gauss coefficients. Arranging the Gauss coefficients into a vector $\underline{g} : = \left(g_1^0, g_1^1, h_1^1, g_2^0, g_2^1, h_2^1, g_2^2, h_2^2\right)^T$, for later application called ideal coefficient vector, the contribution of the internal magnetic field can be rearranged as

$$\underline{B} = -\partial_{\underline{x}}\Phi_i = \underline{\underline{H}}\,\underline{g}, \tag{2}$$

where the terms of the multipole series are arranged in the matrix $\underline{\underline{H}}(r, \theta, \lambda)$. The magnetic field measurements $\underline{B}$ and the underlying model $\underline{\underline{H}}$ are known. The unknown coefficient vector $\underline{g}$ is to be determined. In most applications the number of known magnetic data points is much larger than the number of the expansion coefficients, resulting in an overdetermined inversion problem. Therefore, $\underline{\underline{H}}$ is a rectangular matrix in general and the direct inversion of Equation (2) is impossible. But there exist several inversion methods for estimating $\underline{g}$ [7].

2.2. Least Square Fit (LSF) Method

The most commonly used method for inverse problems is the least square fit method. The method minimizes the quadratic deviation between the disturbed measurements $\underline{B}$ and the model $\underline{\underline{H}}\,\underline{g}$ with respect to the unknown set of coefficients $\underline{g}$ [7]

$$\min_{\underline{g}} \left|\underline{\underline{H}}\,\underline{g} - \underline{B}\right|^2 = \min_{g_l} \left(g_i\, H_{ij}^{\dagger}\, H_{jk}\, g_k - 2\, B_j\, H_{ji}\, g_i + B_i\, B_i\right), \tag{3}$$

providing us

$$\partial_{g_l} \left|\underline{\underline{H}}\,\underline{g} - \underline{B}\right|^2 = 0, \tag{4}$$

where † symbolizes the Hermitian adjunction. The LSF estimator $\underline{g}_L$ realizing the minimal deviation is given by

$$\underline{g}_L = \left[\underline{\underline{H}}^{\dagger}\,\underline{\underline{H}}\right]^{-1} \underline{\underline{H}}^{\dagger}\,\underline{B}. \tag{5}$$

2.3. Capon's Method

Capon's method is based on the construction of a filter matrix $\underline{\underline{w}}$ so that the output power

$$\mathrm{tr}\left[\underline{\underline{w}}^{\dagger}\underline{\underline{M}}\,\underline{\underline{w}}\right] \tag{6}$$

is minimized with respect to $\underline{\underline{w}}$, subject to the distortionless constraint

$$\underline{\underline{w}}^{\dagger}\underline{\underline{H}} = \underline{\underline{I}}, \tag{7}$$

where $\mathrm{tr}\left[\underline{\underline{w}}^{\dagger}\underline{\underline{M}}\,\underline{\underline{w}}\right]$ is the trace of the matrix $\underline{\underline{w}}^{\dagger}\underline{\underline{M}}\,\underline{\underline{w}}$ and $\underline{\underline{I}}$ is the identity matrix. The matrix $\underline{\underline{M}} : = \langle \underline{B} \circ \underline{B} \rangle$ is called the data covariance matrix, where the angular brackets indicate averaging over ensemble, e.g., different samples, realizations, or measurements. The error of the magnetic data is assumed to be Gaussian with variance σ_n and zero mean. In this case, the data covariance matrix can be written as $\underline{\underline{M}} = \langle \underline{B} \rangle \circ \langle \underline{B} \rangle + \sigma_n^2 \underline{\underline{I}}$. Capon's estimator realizing the minimal output power, subject to the distortionless constraint, results in [9]

$$\underline{g}_C = \left[\underline{\underline{H}}^{\dagger}\ \underline{\underline{M}}^{-1}\ \underline{\underline{H}}\right]^{-1} \underline{\underline{H}}^{\dagger}\ \underline{\underline{M}}^{-1}\ \langle \underline{B} \rangle, \tag{8}$$

which has the same structure as the LSF estimator (Equation 4), but with additional weighting by the covariance matrix. This demonstrates that the Capon filter discriminates between preferred and deprived data whereas the LSF treats all data equally. Adding a constant value σ_d^2 to the diagonal of the covariance matrix improves the robustness of Capon's estimator [14]. The diagonal loaded covariance matrix results in

$$\underline{\underline{M}} = \langle \underline{B} \rangle \circ \langle \underline{B} \rangle + \sigma^2 \underline{\underline{I}}, \tag{9}$$

where $\sigma^2 : = \sigma_n^2 + \sigma_d^2$.

3. SIMULATION OF MERCURY'S MAGNETIC FIELD

For the evaluation of Capon's estimator in comparison with the LSF estimator simulated magnetic field data are analyzed. The data are simulated with the hybrid code AIKEF [15], that has successfully been applied to several problems in Mercury's plasma interaction [16]. The internal Gauss coefficients $g_1^0 = -190\,\mathrm{nT}$ and $g_2^0 = -78\,\mathrm{nT}$ [17], defining the non-vanishing components of the ideal coefficient vector $\underline{g}$ (Equation 2), are implemented in the simulation code and the magnetic field resulting from the

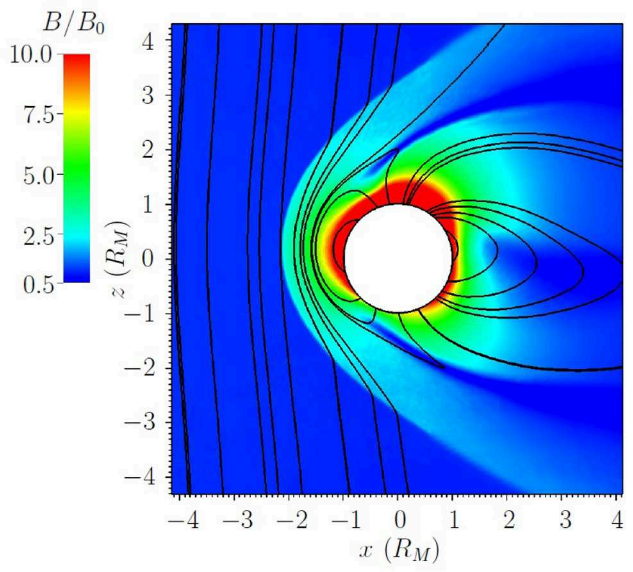

FIGURE 1 | Simulated magnitude of the magnetic field B in multiples of the solar wind magnetic field $B_0 = 20\,\text{nT}$ in the x-z-plane. The black lines describe the magnetic field lines. The white circle of radius $1\,R_M$ symbolizes Mercury. The implemented internal Gauss coefficients are $g_1^0 = -190\,\text{nT}$ and $g_2^0 = -78\,\text{nT}$ [17].

interaction of Mercury with the solar wind is simulated. The solar wind velocity of 400 km/s is orientated parallel to the x-axis and the solar wind magnetic field with $B_0 = 20\,\text{nT}$ is orientated toward the z-axis. The y-axis completes the right hand system. The solar wind density was chosen to $30\,\text{cm}^{-3}$. In **Figure 1**, the simulated magnitude of the magnetic field B is displayed in the x-z-plane (meridional plane).

4. APPLICATION AND DISCUSSION

Now Capon's method is applied to the simulated data for reconstructing the ideal Gauss coefficients implemented in the simulation. The comparison of Capon's estimator $\underline{g}_C$ with the ideal coefficient vector $\underline{g}$ enables the judgement of the method. To classify the role of Capon's method in terms of the diversity of existing inversion methods, Capon's estimator furthermore is compared with the LSF estimator $\underline{g}_L$. The data are evaluated at an ensemble of data points with distance $0.2\,R_M$ from the surface on the night side of Mercury ($x < 0$). The reconstructed Gauss coefficients are presented in **Table 1**.

The underlying model only describes the internal magnetic field $\underline{\underline{H}}\,\underline{g}$. The external parts of the field $\underline{b} := \underline{B} - \underline{\underline{H}}\,\underline{g}$ are not parameterized. Thus, the deviation of the LSF estimator and the ideal coefficient vector is given by

$$\left|\underline{g}_L - \underline{g}\right| = \left|\left[\underline{\underline{H}}^\dagger\,\underline{\underline{H}}\right]^{-1}\underline{\underline{H}}^\dagger\,\underline{b}\right| \approx 32.9\,\text{nT}, \qquad (10)$$

whereas the difference between Capon's estimator and the ideal coefficient vector results in

$$\left|\underline{g}_C - \underline{g}\right| = \left|\left[\underline{\underline{H}}^\dagger\underline{\underline{M}}^{-1}\underline{\underline{H}}\right]^{-1}\underline{\underline{H}}^\dagger\underline{\underline{M}}^{-1}\underline{b}\right| \approx 20.1\,\text{nT}. \qquad (11)$$

TABLE 1 | Capon's and LSF estimators for the internal Gauss coefficients in nT.

Gauss coefficient	Input	Output Capon	Output LSF	MESSENGER [17]
g_1^0	−190.0	−191.6	−215.9	−215.8 to −190.0
g_1^1	0	0.4	0.5	−2.9 to 1.1
h_1^1	0	0.6	0.7	0.8 to 2.7
g_2^0	−78.0	−69.1	−77.9	−83.2 to −57.0
g_2^1	0	16.9	19.0	−1.5 to 3.4
h_2^1	0	5.5	6.2	−1.4 to 0.2
g_2^2	0	−2.8	−3.2	−7.0 to −0.8
h_2^2	0	0.7	0.8	−3.3 to 0.4

In the last column the ranges of Gauss coefficients, reconstructed from MESSENGER data, are shown [17].

To judge the quality of Capon's estimator the comparison of individual coefficients presented in **Table 1** is not a vital metric. For example, the Gauss coefficient g_2^0 reconstructed by the LSF method is in better agreement with the ideal coefficient than the coefficient estimated by Capon's method. But for all coefficients together $\left|\underline{g}_C - \underline{g}\right| < \left|\underline{g}_L - \underline{g}\right|$ holds.

Therefore, Capon's estimator is in better agreement with the ideal coefficient vector than the LSF estimator.

The choice of the diagonal loading parameter σ_d^2 is essential for the difference $\left|\underline{g}_C - \underline{g}\right|$. The diagonal loaded covariance matrix results from the additional quadratic constraint $\text{tr}\left(\underline{\underline{w}}^\dagger\underline{\underline{w}}\right) = T_0$, where $T_0 = \text{const.}$ and σ_d^2 is the corresponding Lagrange multiplier [14]. The choice of T_0 controls the diagonal loading parameter σ_d^2 and defines how the data will be weighted by the filter matrix $\underline{\underline{w}}$. It depends on the underlying model and the evaluated data. **Figure 2** illustrates how σ in principle controls the difference $\left|\underline{g}_C - \underline{g}\right|$. For $\sigma \to 0$ Capon's estimator shows a large deviation to $\underline{g}$. If $\sigma \to \infty$, Capon's estimator approaches the LSF estimator. But if the data are not completely described by the model ($\underline{b} \neq 0$) there exists a parameter $\sigma = \sqrt{\sigma_n^2 + \sigma_d^2} = \sigma_0$, so that for all $\sigma \geq \sigma_0$

$$\left|\underline{g}_C - \underline{g}\right| \leq \left|\underline{g}_L - \underline{g}\right|. \qquad (12)$$

Furthermore it even exists an optimal parameter $\sigma_{opt.}$, that realizes the best agreement between Capon's estimator and $\underline{g}$. For the results presented in **Table 1** this optimal parameter is $\sigma_{opt.} \approx 276\,\text{nT}$.

Since the choice of σ controls $\text{tr}\left(\underline{\underline{w}}^\dagger\underline{\underline{w}}\right)$, the value of the optimal diagonal loading parameter is not directly related with an error of the magnetic measurements. More likely $\sigma_{opt.}$ can be understood as a parameter that measures the model mismatches.

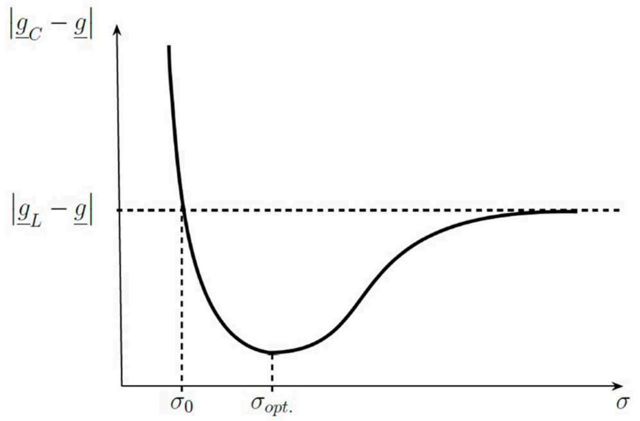

FIGURE 2 | Sketch of the deviation $|\underline{g}_C - \underline{g}|$ between Capon's estimator $\underline{g}_C$ and the ideal coefficient vector $\underline{g}$ subject to σ. For large $\sigma \to \infty$ the deviation converges to the deviation of the least-square-fit estimator $\underline{g}_L$ and the implemented coefficient vector $\underline{g}$. There exists σ_0 so that $|\underline{g}_C - \underline{g}| \leq |\underline{g}_L - \underline{g}|$, for all $\sigma \geq \sigma_0$, and an optimal parameter $\sigma_{opt.}$, that realizes the best agreement between Capon's estimator an the ideal coefficient vector.

When Capon's method is applied to real spacecraft data, the ideal coefficient vector $\underline{g}$ is not available anymore and therefore the deviation $|\underline{g}_C - \underline{g}|$ cannot be used as metric for calculating the optimal diagonal loading parameter. In this case, there exist other methods for estimating $\sigma_{opt.}$, e.g. the L-curve method, that solely depend on the underlying model and the data [18].

5. SUMMARY AND OUTLOOK

In this work Capon's method has been applied to simulated magnetic field data to analyze Mercury's internal magnetic field. The internal field, parameterized by the internal Gauss coefficients, was implemented in the simulation code AIKEF and the magnetic field resulting from the plasma interaction of Mercury and the solar wind was simulated. The comparison of Capon's method and the commonly used least square fit method showed that Capon's estimator is in better agreement with the implemented Gauss coefficients than the least square fit estimator. A helpful procedure is the diagonal loading of the data covariance matrix, that improves the robustness of Capon's estimator. It turns out that there exists an optimal diagonal loading parameter where Capon's estimator is nearest to the ideal coefficient vector.

Since only the internal magnetic field was parameterized, Capon's estimator shows some deviation to the implemented coefficients. Additional parameterizing of the external contributions of the magnetic field, for example by using the paraboloid model for Mercury's magnetosphere [10], may still improve Capon's estimator, especially when data points are collected in some distance above the planetary surface. Moreover, this enables us to reconstruct higher-order terms such as octupole terms. Furthermore, as the Gauss representation is restricted to curl-free regions, the Mie representation (poloidal-toroidal decomposition) would extend the data collection to regions where electrical currents flow.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

All authors contributed conception and design of the study. DH organized the database. ST, YN, and UM performed the statistical analysis and wrote the first draft of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

We acknowledge support by the German Research Foundation and the Open Access Publication Funds of the Technische Universität Braunschweig. This work by YN was supported by the Austrian Space Applications Programme at the Austrian Research Promotion Agency under contract 865967. DH was supported by the German Ministerium für Wirtschaft und Energie and the German Zentrum für Luft- und Raumfahrt under contract 50 QW1501.

ACKNOWLEDGMENTS

The authors are grateful for stimulating discussions and helpful suggestions by Karl-Heinz Glassmeier, Patrick Kolhey, and Alexander Schwenke.

REFERENCES

1. Connerney JEP. The magnetic field of Jupiter: a generalized inverse approach. *J Geophys Res.* (1981) **86**:7679–93. doi: 10.1029/JA086iA09p07679
2. Moore KM, Bloxham J, Connerney JEP, Jørgensen JL, Merayo JMG. The analysis of initial Juno magnetometer data using a sparse magnetic field representation. *Geophys Res Lett.* (2017) **44**:4687–93. doi: 10.1002/2017GL073133
3. Davis L, Smith EJ. New models of Saturn's magnetic field using Pioneer 11 vector helium magnetometer data. *J Geophys Res.* (1986) **91**:1373–80. doi: 10.1029/JA091iA02p01373
4. Cao H, Russell CT, Christensen UR, Dougherty MK, Burton ME. Saturn's very axisymmetric magnetic field: no detectable secular variation or tilt. *Earth Planet Sci.* (2011) **304**:22–8. doi: 10.1016/j.epsl.2011.02.035
5. Jackson A. Intense equatorial flux spots on the surface of the Earth's core. *Nature.* (2003) **424**:760–3. doi: 10.1038/nature01879
6. Capon J. High resolution frequency-wavenumber spectrum analysis. *Proc IEEE.* (1969) **57**:1408–18. doi: 10.1109/PROC.1969.7278
7. Haykin S. *Adaptive Filter Theory. 2nd Edn. Prentice Hall Information and System Science Series.* New Jersey: Prentice-Hall Inc. (1991).
8. Motschmann U, Woodward TI, Glassmeier KH, Southwood DJ, Pinçon JL. Wavelength and direction filtering by magnetic measurements at satellite

arrays: generalized minimum variance analysis. *J Geophys Res.* (1996) **101**:4961–6. doi: 10.1029/95JA03471

9. Narita Y. A note on Capon's minimum variance projection for multi-spacecraft data analysis, *Front Phys.* (2019) **7**:8. doi: 10.3398/fphy.2019.00008
10. Alexeev II, Belenkaya ES, Bobrovnikov S, Slavin Yu, Sarantos JA. Paraboloid model of Mercury's magnetosphere. *JGR.* (2008) **113**: doi: 10.1029/2008JA013368
11. Alexeev II, Belenkaya ES, Slavin JA, Korth H, Anderson BJ, Baker DN, et al. Mercury's magnetospheric magnetic field after the first two MESSENGER flybys. *Icarus*. (2010) **209**:23–39. doi: 10.1016/j.icarus.2010.01.024
12. Johnson CL, Purucker ME, Korth H, Anderson BJ, Winslow RM, Al Asad MMH, et al. MESSENGER observations of Mercury's magnetic field structure. *JGR.* (2012) **117**:E00L14. doi: 10.1029/2012JE004217
13. GaußCF. *Allgemeine Theorie des Erdmagnetismus: Resultate aus den Beobachtungen des Magnetischen Vereins im Jahre 1838.* (1839).
14. Van Trees HL. *Detection, Estimation, and Modulation Theory, Optimum Array Processing.* New York, NY: Wiley (2002).
15. Mueller J, Simon S, Motschmann U, Schuele J, Glassmeier K-H, Pringle GJ. A.I.K.E.F.: adaptive hybrid model for space plasma simulations. *Comput Phys Commun.* (2011) **182**:946–66. doi: 10.1016/j.cpc.2010.12.033
16. Exner W, Heyner D, Liuzzo L, Motschmann U, Shiota D, Kusano K, et al. Coronal mass ejection hits mercury: AIKEF hybrid-code results compared to MESSENGER data. *Planet Space Sci.* (2018) **153**:89–99. doi: 10.1016/j.pss.2017.12.016
17. Wardinski I, Langlais B, Thébault E. Correlated time-varying magnetic field and the core size of mercury. *J Geophys Res.* (2019) **124**:2178–97. doi: 10.1029/2018JE005835
18. Hiemstra JH, Wippert MW, Goldstein JS, Pratt T. Application of the L-curve technique to loading level determination in adaptive beamforming. In: *Conference Record of the Thirty-Sixth Asilomar Conference on Signals, Systems and Computers*. Pacific Grove, CA (2002).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Physics

PERSPECTIVE
published: 21 July 2020
doi: 10.3389/fphy.2020.00221

Lorentzian Entropies and Olbert's κ - Distribution

Rudolf A. Treumann[1,2] *and Wolfgang Baumjohann*[3*]

[1] International Space Science Institute, Bern, Switzerland, [2] Geophysics, Department Geoscience & Environment, Munich University, Munich, Germany, [3] Space Research Institute (IWF), Austrian Academy of Sciences, Graz, Austria

This note derives the various forms of entropy of a systems subject to Olbert distributions (generalized Lorentzian probability distributions known as κ-distributions), which are frequently observed, particularly in high-temperature plasmas. The general expression of the partition function in such systems is given as well in a form similar to the Boltzmann-Gibbs probability distribution, including a possible exponential high-energy truncation. We find the representation of the mean energy as a function of probability, and we provide the implicit form of Olbert (Lorentzian) entropy as well as its high-temperature limit. The relation to phase space density of states is obtained. We then find the entropy as a function of probability, an expression that is fundamental to statistical mechanics and, here, to its Olbertian version. Lorentzian systems through internal collective interactions cause correlations that add to the entropy. Fermi systems do not obey Olbert statistics, while Bose systems might do so at temperatures that are sufficiently far from zero.

Keywords: generalized entropy, Lorentzian systems, Lorentzian countings, Olbert distribution, κ distribution

OPEN ACCESS

Edited by:
Christopher H. K. Chen,
Queen Mary University of London,
United Kingdom

Reviewed by:
Manfred P. Leubner,
University of Innsbruck, Austria
KLaus Scherer,
Ruhr University Bochum, Germany

****Correspondence:***
Wolfgang Baumjohann
Wolfgang.Baumjohann@oeaw.ac.at

Specialty section:
This article was submitted to
Space Physics,
a section of the journal
Frontiers in Physics

Received: *11 April 2020*
Accepted: *22 May 2020*
Published: *21 July 2020*

Citation:
Treumann RA and Baumjohann W
(2020) Lorentzian Entropies and
Olbert's κ - Distribution.
Front. Phys. 8:221.
doi: 10.3389/fphy.2020.00221

1. INTRODUCTION

Many-particle systems not in equilibrium, such as high-temperature plasmas, are usually subject to kinetic theory (cf., e.g., [1]). In equilibrium or stationary quasi-equilibrium, obeying a very large number of degrees of freedom, they can beneficially be treated by the probabilistic methods of statistical mechanics. Conventional textbook knowledge [2] tells us that, for the micro-canonical system under consideration, it being in thermal exchange with a large thermal bath at temperature $T \equiv \beta^{-1}$ (here taken in energy units), the probability p_α of finding it in some particular energy state E_α is proportional to the Boltzmann factor $p_\alpha \propto \exp(-\beta E_\alpha)$. The sum of all un-normalized probabilities of the α states is the partition function $Z = \sum_\alpha p_\alpha = \sum_\alpha \exp(-\beta E_\alpha)$ and the normalized Gibbs probability for the state α becomes

$$P_\alpha = Z^{-1} \exp(-\beta E_\alpha) \tag{1}$$

The partition function $Z \equiv Z(\beta, \{V\})$ is a function of β and all constraining parameters $\{V\}$, which determine the state α-a property that enables calculating a number of thermodynamically interesting average quantities of the system. Varying the constraints $\{V\}$ implies that work is done on the system.

Observations in space plasma physics (for examples, see cf., [3–5]) as well as in other high-temperature systems indicate that the probability distribution of particles (charged or neutral) in a set of energy states E_α deviates from the classical bell (respectively gaussian) shape, frequently exhibiting quasi-stationary power law tails $P_\alpha \propto E_\alpha^{-\kappa}$ for $E_\alpha > \beta^{-1}$, possibly cut off exponentially at large energy. Probability distributions of this kind of family, known as

κ-distributions (introduced[1] by Olbert [7]), have been widely discussed in the literature (for a review see, e.g., [8], and references therein). In the following, we refer to them as *Olbert's κ-distributions* or simply *Olbert distributions*. General physical arguments for their existence as stationary states far from thermal equilibrium were given (first by [20, 24]). Direct weak turbulence calculations of plasma-electron momentum distributions by Hasegawa et al. ([11], in interaction with a photon bath) and by Yoon et al. ([9, 10], accounting for spontaneous and induced emission as well as absorption of Langmuir waves) partially reproduced κ-distributions in the long term limit, suggesting that under quasi-stationary conditions non-linear equilibria can be produced with κ-distributions being their probabilistic signature.

2. LORENTZIAN GENERALIZATION

In generalizing the classical statistical mechanics, we start from a Lorentzian modification of the Boltzmann factor, which leads to the Olbert probability distribution known as κ-distribution.

2.1. Boltzmann-Olbert Distribution

In fact, the Boltzmann factor, being at the heart of Gibbs' normalized probability, is the large κ limit of a more general Lorentzian, the Olbert κ-probability function

$$P_{\kappa\alpha}(E_\alpha,\beta) = Z^{-1}_{\kappa,r}\Big[1+\frac{\beta E_\alpha}{\kappa}\Big]^{-(\kappa+r)}, \qquad \lim_{\kappa\to\infty} P_{\kappa\alpha} \to P_\alpha \quad (2)$$

(with $r = \text{const} \neq 0$), as can easily be confirmed applying l'Hospital's rule. It corresponds to the abovementioned experimentally and frequently confirmed κ-distribution. The resulting Olbert-partition function Z_κ is, in analogy to Gibbs' partition function, defined as

$$Z_{\kappa,r}(\beta) = \sum_\alpha \Big[1+\frac{\beta E_\alpha}{\kappa}\Big]^{-(\kappa+r)} \quad (3)$$

It warrants that the Olbert probabilities of states α are normalized and add up to $\sum_\alpha P_{\kappa\alpha}(E_\alpha,\beta) = 1$. Performing this sum requires knowledge of the different energy states E_α, which, in general, cannot be done easily. In the following, we show that, assuming this form, the rules of classical statistical mechanics can be made applicable to the Olbert-Lorentzian with only weak modifications.

2.2. Remark on Convergence

Before proceeding, we briefly refer to the convergence of Olbert's κ-probability distribution Equation (2).

The Olbert probability converges for arbitrary power $\kappa > 0$. It does, however, for constant κ, not allow the calculation of arbitrarily high average moments, for instance if one is interested in fluid descriptions. In principle, at this stage, $\kappa(\beta, E_\alpha)$, being a function of temperature and/or even energy states E_α, is not excluded; in the latter case one would, however, require that its dependence is weak in order to maintain the above summation procedure as simple as possible. Such a dependence is implicit to the non-linear calculations of Hasegawa et al. [11] and Yoon et al. [9]. The additional freedom introduced by the constant r just adjusts for the mean energy in an ideal gas (see, e.g., [26], and references therein). In general, however, the number of moments that can be calculated is limited. In a fluid approach, it requires artificially truncating the chain of moments, for instance by applying a water-bag model for $\kappa(\beta, E_\alpha)$ of the kind $\kappa = \text{const}, E_\alpha \leq E_c$, and $\kappa \to \infty, E_\alpha > E_c$ (implicitly assumed in [12]). Truncation may be justified via additional assumptions on the underlying physics, like suppression of higher moments than heat flows and similar conditions. Physically, this may not be unreasonable. From a formal point of view, brute force truncation is not satisfactory. However, this restriction can easily be circumvented (see e.g., [12–14]) when introducing an exponential cut-off energy $\beta E_c \gg 1$ through

$$P_{\kappa\alpha} = \frac{e^{-E_\alpha/E_c}}{Z_{\kappa,r}}\Big[1+\frac{\beta E_\alpha}{\kappa}\Big]^{-(\kappa+r)}, \qquad \beta E_c \gg 1 \quad (4)$$

which warrants convergence of all moments for arbitrary $\kappa > 0$ [13, 14]. The chain of physically interesting moments is discussed in these papers. In the Olbert partition function, the energy cut-off simply appears as a truncator

$$Z_{\kappa,r} = \sum_\alpha e^{-E_\alpha/E_c}\Big[1+\frac{\beta E_\alpha}{\kappa}\Big]^{-(\kappa+r)} \quad (5)$$

not having any further effect on the determination of averages and/or any other thermodynamic quantities other than warranting the convergence of the chain of moments. The independence of the exponential cut-off on temperature and β guaranties that, in all derivatives or integrals with respect to β, it appears as an energy dependent factor. An example has been given [15] by application to the Cosmic Ray energy spectrum, where the cut-off is, for quantum physical reasons, found in the GZ-energy spectral limit. Its inclusion, if necessary, does not cause any principal problems. In the following, we therefore suppress it in order to not unnecessarily complicate the expressions.

3. OLBERT-LORENTZIAN STATISTICS

It is reasonable to assume that, given the above definition of the probability, ensemble averages can be calculated as linear mean values, with the probability $P_{\kappa\alpha}$ determining the weight each energy level contributes. This is the basic probability assumption. One may argue that this may not necessarily be true if the probability of the states are not independent. Such arguments have been put forward in some entropy definitions

[1] Stanislaw (Stan) Olbert (1923–2017, of Polish origin, after WW II a graduate of Arnold Sommerfeld in Munich, and, since 1957, Professor of Physics at MIT, working on the American Space Program with Bruno Rossi, the main discoverer of the X-ray sky and, together with Riccardo Giacconi, who later was awarded the Nobel Prize for this, founder of X-ray astronomy) invented the κ-probability distribution to fit observed IMP spacecraft particle spectra. He suggested its application to electron fluxes measured by the OGO spacecraft to Vasyliunas [6] whose publication became one of the most referenced papers in space physics for no other reason than the first refereed formal appearance of Olbert's κ distribution in the literature.

(see, e.g., [16]), and some of them, like Renyi and Tsallis q-entropies (cf. [17–19], for their invention) are used in chaotic theory [2]. However, as long as there is no need to worry about, the κ-generalization of the probabilities already accounts for a particular kind of internal correlations among the occupations of the different states. The states are physically ordered as in Boltzmann-Gibbs theory, while the probabilities of their occupations have become not completely independent. In this spirit the mean energy is defined as,

$$U(\beta) \equiv \langle E \rangle = Z_{\kappa,r}^{-1} \sum_\alpha E_\alpha \Big[1 + \frac{\beta E_\alpha}{\kappa}\Big]^{-\kappa-r} \tag{6}$$

3.1. Mean Energy

In full generality, the sum can formally be done in two ways when observing the properties of the partition function. The first way completes the energy, and one easily finds that

$$U(\beta) = \frac{\kappa}{\beta}\Big[\frac{Z_{\kappa,r-1}(\beta)}{Z_{\kappa,r}(\beta)} - 1\Big] \tag{7}$$

also showing the importance of having made use of the freedom of introducing the arbitrary constant $r \neq 0$.

A second form, resembling that of conventional statistical mechanics, takes advantage of the differential property

$$\frac{\partial Z_{\kappa,r-1}}{\partial\beta} = -\frac{\kappa + r - 1}{\kappa} Z_{\kappa,r} U(\beta) \tag{8}$$

of the partition function, yielding

$$U(\beta) = -\frac{\kappa}{\kappa + r - 1}\frac{Z_{\kappa,r-1}}{Z_{\kappa,r}}\Big(\frac{\partial \log Z_{\kappa,r-1}}{\partial\beta}\Big)_{\{V\}} \tag{9}$$

Here, generalization to Olbert-Lorentzian distributions introduces the (inconvenient) partition function ratio of different indices. It again shows the need for the additional constant $r \neq 1$, which depends on the assumptions on an underlying model. For instance, under classical ideal gas conditions with continuously distributed energy states, the average thermal energy (in three dimensions and isotropy) is $\beta U = \frac{3}{2}$. On switching to momentum $\mathbf{p}$ with $E_\alpha \to p^2/2m$ and integrating over momentum space, one obtains that $r = \frac{5}{2}$ in this particular case ([26], and elsewhere; see references therein). This is not necessarily true, however, for discrete energy levels E_α in more general non-ideal or quantum conditions. There r must be chosen differently and a general prescription for its choice cannot be given *a priori*.

Both the above forms apply to any micro-canonical κ-system. Generalization to canonical systems is easily done in the same way as in statistical mechanics (cf., e.g., [2]) via introducing the dependence on (possibly variable) particle number N. It requires reference to Lagrange multipliers μ playing the role of chemical potentials for each subsystem and transforming $E_\alpha \to E_\alpha - \mu$ in the probabilities and partition functions. Clearly, all μ must become equal in thermal equilibrium.

The two Equations (7, 9) allow for the determination of the ratio of the partition functions by eliminating $U(\beta)$

$$\frac{Z_{\kappa,r-1}}{Z_{\kappa,r}} = \frac{\kappa + r - 1}{\kappa + r - 1 + \beta\Big(\partial \log Z_{\kappa,r-1}/\partial\beta\Big)_{\{V\}}} \tag{10}$$

This is a recursive relation between the κ partition functions. Combined with Equation (9), it gives a final expression for the mean energy

$$U(\beta) = -\frac{\kappa\Big(\partial \log Z_{\kappa,r-1}/\partial\beta\Big)_{\{V\}}}{\kappa + r - 1 + \beta\Big(\partial \log Z_{\kappa,r-1}/\partial\beta\Big)_{\{V\}}} \tag{11}$$

which contains just the r-reduced partition function. Like in ordinary statistical mechanics, $U(\beta)$ is determined as a derivative form of the partition function. This shows that all other statistical mechanical quantities can be derived solely from the partition function, which therefore contains all the physics of the micro-canonical system. Still being quite involved, this form, as expected for very large κ, coincides with the expression $U = -\big[\partial(\log Z)/\partial\beta\big]_{\{V\}}$ of the mean energy in Boltzmann-Gibbs statistical mechanics. It is thus consistent with the expectations. Moreover, at increased temperatures $\beta \to 0$, the mean classical energy becomes

$$U(\beta) = -\frac{\kappa}{\kappa + r - 1}\Big[\frac{\partial(\log Z_{\kappa,r-1})}{\partial\beta}\Big]_{\{V\}} \qquad T \gg 0 \tag{12}$$

The general second last equation (11), which holds for arbitrary $\beta < \infty$, resolved for the derivative of the partition function as

[2] Historically, Renyi's proposal of a q-entropy (for the complete theory see [18]) came first (in a badly accessible publication by Balatoni and Renyi [17] in very general form, which implicitly already contained Tsallis' entropy as a particular case). This might have been known to Stan Olbert (who probably was familiar with the Hungarian literature). He used the property that for large parameter $q \to \infty$, as proposed by Renyi, and the modified mathematical expression agreed with Boltzmann's exponential. Olbert, however, tried a substantially simpler analytical form, calling the free parameter κ instead of q to distinguish it from Renyi's logarithm, as, in fact, it has a different meaning. Two decades later, Tsallis [19] used the property of Olbert's function, presumably not knowing Olbert's or Vasyliunas' much earlier papers and probably also not those by Balatoni and Renyi; Renyi, however, referred to the latter in his book, which Tsallis should have been familiar with because, at that time, Renyi's q-entropy was already highly celebrated in the then blossoming chaos theory. In contrast to Olbert, however, Tsallis did not apply it to the probability distribution. Rather, following Renyi's logarithmic approach, he used it in the entropy definition, arriving at his analytically simpler modified q-entropy. The two approaches of Olbert and Tsallis thus differ in the way of how the substitution for the exponential is used. As with ours, Olbert's interest was in the observed probability or momentum space distribution, and it was thus manifestly practical. Renyi's interest and later that of Tsalli was theoretica, and it was thus directed at entropy. Tsallis' led, and consequently developed, to his thermostatistics. In contrast, in an attempt to justify Olbert's distribution, we arrived originally at a κ-distribution from a consequent reference to kinetic theory [20], not yet recognizing, however, the important role of the constant r. As it turns out, both approaches are indeed rather different, even though a formal relation between the parameters κ and q can easily be construed while maintaining their different meanings, which is frequently overlooked when identifying q and κ statistics, as these have little in common.

function of mean energy $U(\beta)$, yields

$$\Big(\frac{\partial\log Z_{\kappa,r-1}}{\partial\beta}\Big)_{\{V\}}=\Big(1+\frac{r-1}{\kappa}\Big)U(\beta)\Big(1+\frac{\beta U(\beta)}{\kappa}\Big)^{-1} \tag{13}$$

$$=\Big(1+\frac{r-1}{\kappa}\Big)\Big[1-\Big(1+\frac{\beta U(\beta)}{\kappa}\Big)^{-1}\Big] \tag{14}$$

an expression that can be made use of later. At high temperatures, i.e., small β, the first version shows that the derivative of the partition function yields the mean energy, the usual Boltzmann-Gibbs result.

At very low temperature $\beta\gg 1$, the mean energy drops out, which contradicts the physical intuition showing that the theory in this form applies only to temperatures far from zero. The logarithm of the partition function is the integral

$$\log Z_{\kappa+r-1}=\Big(1+\frac{r-1}{\kappa}\Big)\Big[\beta-\int\frac{d\beta}{1+\beta U(\beta)/\kappa}\Big]+G_\kappa(\{V\}) \tag{15}$$

with $G_\kappa(\{V\})$ a function of the constraints alone. Olbert-Lorentzian statistical mechanics in the above form applies to micro-canonical systems at high temperature only. It does, in this form, not describe quantum systems consisting of many components–a conclusion we had drawn already from different reasoning. This conclusion may, however, be circumvented when large external potential fields Φ are imposed, for instance, strong electric (cf., e.g., [21], who tried an application to high temperature non-ideal quantum systems) or gravitational potential fields (an example would be the region around the black hole horizon), in which case the difference $U(\beta)-\Phi>0$ may become positive for $-\Phi>\kappa/2\beta$.

3.2. Entropy

The most important quantity in statistical mechanics is the entropy. Differentiating the energy $U(\beta)$ with respect to temperature $k_B\beta^{-1}$ while fixing the set of constraints $\{V\}$ gives the heat capacity

$$C_{\{V\}\kappa}=-k_B\beta^2\Big(\frac{\partial U(\beta)}{\partial\beta}\Big)_{\{V\}} \tag{16}$$

With entropy S, one has quite generally $TdS=C_{\{V\}}dT$ holding in the micro-canonical ensembles where the volume is fixed, $dV=0$. Hence $C_{\{V\}}=-\beta(\partial S/\partial\beta)_{\{V\}}$. Keeping the constraints fixed, these relations usually lead to

$$\Big(\frac{\partial S}{\partial\beta}\Big)_{\{V\}}=-k_B\beta\Big(\frac{\partial U(\beta)}{\partial\beta}\Big)_{\{V\}} \tag{17}$$

Integration with respect to β then yields in full generality the well-known formal expression for the wanted Olbert entropy S_κ of the micro-canonical system

$$\frac{S}{k_B}=-\beta U(\beta)+\int d\beta\,U(\beta)+G_S(\{V\}) \tag{18}$$

as the integral over the mean energy $U(\beta)$, where $G_S(\{V\})$ is an arbitrary function of the constraints alone. In classical statistics, this formula yields the well-known closed analytical expression of the entropy. Unfortunately, in Olbert's case the mean energy Equation (11) is not as simple as in Boltzmann-Gibbs statistical mechanics[3]. We are thus stuck for the moment. Nevertheless, taking the derivative of the mean energy with respect to β, one obtains formally

$$\Big(\frac{\partial S}{\partial\beta}\Big)_{\{V\}}=-\kappa k_B\beta\frac{\partial}{\partial\beta}\Big[\frac{\partial(\log Z_{\kappa,r-1})/\partial\beta}{\kappa+r-1+\beta\partial(\log Z_{\kappa,r-1})/\partial\beta}\Big]_{\{V\}} \tag{19}$$

as an implicit expression for the derivative of the entropy S as functional of the partition function. It replaces the corresponding relation in classical Boltzmann-Gibbs statistical mechanics, which applies to any purely stochastic many particle system–in particular to high-temperature plasmas.

Equation (18) is the entropy of a micro-canonical κ system. It is a quite involved form whose properties cannot be easily inferred. Its discussion requires the complete knowledge of the set of energy levels of the micro-canonical system. As discussed above, its extension to canonical systems is straightforward, as well as the inclusion of an exponential "ultraviolet" truncation of the distribution at high energy $E_c>U$. All interesting statistical mechanical properties of the κ ensemble can *in principle* be deduced from this entropy respectively the partition function $Z_{\kappa,r-1}$.

3.3. High-Temperature Limit

In the high temperature small β limit, one neglects the derivative in the denominator in the second last equation. In this case, the entropy becomes a κ-modified (Boltzmann-Olbert-Lorentzian) entropy

$$\Big(\frac{\partial S}{\partial\beta}\Big)_{\{V\}}=-\frac{\kappa k_B\beta}{\kappa+r-1}\frac{\partial}{\partial\beta}\Big[\frac{\partial(\log Z_{\kappa,r-1})}{\partial\beta}\Big]_{\{V\}},\qquad T\gg0 \tag{20}$$

No zero-temperature expression exists, while the role of the partition function is played by the sum of the probabilities of the states indexed by the constant power $r-1$ instead of r. For a three-dimensional ideal gas with continuous energy spectrum one has $r=\frac{5}{2}$, and its high-temperature classical κ-partition function is

$$Z_{\kappa+\frac{3}{2}}(\beta)=\sum_\alpha p_{\alpha,\kappa+\frac{3}{2}}\equiv\sum_\alpha\Big(1+\frac{\beta E_\alpha}{\kappa}\Big)^{-\kappa-\frac{3}{2}},\qquad T\gg0 \tag{21}$$

With this partition function and the definition of the high-temperature mean energy (12), we are in the position to obtain

[3] At this point κ and q statistics differ for the simple fact that in the latter the entropy is analytically prescribed in the form of a rational function with real q, and the complication is transferred to the construction of the distribution. Here, instead, the starting point is the observed distribution, which, naturally, leads to complications in finding the entropy, as it is the entropy which contains the complicated physics; this then leads to the measured probability distribution. One should also note that the combined entropies of two systems in both cases, q and Olbert statistical mechanics, though the two theories are different and describe different physics, are super-additive, sometimes called non-extensive. They contain an additional mixed term which contributes to the entropy, as criticized by Nauenberg [22]. This, however, does not mean that the theory has no physical meaning. It just implies that the theory describes statistical quasi-equilibria far from thermal equilibrium, i.e., slowly variable quasi-stationary states which pass through several equilibria, typical for non-equilibrium statistical mechanics [23].

the high-temperature entropy in the form in which it applies to fluids and plasmas:

$$\frac{S_\kappa}{k_{B\kappa}} = -\beta\Big[\frac{\partial \log Z_{\kappa+r-1}(\beta)}{\partial\beta}\Big]_{\{V\}} + \log Z_{\kappa+r-1}(\beta) \tag{22}$$

where we left r undetermined available for application to any non-ideal systems and dropped the arbitrary function G_S of the constraints, which can be added when needed–for instance to account for boundary conditions. Except for the modification of the partition function, the entropy at high temperatures is measured in units of a κ-reduced Boltzmann constant

$$k_{B\kappa} = k_B\Big(1+\frac{r-1}{\kappa}\Big)^{-1}, \qquad 0 \nleq \kappa < \infty \tag{23}$$

which in a three-dimensional ideal plasma becomes $k_{B\kappa} = k_B/(1+3/2\kappa)$.

3.4. Phase Space Density of States

As in ordinary Boltzmann-Gibbs statistical mechanics, the Olbert partition function for large numbers of states (energy levels) Ω_κ, which is the volume of the phase space, is well-approximated by

$$Z_{\kappa,r-1} \approx \Omega_{\kappa,r-1}\Big(1+\frac{\beta U(\beta)}{\kappa}\Big)^{-\kappa-r+1} \tag{24}$$

the product of the phase space volume $\Omega_{\kappa,r-1}$ and the probability of the most probable state, which is the state of mean energy $U(\beta)$. This holds, in particular for the exponentially truncated distribution, because the high energy states contribute very little if only the energy fluctuations are not overwhelmingly large. These fluctuations become large only in systems containing very small numbers of particles, which is barely given at the assumed high temperatures in a plasma.

Taking advantage of the dependence of the ratio of the partition functions $Z_{\kappa,r-1}/Z_{\kappa,r}$ on the average energy $U(\beta)$, which does not depend on r, one finds that

$$\Omega_{\kappa,r-1} = \Omega_{\kappa,r} \equiv \Omega_\kappa \tag{25}$$

At high temperatures $\beta \ll 1$, we have

$$S_\kappa \approx k_{B\kappa}\log Z_{\kappa,r-1} + k_{B\kappa}(\kappa+r-1)\log\Big(1+\frac{\beta U(\beta)}{\kappa}\Big) \approx k_{B\kappa}\log\Omega_\kappa \tag{26}$$

which can also be written

$$P_\kappa \sim \Omega_\kappa = \exp\Big(S_\kappa/k_{B\kappa}\Big) \tag{27}$$

In classical high-temperature micro-canonical systems (many-particle plasmas) this closes the circle, as we have shown [24] that from this equation it follows by standard methods that the probability distribution is given by Equation (2). Generalization to the canonical system of N particles is straightforward. Notably, it generalizes to κ-systems Einstein's prescription [25] of the dependence of the phase space density on entropy in his proof of the stochastic nature of the diffusion in Brownian motion, though with substantially more complicated expression for the entropy.

3.5. Approximation

The Olbert κ-distribution maintains the structure of statistical mechanics at high temperatures while it substantially modifies it at moderate and low temperatures with no zero-temperature limit existing. We have argued previously that this is quite reasonable whenever internal correlations come into play causing κ to deviate strongly from $\kappa = \infty$. Classically, this can happen only at large T and is due to non-linear interactions that violate ideal stochasticity and cause anomalous effects like anomalous diffusivity. Nevertheless, in the range

$$\log Z_{\kappa,r-1} \gg (\kappa+r-1)\log\beta \quad \text{or} \quad Z_{\kappa,r-1} \gg \beta^{\kappa+r-1} \tag{28}$$

which holds for sufficiently large β, the equation for the Olbert entropy simplifies. In this case the derivatives of the logarithms of the partition function cancel, and the entropy equation becomes

$$\Big(\frac{\partial S}{\partial\beta}\Big)_{\{V\}} \approx -\kappa k_B\beta\frac{\partial}{\partial\beta}\frac{1}{\beta}, \qquad 1 \ll \beta < \infty \tag{29}$$

which of course holds for finite β only. Integration then yields that in this β range

$$\frac{S}{k_B} \propto \kappa\log\beta \ll \frac{\kappa}{\kappa+r-1}\log Z_{\kappa,r-1} \tag{30}$$

or otherwise

$$Z_{\kappa,r-1} \gg \exp\frac{S}{k_{B\kappa}} \tag{31}$$

In the moderate temperature range where $0 \ll \beta \ll \infty$ no closed forms for either the energy nor the entropy are obtained. For those values of β, the full expression (19) for the derivative of the entropy applies. A correction to this equation follows when taking the first next term of the expansion of the denominator

$$\Big(\frac{\partial S}{\partial\beta}\Big)_{\{V\}} = -\frac{\kappa k_B}{\kappa+r-1}\beta\frac{\partial}{\partial\beta}\Big\{\Big(\frac{\partial\log Z_{\kappa,r-1}}{\partial\beta}\Big)_{\{V\}} - \frac{\beta(\partial\log Z_{\kappa,r-1}/\partial\beta)^2_{\{V\}}}{\kappa+r-1} + \text{h.o.t.}\Big\} \tag{32}$$

The first term yields, when integrated, the above high-temperature entropy. The second term is quadratic and hence remains to be negative. It subtracts from the first term. Reducing the temperature, i.e., increasing β, obviously diminishes the derivative of the entropy because the last quadratic term is always positive. It seems that the derivative of the entropy as function of temperature in a κ-system flattens out when the temperature drops into the intermediate range. Any $\kappa \neq \infty$ affects the increase in entropy.

In principle, the last equation can be solved iteratively for the entropy, which then retains the effects of the parameter κ outside the ranges of very large and small β.

4. ENTROPY AS FUNCTIONAL OF PROBABILITY

Boltzmann defined the micro-canonical entropy $S_{B\alpha} \propto \log p_\alpha$ as a functional of probability. The average measured entropy is its expectation value, the sum $\langle S_B\rangle \propto \sum_\alpha p_\alpha \log p_\alpha$ of all probability-weighted contributions of the states to the entropy. For a continuous distribution of states, this is as usually defined as the probability integral taken over the micro-canonical entropy. For arbitrary temperatures, the above expression cannot be integrated to provide a general analytical form for the entropy comparable to conventional Boltzmann-Gibbs statistical mechanics. This was possible only at high temperatures. One can, however attempt to find an expression for the functional dependence of the entropy on the probability in order to have an equivalent representation to the Boltzmann-Gibbs entropy when dealing with the Olbert entropy.

4.1. Reformulation of Mean Energy and Entropy

For this to achieve, the mean energy (7) must be rewritten in terms of the probability $p_{\kappa\alpha}(\beta)$. This can, indeed, be done, and the corresponding expression reads

$$U\{p_{\kappa\alpha}(\beta)\} = \frac{\kappa}{\beta} Z_\kappa^{-1}\{p_{\kappa\alpha}(\beta)\} \sum_\alpha \Big[p_{\kappa\alpha}^{1-\gamma}(\beta) - p_{\kappa\alpha}\Big] \tag{33}$$

where we introduced the exponent

$$\gamma \equiv \frac{1}{\kappa + r} \tag{34}$$

The braces indicate the functional dependence on $p_{\kappa\alpha}$. In fact, Equation (33) is identical to the inverse function that has been proposed [20, 26] in the particular case of the Olbert-Lorentzian probability distribution[4]. On use of this functional dependence in the expression for the β-derivative we have

$$\frac{1}{\kappa k_B}\Big(\frac{\partial S}{\partial \beta}\Big)_{\{V\}} \\ = -\sum_\alpha \beta \Big\{\frac{\partial}{\partial\beta}\frac{1}{\beta} Z_\kappa^{-1}\{p_{\kappa\alpha}(\beta)\}\Big[p_{\kappa\alpha}^{1-\gamma}(\beta) - p_{\kappa\alpha}(\beta)\Big]\Big\}_{\{V\}} \tag{35}$$

which shows that the derivative of the micro-canonical entropy with respect to β respectively temperature T is the sum over all states of the particular entropies

$$\frac{1}{\kappa k_B}\Big(\frac{\partial S_\alpha}{\partial \beta}\Big)_{\{V\}} = -\beta \Big\{\frac{\partial}{\partial\beta}\frac{\big[p_{\kappa\alpha}^{1-\gamma}(\beta) - p_{\kappa\alpha}(\beta)\big]}{\beta Z_\kappa\{p_{\kappa\alpha}(\beta)\}}\Big\}_{\{V\}} \tag{36}$$

of the α states. This expression replaces Boltzmann's definition to become Olbert's micro-canonical entropy, and it follows that

$$\frac{S_\alpha\{p_{\kappa\alpha}(\beta)\}}{\kappa k_B} = -\frac{p_{\kappa\alpha}(\beta)}{Z_\kappa\{p_{\kappa\alpha}(\beta)\}}\big[p_{\kappa\alpha}^{-\gamma}(\beta) - 1\big] \\ + \int d\beta \frac{p_{\kappa\alpha}(\beta)}{Z_\kappa\{p_{\kappa\alpha}(\beta)\}}\frac{\big[p_{\kappa\alpha}^{-\gamma}(\beta) - 1\big]}{\beta} \tag{37}$$

The factor $p_{\kappa\alpha}/Z_\kappa = P_{\kappa\alpha}$ is the normalized Olbert-Gibbs distribution, and the first term becomes its product with a function

$$R_{\kappa\alpha} = 1 - p_{\kappa\alpha}^{-\gamma} \tag{38}$$

This function also appears under the integral sign, such that we can write the latter in an abbreviated version

$$\frac{S_\alpha\{p_{\kappa\alpha}(\beta)\}}{\kappa k_B} = P_{\kappa\alpha}R_{\kappa\alpha} - \int \frac{d\beta}{\beta} P_{\kappa\alpha}R_{\kappa\alpha} \tag{39}$$

This is the relation between the probabilities of states α and their corresponding entropies. The sum over all α states gives the total entropy

$$\frac{S_\kappa(\beta)}{\kappa k_B} = 1 - \log(\beta U_0) - \big\langle p_{\kappa\alpha}^{-\gamma}(\beta)\big\rangle + \int \frac{d\beta}{\beta}\big\langle p_{\kappa\alpha}^{-\gamma}(\beta)\big\rangle + G(\{V\}) \tag{40}$$

in terms of the average probability, i.e., the expectation value of the probability raised to the power $-\gamma$. Again, the angular brackets indicate the probability weighted average over all states α. U_0 is some normalizing thermal energy which to chose is arbitrary. The term containing it is of little importance.

This entropy is substantially more complicated than in ordinary classical statistical mechanics. Nevertheless, it exhibits the relation between entropy and probability. It distinguishes the Olbert-Lorentzian entropy from Boltzmann-Gibbs-Shannon.

4.2. Boltzmann-Gibbs Like Form of the Olbert Entropy

Some insight can be obtained when considering the functional R, writing it

$$R_{\kappa\alpha} = 1 - \exp\big(-\gamma \log p_{\kappa\alpha}\big) \tag{41}$$

Expanding the exponential yields to first order

$$R_{\kappa\alpha} = \gamma \log p_{\kappa\alpha} + \text{higher order terms} \tag{42}$$

Except for the factor γ this is just Boltzmann's micro-canonical entropy which, after multiplication with the probability and summation respectively integration yields the classical expression for the average entropy. From this equivalence, we conclude that, in the Olbert entropy $S_\alpha\{p_{\kappa\alpha(\beta)}\}$, the functional $R_{\kappa\alpha}$ plays exactly the role of Boltzmann's micro-canonical entropy. However, in Boltzmann theory, the logarithm of the probability is just the inverse of the Boltzmann factor of the energy of state α, with the energy E_α expressed in terms of the probability p_α. This is also exactly the meaning of the functional $R_{\kappa\alpha}\{p_{\kappa\alpha}\}$, which enables us to formulate the general

Theorem

Let $p_\alpha(\beta, E_\alpha) = f_\alpha(\beta, E_\alpha)$ be the properly defined probability of a micro-canonical state E_α, and $F\{p_\alpha\} = E_\alpha(p_\alpha)$ the inverse of f_α. Then, up to some numerical factors, the micro-canonical entropy S_α of the state α, expressed in terms of the probability p_α, is defined

[4] In other approaches this inverse appears as a mysterious "escort distribution," which plays the role of some integration condition when forming lowest order moments. In fact, it is nothing but an inverse function as was proposed already [26] and, in other choices of the probability distribution, would be obtained in the same way by inversion.

as $S_\alpha\{p_\alpha\} \propto F\{p_\alpha\}$; and the mean entropy $S \equiv \langle S_\alpha\rangle$ of the micro-canonical system, given by the sum over all probability-weighted states α, is obtained in the form

$$S \propto \sum_\alpha p_\alpha S_\alpha\{p_\alpha\} \propto \sum_\alpha p_\alpha F_\alpha\{p_\alpha\} \tag{43}$$

if only the inverse functional $F\{p_\alpha\}$ exists and can be given either analytically or numerically. This formula is the general prescription of calculating the entropy in the micro-canonical state.

Let us, for convenience, discuss just the leading first order terms in the above expression for the Olbert entropy, assuming for our purposes of understanding that the higher order terms do not substantially contribute, which in general might not always be true. It then follows from Equation (37) that

$$\begin{aligned}\frac{S_\alpha\{p_{\kappa\alpha(\beta)}\}}{\gamma\kappa k_B} &= P_{\kappa\alpha}(\beta)\big(\log P_{\kappa\alpha} + \log Z_\kappa\big)\\ &\quad - \int \frac{d\beta}{\beta} P_{\kappa\alpha}\big(\log P_{\kappa\alpha} + \log Z_\kappa\big) + \ldots\end{aligned} \tag{44}$$

Except for the difficulty with the integral term, the first terms look about familiar. However, interestingly, this holds for the unsummed entropy. Summation then leads to the average entropy

$$\frac{S_\kappa}{|\gamma|\kappa k_B} = \langle\log P_\kappa(\beta)\rangle + \langle\log Z_\kappa\rangle - \int\frac{d\beta}{\beta}\big[\langle\log P_\kappa(\beta)\rangle + \langle\log Z_\kappa\rangle\big] \tag{45}$$

It reproduces the logarithmic dependence on the mean logarithm of the partition function in the second term. The first term also reproduces the classical dependence on the logarithm of the mean probability P_κ. Further discussion is, however, less transparent, and the role of any higher order terms in the expansion as well as the structure of the integral term obscure its interpretation. In this form, however, we may conclude that the κ-generalization of classical statistical mechanics maintains its basic structure at least to lowest order. In any case, it becomes clear that the Olbert-Lorentzian generalization can be justified in its application to micro-canonical and, after proper extension to include the dependence on particle number, also to canonical systems. This is very satisfactory, as it gives Olbert-Lorentzian statistical mechanics and the resulting Olbert-κ distribution a physically justified place in the treatment of many-particle systems like high temperature plasmas. The different expressions for the entropies are then available for the proper description of the evolution of such states in thermal equilibrium as well as in non-equilibrium.

4.3. Quantum Considerations

In this subsection we, for completeness, though just briefly, touch on the quantum extensions of Lorentzian entropies. We argued above that there is no zero-temperature limit of the Lorentzian statistics. This holds generally. Fermi statistics in addition inhibits correlations in the sense that any states α could be occupied by more than one particle. Hence, correlations involved in κ can only be of the nature of entanglements, and, in addition to our finding that the state $T = 0$ is principally excluded, this additional restriction categorically excludes application to Fermi systems other than entanglement of two particles of opposing spins. Below we briefly consider this case.

What concerns Bose statistics, the latter restriction is relaxed. States can obey arbitrary occupation numbers. Hence, high energy states can exist. Then, one will be able to find an appropriate expression for the Bose entropy, which we will provide in a follow-up communication, as this requires another lengthy derivation which goes beyond the present note.

We just briefly mention another interesting quantum case resulting in Fermi systems, the entanglement, or von Neumann entropy [27]. It is defined as

$$S_{vN} = \mathrm{Trace}\big(\rho\log\rho\big) \tag{46}$$

where $\rho = \sum_\alpha |\psi_\alpha\rangle\langle\psi_\alpha|$ is the average scattering matrix in a quantum system, and Trace is its trace. Clearly, if all ψ_α are true eigenstates of the entire system, $\rho = 0$, then the system is in its own eigenstate, and no entropy is produced. Otherwise, the entropy results from superposing all eigenstates $|\alpha\rangle$ of its components, yielding $\rho = \sum_\alpha \eta_\alpha|\alpha\rangle\langle\alpha|$, which contains all the irreversible interactions encoded in the superpositions of eigenstates of the components which contribute to the common wave function of the entire many particle system. Intuitively this is clear because all the different phases of the components will mix; the common wave function, being the superposition of all individual or grouped particle wave functions, will by no means become an eigenstate of the system. This is very frequently misunderstood when talking about fluid models of quantum theory and identifying the density with the expectation value of the wave function. In a quantum mechanical Olbert κ system, where the particles are correlated and by some interaction mechanism are bunched together one may even expect that the scattering matrix contains non-diagonal terms indicating dissipation. One such mechanism is entanglement between two prepared Fermions of opposite spin. By it, two particles (electrons in the same state but of different spin) are bound together in their common behavior. They are subject to von Neumann's entropy. If the entanglement can be encoded into a parameter κ, then its entropy may be conjectured to become

$$S_{vN\kappa} \sim \mathrm{Trace}\big(\rho_\kappa\log\rho_\kappa^{1-\gamma}\big) = (1-\gamma)\mathrm{Trace}\big(\rho_\kappa\log\rho_\kappa\big) \tag{47}$$

and one has $\rho_\kappa = \sum_\alpha |\psi_{\kappa\alpha}\rangle\langle\psi_{\kappa\alpha}| = \sum_\alpha \eta_{\kappa\alpha}|\alpha_\kappa\rangle\langle\alpha_\kappa|$. The κ wave function might, however, not be known a priori. Since entanglement applies to electrons, or in general Fermions, which by our above reasoning are not subject to Lorentzian statistics, then, in κ-statistics, it would apply to the bosonic property of paired electrons of opposite spin and must thus somehow, though not in an elucidated manner, relate to Boson-Lorentzian entropy of collectively grouped pairs like in superconductivity. If true, the parameter κ appearing in the von Neumann entropy then contains the physics of group entanglement. Otherwise, κ statistics do not apply in any manner to any entanglement, and no von Neumann-Lorentzian entropy exists.

5. CONCLUSIONS

In the present paper, we have undertaken the task of trying to understand what physically would be behind Olbert-Lorentzian statistics. As the Olbert-κ distribution function that belongs to it is well-confirmed from a large number of observations mainly in space plasmas, this effort is needed to give a clue on its foundations. Applying statistical mechanical reasoning we have obtained expressions for the entropy as a functional of energy and also as functional of probability of states. What is most interesting in such an approach is that the Olbert entropy S_κ has an equivalent form to that in ordinary non-equilibrium statistical mechanics. Olbert entropy, however, contains additional terms which can be calculated in an iterative perturbation theoretical way. It is for this reason super-additive (or super-extensive if wanted), a property that it has in common with q-statistics though being rather different. We have elucidated the main difference here. This means that in κ-systems, i.e., for instance, in high-temperature plasmas exhibiting Olbert-distributions, the particles are correlatively grouped together to behave collectively, thereby providing the collective contribution to entropy. Such correlations are implicit to the index κ and indicate strong non-linear couplings, which are provided by interaction potentials which are mediated not by collisions but by excitation of waves. It is thus not surprising if κ-distributions are found in turbulent dilute high temperature plasmas like the solar wind [28], near collisionless shock waves [29], Earth's bow shock [30], the magnetosheath [31], at the boundaries of the heliosphere and astrospheres [32], where various types of waves can be excited as both, eigenmodes or sidebands, which even occupy the evanescent branches of the dielectric response function causing a continuous almost featureless power spectrum of fluctuations, which is typical for well-developed turbulence. One may, therefore, expect that the statistical mechanics underlying well-developed collisionless turbulence will become kind of Olbert-Lorentzian in terms of the probability distribution. The precise relation between these interactions and the particular value of the parameter κ is still open to investigation. The consideration of entropy given here only shows its micro-canonical statistical-mechanical effect.

AUTHOR CONTRIBUTIONS

Both authors contributed equally to this work, and approved it for publication.

ACKNOWLEDGMENTS

We acknowledge valuable discussions with J. Geiss, C. H. Jaroschek, K.-H. Glassmeier, M. Hoshino, M. P. Leubner, R. Nakamura, Y. Narita, Z. Vörös, P. H. Yoon, and L. Zelenyi.

REFERENCES

1. Kimontovich YL. *The Statistical Theory of Non-equilibrium Processes in a Plasma*. Cambridge, MA: The M.I.T. Press (1967).
2. Kittel C, Kroemer H. *Thermal Physics*. New York, NY: W.H.Freeman Comp. (1980).
3. Christon SP, Mitchell DG, Williams DJ, Frank LA, Huang CY, Eastman TE. Energy spectra of plasma sheet ions and electrons from ∼ 50 eV/e to ∼ 1 MeV during plasma temperature transitions. *J Geophys Res.* (1988) **93**:2562–72. doi: 10.1029/JA093iA04p02562
4. Christon SP, Williams DJ, Mitchell DG, Frank LA, Huang CY. Spectral characteristicds of plasma sheet ion and electron populations during undisturbed geomagnetic conditions. *J Geophys Res.* (1989) **94**:13409–24. doi: 10.1029/JA094iA19p13409
5. Christon SP, Williams DJ, Mitchell DG, Huang CY, Frank LA. Spectral characteristicds of plasma sheet ion and electron populations during disturbed geomagnetic conditions. *J Geophys Res.* (1991) **96**:1–22. doi: 10.1029/90JA01633
6. Vasyliunas VM. A survey of low-energy electrons in the evening sector of the magnetosphere with OGO 1 and OGO 3. *J Geophys Res.* (1968) **73**:2839–84. doi: 10.1029/JA073i009p02839
7. Olbert S. Summary of experimental results from M.I.T. detector on IMP-1. In:.Carovillano DL, McClay JF, editors. *Physics of the Magnetosphere, Proceeding of a Conference at Boston College, Astrophysics and Space Science Library* 40. Dordrecht: Reidel Publ. (1968). p. 641.
8. Livadiotis G, McComas DJ. Understanding kappa distributions: a toolbox for space science and astrophysics. *Space Sci Rev.* (2013) **175**:183–214 doi: 10.1007/s11214-013-9982-9
9. Yoon PH, Rhee T, Ryu C-M. Self-consistent generation of superthermal electrons by beam-plasma interaction. *Phys Rev Lett.* (2005) **95**:215003. doi: 10.1103/PhysRevLett.95.215003
10. Yoon PH, Rhee T, Ryu C-M. Self-consistent generation of electron κ distribution: 1. Theory. *J Geophys Res.* (2006) **111**:A09106. doi: 10.1029/2006JA011681
11. Hasegawa A, Mima K, Duong-van M. Plasma distribution function in a superthermal radiation field. *Phys Rev Lett.* (1985) 54:2608–10. doi: 10.1103/PhyRevLett.54.2608
12. Treumann RA, Jaroschek CH, Scholer M. Stationary plasma states far from equilibrium. *Phys Plasmas*. (2004) **11**:1317–25. doi: 10.1063/1.1667498
13. Scherer K, Fichtner H, Lazar M. Regularized κ-distributions with non-diverging moments. *Europhy Lett.* (2017) **120**:50002 doi: 10.1209/0295-5075/120/50002
14. Lazar M, Scherer K, Fichtner H, Pierrard V. Toward a realistic macroscopic parametrization of space plasmas with regularized κ-distribution. *Astron Astrophys.* (2020) **643**:A20. doi: 10.1051/0004-6361/201936861
15. Treumann RA, Baumjohann W. The differential cosmic ray energy flux in the light of an ultrarelativistic generalized Lorentzian thermodnamics. *Astrophys Space Sci.* (2018) **363**:37. doi: 10.1007/s10509-018-3255-8
16. Wehrl A. General properties of entropy. *Rev Mod Phys.* (1978) **50**:221–60. doi: 10.1103/RevModPhys.50.221
17. Balatoni J, Renyi A. Publications of the Mathematical Institute of the Hungarian Academy of Sciences 1:9 (1956).
18. Renyi, A. *Probability Theory*. North-Holland Publishing Company (1970).
19. Tallis C. Possible generalization of Boltzmann-Gibbs statistics. *J Stat Phys.* (1988) **52**:479–87. doi: 10.1007/BF01016429
20. Treumann RA. Kinetic theoretical foundation of Lorentzian statistical mechanics. *Phys Script.* (1999) **59**:19–26. doi: 10.1238/Physica.Regular.059a00019
21. Domenech-Garret JL, Tierno SP, Conde L. Non-equilibrium thermionic electron emission for metals at high temperatures. *J Appl Phys.* (2015) 118:074904. doi: 10.1063/1.49929150
22. Nauenberg M. Critique of q-entropy for thermal statistics. *Phys Rev E.* (2003) **67**:036114. doi: 10.1103/PhysRevE.67.036114
23. Landau LD, Lifshitz EM. *Statistical Physics, Part 1*. Oxford: Perhgamon Press (1980).
24. Treumann RA, Jaroschek CH. Gibbsian theory of power law distributions. *Phys Rev Lett.* (2008) **100**:155005. doi: 10.1103/PhysRevLett.100.155005

25. Einstein, A. Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Ann Phys*. (1905) **322**:549–60. doi: 10.1002/andp.19053220806
26. Treumann RA, Baumjohann W. Beyond Gibbs-Boltzmann-Shannon: general entropies – The Gibbs-Lorentzian example. *Front Phys*. (2014) **2**:49. doi: 10.3389/fphy.2014.00049
27. von Neumann J. *Mathematical Foundations of Quantum Mechanics*. Princeton, NJ: Princeton University Press (1955).
28. Goldstein ML, Eastwood JP, Treumann RA, Lucek EA, Pickett J, Décréau P. The near-earth solar wind. *Space Sci Rev*. (2005) **118**:7–39. doi: 10.1007/s11214-005-3823-4
29. Balogh A, Treumann RA. *Physics of Collisionless Shocks*. New York, NY: Springer Media (2013). doi: 10.1007/978-1-4614-6099-2
30. Eastwood JP, Lucek EA, Mazelle C, Meziane K, Narita Y, Pickett J, et al. The foreshock. *Space Sci Rev*. (2005) **118**:41–94. doi: 10.1007/s11214-005-3824-3
31. Lucek EA, Constantinescu D, Goldstein ML, Pickett J, Pinccon JL, Sahraoui F, Treumann RA, et al. The magnetosheath. *Space Sci Rev*. (2005) **118**:95–112. doi: 10.1007/s11214-005-3825-2
32. Scherer K, Baalmann LR, Fichtner H, Kleinmann J, Bomans, DJ, Weis K, et al. MHD-shock structures of astrospheres: λ Cephei-like astrospheres. *Mon Not R Astron Soc*. (2020) **493**:4172–85. doi: 10.1093/mnras/staa497

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

One-Way Pedestrian Traffic Is a Means of Reducing Personal Encounters in Epidemics

*Bernardo A. Mello**

Department of Physics Institute, University of Brasilia, Brasilia, Brazil

Minimizing social contact is an important tool to reduce the spread of diseases but harms people's well-being. This and other more compelling reasons urge people to walk outside periodically. The present work explores how organizing the traffic of pedestrians affects the number of walking or running people passing by each other. By applying certain rules, this number can be significantly reduced, potentially reducing the contribution of person-to-person contagion to the basic reproductive number, R_0. One example is the traffic of pedestrians on sidewalks. Another is the use of walking or running tracks in parks. It is obtained here that the number of people encountering each other can be drastically reduced if one-way traffic is enforced and runners are separated from walkers.

Keywords: epidemics, pedestrians, jogging, urban mobility, basic reproductive number

OPEN ACCESS

Edited by:
Víctor M. Eguíluz,
Institute of Interdisciplinary Physics
and Complex Systems (IFISC), Spain

Reviewed by:
Nicolas Francisco Lori,
University of Minho, Portugal
George Yannis,
National Technical University of
Athens, Greece

***Correspondence:**
Bernardo A. Mello
bernardomello@unb.br

Specialty section:
This article was submitted to
Interdisciplinary Physics,
a section of the journal
Frontiers in Physics

Received: 11 May 2020
Accepted: 03 August 2020
Published: 10 September 2020

Citation:
Mello BA (2020) One-Way Pedestrian
Traffic Is a Means of Reducing
Personal Encounters in Epidemics.
Front. Phys. 8:376.
doi: 10.3389/fphy.2020.00376

1. INTRODUCTION

Contagious epidemics, as the Covid-19 pandemic, often demand limiting physical interactions among people in order to reduce the contagious rate. Governmental measures to reduce physical contact range from the closing of public facilities and schools to restrictions on mobility, lockdowns, quarantines, and curfews. These extreme measures, though necessary, should be used as last resources, due to their economic and personal negative impacts. Of great help in these situations are the physical and psychological benefits of physical exercises, walking included [1, 2]. On the other hand, physical contact and proximity should be avoided to reduce the spread of pathogens such as SARS-CoV-2 [3].

Measures that reduce the physical interaction with minimal disruption in the daily activities, as the ones proposed here, must be adopted whenever possible. For example, if sidewalks and crosswalks at intersections are made one-way, with walking only allowed on the right-hand ones (the street must be at pedestrians' left), the major inconvenience would be one more block of walking for pedestrians when leaving their starting points or reaching their destinations.

Pedestrian behavior depends on internal and external aspects such as urban environment, contingent individual situation, and crowd behavior [4–6], which must all be considered when planning traffic interventions. Sophisticated methods have been proposed to study pedestrian motion [7–9] and used, for example, to study the efficiency of pedestrian mobility [10]. Mostly, the studies on urban mobility have been concentrated on efficiency, well-being, safety, and other relevant aspects of daily life [11–13].

The present work addresses a completely different goal, which is only justifiable in abnormal circumstances, such as epidemics: minimizing the encounters among pedestrians. The conceptual problem is much simpler, since the interpersonal effect on mobility is not relevant due to the low density of people, justifying the use of a more "pedestrian" mathematical model.

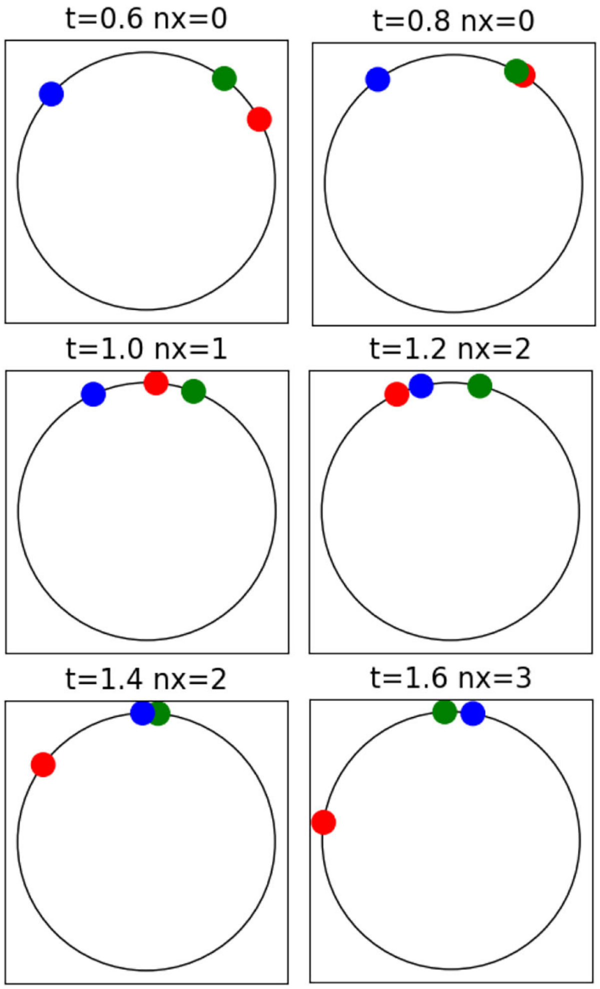

FIGURE 1 | Example of encounters between walkers on a circular track. The blue walker moves in the clockwise direction and the green and the red walkers move in the counter-clockwise direction, the red walker moves faster than the green. The time and the number of encounters up to that instant are shown above each figure.

2. METHODS

2.1. Modeling the Encounters

Figure 1 illustrates the movement of three walkers on a circular track. The first encounter occurs between $t = 0.8$ and 1.0, involving two walkers moving in the same direction, the faster red overtaking the slower green. The second encounter occurs between $t = 1.0$ and 1.2, between the blue and the red walkers that move in opposite directions. Video 1, available in the **Supplementary Material**, features the movement of these walkers.

The encounter of walkers moving in the same direction is due to different speeds, and its frequency is reduced if the walkers walk at a similar rate, regardless of fast or slow. On the other hand, the encounter frequency of walkers moving in opposite directions is proportional to the average of their absolute speeds, regardless of the difference in their absolute values.

To represent the movement of a crowd, the initial position is supposed to be random and uniformly distributed over the track. Measures of walking speed in several conditions presented in [14] were used to adopt the mean value of 1.4 m/s with the standard deviation of 0.25 m/s for the walking speed. From [15], the running speed is assumed to be twice these values, i.e., an average of 2.8 m/s and a standard deviation of 0.5 m/s. Random and constant speeds with normal distribution are assigned to the walkers and runners, with the corresponding average and standard deviation. When minimum or maximum speed were imposed, the speed of individuals under or above these limits was redefined as equal to the boundary values.

Two measures are proposed to evaluate the number of encounters per person. One is the number of encounters per minute, that is suitable to evaluate the encounters of a person who goes out for a given amount of time, for example, for jogging. The other is the number of encounters when a person walks along 100 m, typically, the length of one block. This is suitable for the analysis of a person who goes out to reach a certain place.

2.2. The Frequency of Encounters

Consider two people with constant speeds v and v' moving around a closed path of length L. If they stay on the track long enough, they will cross each other with a common temporal frequency

$$f_t(v, v') = f_t(v', v) = \frac{|v - v'|}{L}. \tag{1}$$

If N people with constant speeds distribution $p(v)$ share the same track, the average frequency of encounters of a person with speed v is

$$\langle f_t(v) \rangle_{v'} = \frac{N-1}{L} \int_{-\infty}^{\infty} |v - v'| p(v')\, dv', \tag{2}$$

and the population average is

$$\langle f_t \rangle_{vv'} = \int_{-\infty}^{\infty} \langle f_t(v) \rangle_{v'} p(v)\, dv. \tag{3}$$

Another relevant quantity is the spatial frequency, the number of encounters by unit of length of the distance traveled by the particle with speed v, given by

$$f_s(v, v') = \frac{f_t(v, v')}{|v|}. \tag{4}$$

From this we can define the person and the population averages,

$$\langle f_s(v) \rangle_{v'} = \frac{\langle f_t(v) \rangle_{v'}}{|v|}, \tag{5}$$

$$\langle f_s \rangle_{vv'} = \int_{-\infty}^{\infty} \langle f_s(v) \rangle_{v'} p(v)\, dv. \tag{6}$$

2.3. Probability of Diverse Encounters

The number of encounters may not be the best quantity to evaluate the probability of contagion if two people encounter more than once when traveling along a short track. Encountering the same person twice is not the same as encountering two people, one time each. Despite the same number of encounters, in the latter case the probability of meeting a contagious person is higher.

In the situation analyzed here, all people spend the same time τ on the track, entering and leaving it at the same rate, N/τ, which leads to the mean occupancy N. The number of people who visit the track at a time window $[t_0, t_0 + \tau]$ is $2N$, consisting of the people who arrive in the time interval $[t_0 - \tau, t_0 + \tau]$. Of the time that each of those people stay on the track, a time τ' is spent inside the time window $[t_0, t_0 + \tau]$, with the uniform probability density τ^{-1} in the interval $0 < \tau' < \tau$. Accordingly, a person on the track will share it with $2(N - 1)$ other people during the total time of his/her visit, and the time shared with those people will be in the interval $0 < \tau' < \tau$ with probability density τ^{-1}.

If two people with speeds v and v' share the same track for a time τ', the probability that they will meet each other at least once is

$$\pi(\tau', v, v') = \begin{cases} 1 & \text{if } L/\tau' \leq |v - v'|, \\ |v - v'|\,\tau'/L & \text{if } |v - v'| < L/\tau'. \end{cases} \tag{7}$$

For any two people on the track, the probability of meeting at least once is

$$\langle \pi(v, v') \rangle_{\tau'} = \frac{1}{\tau} \int_0^\tau \pi(\tau', v, v')\, d\tau', \tag{8}$$

$$= \begin{cases} 1 - \dfrac{L/\tau}{2|v - v'|} & \text{if } L/\tau \leq |v - v'|, \\ \dfrac{|v - v'|}{2}\,\tau/L & \text{if } |v - v'| < L/\tau. \end{cases} \tag{9}$$

From the above expression, the diversity coefficient of encounters is defined as the fraction of non-repeated encounters,

$$\phi(v, v') = \frac{2\langle \pi(v, v') \rangle_{\tau'}/\tau}{f_t(v, v')} \tag{10}$$

$$= \begin{cases} 2\dfrac{L/\tau}{|v - v'|} - \dfrac{L^2/\tau^2}{|v - v'|^2} & \text{if } L/\tau \leq |v - v'|, \\ 1 & \text{if } |v - v'| < L/\tau. \end{cases} \tag{11}$$

The factor 2 in Equation (10) comes from the above discussion about the number of different people that enter or leave the track in the course of the time τ. For the population, the fraction of non-repeated encounters is

$$\langle \phi \rangle_{vv'} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi(v, v') p(v) p(v')\, dv dv'. \tag{12}$$

2.4. Monte Carlo Simulations

Complementary to the analytical approach presented in section 2.2, Monte Carlo simulation was used to describe a closed track with N individuals, and a random generator with a uniform distribution over the track length was used to define their position. The time evolution was performed by Euler integration with constant Δt. From the number of encounters of each individual, $\times_i$, the mean number of encounters per minute and per 100 m were, respectively, calculated as

$$\times_{\text{minute}} = \frac{\langle \times_i \rangle_{\text{pop}}}{T_{\text{sim}}} \tag{13}$$

and

$$\times_{100\text{ m}} = \left\langle \frac{\times_i}{L_i} \right\rangle_{\text{pop}}, \tag{14}$$

where T_{sim} is the total simulation time in minutes, and L_i is the distance traveled by the individual i, in hectometers. The expected values (ensemble average) of these quantities are equal to the frequencies Equations (3) and (6),

$$\langle \times_{\text{minute}} \rangle_{\text{ens}} \equiv \langle f_t \rangle_{vv'}, \qquad \langle \times_{100\text{m}} \rangle_{\text{ens}} \equiv \langle f_s \rangle_{vv'}, \tag{15}$$

with the proper units of time and distance.

3. RESULTS

As a concrete case of the analysis presented here, one can consider a 5,000 m track shared by 500 people. The results are the average number of encounters per person at each minute or each block (100 m), calculated by numerical integration of the expressions from section 2.2. These two quantities are proportional to the average density of people on the track, in this case, 1 person at every 10 m, and do not depend on the simulation time or the track length and shape. Proportionality may be used to determine these quantities for other densities.

We start with equal numbers of people moving in both directions of a track, 80% of them walking and 20% running. These percentages were arbitrarily chosen but the results are qualitatively equivalent for other values. The distribution of their speeds is shown in **Figure 2A**, together with the number of encounters per minute and per 100 m. It can be seen that the fast runners run by several people per minute, but have a minimal number of encounters along 100 m since they cover that distance quickly. A slow walker, on the other hand, exhibits the opposite behavior, for reverse reasons.

To reduce the number of encounters observed in **Figure 2A**, we enforce unidirectional movement, illustrated by **Figure 2B**. To evaluate the effect on the crowd, we refer to the average values, Equations (3) and (6), shown in **Table 1**. By comparing the bidirectional and the unidirectional columns of line 3 in **Table 1**, it can be seen that the number of encounters per minute and per 100 m are both reduced by 68%, to around one-third of their values in bidirectional traffic.

The encounters between people moving unidirectionally are due to their heterogeneous speed, which can be made more homogeneous by separating runners and walkers. Encounters of walkers are shown in **Figure 2C**. By comparing the columns of unidirectional encounters of lines 3 and 5 of **Table 1**, we find a contraction to 43% and 54%, respectively, of the number of encounters per minute and per 100 m.

Line 1 of **Table 1** shows the average number of encounters for a track where only runners are allowed. As predicted by

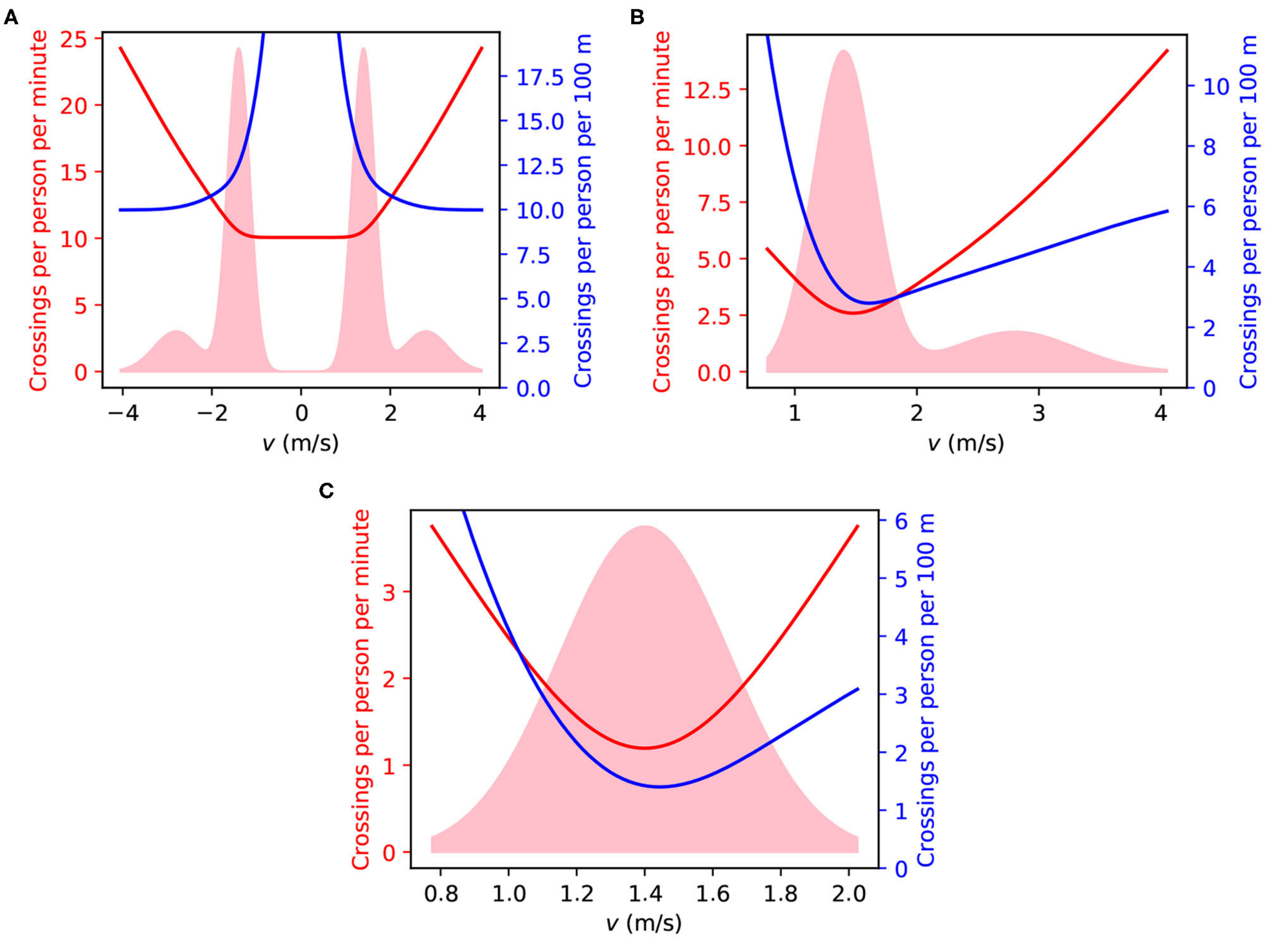

FIGURE 2 | (A) Analysis of a track with 100 people per kilometer, 80% of them walking and 20% of them running, in both directions. The speed distribution is shown as the pink area in arbitrary units. The red curve is the average number of encounters per minute for people moving at a certain speed, given by Equation (2). The blue curve is the average number of encounters per 100 m for people moving at a certain speed, given by Equation (5). **(B)** Same as **(A)**, with one-way enforced. **(C)** Same as **(A)** with one-way enforced and running forbidden. For illustration, Videos 2–4, representing 50 people in a 5,000 m track under the same rules can be found in the **Supplementary Material**.

Equations (3) and (6), comparison of the runners on line 1 with the walkers on line 5 of **Table 1** shows that the number of encounters per 100 m is the same for walkers and runners, but the number of encounters per minute is twice bigger for runners. Therefore, the density on a track of runners must be half of that of walkers to produce the same number of encounters per minute, but equal densities produce identical numbers of encounters per 100 m in tracks exclusive for runners or walkers.

Further reduction in the encounter rates can be achieved by imposing minimum and maximum walking speeds, for example, to no more and no less than one standard deviation of the average speed. Reductions of 30% are obtained, as can be seen by comparing the unidirectional encounters of lines 5 and 8 in **Table 1**. The modest improvement and the practical difficulties of imposing such measures suggest that they are not workable.

The fraction of non-repeated encounters between two people is measured by the diversity coefficient, Equation (11), plotted in **Figure 3** for a runner and a walker. If the track is large enough to prevent multiple encounters, $|v - v'| < L/\tau$, the diversity coefficient is 1, indicating that every encounter of a person involves different people. For shorter tracks, each pair of people may meet each other several times, and the coefficient of diversity is reduced.

Figure 3 shows that the diversity coefficient is bigger for a crowd equally divided between runners and walkers than for one runner and one walker, because the presence of people with similar speed diminishes the chances of repeated encounters. As a general rule, the analysis of **Figure 3** demonstrates that more uniform speed distributions lead to higher diversity coefficients. However, if the track is long enough, the number of repeated encounters vanishes and the diversity coefficient reaches its maximum value, $\langle\phi_{vv'}\rangle = 1$.

4. DISCUSSION

Transmission of diseases involves a multitude of aspects and this work addresses just one of them, the number of people met. This information is important to evaluate the probability

TABLE 1 | The average number of encounters per minute and per 100 m for a track with 100 people per kilometer subjected to different conditions, as defined by Equations (3) and (6).

	$v_{\min}$	$v_{\max}$	Fraction of runners	$\langle f_t \rangle_{vv'}$ (min^{-1}) Bidir.	$\langle f_t \rangle_{vv'}$ (min^{-1}) Unidir.	$\langle f_s \rangle_{vv'}$ (hm^{-1}) Bidir.	$\langle f_s \rangle_{vv'}$ (hm^{-1}) Unidir.
1	–	–	100%	18.5	3.38	11.2	2.12
2	–	–	50%	15.3	5.46	13.3	4.94
3	–	–	20%	12.0	3.90	12.5	3.93
4	–	–	10%	10.7	2.91	12.0	3.14
5	–	–	0%	9.23	1.69	11.2	2.12
6	–	1.65	0%	8.99	1.46	11.1	1.89
7	1.15	–	0%	9.24	1.46	11.0	1.72
8	1.15	1.65	0%	9.00	1.23	10.8	1.50

The v_{min} and v_{max} are the minimum and the maximum acceptable speeds for people on the track. The third column is the fraction of runners on the crowd.

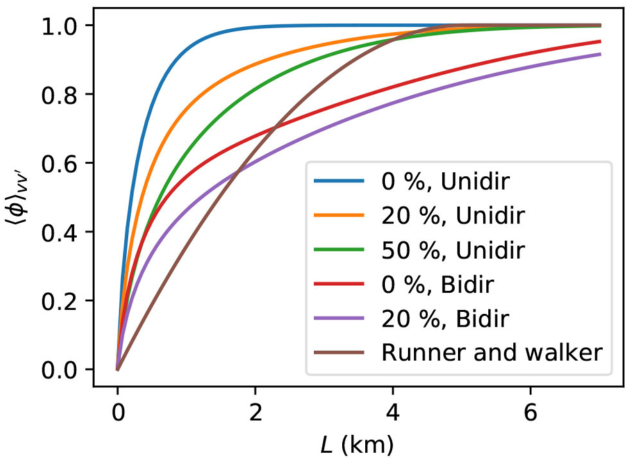

FIGURE 3 | Coefficient of diversity of the encounters, defined by Equation (12), with the fraction of runners and the direction indicated in the legend. The curve for one runner and one walker, with velocities respectively equal to 2.8 and 1.4 m/s in the same direction, uses Equation (11).

of encounters with infected people, of physical contacts, and of meeting and interacting with acquaintances.

Recent results on aerodynamics indicate that a trail of potentially contagious droplets is left behind walking and running people [16]. The probability of contagion depends on the level of exposure to the pathogen [17], which is proportional to the density of pathogens on that cloud and to the time spent inside the trail of droplets. This aspect is not covered in this paper and will be addressed in a subsequent work.

Regarding the organization of pedestrian traffic, if one-way movement and walking-only rules are imposed on bidirectional tracks shared by walkers (80%) and runners (20%), the number of people encountering each other per minute is reduced to one-seventh of its original value and the number of encounters per 100 m is reduced to one-sixth of its original value. If one-way movement is imposed on a walking-only walkway, or sidewalks, for example, the number of encounters is reduced to one-fifth of its original value. The improvements are also significant for running-only tracks. Therefore, establishing one-way walkways and separating runners from walkers are effective measures to reduce the physical encounter of people during contagious epidemics.

To avoid complicating the discussion, this paper focuses on closed paths. Open paths, such as trajectories of pedestrians reaching a destination in cities, or of joggers in roads, may be considered very large closed paths partially traveled by each individual. The frequency of encounters discussed here depends only on the density and the conclusion regarding these quantities hold also for open tracks. For open or very large tracks, the chance of multiple encounters is zero, the diversity coefficient is one, and only the frequency of encounters matters.

From the biological point of view, contagion is also related to the variability of the host [18] and pathogen [19], which depends on the number of people met. On short tracks, measures that reduce the number of crossings also increase the diversity coefficient of the encounters. The product of these two quantities provides the number of different people met. On the one hand, higher diversity increases the chance of meeting someone infected. On the other hand, repeated encounters with the same person increase the pathogen doses received from infected people. This paper provides the tools to quantify these two effects, but proper decisions about the suitability of each measure can only be made considering the dynamics of each disease.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

BM proposed the work, developed the model, wrote the code, ran the simulations, did the analysis, and wrote the paper.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphy.2020.00376/full#supplementary-material

Supplementary Video 1 | Simulation of two walkers (green and blue) and a runner (red) in a 5,000 m long circular track for one hour. The simulation time and the total number of encounters are shown.

Supplementary Video 2 | Simulation of 50 walkers and runners in a 5,000 m long circular track for one hour, moving in both directions. The simulation time and the total number of encounters are shown.

Supplementary Video 3 | Simulation of 50 walkers and runners in a 5,000 m long circular track for one hour, moving only in the counter-clockwise direction. The simulation time and the total number of encounters are shown.

Supplementary Video 4 | Simulation of 50 walkers in a 5,000 m long circular track for one hour, moving only in the counter-clockwise direction. The simulation time and the total number of encounters are shown.

REFERENCES

1. Lee IM, Buchner DM. The importance of walking to public health. *Med Sci Sports Exerc*. (2008) **40**:S512–8. doi: 10.1249/MSS.0b013e31817c65d0
2. Roe J, Aspinall P. The restorative benefits of walking in urban and rural settings in adults with good and poor mental health. *Health Place*. (2011) **17**:103–13. doi: 10.1016/j.healthplace.2010.09.003
3. Sohrabi C, Alsafi Z, O'Neill N, Khan M, Kerwan A, Al-Jabir A, et al. World Health Organization declares global emergency: a review of the 2019 novel coronavirus (COVID-19). *Int J Surg*. (2020) **76**:71–76. doi: 10.1016/j.ijsu.2020.02.034
4. Zacharias J. Pedestrian behavior pedestrian behavior and perception in urban walking environments. *J Plann Literat*. (2001) **16**:3–18. doi: 10.1177/08854120122093249
5. Seitz MJ, Bode NW, Köster G. How cognitive heuristics can explain social interactions in spatial movement. *J R Soc Interface*. (2016) **13**:20160439. doi: 10.1098/rsif.2016.0439
6. von Sivers I, Templeton A, Köster G, Drury J, Philippides A. Humans do not always act selfishly: social identity and helping in emergency evacuation simulation. *Transport Res Proc*. (2014) **2**:585–93. doi: 10.1016/j.trpro.2014.09.099
7. Von Sivers I, Köster G. Dynamic stride length adaptation according to utility and personal space. *Transport Res B Methodol*. (2015) **74**:104–17. doi: 10.1016/j.trb.2015.01.009
8. Seitz MJ, Köster G. Natural discretization of pedestrian movement in continuous space. *Phys Rev E*. (2012) **86**:046108. doi: 10.1103/PhysRevE.86.046108
9. Feliciani C, Nishinari K. An improved Cellular Automata model to simulate the behavior of high density crowd and validation by experimental data. *Phys A Stat Mech Appl*. (2016) **451**:135–48. doi: 10.1016/j.physa.2016.01.057
10. Davidich M, Köster G. Predicting pedestrian flow: a methodology and a proof of concept based on real-life data. *PLoS ONE*. (2013) **8**:e83355. doi: 10.1371/journal.pone.0083355
11. Goodwin P. Transformation of transport policy in Great Britain. *Transport Res A Policy Pract*. (1999) **33**:655–69. doi: 10.1016/S0965-8564(99)00011-7
12. Middleton J. Sense and the city: exploring the embodied geographies of urban walking. *Soc Cult Geogr*. (2010) **11**:575–96. doi: 10.1080/14649365.2010.497913
13. St-Louis E, Manaugh K, van Lierop D, El-Geneidy A. The happy commuter: a comparison of commuter satisfaction across modes. *Transport Res F Traffic Psychol Behav*. (2014) **26**:160–70. doi: 10.1016/j.trf.2014.07.004
14. Chandra S, Bharti AK. Speed distribution curves for pedestrians during walking and crossing. *Proc Soc Behav Sci*. (2013) **104**:660–7. doi: 10.1016/j.sbspro.2013.11.160
15. Smyth B. Fast starters and slow finishers: a large-scale data analysis of pacing at the beginning and end of the marathon for recreational runners. *J Sports Anal*. (2018) **4**:229–42. doi: 10.3233/JSA-170205
16. Blocken B, Malizia F, van Druenen T, Marchal T. Towards aerodynamically equivalent COVID19 1.5 m social distancing for walking and running. *Preprint*. (2020). Available online at: https://api.semanticscholar.org/CorpusID:215752493
17. Hall CB, Douglas R, Schnabel KC, Geiman JM. Infectivity of respiratory syncytial virus by various routes of inoculation. *Infect Immun*. (1981) **33**:779–83. doi: 10.1128/IAI.33.3.779-783.1981
18. Ajelli M, Moise IK, Hutchings TCS, Brown SC, Kumar N, Johnson NF, et al. Host outdoor exposure variability affects the transmission and spread of Zika virus: Insights for epidemic control. *PLoS Negl Trop Dis*. (2017) **11**:e0005851. doi: 10.1371/journal.pntd.0005851
19. Sankalé JL, de la Tour RSA, Renjifo B, Siby T, Mboup S, Marlink RGG, et al. Intrapatient variability of the human immunodeficiency virus type 2 envelope V3 loop. *AIDS Res Hum retroviruses*. (1995) **11**:617–23. doi: 10.1089/aid.1995.11.617

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 06 November 2020
doi: 10.3389/fphy.2020.00265

Floating-Point Calculations on a Quantum Annealer: Division and Matrix Inversion

Michael L. Rogers[1*†] *and Robert L. Singleton Jr.*[2*†]

[1] SavantX, Santa Fe, NM, United States, [2] School of Mathematics, University of Leeds, Leeds, United Kingdom

OPEN ACCESS

Edited by:
Marcelo Silva Sarandy,
Fluminense Federal University, Brazil

Reviewed by:
Marcos César de Oliveira,
Campinas State University, Brazil
Alexandre M. Souza,
Centro Brasileiro de Pesquisas
Físicas, Brazil

***Correspondence:**
Robert L. Singleton Jr.
r.singleton@leeds.ac.uk
Michael L. Rogers
michael.rogers@savantx.com

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to
Quantum Computing,
a section of the journal
Frontiers in Physics

Received: 30 January 2019
Accepted: 12 June 2020
Published: 06 November 2020

Citation:
Rogers ML and Singleton RL Jr (2020)
Floating-Point Calculations on a
Quantum Annealer: Division and
Matrix Inversion. Front. Phys. 8:265.
doi: 10.3389/fphy.2020.00265

Systems of linear equations are employed almost universally across a wide range of disciplines, from physics and engineering to biology, chemistry, and statistics. Traditional solution methods such as Gaussian elimination are very time consuming for large matrices, and more efficient computational methods are desired. In the twilight of Moore's Law, quantum computing is perhaps the most direct path out of the darkness. There are two complementary paradigms for quantum computing, namely, circuit-based systems and quantum annealers. In this paper, we express floating point operations, such as division and matrix inversion, in terms of a *quadratic unconstrained binary optimization* (QUBO) problem, a formulation that is ideal for a quantum annealer. We first address floating point division, and then move on to matrix inversion. We provide a general algorithm for any number of dimensions, and, as a proof-of-principle, we demonstrates results from the D-Wave quantum annealer for 2 × 2 and 3 × 3 general matrices. In principle, our algorithm scales to very large numbers of linear equations; however, in practice the number is limited by the connectivity and dynamic range of the machine.

Keywords: quantum computing, matrix inversion, quantum annealing algorithm, linear algebra algorithms, D-wave

1. INTRODUCTION

Systems of linear equations are employed almost universally across a wide range of disciplines, from physics and engineering to biology, chemistry, and statistics. An interesting physics application is computational fluid dynamics (CFD), which requires inverting very large matrices to advance the state of the hydrodynamic system from one time step to the next. An application of importance in biology and chemistry would include the protein folding problem. For large matrices, Gaussian elimination and other standard techniques becomes too time consuming, and faster computational methods are therefore desired. As Moore's Law draws to a close, quantum computing offers the most direct path forward; it is also perhaps the most radical path. In a nutshell, quantum computers are physical systems that exploit the laws of quantum mechanics to perform arithmetic and logical operations much faster than a conventional computer. In the words of Harrow, Hassidim, and Lloyd (HHL) [1], "quantum computers are devices that harness quantum mechanics to perform computations in ways that classical computers cannot." There are currently two complementary paradigms for quantum computing, namely, circuit-based systems and quantum annealers. Circuit-based systems exploit the deeper properties of quantum mechanics such coherence, entanglement, and non-locality, while quantum annealers mainly take advantage of tunneling between metastable states and the ground state. In [1], HHL introduces a circuit-based method by which the inverse of a matrix can be computed, and [2, 3] provide implementations of the algorithm to invert 2 × 2 matrices. Circuit-based methods are limited by the relatively small number of qubits that can be entangled into a fully coherent quantum state, currently of order 50 or so. An alternative approach to quantum computing is the quantum annealer [4], which takes advantage of quantum tunneling between metastable states and the ground state. The D-Wave Quantum Annealers have reached capacities of 2000+ qubits, which suggests that quantum annealers could be quite effective for linear algebra with hundreds to thousands of degrees of freedom. In this paper, we express floating point operations such as division and matrix inversion as *quadratic unconstrained binary optimization* (QUBO) problems, which are ideal for a quantum annealer. We should mention that our algorithm provides the full solution the matrix problem, while HHL provides only an expectation value. Furthermore, our algorithm places no constraints on the matrix that we are inverting, such as a sparsity condition.

The first step in mapping a general problem to a QUBO problem begins with constructing a Hamiltonian that encodes the problem in terms of a set of "logical" qubits. Next, because of the limited connectivity of the D-Wave chip, it will be necessary to "embed" the problem onto the chip, first by mapping each logical qubit to a collection or "chain" of physical qubits and then by determining parameter settings for all the physical qubits, including the chain couplings. We have implemented our algorithms on the D-Wave 2000Q and 2X chips, illustrating that division and matrix inversion can indeed be performed on an existing quantum annealer. The algorithms that we propose should ideally scale well for large numbers of equations, and should be applicable to a matrix inversion of relatively high order (although probably not exponentially higher order as in HHL). Currently, the scaling that may be achieved is limited by the connectivity and dynamic range of the chip.

Before examining the various algorithms, it is useful to review the basic formalism and to establish some notation. The general problem starts with a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the vertex set and $\mathcal{E}$ is the edge set. The QUBO *Hamiltonian* on $\mathcal{G}$ is defined by

$$H_{\mathcal{G}}[Q] = \sum_{r\in\mathcal{V}} A_r\, Q_r + \sum_{rs\in\mathcal{E}} B_{rs}\, Q_r Q_s \tag{1.1}$$

with $Q_r \in \{0, 1\}$ for all $r \in \mathcal{V}$. The coefficient A_r is called the *weight* at vertex r, while the coefficient B_{rs} is called the *strength* between vertices r and s. It might be better to call (1.1) the *objective function* rather than the Hamiltonian, as $H_{\mathcal{G}}$ is a real-valued function and not an operator on a Hilbert-space. However, it is easy to map (1.1) in an equivalent Hilbert space form,

$$\hat{H}_{\mathcal{G}} = \sum_{r\in\mathcal{V}} A_r\, \hat{Q}_r + \sum_{rs\in\mathcal{E}} B_{rs}\, \hat{Q}_r \hat{Q}_s \tag{1.2}$$

where $\hat{Q}_r|Q\rangle = Q_r|Q\rangle$ for all $r \in \mathcal{V}$, and $|Q\rangle \in \mathcal{H}$ for Hilbert space $\mathcal{H}$. The hat denotes an operator on the Hilbert space, and Q_r is the corresponding Eigenvalue of $\hat{Q}_r$ with Eigenstate $|Q\rangle$. Consequently, we can write

$$\hat{H}_{\mathcal{G}}|Q\rangle = H_{\mathcal{G}}[Q]\, |Q\rangle \tag{1.3}$$

and we use the terms *Hamiltonian* and *objective function* interchangeably. By the *QUBO problem*, we mean the problem of finding the lowest energy state $|Q\rangle$ of the Hamiltonian (1.2), which corresponds to minimizing Equation (1.1) with respect to the Q_r. This is an NP-hard problem uniquely suited to a quantum annealer. Rather than sampling all $2^{\#\mathcal{V}}$ possible states, quantum tunneling finds the *most likely* path to the ground state by minimizing the Euclidian action. In the case of the D-Wave 2X chip, the number of distinct quantum states is of order the very large number 2^{1000}, and the ground state is selected from this jungle of quantum states by tunneling to those states with a smaller Euclidean action.

The Ising model [5] is perhaps the quintessential physical example of a QUBO problem, and, indeed, it is one of the most studied systems in statistical physics. The Ising model consists of a square lattice of spin-1/2 particles with nearest neighbor spin–spin interactions between sites r and s, and when the system is immersed in a nonuniform magnetic field, this introduces coupling terms at individual sites r, thereby producing a Hamiltonian of the form

$$H_{\mathcal{G}}[J] = \sum_{r\in\mathcal{V}} B_r\, \Sigma_r + \sum_{rs\in\mathcal{E}} J_{rs}\, \Sigma_r \Sigma_s \tag{1.4}$$

where $\Sigma_r = \pm 1/2$. The Ising problem is connected to the QUBO problem by $\Sigma_r = Q_r - 1/2$.

For floating point division to R bits of resolution, the graph $\mathcal{G}$ is in fact just the fully connected graph $K_{\rm R}$. In terms of vertex and edge sets, we write $K_{\rm R} = (\mathcal{V}_{\rm R}, \mathcal{E}_{\rm R})$, and **Figure 1** illustrates K_8 and K_4. The left panel shows the completely connected graph K_8, with vertex and edge sets

$$\mathcal{V}_8 = \{0, 1, 2, \cdots, 7\} \tag{1.5}$$

$$\mathcal{E}_8 = \{\{0,1\}, \{0,2\}, \cdots, \{0,7\}, \{1,2\}, \cdots, \{1,7\}, \cdots, \{6,7\}\} \tag{1.6}$$

while the right panel shows the K_4 graph,

$$\mathcal{V}_4 = \{0, 1, 2, 3\} \tag{1.7}$$

$$\mathcal{E}_4 = \{\{0,1\}, \{0,2\}, \{0,3\}, \{1,2\}, \{1,3\}, \{2,3\}\}. \tag{1.8}$$

Just as 8-bit is called a *word*, 4-bit is called a *nibble*. As we will also see, the dynamic range of the D-Wave is most directly suitable to K_4, and the connectivity of K_4 consequently gives a quantum nibble.

Let us remark about our summation conventions. Rather than summing over the edges,

$$H[Q] = \sum_{r\in\mathcal{V}_{\rm R}} A_r\, Q_r + \sum_{rs\in\mathcal{E}_{\rm R}} B_{rs}\, Q_r Q_s \tag{1.9}$$

$$= \sum_{r=0}^{R-1} A_r\, Q_r + \sum_{r=0}^{R-1}\sum_{s>r}^{R-1} B_{rs}\, Q_r Q_s \tag{1.10}$$

we find it convenient to sum over all values of r and s taking B_{rs} to be symmetric. In this case, the double sum differs by a factor of two relative to summing over the edge set of the graph,

$$H[Q] = \sum_{r=0}^{R-1} A_r\, Q_r + \sum_{r=0}^{R-1}\sum_{s=0}^{R-1} \frac{1}{2}\, B_{rs}\, Q_r Q_s. \tag{1.11}$$

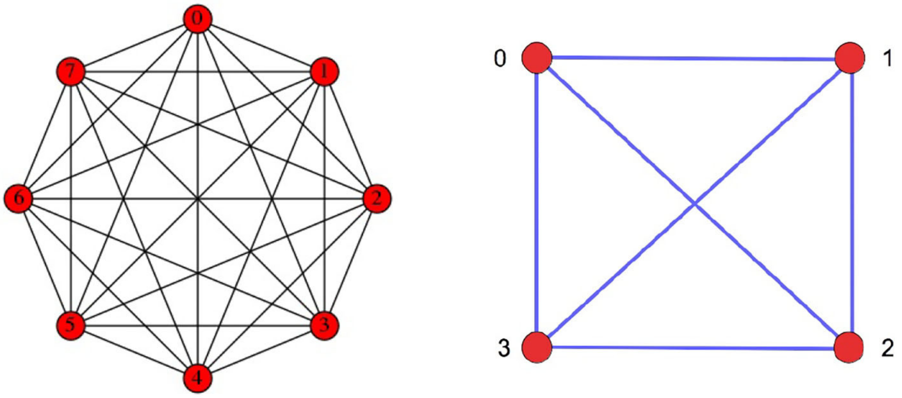

FIGURE 1 | The left panel shows the fully connected graph K_8, and the right panel shows the corresponding graph K_4. To perform a calculation to 8-bit accuracy requires the connectivity of K_8. We take the vertex and edge sets of K_8 to be $\mathcal{V}_8 = \{0, 1, 2, \cdots, 7\}$ and $\mathcal{E}_8 = \{\{0, 1\}, \{0, 2\}, \cdots, \{0, 7\}, \{1, 2\}, \{1, 3\}, \cdots, \{6, 7\}\}$. To perform a calculation to 4-bit accuracy requires K_4 connectivity, and, similarly, the vertex and edge sets for K_4 are $\mathcal{V}_4 = \{0, 1, 2, 3\}$ and $\mathcal{E}_4 = \{\{0, 1\}, \{0, 2\}, \{0, 3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}\}$.

Furthermore, for $r = s$, there will be a linear contribution from the idempotency condition $Q_r^2 = Q_r$, so that

$$H[Q] = \sum_{r=0}^{R-1} \left[A_r + \frac{1}{2}\, B_{rr} \right] Q_r + \sum_{r=0}^{R-1} \sum_{s \neq r, s=0}^{R-1} \frac{1}{2}\, B_{rs}\, Q_r Q_s. \quad (1.12)$$

We can write this as

$$H[Q] \;=\; \sum_{r=0}^{R-1} \tilde{A}_r\, Q_r + \sum_{r=0}^{R-1} \sum_{s \neq r, s=0}^{R-1} \tilde{B}_{rs}\, Q_r Q_s. \qquad (1.13)$$

2. FLOATING POINT DIVISION ON A QUANTUM ANNEALER

2.1. Division as a QUBO Problem

In this section we present an algorithm for performing floating point division on a quantum annealer. Given two input parameters m and y to R-bits of resolution, the algorithm calculates the ratio y/m to R bits of resolution. The corresponding division problem can be represented by the linear equation

$$m \cdot x - y = 0, \qquad (2.1)$$

which has the unique solution

$$x = y/m. \qquad (2.2)$$

Solving (2.1) on a quantum annealer amounts to finding an objective function $H(x)$ whose minimum corresponds to the solution of Equation (2.2). Although the form of $H(x)$ is not unique, for this work we employ the simple real-valued quadratic function

$$H(x; m, y) = \big(m\, x - y\big)^2 \qquad (2.3)$$

where m and y are continuous parameters. For an ideal annealer, we do not have to concern ourselves with the numerical range and resolution of the parameters m and y; however, for a real machine such as the D-Wave, this is an important consideration. For a well-conditioned matrix, we require that the parameters m and y possess a numerical range that spans about an order of magnitude, from approximately 0.1–1.0. This provides about 3–4 bits of resolution: $1/2^0 = 1$, $1/2^1 = 0.5$, $1/2^2 = 0.25$, and $1/2^3 = 0.125$. The dynamic range and the connectivity both impact the resolution of a calculation.

To proceed, let us formulate floating point division as a *quadratic unconstrained binary optimization* (QUBO) problem. The algorithm starts by converting the real-valued number x in (2.3) into an R-bit binary format, while the numbers m and y remain real valued parameters of the objective function. For any number $\chi \in [0, 2)$, the binary representation accurate to R bits of resolution can be expressed by $[Q_0.Q_1 Q_2 \cdots Q_{R-1}]_2$, where $Q_r \in \{0, 1\}$ is value of the r-th bit, and the square bracket indicates the binary representation[1]. It is more algebraically useful to express this in terms of the power series in 2^{-r},

$$\chi \;=\; \sum_{r=0}^{R-1} 2^{-r}\, Q_r. \qquad (2.4)$$

In order to represent negative number, we perform the binary offset

$$x \;=\; 2\chi - 1 \qquad (2.5)$$

where $x \in [-1, 3)$. The objective function now takes the form

$$H(\chi) \;=\; 4m^2 \chi^2 - 4m(m + y)\chi + (m + y)^2. \qquad (2.6)$$

The constant term $(m + y)^2$ can be dropped when finding the minimum of (2.6), but we choose to keep it for completeness.

[1] Since the infinite geometric series $\sum_{r=0}^{\infty} 2^{-r}$ sums to 2, the finite series is <2. In binary form we have $[1.11 \cdots]_2 = 2$ and $[1.11 \cdots 1]_2 < 2$. Working to resolution R is like calculating the R-th partial sum of an infinite series.

Equation (2.4) provides a change of variables $\chi = \chi[Q]$ (where Q is the collection of the Q_r), and this allows us to express (2.3) in the form

$$H[Q] = \sum_{r=0}^{R-1} A_r\, Q_r + \sum_{r=0}^{R-1} \sum_{s\neq r, s=0}^{R-1} B_{rs}\, Q_r Q_s. \tag{2.7}$$

In the notation of graph theory, we would write

$$H[Q] = \sum_{r\in\mathcal{V}_\mathrm{R}} A_r\, Q_r + \sum_{rs\in\mathcal{E}_\mathrm{R}} \frac{1}{2}\, B_{rs}\, Q_r Q_s \tag{2.8}$$

where $\mathcal{V}_\mathrm{R} = \{0, 1, 2, \cdots, R-1\}$ is the vertex set, and $\mathcal{E}_\mathrm{R}$ is the edge set. We often employ an abuse of notation and write $rs \in \mathcal{E}_\mathrm{R}$ to mean $\{r, s\} \in \mathcal{E}_\mathrm{R}$. Thus, instead of $B_{\{r,s\}}$, we write B_{rs}. Since the order of the various elements of a set are immaterial, we require B_{rs} to be symmetric in r and s. Rather than summing over the edge sets $rs \in \mathcal{E}_\mathrm{R}$, we employ the double sum $\sum_{r\neq s}$, which introduces a relative factor of two in the convention for the strengths B_{rs}. The goal of this section is to find A_r and B_{rs} in terms of m and y.

Note that we can generalize the simple binary offset (2.4) if we scale and shift $\chi \in [0, 2)$ by

$$x = c\chi - d \tag{2.9}$$

so that $x \in [-d, 2c - d)$. When $d > 0$ and $c > d/2$, the domain of x always contains a positive and negative region, and the precise values for d and c can be chosen based on the specifics of the problem. For Equation (2.9), the objective function takes the form

$$H(\chi) = 4m^2c^2\, \chi^2 - 4mc\,(m + y)\chi + (md + y)^2. \tag{2.10}$$

For simplicity of notation, this paper employs the simple binary offset (2.5), although our Python interface to the D-Wave quantum annealer employs the generalized form (2.10).

Equation (2.4) allows us to express the quadratic term in χ as

$$\chi^2 = \sum_{r=0}^{R-1}\sum_{s=0}^{R-1} 2^{-r-s} Q_r Q_s = \sum_{r=0}^{R-1} \sum_{s\neq r,s=0}^{R-1} 2^{-r-s} Q_r Q_s + \sum_{r=0}^{R-1} 2^{-2r} Q_r \tag{2.11}$$

where we have used the idempotency condition $Q_r^2 = Q_r$ along the diagonal in the last term of (2.11). Substituting the forms (2.4) and (2.11) into (2.6) yields the Hamiltonian

$$H[Q] = \sum_{r=0}^{R-1} 4m\, 2^{-r}\Big[m\, 2^{-r} - (y + m)\Big] Q_r + \sum_{r=0}^{R} \sum_{s\neq r,s=0}^{R-1} 4m^2\, 2^{-r-s}\, Q_r Q_s \tag{2.12}$$

and the Ising coefficients in (2.7) can be read off:

$$A_r = 4m\, 2^{-r}\Big[m\, 2^{-r} - (y + m)\Big] \tag{2.13}$$

$$B_{rs} = 4m^2\, 2^{-r-s} \quad r \neq s. \tag{2.14}$$

Because of the double sum over r and s in the objective function in (2.12), the algorithm requires a graph of connectivity K_R. The special cases of K_8 and K_4 have been illustrated in **Figure 1**. To obtain higher accuracy than the K_R graph allows, we can iterate this procedure in the following manner. Suppose we start with $y_0 = y$ and are given a value y_{n-1} with $n > 1$; we then advance the iteration to y_n in the following manner:

$$\text{solve } \; mx_n = y_{n-1} \;\text{ for } x_n \text{ to } R \text{ bits} \tag{2.15}$$

$$\text{calculate the error } \; y_n = y_{n-1} - mx_n. \tag{2.16}$$

Now that we have the value y_n, we can repeat the process to find y_{n+1}, and we can stop the iterative procedure when the desired level of accuracy has been achieved.

2.2. Embedding K_R Onto the D-Wave Chimera Architecture

The D-Wave Chimera chip consists of coupled bilayers of micro rf-SQUIDs overlaid in such a way that, while relatively easy to fabricate, results in a fairly limited set of physical connections between the qubits. However, by *chaining* together well chosen qubits in a positively correlated manner, this limitation can largely be overcome. The process of chaining requires that we (i) embed the logical graph onto the physical graph of the chip (for example K_4 into C_8) and that we (ii) assign weights and strengths to the physical graph embedding in such as a way as to preserve the ground state of the logical system. These steps are called graph embedding and Hamiltonian embedding, respectively.

Let us explore the connectivity of the D-Wave Chimera chip in more detail. The D-Wave architecture employs the C_8 bipartate Chimera graph as its most basic unit of connectivity. This *unit cell* is illustrated in **Figure 2**, and it consists of 8 qubits connected in a 4 × 4 bipartate manner. The left panel of the figure uses a *column* format in laying out the qubits, and the right panel illustrates the corresponding qubits in a *cross* format, where the gray lines represent the direct connections between the qubits. The cross format is useful since it minimizes the number intersecting connections. The complete two-dimensional chip is produced by replicating C_8 along the vertical and horizontal directions, as illustrated in **Figure 3**, thereby providing a chip with thousands of qubits. The connections between qubits are limited in two ways: (i) by the connectivity of the basic unit cell C_8 and (ii) by the connectivity between the unit cells across the chip. The bipartate graph $C_8 = (\mathcal{V}_8, \mathcal{B}_8)$ is formally defined by the vertex set $\mathcal{V}_8 = \{1, 2, \cdots, 8\}$, and the edge set

$$\mathcal{B}_8 = \Big\{\{1,5\},\{1,6\},\{1,7\},\{1,8\},\{2,5\},\{2,6\},\{2,7\},\{2,8\}, \{3,5\},\{3,6\},\{3,7\},\{3,8\},\{4,5\},\{4,6\},\{4,7\},\{4,8\}\Big\}. \tag{2.17}$$

The set $\mathcal{B}_8$ represents the connections between a given red qubit and the corresponding blue qubits in the figures. The red and blue dots illustrate the bipartate nature of C_8, as every red dot is connected to every blue dot, while none of the blue and red dots are connected to one another.

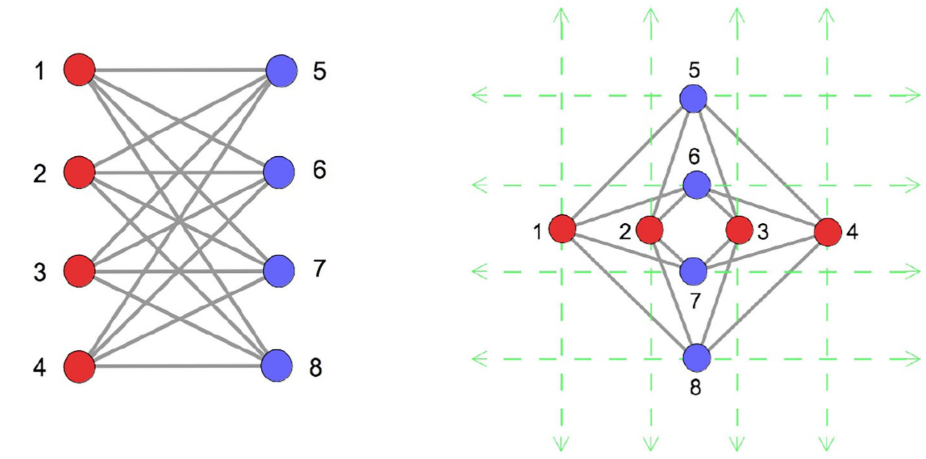

FIGURE 2 | The left panel illustrates the bipartite graph C_8 in *column* format, while the right panel illustrates the corresponding graph in *cross* format, often called a Chimera graph. The gray lines represent direct connections between qubits. The cross format is useful since it minimizes the number intersecting connections. The use of red and blue dots emphasize the bipartate nature of C_8, as every red dot is connected to every blue dot, while none of the red and blue dots are connected to one another. The vertex set of C_8 is taken to be $\mathcal{V}_8 = \{1, 2, \cdots, 8\}$ and edge set is $\mathcal{B}_8 = \{\{1, 5\}, \{1, 6\}, \{1, 7\}, \{1, 8\}, \{2, 5\}, \{2, 6\} \cdots \{7, 8\}\}$.

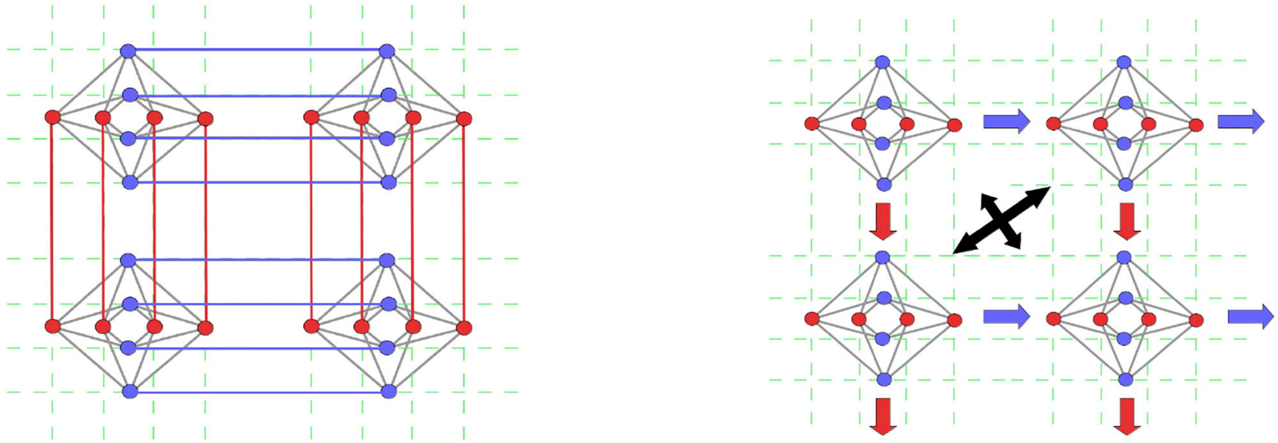

FIGURE 3 | The left panel shows the connectivity between four C_8 bipartate Chimera zones, and the right panel illustrates how multiple C_8 graphs are stitched together along the vertical and horizontal directions to provide thousands of possible qubits. A limitation of this connectivity strategy is that red and blue zones cannot communicate directly with one another, as indicated by the black crossed arrows. The purpose of *chaining* is to allow communication between the red and blue qubits.

We will denote the *physical* qubits on the D-Wave chip by q_ℓ. For the D-Wave 2000Q there is a maximum of 2,048 qubits, while the D-Wave 2X has 1,152 qubits. For the example calculation in this text, we only use 10–50 qubits. The physical Hamiltonian or objective function takes the form

$$H[q] = \sum_\ell a_\ell \, q_\ell + \sum_{\ell \neq m} 2b_{\ell m} \, q_\ell q_m \tag{2.18}$$

where we have introduced a factor of 2 in the strength to account for the symmetric summation over r and s. We will call the qubits Q_r of the previous section the *logical qubits*. To write a program for the D-Wave means finding an embedding of the problem for logical qubits onto the physical collection of qubits q_ℓ. If the connectivity of the Chimera graphs were large enough, then the logical qubits would coincide exactly with the physical qubits. However, since the graph C_8 possesses less connectivity than K_4, we must resort to chaining on the D-Wave, even for 4-bit resolution. **Figure 4** illustrates the K_4 embedding used by our algorithm, where, as before, the left panel illustrates the bipartate graph in column format, and the right panel illustrates the corresponding graph in cross format.

In **Figure 4**, we have labeled the physical qubits by $\ell = 1, 2, 3 \cdots 8$, and we wish to map the problem involving logical qubits $Q_r \, Q_s \, Q_t$ onto the four physical qubits $q_5 \, q_1 \, q_6 \, q_2$. The embedding requires that we *chain* together the two qubits 1-6 and 3-8, respectively. We may omit qubits 4 and 7 entirely. As illustrated in **Figure 5**, the physical qubits q_1 and q_6 are *chained*

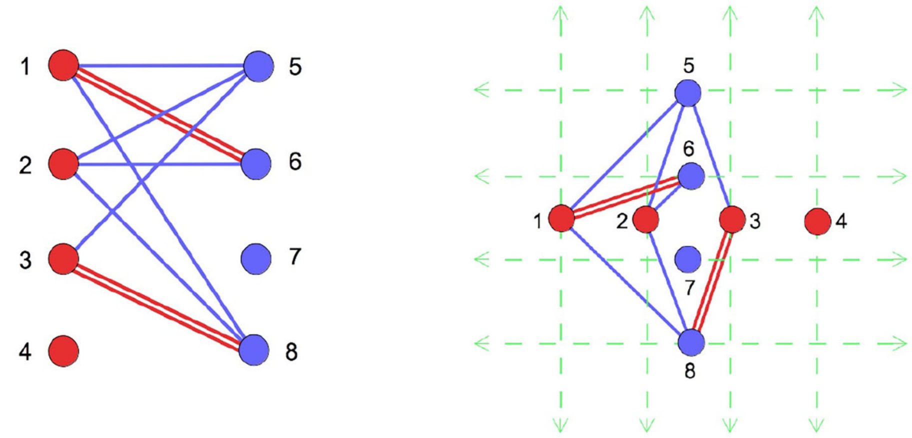

FIGURE 4 | The K_4 embedding onto C_8 used in our implementation of 4-bit of division on the D-Wave. The blue lines represent normal connections between qubits, while the red double-lines represent chained qubits, that is to say, qubits that are strictly correlated (and can thereby represent a single logical qubit at a higher level of abstraction). The qubits 1–6 are chained together, as are the qubits 3–8.

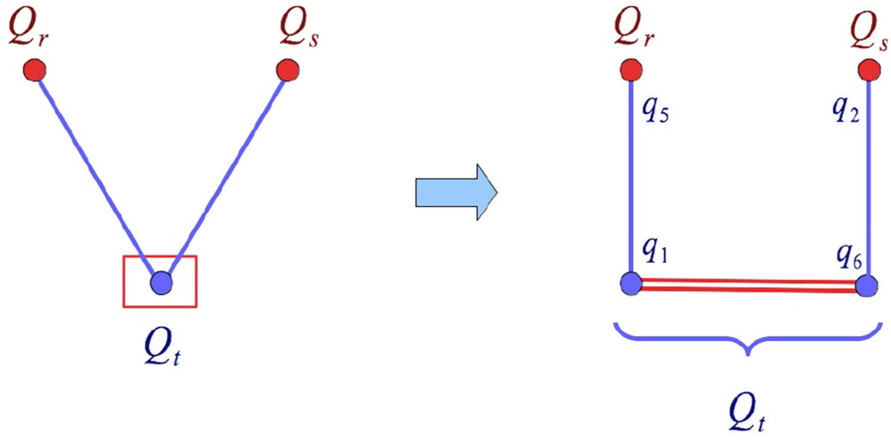

FIGURE 5 | The left panel shows three logical qubits Q_r, Q_s, Q_t with connectivity between *r*-*t* and *t*-*s*. The box surrounding qubit *t* means that it will be modeled by a linear chain of physical qubits, as illustrated in the right panel. The labeling is taken from **Figure 4** for qubits 5-1-6-2, where Q_r is mapped to q_5, Q_s is mapped to q_2, and Q_t is split between q_1 nd q_6. Qubits q_1 and q_6 are chained together to simulate the single logical qubit Q_t, while qubits Q_r and Q_s map directly onto physical qubits q_5 and q_2.

together to simulate a single logical qubit Q_t, while qubits q_5 and q_2 are mapped directly to the logical qubits Q_r and Q_s, respectively. Qubit q_5 is assigned the weight $a_5 = A_r$ and the coupling between q_5 and q_1 is assigned the value $b_{51} = B_{rt}$. Similarly for qubit q_2, the vertex is assigned weight $a_2 = A_s$, and strength between q_2 and q_6 is $b_{26} = B_{st}$. We must now distribute the logical qubit Q_t between q_1 and q_6 by assigning the values a_1, a_6 and b_{16}. We distribute the weight A_t uniformly between qubits q_1 and q_2, giving $a_1 = A_t/2$ and $a_6 = A_t/2$. We must now choose b_{16}. To preserve the energy spectrum, we must shift the values of the weights a_1 and a_6. We can do this by adding a counter-term Hamiltonian

$$H^{\text{CT}} = a\,q_1 + a\,q_6 + 2b_{16}\,q_1 q_6 \qquad (2.19)$$

to the physical Hamiltonian. The double lines in **Figures 4**, **5** indicate that two qubits are chained together. This means that the qubits are strictly correlated, i.e., when q_1 is up then q_6 is up, and when q_1 is down then q_6 is down. This is accomplished by choosing the coupling strength b_{16} to favor a strict correlation; however, to preserve the ground state energy, this also requires shifting the weights for q_1 and q_6. For $q_1 = q_6 = 0$ we have $H^{\text{CT}} = 0$. We wish to preserve this condition when $q_1 = q_6 = 1$, which means $2a + 2b = 0$. Furthermore, the state $q_1 = 1$ and $q_6 = 0$ must have positive energy, which means $a > 0$. Similarly for $q_1 = 0$ and $q_6 = 1$. We therefore choose $a_1 = a_6 = \alpha > 0$ and $b_{16} = -\alpha$, where α is an arbitrary parameter. This is illustrated in **Table 1**. A more complicated case is the linear chain of N qubits as shown in **Figure 6**. The counter-term Hamiltonian is taken to be

$$H^{\rm CT} = \sum_{m=1}^{N} a^t_m \, q^t_m + \sum_{m=1}^{N-1} b^t_{m,m+1} \, q^t_m q^t_{m+1}. \tag{2.20}$$

Note that $H^{\rm CT}$ vanishes when $q_m = 0$ for all $m = 1 \cdots N$. And conversely, we must arrange the counter-term to vanish when $q_m = 1$. The simplest choice is to take all weights to be the same and all couplings to be identical. Then, to preserve the ground state when the $q_r = 1$, we impose

$$a^t_r = \frac{A_t}{N} + \frac{2(N-1)}{N}\,\alpha \tag{2.21}$$

$$b^t_{r,r+1} = -\alpha \tag{2.22}$$

with $\alpha > 0$ and $r = 1 \cdots N$. The first term in a^t_r distributes the weight A_t uniformly across all N nodes in the chain. The second set of terms $b^t_{r,r+1}$ ensures that the qubits of the chain are strictly correlated. The counter-term energy contribution is positive when the linearly chained qubits are not correlated, therefore anti-correlation is always penalized. **Table 2** illustrates the spectrum of the counter-term Hamiltonian for three qubits. We may need to choose large values of α, of order 20 or more, to sufficiently separate the states. The uniform spectrum of 4 states with $H^{\rm CT} = a$ in **Table 2** arises from a permutation symmetry in q_1, q_2, q_3.

To review, note that a linear counter-term is represented in **Figure 6**. We add a counter-term to break the logical qubits into a chain of physical qubits that preserve the ground state. Let us consider the conditions that we place on the Hamiltonian to ensure strict correlation between the chained qubits. We adjust the values of A_r and B_{rs} to ensure that spin alignment is energetically favorable. By slaving several qubits together, we can overcome the limitations of the Chimera connectivity. As a more complex example, consider the four logical qubits of **Figure 7** connected in a circular chain by strengths B_{12}, B_{24}, B_{43}, and B_{31}. Suppose the weights are A_1, A_2 A_3, and A_4. **Figure 8** provides an example in which each logical qubit is chained in a linear fashion to the physical qubits.

TABLE 1 | For two qubits the counter-term Hamiltonian is $H^{\rm CT}(q_1, q_6) = a\,q_1 + a\,q_6 + 2b\,q_1 q_6$.

q_1	q_6	$H^{\rm CT}$
0	0	0
0	1	α
1	0	α
1	1	0

The lowest energy state is preserved for $b = -\alpha$ and $a = \alpha$ where $\alpha > 0$. We will split the weight A_t uniformly across the N chained physical qubits, thereby giving a contribution to the physical Hamiltonian $H^t_{16} = A_t/2 + \alpha\, q_1 + \alpha\, q_6 - 2\alpha\, q_1 q_6$. The energy spectrum ensures that the two qubits are strictly correlated.

3. MATRIX INVERSION AS A QUBO PROBLEM

In this section we present an algorithm for solving a system of linear equations on a quantum annealer. To precisely define the mathematical problem, let $\boldsymbol{M}$ be a nonsingular $N \times N$ real matrix, and let $\mathbf{Y}$ be a real N dimensional vector; we then wish to solve the linear equation

$$\boldsymbol{M} \cdot \mathbf{x} = \mathbf{Y}. \tag{3.1}$$

The linearity of the system means that there is a unique solution,

$$\mathbf{x} = \boldsymbol{M}^{-1} \cdot \mathbf{Y} \tag{3.2}$$

and the algorithm is realized by specifying an objective function whose ground state is indeed (3.2). The objective function is not unique, although it must be commensurate with the architecture of the hardware. If the inverse matrix itself is required, it can be constructed by solving (3.1) for each of the N linearly independent basis vectors for $\mathbf{Y}$. It is easy to construct a quadratic objective $H(\mathbf{x})$ whose minimum is (3.2), namely,

$$H(\mathbf{x}) \;=\; \big(\boldsymbol{M}\mathbf{x} - \mathbf{Y}\big)^2 = \big(\boldsymbol{M}\mathbf{x} - \mathbf{Y}\big)^{\rm T} \cdot \big(\boldsymbol{M}\mathbf{x} - \mathbf{Y}\big). \tag{3.3}$$

In terms of matrix components, this can be written as

$$H(\mathbf{x}) \;=\; \mathbf{x}^{\rm T}\boldsymbol{M}^{\rm T}\boldsymbol{M}\mathbf{x} - \mathbf{x}^{\rm T}\boldsymbol{M}^{\rm T}\mathbf{Y} - \mathbf{Y}^{\rm T}\boldsymbol{M}\mathbf{x} + \mathbf{Y}^{\rm T}\mathbf{Y}$$

TABLE 2 | For a three qubit chain the counter-term Hamiltonian is $H^{\rm CT}(q_1, q_2, q_3) = a\,q_1 + a\,q_2 + a\,q_3 + 2b\,q_1 q_2 + 2b\,q_2 q_3$, where $a = 4\alpha/3$ and $b = -\alpha$.

q_1	q_2	q_3	$H^{\rm CT}$	
0	0	0	0	0
0	0	1	$4\alpha/3$	a
0	1	0	$4\alpha/3$	a
0	1	1	$2\alpha/3$	$a/2$
1	0	0	$4\alpha/3$	a
1	0	1	$8\alpha/3$	$2a$
1	1	0	$2\alpha/3$	a
1	1	0	0	0

The degeneracy in energy of value a arises from a permutation symmetry in $q_1 \rightarrow q_2 \rightarrow q_3$ that preserves the form of the counter-term Hamiltonian.

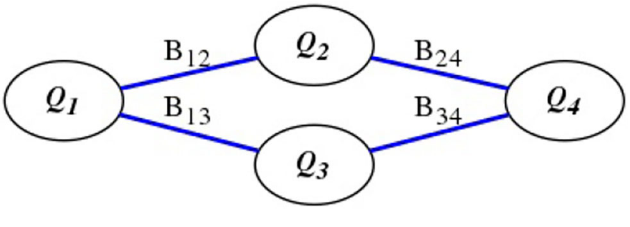

FIGURE 7 | Four logical qubits Q_1, Q_2, Q_3, Q_4 in a circular loop with connection strengths B_{12}, B_{24} B_{43}, and B_{31}.

$$= \sum_{ijk=1}^{N} M_{ki}M_{kj}\, x^i x^j - 2\sum_{ij=1}^{N} Y_j M_{ji}\, x^i + \|\mathbf{Y}\|^2. \quad (3.4)$$

Note that $\|\mathbf{Y}\|^2$ is just a constant, which will not affect the minimization. In principle all constants can be dropped from the objective function, although we choose to keep them for completeness. One may obtain a floating point representation of each component of $\mathbf{x} = (x^1, \cdots, x^{\mathrm{N}})^{\mathrm{T}}$ by expanding in powers of 2 multiplied by Boolean-valued variables $q_r^i \in \{0, 1\}$,

$$\chi^i = \sum_{r=0}^{R-1} 2^{-r} q_r^i \quad (3.5)$$

$$x^i = 2\chi^i - 1. \quad (3.6)$$

As before, the domains are given by $\chi^i \in [0, 2)$ and $x^i \in [-1, 3)$, and upon expressing $\mathbf{x}$ as a function q_r^i, we can recast (3.4) in the form

$$H[q] = \sum_{i=1}^{N}\sum_{r=0}^{R-1} a_r^i\, q_r^i + \sum_{i=1}^{N}\sum_{i\neq j=1}^{N}\sum_{r=0}^{R-1}\sum_{s=0}^{R-1} b_{rs}^{ij}\, q_r^i q_s^j. \quad (3.7)$$

The coefficients a_r^i are called the *weights* and the coefficients b_{rs}^{ij} are the interaction *strengths*. Note that the algorithm requires a connectivity of K_{NR} for arbitrary matrices.

Let us first calculate the product $x^i x^j$ in (3.4). From (3.5) and (3.6), we find

$$x^i x^j = \left(2\sum_{r=0}^{R-1} 2^{-r} q_r^i - 1\right)\left(2\sum_{r'=0}^{R-1} 2^{-r'} q_{r'}^i - 1\right)$$

$$= 4\sum_{rr'} 2^{-(r+r')} q_r^i q_{r'}^j - 4\sum_r 2^{-r} q_r^i + 1 \quad (3.8)$$

$$= 4\sum_{r\neq r'} 2^{-(r+r')} q_r^i q_{r'}^j + 4\sum_r 2^{-2r} q_r^i - 4\sum_r 2^{-r} q_r^i + 1 \quad (3.9)$$

where we have used the idempotency condition $(q_r^i)^2 = q_r^i$ in the second term of (3.9). While the second form is one used by the code, it is more convenient algebraically to use the first form. Substituting (3.8) into the first term in (3.4) gives

$$H_1 \equiv \sum_{ijk} M_{ki}M_{kj}\, x_i x_j \quad (3.10)$$

$$= \sum_{ijk} M_{ki}M_{kj} \left\{ 4\sum_{rr'} 2^{-(r+r')} q_r^i q_{r'}^j - 4\sum_r 2^{-r} q_r^i + 1 \right\} (3.11)$$

$$= 4\sum_{ir}\sum_{js}\sum_{k} 2^{-r-s} M_{ki}M_{kj}\, q_r^i q_s^j - 4\sum_{ir}\sum_{k} 2^{-r} M_{ki}M_{ki}\, q_r^i$$

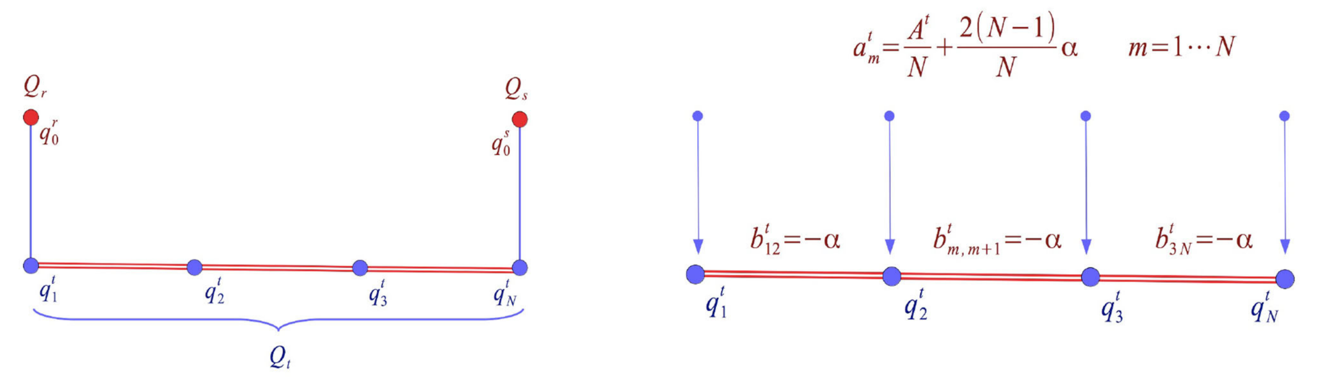

FIGURE 6 | Generalization of **Figure 5** to a chain of *N* linear qubits. The right panel illustrates the chain coupling parameters used to create strict correlations of the physical qubits within the chain.

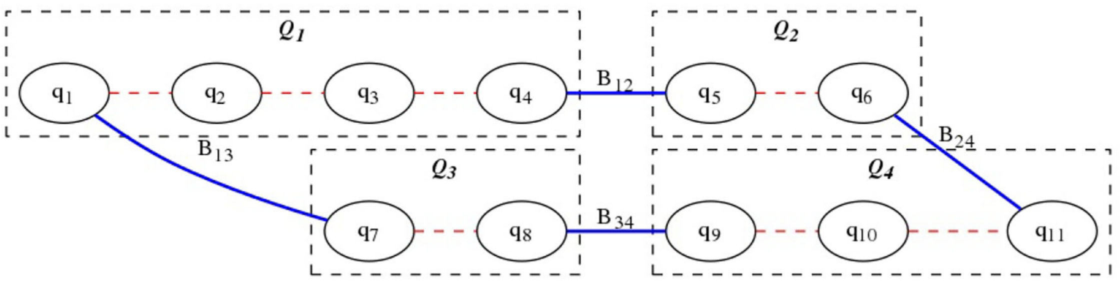

FIGURE 8 | A possible mapping of the logical qubits in **Figure 7** onto the physical device. Each logical qubit is modeled by a linear chain of strictly correlated qubits.

$$+\sum_{ijk} M_{ki}M_{kj}. \tag{3.12}$$

The second term in (3.4) can be expressed as

$$H_2 \equiv -2\sum_{ij} Y_j M_{ji}\, x_i = -2\sum_{ij} Y_j M_{ji}\left(2\sum_r 2^{-r} q_r^i - 1\right) \tag{3.13}$$

$$= -4\sum_{ij}\sum_r 2^{-r} M_{ji} Y_j\, q_r^i + 2\sum_{ij} Y_j M_{ji}. \tag{3.14}$$

Adding H_1 and H_2 gives

$$H = 4\sum_{ir}\sum_{js}\sum_k 2^{-r-s} M_{ki}M_{kj}\, q_r^i q_s^j - 4\sum_{ir}\sum_k 2^{-r} M_{ki}M_{ki}\, q_r^i \tag{3.15}$$

$$-4\sum_{ij}\sum_r 2^{-r} M_{ji} Y_j\, q_r^i + 2\sum_{ij} Y_j M_{ji} + \sum_{ijk} M_{ki}M_{kj}. \tag{3.16}$$

The QUBO coefficients for logical qubits are therefore

$$a_r^i = 4\cdot 2^{-r}\sum_k M_{ki}\left\{2^{-r}M_{ki} - \left(Y_k + \sum_j M_{kj}\right)\right\} \tag{3.17}$$

$$b_{rs}^{ij} = 4\cdot 2^{-(r+s)}\sum_k M_{ki}\, M_{kj}. \tag{3.18}$$

In the programming interface, the coefficients are addressed with a 1-dimensional linear index, while the parameters in 3.17 and 3.18 are defined in terms of the 2-dimensional indices i and r, where $i = 0, 1, \cdots, N-1$ and $r = 0, 1, \cdots, R-1$. Now, we define a 1-1 mapping between these indices and the linear index $\ell = 0, 1, \cdots, N\cdot R - 1$. This is just an ordinary linear indexing for 2-dimensional matrix elements, so we choose the usual row-major linear index mapping,

$$\ell(i, r) = i\cdot R + r \tag{3.19}$$

$$M_\ell = M_{ir}. \tag{3.20}$$

The inverse mapping gives the row and column indices as below,

$$i_\ell = \lfloor \ell/R \rfloor \tag{3.21}$$

$$r_\ell = \ell \bmod R \tag{3.22}$$

where $\lfloor n \rfloor$ is the greatest integer less than or equal to n. The expression "ℓ mod R" is ℓ modulo R. This is a 1-1, invertible mapping between each pair of values of i and r in the matrix index space to every value of ℓ in the linear qubits index space. We can simply replace sums over all index pairs i, r by a single sum over ℓ, provided we also rewrite any isolated indices in i and r as functions of ℓ via their inverse mapping.

We may summarize this observation in the following formal identity. Given some arbitrary quantity, A, that depends functionally upon the tuple (i, r), and possibly upon the individual indices i and r, it is trivial to verify that

$$A[(i, r), i, r] = \sum_{\ell=0}^{N\cdot R-1} A[\ell, i_\ell, r_\ell]\, \delta_{i,i_\ell}\delta_{rr_\ell} \tag{3.23}$$

where ℓ, i_ℓ, and r_ℓ are related as in Equations (3.19)–(3.22). This identity is useful for formal derivations. For example, we may use it to quickly derive the binary expansion of x_i in terms of logical qubits. Inserting (3.23) into (3.6) gives,

$$x_i = 2\left(\sum_{r=0}^{R-1} 2^{-r}\sum_{\ell=0}^{N\cdot R-1} q_\ell\, \delta_{i\,i_\ell}\delta_{r\,r_\ell}\right) - 1$$

$$= 2\sum_{\ell=0}^{N\cdot R-1} 2^{-r_\ell}\, q_\ell\, \delta_{i,i_\ell} - 1. \tag{3.24}$$

Clearly, x_i has non-zero contributions only for those indices corresponding to $i = i_\ell = \lfloor \ell/R \rfloor$, that is, only from those qubits within a row in the q_r^i array. Also, those contributions are summed along that row, i.e.,, over $r_\ell = \ell \bmod R$. This equation will be used to reconstruct the floating-point solution, **x**, from the components q_ℓ of the binary solution returned from the D-Wave annealing runs. The weights and strengths now become

$$a_\ell = 4\cdot 2^{-r_\ell}\sum_k M_{k\,i_\ell}\left\{2^{-r_\ell}M_{k\,i_\ell} - (Y_k + \sum_j M_{kj})\right\} \tag{3.25}$$

$$b_{\ell m} = 4\cdot 2^{-(r_\ell + r_{\ell'})}\sum_k M_{k\,i_\ell}\, M_{k\,i_m}. \tag{3.26}$$

For a 2 × 2 matrix to 4-bit accuracy, we need K_8 ($4\times 2 = 8$), and to 8-bit accuracy we need K_{16} ($8\times 2 = 16$). We have inverted matrices up to 3×3 to 4-bit accuracy, which requires K_{12} ($3\times 4 = 12$). For an $N\times N$ matrix with R bits of resolutions, we must construct linear embeddings of K_{RN}. We could generalize this procedure for complex matrices.

4. CALCULATIONS

4.1. Implementation

The methods above were implemented using D-Wave's Python SAPI interface and tested on a large number of floating-point calculations. Initially, we performed floating-point division on simple test problems with a small resolution. Early on, we discovered that larger graph embeddings tended to produce noisier results. To better understand what was happening we started with a K_8 graph embedding to represent two floating-point numbers with only four bits of resolution. Since the D-Wave's dynamic range is limited to about a factor of 10 in the scale of the QUBO parameters, we determined that we could expect no more than 3–4 bits of resolution from any one calculation in any event. However, our binary offset representation (3.5) implies that we should expect no more than 3 bits of resolution in any single run. Indeed, using the K_8 embedding, we were able to get *exact* solutions from the annealer for any division problems that had answers that were multiples of 0.25 between −1.0 and 1.0. Problems in this range that had solutions that were *not* exact multiples of 0.25 resulted in approximate solutions, effectively "rounded" to the nearest of ±0.25 or ±0.75. At this point we implemented an iterative scheme that uses the current error, or residual, as a new input, keeping track of the accumulated floating-point solution.

The iteration method has been implemented and tested for floating-point division, but we have not yet implemented iteration for matrix inversion. That can be done by using the previous residual (error) as the new inhomogeneous term in the matrix equation. We plan to implement an iterative method in the matrix inversion code eventually. However, we already have good preliminary results on matrix inversion that suggests that this should work reasonably well, at least for well-conditioned matrices. Currently, we are able to solve 2×2 and 3×3 linear equations involving floating-point numbers up to a resolution of 4 bits, and having well-conditioned matrices, *exactly* for input vectors with elements defined on $[-1, 1]$ and that are multiples of 0.25. Using an example matrix that is poorly-conditioned, we find that it is generally not possible to get the right answer without first doing some sort of preconditioning to the matrix. But, more importantly, we were able to obtain some insight about why ill-conditioned matrices can be difficult to solve as QUBO problems on a quantum annealer, which gives some hints about how to ameliorate the problem. We will discuss these results, and the effects of ill-conditioning on the QUBO matrix inverse problem below.

4.1.1. Note on Solution Normalization and Iteration

Allowing both the division and linear equation QUBO solvers to work for arbitrary floating-point numbers, and to allow for iterative techniques, requires normalizing the ratio of the current dividend and the divisor, or the residual and matrix, to a value in a range between −1 and 1. For the division problems, we wanted to avoid "dividing in order to divide," so we normalized each ratio using the difference between the binary exponents of $\lfloor \text{divisor} \rfloor$ and $\lfloor \text{dividend} \rfloor$. These can be found just by using order comparisons, with no explicit divisions. Adding 1 to this yields an "offset"—the largest binary exponent of the ratio—to within a factor of 2 (±1 in the offset), which is sufficient for scaling our QUBO parameters as needed. The fact that our QUBO solutions are always returned in binary representation provides a simple way to bound the solution into a range solvable with the annealer by simply shifting the binary representation of the current dividend by a few bits (using the current offset), which is why we refer to the solution exponent as an "offset." In this way, the solution can easily be guaranteed to be in the correct range without having to perform any divisions in Python. The "offset" is accumulated and used to construct the current approximation to the floating-point solution on each iteration. The iteration process continues until the error of the approximate solution is less than some tolerance specified by the user.

4.2. Results for Division

First, we present some examples for division without iteration. We used a K_4 graph embedding for expanding the unknown x up to a resolution of four bits. However, using the binary offset representation we can only get a true precisions of three bits. We solved the simple division problem,

$$x = \frac{y}{m}. \tag{4.1}$$

4.2.1. Basic Division Solver

Table 3 gives an extensive list of tested exact solutions returned by the floating-point annealing algorithm on the D-Wave machine using the K_4 graph embedding with an effective binary resolution of 3, corresponding to the multiples of 0.25 in the range $[-1, 1]$. The "Ground State" column lists the raw binary vector solutions, corresponding to the expansion in Equation (2.5). It is easy to check from Equations (2.5) and (2.4) that these give the floating-point solutions found in the corresponding "D-Wave Solution" column. In all of these cases, values of $\alpha \geq 0.5$ yielded the solution exactly; however, α is set to 20.0 here because that gives a better approximate solution for the inexact divisions, and faster convergence for the iterated divisions. It does not change the solutions for the exact cases.

Table 4 lists some illustrative division problems on $[-1, 1]$ that do not have solutions which are multiples of ±0.25, and they are therefore not solved exactly by the quantum annealing algorithm to 3 bits of resolution. Note that the energies are different for the ground states because the Hamiltonians are somewhat different for these problems. The "rounding"

TABLE 3 | Exact quantum annealed division problems to 3-bit resolution.

y	m	x, Exact	x, D-Wave	Ground state	Energy	α
DIVISION PROBLEMS WITH EXACT D-WAVE SOLUTIONS						
1.00	1.0	1.00	1.00	[1,0,0,0]	−2.0	20.0
0.50	0.5	1.00	1.00	[1,0,0,0]	−2.0	20.0
1.00	−1.0	−1.00	−1.00	[0,0,0,0]	0.0	20.0
−1.00	1.0	−1.00	−1.00	[0,0,0,0]	0.0	20.0
0.50	−0.5	−1.00	−1.00	[0,0,0,0]	0.0	20.0
−0.50	0.5	−1.00	−1.00	[0,0,0,0]	0.0	20.0
0.75	1.0	0.75	0.75	[0,1,1,1]	−1.53125	20.0
−0.75	1.0	−0.75	−0.75	[0,0,0,1]	−0.03125	20.0
0.75	−1.0	−0.75	−0.75	[0,0,0,1]	−0.03125	20.0
0.50	1.0	0.50	0.50	[0,1,1,0]	−1.125	20.0
−0.50	1.0	−0.50	−0.50	[0,0,1,0]	−0.125	20.0
0.50	−1.0	−0.50	−0.50	[0,0,1,0]	−0.125	20.0
0.25	1.0	0.25	0.25	[0,1,0,1]	−0.78125	20.0
−0.25	1.0	−0.25	−0.25	[0,0,1,1]	−0.28125	20.0
0.25	−1.0	−0.25	−0.25	[0,0,1,1]	−0.28125	20.0
0.25	0.5	0.50	0.50	[0,1,1,0]	−1.125	20.0
−0.25	0.5	−0.50	−0.50	[0,0,1,0]	−0.125	20.0
0.25	−0.5	−0.50	−0.50	[0,0,1,0]	−0.125	20.0
0.00	± 1.00	0.00	0.00	[0,1,0,0]	−0.5	20.0
0.00	± 0.75	0.00	0.00	[0,1,0,0]	−0.5	20.0
0.00	± 0.50	0.00	0.00	[0,1,0,0]	−0.5	20.0
0.00	± 0.25	0.00	0.00	[0,1,0,0]	−0.5	20.0

TABLE 4 | "Rounded" quantum annealed division solutions to 3-bit resolution.

y	m	x, Exact	x, D-Wave	Ground state	Energy	α
DIVISION PROBLEMS WITH APPROXIMATE D-WAVE SOLUTIONS						
0.90	1.0	0.90	1.00	[1,0,0,0]	−1.8	20.0
−0.90	1.0	−0.90	−1.00	[0,0,0,0]	0.0	20.0
0.80	1.0	0.80	0.75	[0,1,0,0]	−1.6875	20.0
−0.80	1.0	−0.80	−0.75	[0,0,0,1]	−0.01875	20.0
0.70	1.0	0.70	0.75	[0,1,0,0]	−1.44375	20.0
−0.70	1.0	−0.70	−0.75	[0,0,0,1]	−0.04374	20.0
0.60	1.0	0.60	0.50	[0,1,1,0]	−1.275	20.0
−0.60	1.0	−0.60	−0.50	[0,0,1,0]	−0.075	20.0
0.40	1.0	0.40	0.50	[0,1,1,0]	−0.975	20.0
−0.40	1.0	−0.40	−0.50	[0,0,1,0]	−0.175	20.0
0.30	1.0	0.30	0.25	[0,1,0,1]	−0.84375	20.0
−0.30	1.0	−0.30	−0.25	[0,0,1,1]	−0.24375	20.0
0.20	1.0	0.20	0.25	[0,1,0,1]	−0.71875	20.0
−0.20	1.0	−0.20	−0.25	[0,0,1,1]	−0.31875	20.0
0.10	1.0	0.10	0.00	[0,1,0,0]	−0.6	20.0
−0.10	1.0	−0.10	0.00	[0,0,1,1]	−0.4	20.0
0.30	0.9	$0.\bar{3}$	0.25	[0,1,0,1]	−0.88542	20.0
−0.30	0.9	$-0.\bar{3}$	−0.25	[0,0,1,1]	−0.21875	20.0
1.0	7.0	$0.14\bar{2}875$	0.25	[0,1,0,1]	−0.64732	20.0
−1.0	7.0	$-0.14\bar{2}875$	−0.25	[0,0,1,1]	−0.36161	20.0

TABLE 5 | Iterated quantum annealed division problems to resolution 1.0×10^{-6}.

y	m	x, Exact	x, D-Wave	α	Iterations
ITERATED DIVISION PROBLEMS ON THE D-WAVE ANNEALER					
0.25	1.0	0.25	0.25	20.0	1
−0.25	1.0	−0.25	−0.25	20.0	1
0.50	1.0	0.50	0.50	20.0	1
−0.50	1.0	−0.50	−0.50	20.0	1
0.75	1.0	0.75	0.75	20.0	1
−0.75	1.0	−0.75	−0.75	20.0	1
0.80	1.0	0.80	0.799999	20.0	5
−0.80	1.0	−0.80	−0.799999	20.0	5
0.70	1.0	0.70	0.700000	20.0	5
−0.70	1.0	−0.70	−0.700000	20.0	5
0.10	1.0	0.10	0.999999	20.0	5
−0.10	1.0	−0.10	−0.999999	20.0	5
0.30	0.9	$0.\bar{3}$	0.333333	20.0	10
−0.30	0.9	$-0.\bar{3}$	−0.333333	20.0	10
1.0	7.0	$0.14\bar{2}875$	0.1248751	20.0	7
−1.0	7.0	$-0.14\bar{2}875$	−0.1248751	20.0	7

here occurs naturally in the quantum annealing algorithm as the annealer settles into the lowest energy ground state that approximates the solution. The last four problems are "challenge" problems for the iterated division solver.

4.2.2. Iterated Division Solver

Table 5 lists a few example division problems returned from the iterated quantum annealing solver. These are problems selected from both **Tables 3**, **4** to illustrate the nature of the solutions returned for both cases. These problems were iterated to an error tolerance of 1.0×10^{-6}. The four "challenge" problems from **Table 4** can now be solved with the iterative method. The ground state is no longer given since the solution is generally the concatenation of multiple binary vectors for every iteration. Instead, the number of iterations is listed in the last column. Note that some of the energies are the same for the solutions of different problems. We have also left out an "Energy" column since it was only calculated for the partial solution from the last iteration.

4.3. Results for Matrix Equations

Note that we have occasionally been somewhat loose in calling this "matrix inversion" since we are technically solving the linear equation, rather than directly inverting the matrices. However, for the problems considered here, we may simply obtain the solutions to the equations using trivial orthonormal eigenvectors, such as $(1, 0)$ and $(0, 1)$, in which case the inverse of the matrix will just be the matrix having those solutions as columns.

The linear equation algorithm was implemented and used to solve several 2 × 2 and 3 × 3 matrices on the D-Wave quantum annealer. Floating-point numbers are represented using the same offset binary representation as was used for the division problems. There are thus 4 qubits per floating-point number. As in the previous section, this gives an effective resolution of 3 bits for floating-point numbers defined on $[-1, 1]$. In this case, however, we employed the normalization technique discussed in the division iteration to allow solutions with positive and negative floating-point numbers with larger magnitudes than 1. But, in these matrix problems we still use solution values with relatively small magnitudes and within an order of magnitude of each other for all solution vector elements. All of the cases shown here are matrix equations with exact solutions, in which case the values of the solution vector elements are multiples of 0.25. We are therefore optimistic that the iterative solver for the matrix inversion could be implemented fairly quickly.

In general, every qubit representing part of a floating-point number may be coupled to every other qubit representing part of the same number. In turn, every logical qubit may be connected to every other logical qubit, which implies that every qubit in the logical qubit representation of the problem, may be coupled to every other logical qubit in the problem. Therefore, the linear solution algorithm is implemented using a K_8 graph embedding to solve 2 × 2 matrix equations, having a 2 dimensional solution vector with 4 qubits per element, and using a K_{12} graph embedding to solve 3 × 3 matrix equations, having a 3 dimensional solution vector with 4 qubits per element.

Most of these solutions involve well-conditioned matrices; however, one does not generally find a feasible solution when using an ill-conditioned matrix. This is illustrated in two cases, one with a of a 2 × 2 matrix another with a 3 × 3 matrix. We were able to obtain the correct solutions by preconditioning these matrices before converting to QUBO form, however the 3 × 3 matrix, still had a nearly degenerate ground state and required a very large chaining penalty α to get the correct solution. This is analyzed and discussed in detail below.

4.3.1. Simple Analytic Problem

Recalling equation (3.1), we shall obtain solutions $\mathbf{x}$ of the following matrix equation,

$$\boldsymbol{M} \cdot \mathbf{x} = \mathbf{Y}, \tag{4.2}$$

using values of $\boldsymbol{M}$ and $\mathbf{Y}$ listed in Matrix Test Problems. Here we present the first two tests as an example. Consider the following matrix:

$$\boldsymbol{M} = \begin{pmatrix} \frac{1}{2} & \frac{3}{2} \\ \frac{3}{2} & \frac{1}{2} \end{pmatrix}. \tag{4.3}$$

We can solve Equation (3.1) for $\boldsymbol{M}$, with the following two Y vectors:

$$\mathbf{Y_1} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \ \mathbf{Y_2} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{4.4}$$

The exact solutions are

$$\mathbf{x_1} = \begin{pmatrix} -\frac{1}{4} \\ \frac{3}{4} \end{pmatrix}, \ \mathbf{x_2} = \begin{pmatrix} \frac{3}{4} \\ -\frac{1}{4} \end{pmatrix}. \tag{4.5}$$

We may obtain M^{-1} simply as

$$\boldsymbol{M^{-1}} = \begin{pmatrix} -\frac{1}{4} & \frac{3}{4} \\ \frac{3}{4} & -\frac{1}{4} \end{pmatrix} \tag{4.6}$$

In the next section we summarize all of the solutions obtained by the DWave for all of our test problems.

4.3.2. QUBO Solution Results

The solutions for the 2 × 2 linear solves are presented in **Table 6**. Notice that all of the test problems are presented with $\alpha = 20.0$ except for the last two. This was done to illustrate the effect of preconditioning for the ill-conditioned case. However, for this example, the difference disappeared above $\alpha = 2.0$, and both began to give incorrect answers below $\alpha = 1.5$. This is in contrast to the 3 × 3 matrix solution cases, which are evidently more sensitive to condition number than the 2 × 2 tests.

The 3 × 3 matrix solutions are presented in **Table 7**. Note that we have not included the 12 digit binary ground states here because they take up too much room in the table and are not particularly illuminating. Problems 2(f) and 2(g) are the ill-conditioned matrix test and its preconditioned equivalent. For $\alpha = 20.0$ both versions of the poorly-conditioned problem gave only D-Wave solutions with broken chains. One only begins to get solutions with unbroken chains at a value of α above 1000, but those solutions are generally wrong and basically random until one gets to a very high α. We discuss this in greater detail in the following section.

TABLE 6 | 2 × 2 Matrix equation solutions to 3-bit resolution.

Test	Exact solution	D-wave solution	Ground state	Energy	α
DIVISION PROBLEMS WITH APPROXIMATE D-WAVE SOLUTIONS					
1(*a*)	(−0.25, 0.75)	(−0.25, 0.75)	[0,0,1,1,0,1,1,1]	−2.167	20.0
1(*b*)	(0.75, −0.25)	(0.75, −0.25)	[0,1,1,1,0,0,1,1]	−2.167	20.0
1(*c*)	(1.00, 1.00)	(1.00, 1.00)	[1,0,0,0,1,0,0,0]	−0.444	20.0
1(*d*)	(−1.00, 1.00)	(−1.00, 1.00)	[0,0,0,0,1,0,0,0]	−1.889	20.0
1(*e*)	(1.00, −1.00)	(1.00, −1.00)	[1,0,0,0,0,0,0,0]	−1.650	20.0
1(*f*)	(1.00, 0.00)	(1.00, 0.00)	[0,0,0,0,0,1,0,0]	−2.125	20.0
1(*g*)	(0.25, −0.50)	(0.25, −0.50)	[0,1,0,1,0,0,1,0]	−0.925	20.0
1(*h*)	(0.25, 0.25)	(0.25, 0.25)	[0,1,0,1,0,1,0,1]	−2.03125	20.0
1(*i*)	(2.00, 1.00)	(2.00, 1.00)	[1,1,0,0,1,0,0,0]	−2.450126	20.0
1(*j*)	(2.00, 1.00)	(2.00, 1.00)	[1,1,0,0,1,0,0,0]	−2.532545	20.0
1(*i*)	(2.00, 1.00)	(2.50, 0.75)	[1,1,1,0,0,1,1,1]	−2.887689	1.5
1(*j*)	(2.00, 1.00)	(2.00, 1.00)	[1,1,0,0,1,0,0,0]	−2.951557	1.75

TABLE 7 | 3 × 3 matrix equation solutions to 3-bit resolution.

Test	Exact solution	D-wave solution	Energy	α
DIVISION PROBLEMS WITH APPROXIMATE D-WAVE SOLUTIONS				
2(*a*)	(0.25, −0.5, 1.0)	(0.25, −0.5, 1.0)	−15.5625	20.0
2(*b*)	(0.25, −0.5, 0.0)	(0.25, −0.5, 0.0)	−12.5625	20.0
2(*c*)	(0.25, 0.0, −0.5)	(0.25, 0.0, −0.5)	−13.5	20.0
2(*d*)	(1.0, 0.25, −0.5)	(1.0, 0.25, −0.5)	−15.6875	20.0
2(*e*)	(0.0, 0.25, −0.5)	(0.0, 0.25, −0.5)	−12.75	20.0
2(*f*)	(0.0, 0.25, −0.75)	broken chains	N/A	20.0
2(*g*)	(0.0, 0.25, −0.75)	broken chains	N/A	20.0
2(*f*)	(0.0, 0.25, −0.75)	(1.75, 1.25, 0.75)	−58.188	2200.0
2(*g*)	(0.0, 0.25, −0.75)	(0.0, 0.25, −0.75)	−557.437	2200.0

4.4. Discussion

The algorithms described here generally worked quite well for these small test cases, with the exception of the ill-conditioned 3 × 3 matrix. The ill-conditioned cases clearly demonstrate not only the limitations of quantum annealing applied solving linear equations, but the limitations of quantum annealing in general. Consider the two ill-conditioned tests presented here. When translated to a QUBO problem, the Hamiltonian spectra for these tests contain many energy eigenvalues very close to the ground state energy. When these are embedded within a larger graph of physical qubits they result in a very nearly degenerate ground state, typically with thousands of states having energies within the energy uncertainty of the ground state over the annealing time, τ, given by

$$\Delta E = \frac{\hbar}{\tau}. \tag{4.7}$$

Consider a set of excited states with energy, E_n for $n > 0$, with $n = 0$, corresponding to the ground state with energy denoted by E_0, and with E_n ordered by energy. The quantum annealer near the ground state evolves adiabatically whenever

$$E_1 - E_0 \gg \frac{\hbar}{\tau} \tag{4.8}$$

This is the adiabatic condition for quantum time evolution [6]. However, when this condition is badly violated, which can occur dynamically since the instantaneous energies (eigenvalues of H) are time dependent, the time evolution for the system near the ground state deviates significantly from adiabatic behavior, resulting in a highly mixed superposition of those eigenstates states close in energy to the ground state. Now, the energy spectra corresponding to poorly conditioned matrices have a large number of eigenstates sufficiently near the ground state to strongly violate the adiabaticity condition. Furthermore, these states, in general, will have no correlation to the solution encodings for any particular problem (e.g., the offset binary floating-point representation). For example, they are not, in general, related in any meaningful way to Hamming distance. Therefore, these problems effectively cannot resolve the true ground state and tend to give nearly random lowest energy "solutions" when the final state is measured on any individual annealing run. Since there are so many of these states for ill-conditioned problems, a very large number of "reads" (annealing runs) may be have to be specified to sufficient sample the solution space to find the true ground state. Thus, we determined that the condition number of a matrix has a strong effect on the ability to solve a linear equation using a quantum annealer, as it influences the shape of the energy surface near the ground state.

The preconditioning method we used was very simple and was probably too crude to be practical for arbitrary matrices. However, the intent here was simply to test the effects of preconditioning on the quantum annealing solutions. We have been studying this issue and believe it may be possible to precondition such problems to solve them more efficiently on a quantum annealing machine. We suspect that a related preconditioning method may be applicable to more general QUBO problems suffering from similar spectral density pathologies in order to better separate the ground state energy, thereby allowing more practical solution on a quantum annealer. This is work in progress. We plan to further develop and test those ideas in the future.

MATRIX TEST PROBLEMS

The problems we solved to test our quantum annealing algorithm to solve equation (3.1) are listed below. Note that, although the QUBO B_{ij} matrix is symmetric by construction, the matrix M need not be symmetric.

1. Test Problems with 2 × 2 Matrices

Test 1(a):

$$\boldsymbol{M} = \begin{pmatrix} 0.5 & 1.5 \\ 1.5 & 0.5 \end{pmatrix}, \ \mathbf{Y} = \begin{pmatrix} 1.0 \\ 0.0 \end{pmatrix}, \ \mathbf{x} = \begin{pmatrix} -0.25 \\ 0.75 \end{pmatrix}$$

Test 1(b):

$$\boldsymbol{M} = \begin{pmatrix} 0.5 & 1.5 \\ 1.5 & 0.5 \end{pmatrix}, \ \mathbf{Y} = \begin{pmatrix} 0.0 \\ 1.0 \end{pmatrix}, \ \mathbf{x} = \begin{pmatrix} 0.75 \\ -0.25 \end{pmatrix}$$

Test 1(c):

$$\boldsymbol{M} = \begin{pmatrix} 2.0 & -1.0 \\ -0.5 & 0.5 \end{pmatrix}, \ \mathbf{Y} = \begin{pmatrix} 1.0 \\ 0.0 \end{pmatrix}, \ \mathbf{x} = \begin{pmatrix} 1.0 \\ 1.0 \end{pmatrix}$$

Test 1(d):

$$\boldsymbol{M} = \begin{pmatrix} 1.0 & 2.0 \\ 0.5 & 0.5 \end{pmatrix}, \ \mathbf{Y} = \begin{pmatrix} 1.0 \\ 0.0 \end{pmatrix}, \ \mathbf{x} = \begin{pmatrix} -1.0 \\ 1.0 \end{pmatrix}$$

Test 1(e):

$$\boldsymbol{M} = \begin{pmatrix} 3.0 & 2.0 \\ 2.0 & 1.0 \end{pmatrix}, \ \mathbf{Y} = \begin{pmatrix} 1.0 \\ 1.0 \end{pmatrix}, \ \mathbf{x} = \begin{pmatrix} 1.0 \\ -1.0 \end{pmatrix}$$

Test 1(f):

$$\boldsymbol{M} = \begin{pmatrix} 1.0 & 0.5 \\ 1.0 & -0.5 \end{pmatrix}, \ \mathbf{Y} = \begin{pmatrix} 1.0 \\ 1.0 \end{pmatrix}, \ \mathbf{x} = \begin{pmatrix} 1.0 \\ 0.0 \end{pmatrix}$$

Test 1(g):

$$\boldsymbol{M} = \begin{pmatrix} 0.0 & -2.0 \\ -2.0 & -1.5 \end{pmatrix}, \ \mathbf{Y} = \begin{pmatrix} 1.0 \\ 0.25 \end{pmatrix}, \ \mathbf{x} = \begin{pmatrix} 0.25 \\ -0.5 \end{pmatrix}$$

Test 1(h):

$$\boldsymbol{M} = \begin{pmatrix} 0.0 & -2.0 \\ -2.0 & -1.5 \end{pmatrix}, \ \mathbf{Y} = \begin{pmatrix} -0.5 \\ -0.875 \end{pmatrix}, \ \mathbf{x} = \begin{pmatrix} 0.25 \\ 0.25 \end{pmatrix}$$

Test 1(i): Ill-conditioned problem with $\kappa \approx 25$

$$\boldsymbol{M} = \begin{pmatrix} 1.0 & 2.0 \\ 2.0 & 3.999 \end{pmatrix}, \ \mathbf{Y} = \begin{pmatrix} 4.0 \\ 7.999 \end{pmatrix}, \ \mathbf{x} = \begin{pmatrix} 2.0 \\ 1.0 \end{pmatrix}$$

Test 1(j): Pre-conditioned version of 1(i) with $\kappa = 5.0$

$$\boldsymbol{M} = \begin{pmatrix} 1.80026 & 1.6019 \\ 1.6019 & 4.19974 \end{pmatrix}, \ \mathbf{Y} = \begin{pmatrix} 5.2007 \\ 7.40013 \end{pmatrix}, \ \mathbf{x} = \begin{pmatrix} 2.0 \\ 1.0 \end{pmatrix}$$

2. Matrix Problems with 3 × 3 Matrices

Test 2(a):

$$\boldsymbol{M} = \begin{pmatrix} 0.0 & -2.0 & 0.0 \\ -2.0 & 1.5 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{pmatrix}, \ \mathbf{Y} = \begin{pmatrix} 1.0 \\ 0.25 \\ 1.0 \end{pmatrix}, \ \mathbf{x} = \begin{pmatrix} 0.25 \\ -0.5 \\ 1.0 \end{pmatrix}$$

Test 2(b):

$$\boldsymbol{M} = \begin{pmatrix} 0.0 & -2.0 & 0.0 \\ -2.0 & 1.5 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{pmatrix}, \ \mathbf{Y} = \begin{pmatrix} 1.0 \\ 0.25 \\ 0.0 \end{pmatrix}, \ \mathbf{x} = \begin{pmatrix} 0.25 \\ -0.5 \\ 0.0 \end{pmatrix}$$

Test 2(c):

$$\boldsymbol{M} = \begin{pmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & -2.0 \\ 0.0 & -2.0 & -1.5 \end{pmatrix}, \ \mathbf{Y} = \begin{pmatrix} 1.0 \\ 0.0 \\ 0.25 \end{pmatrix}, \ \mathbf{x} = \begin{pmatrix} 0.25 \\ 0.0 \\ -0.5 \end{pmatrix}$$

Test 2(d):

$$\boldsymbol{M} = \begin{pmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & -2.0 \\ 0.0 & -2.0 & -1.5 \end{pmatrix}, \ \mathbf{Y} = \begin{pmatrix} 1.0 \\ 1.0 \\ 0.25 \end{pmatrix}, \ \mathbf{x} = \begin{pmatrix} 1.0 \\ 0.25 \\ -0.5 \end{pmatrix}$$

Test 2(e):

$$\boldsymbol{M} = \begin{pmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & -2.0 \\ 0.0 & -2.0 & -1.5 \end{pmatrix}, \ \mathbf{Y} = \begin{pmatrix} 0.0 \\ 1.0 \\ 0.25 \end{pmatrix}, \ \mathbf{x} = \begin{pmatrix} 0.0 \\ 0.25 \\ -0.5 \end{pmatrix}$$

Test 2(f): Ill-conditioned problem with $\kappa \approx 78$

$$\boldsymbol{M} = \begin{pmatrix} -4.0 & 6.0 & 1.0 \\ 8.0 & -11.0 & -2.0 \\ -3.0 & 4.0 & 1.0 \end{pmatrix}, \ \mathbf{Y} = \begin{pmatrix} 0.75 \\ -1.25 \\ 0.25 \end{pmatrix},$$

$$\mathbf{x} = \begin{pmatrix} 0.0 \\ 0.25 \\ -0.75 \end{pmatrix}$$

Test 2(g): Pre-conditioned version 2(g) with $\kappa \approx 1$

$$\boldsymbol{M} = \begin{pmatrix} 6.1795 & 11.8207 & 2.0583 \\ 15.673 & -7.56717 & -3.8520 \\ -5.6457 & 7.96872 & 15.9418 \end{pmatrix}, \ \mathbf{Y} = \begin{pmatrix} 1.4114 \\ 0.9972 \\ 9.9643 \end{pmatrix},$$

$$\mathbf{x} = \begin{pmatrix} 0.0 \\ 0.25 \\ -0.75 \end{pmatrix}$$

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

We received funding for this work from the *ASC Beyond Moore's Law Project* at LANL.

ACKNOWLEDGMENTS

We would like to thank Andrew Sornborger, Patrick Coles, Rolando Somma, and Yiğit Subaşı for a number of useful conversations.

REFERENCES

1. Harrow A, Hassidim A, Lloyd S. Quantum algorithm for linear systems of equations. *Phys Rev Lett.* (2009) 103:150502. doi: 10.1103/PhysRevLett.103.150502
2. Barz S, Kassal I, Ringbauer M, Ole Lipp Y, Dakic B, Aspuru-Guzik A, et al. Solving systems of linear equations on a quantum computer. *Sci. Rep.* (2014) 4:115. doi: 10.1038/srep06115
3. Cai XD, Weedbrook C, Su ZE, Chen MC, Gu M, Zhu MJ, et al. Experimental quantum computing to solve systems of linear equations. *arXiv:1302.4310v2* (2013). doi: 10.1103/PhysRevLett.110.230501
4. Farhi E, Goldstone J, Gutmann S, Sipser M. Quantum computation by adiabatic evolution. *arXiv:000110* (2000).
5. Ising E. Beitrag zur Theorie des Ferromagnetismus *Z Phys.* (1925) 31:253–18.
6. Albert M. *Chapter XVII: Quantum Mechanics.* New York, NY: Dover Publications (1999).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Physics

BRIEF RESEARCH REPORT
published: 16 November 2020
doi: 10.3389/fphy.2020.525731

The Quantitative Comparison Between the Neuronal Network and the Cosmic Web

F. Vazza[1,2,3]* *and A. Feletti*[4,5]

[1]Dipartimento di Fisica e Astronomia, Universitá di Bologna, , Bologna, Italy, [2]Hamburger Sternwarte, Hamburg, Germany, [3]Istituto di Radio Astronomia, INAF, Bologna, Italy, [4]Institute of Neurosurgery, Department of Neurosciences, Biomedicine, and Movement Sciences, University of Verona, Verona, Italy, [5]Azienda Ospedaliera-Universitaria di Modena, Modena, Italy

We investigate the similarities between two of the most challenging and complex systems in Nature: the network of neuronal cells in the human brain, and the cosmic network of galaxies. We explore the structural, morphological, network properties and the memory capacity of these two fascinating systems, with a quantitative approach. In order to have an homogeneous analysis of both systems, our procedure does not consider the true neural connectivity but an approximation of it, based on simple proximity. The tantalizing degree of similarity that our analysis exposes seems to suggest that the self-organization of both complex systems is likely being shaped by similar principles of network dynamics, despite the radically different scales and processes at play.

Keywords: cosmology: theory, neuroscience, network analysis, complex systems, large-scale structure formation

OPEN ACCESS

Edited by:
Francesco Caravelli,
Los Alamos National Laboratory (DOE), United States

Reviewed by:
Mark Neyrinck,
University of the Basque Country, Spain
Marius Cautun,
Leiden University, Netherlands
Michael Silver,
University of California, Berkeley, United States

****Correspondence:***
F. Vazza
franco.vazza2@unibo.it

Specialty section:
This article was submitted to Interdisciplinary Physics, a section of the journal Frontiers in Physics

Received: *16 January 2020*
Accepted: *28 September 2020*
Published: *16 November 2020*

Citation:
Vazza F and Feletti A (2020) The Quantitative Comparison Between the Neuronal Network and the Cosmic Web. Front. Phys. 8:525731. doi: 10.3389/fphy.2020.525731

INTRODUCTION

Central to our vision of Nature are two fascinating systems: the network of neurons in the human brain and the cosmic web of galaxies.

The human brain is a complex temporally and spatially multiscale structure in which cellular, molecular and neuronal phenomena coexist. It can be modeled as a hierarchical network (i.e., “the human connectome” [1]), in which neurons cluster into circuits, columns, and different interconnected functional areas. The structure of the neuronal network allows the linking between different areas, all devoted to process specific spatiotemporal activities over their neurons, forming the physical and biological basis of cognition [e.g., Ref. 2]. Some of major challenges of contemporary neuroscience are to disentangle the structure of the connectome (e.g., the complete map of the neural connections in a brain), to understand how this structure can produce complex cognitive functions, and to define the role of glial cells and of the microenvironment in the interneuronal physiology.

The Universe, according to the large collection of telescope data gathered over many decades, seems to be reasonably well described by a “consensus” physical model called the ΛCDM model (Lambda Cold Dark Matter), which accounts for gravity from ordinary and dark matter (i.e., very weakly interacting particles), for the expanding space-time described by General Relativity, and for the anti-gravitational energy associated to the empty space, called the “dark energy”. Such model presently gives the best picture of how cosmic structures have emerged from the expanding background and have formed the cosmic web [e.g., Refs. 3 and 4]. The most important building blocks of the cosmic web are self-gravitating dark matter dominated halos, in which ordinary matter has collapsed to form galaxies (and all stars within them). The initial distribution of matter density fluctuations was early amplified by the action of gravity, and has developed into larger groups or clusters of galaxies, filaments, matter sheets, and voids, in a large-scale web in all directions in space.

Among the main challenges that cosmology still faces, are the physical nature of dark energy, the composition of dark matter (or the realm of alternative scenarios for it), the apparent tension between different measurements of the expansion rate of the Universe, the exact sequence of processes responsible for the variety of galaxy morphology and their co-evolution with supermassive black holes [e.g., Ref. 5, for a recent review].

Although the relevant physical interactions in the above two systems are completely different, their observation through microscopic and telescopic techniques have captured a tantalizing similar morphology, to the point that it has often been noted that the cosmic web and the web of neurons look alike [e.g., Refs. 6 and 7].

In this work, we apply methods from cosmology, neuroscience, and network analysis to explore this thought-provoking question quantitatively for the first time, to our knowledge.

MATERIALS AND METHODS

Immunohistochemistry and Microscopy

We analyzed several independent samples of cerebral and of cerebellar human cortex were formalin-fixed and paraffin-embedded [8], sampling slices of depth 4 μm, with magnification factors of 4×, 10× and 40×. Neurofilaments were labeled using the Neurofilament (2F11) Mouse Monoclonal Antibody (Ventana/CellMarque/Roche). Samples were automatically processed by Ventana BenchMark Ultra Immunostainers. A Nikon eclipse 50i microscope was then used to visualize the samples. Magnifications larger than 40× was avoided in order to obtain a better optical depth resolution, as well as to minimize the non-linear response of the optic microscopy.

Cosmological Simulations

We used synthetic samples of the cosmic web from a high-resolution (2400^3 cells and dark matter particles) simulation of a cubic 100^3 Mpc^3 cosmic volume (1 Mpc = $3.085 \cdot 10^{24}$ cm), performed with the grid code ENZO [9] as detailed in Ref. 10. The simulation produces a realistic distribution of dark matter, ordinary matter, and magnetic fields at the present epoch. In order to mimic the "slicing" procedure of brain tissues, we produced 12 different thin slices (with thickness 25 Mpc) from the simulated volume, by extracting four slices in perpendicular directions with respect to each of the independent axes of the simulation. We give public access to our cosmic web images, as well as to the brain samples and to the images of other natural networks discussed below at this URL https://cosmosimfrazza.myfreesites.net/cosmic-web-and-brain-network-datasets.

RESULTS

Absolute Numbers, Internal Proportions, and Composition

We first quote data available from the literature, which allow us a first sketchy comparison of the absolute sizes of both systems. The radius of the observable Universe is $R_U \sim 13.9$ Gpc [11]. The extrapolation of recent observations posits that a total of $N_g \sim 2.6 \cdot 10^{12}$ galaxies may be present in within the sphere of the observable Universe [12], with up to $\sim 5 \cdot 10^{10}$ galaxies with masses equal or larger to the one of the Milky Way. The largest clusters of galaxies total a mass exceeding 10^{15} solar masses (1 solar mass = $1.989 \cdot 10^{33}$ g). Long filaments of ordinary and dark matter, with extension of several tens of Megaparsecs, connect clusters and groups of galaxies and are separated by mostly empty space [e.g., Ref. 4].

According to recent estimates, the adult human brain contains $N_{\rm neu} \approx 8.6 \pm 0.8 \cdot 10^{10}$ neurons in total, and almost an equal number of non-neuronal cells. Only $\sim 20 - 25\%$ of all neurons are located in the cortical gray matter (representing $\sim 80\%$ percent of brain mass), while the cerebellum has about $\sim 6.9 \cdot 10^{10}$ neurons ($\sim 80\%$ of brain neurons) [13, 14].

It can be noticed that the two systems are organized in well defined networks, with $\sim 10^{10} - 10^{11}$ nodes interconnected through filaments (if we consider as nodes all galaxies with masses comparable or larger to that of the Milky Way, see above), whose typical extent is only a tiny fraction ($\leq 10^{-3}$) of their host system size. Also, galaxies and neurons have a typical scale radius, which is only a very small fraction of the typical length of filaments they are connected to. Moreover, available data suggest that the flow of information and energy in the two networks is mostly confined to $\leq 25\%$ of the mass/energy content of each system.

In the case of the Universe we refer to the present-day composition, based on Planck Collaboration et al. [15], as the relative energy distribution is a function of time in the ΛCDM cosmological model; for the human brain, we referred to the published researches about human brain composition [e.g., Refs. 16 and 17].

In summary: 1) the brain is composed by water ($77 - 78\%$), lipids ($10 - 12\%$), proteins (8%), carbohydrates (1%), soluble organic substances (2%), salt (1%); 2) the Universe is made for a $73 \sim \%$ by Dark Energy (a scalar energy field of the empty space), for a 22.5% by Dark Matter, for 4.4% by ordinary baryonic matter and for less than $\leq 0.1\%$ by photons and neutrinos.

Strikingly, in both cases $\sim 75\%$ of the mass/energy distribution is made of an apparently passive material, that permeates both systems and has an only indirect role in their internal structure: water in the case of the brain, and dark energy in cosmology, which to a large extent does not affect the internal dynamics of cosmic structures [e.g., Ref. 18].

Morphological Comparison

Small samples of the human cerebral and cerebellar cortex were harvested during corticectomy to approach subcortical tumors (**Section 2.1**). The neuronal cells have been then stained with clone 2F11 monoclonal antibody against neurofilaments, which are neuron-specific intermediate filaments in the cytoplasm of neurons that provide structural support to the neuronal cytoskeleton, along with microtubules and microfilaments. It has been shown that the number, spacing, and areal density of neurofilaments in neurons are measures with a strong dependency on axon caliber [e.g., Refs. 19–21]. Although also microtubules density depends on axon caliber, it has been shown

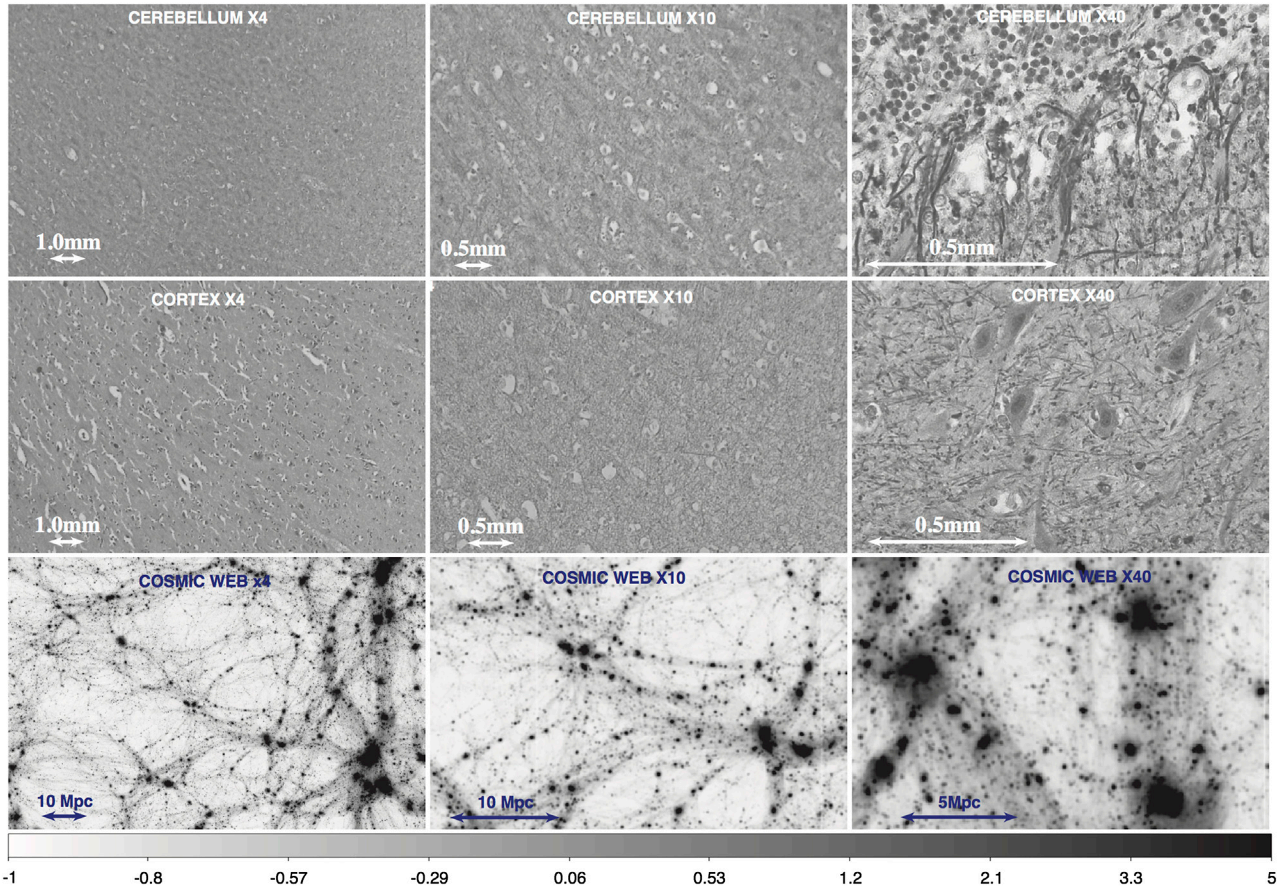

FIGURE 1 | Maps of normalized density contrast, $\delta\rho$ for examples of slices of the cerebellum (top row), of the brain cortex (middle row) and of the dark matter distribution of the cosmic web (lower row), for various levels of magnification. The spectral and network analysis presented in this work will mostly focus on 40× samples as in the right panels.

that microtubules often form small clusters in the vicinity of membranous organelles [22]. For this reason we consider neurofilaments might be more homogeneously arranged in the neuron, and likely to be a better target to visualize the spatial distribution of neurons in the slices. For the cosmic web, we analyzed each one of the 12 thin slices from the simulated volume (**Section 2.2**), to assess the effect of cosmic variance. Such 2-dimensional approach mimics what is done with brain samples, and due to the large degree of isotropy of the cosmic web on such large scales this approach can also be used to readily translate our statistics into the 3-dimensional case.

Figure 1 gives an overview of the details of structures observed at various scales (from 4×, 10× and 40× magnifications in the case of brain tissues, and on corresponding steps in zoom in the case of the cosmic web) in our dataset. Especially on large scales, the various samples do not show any spectacular degree of similarity. In particular, the predominant role of the large overdensities marked by clusters of galaxies is evident in the cosmic web sample, while the finer structure of neurofilaments in the brain samples produces richer small-scale patterns. At the highest magnification achieved in our brain slices, however, the refined network of neuronal bodies and of filaments start presenting some similarity with the cosmic web on ≤ 20 Mpc scales. When focusing on histological slices, some variability can be noticed that depends on different neuronal subtypes in brain and cerebellar cortex. In the first slice, small neurons in the granular layer are shown, along with the transition to the gangliar layer with some Purkinje cells at the bottom of the picture. Conversely, the second slice depicts large pyramidal cells interspersed with small neuronal cells of the brain gray matter (granular cells).

We will use in this work statistical tools to 1) compare the distribution of structure across the entire continuum range of spatial scales of our samples, also compared to other natural complex systems (**Section 3.3**); 2) measure the properties of connectivity between nodes in the network, estimating the tendency to form highly clustered configurations (see **Section 3.4**).

Spectral Analysis

We used here a statistics commonly used in cosmology: the density power spectrum, $P(k)$, which directly measures the

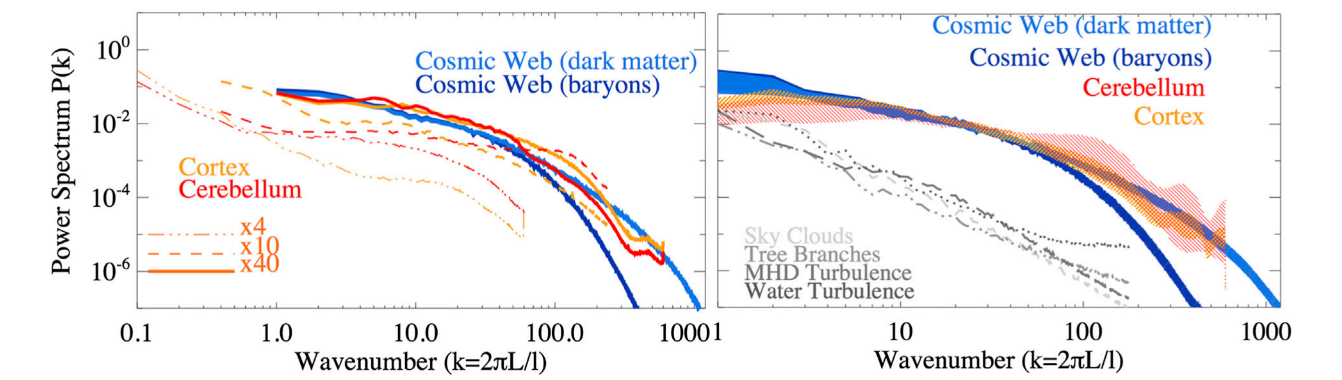

FIGURE 2 | Left: power spectra of density fluctuations in all samples (in the case of brain samples, spectra from slices of different size and magnification factors are shown). Right: power spectra of density fluctuations (in the case of brain tissues, only the 40× slices are used here), with shaded areas enclosing the entire distribution of spectra for all samples (12 slices for cosmic web samples and four for the brain samples). Comparison power spectra from other samples natural systems are shown with gray lines.

contributions of different spatial frequencies, $k = 2\pi L/l$ (where l is the spatial scale and L is the maximum scale of each system), to the density contrast, defined as $\delta\rho = \rho/\overline{\rho} - 1$, where ρ is the density and $\overline{\rho}$ is the average density of each sample. We measured $P(k)$ for our 2-dimensional samples, by decomposing $\delta\rho$ into a series of discrete spatial frequencies, $\delta(\overrightarrow{k})$: $\langle\delta(\overrightarrow{k})\delta(\overrightarrow{k'})\rangle = 2\pi^3 P(k)\delta_D^2(\overrightarrow{k} + \overrightarrow{k'})$, where δ_D^2 is the 2-dimensional Dirac delta function.

In the case of the cosmological simulation, $\langle\rho\rangle$ is uniquely constrained by the initial conditions of the simulation, while in the sample of the cortex and cerebellum we define it based on the average measured within the sample itself. While accurately knowing the local density contrast is trivial in the simulation, it shall be noticed that a precise mapping of the recorded pixel intensity to a projected matter density is far from trivial in microscope observation, due to the non-linear response of the microscopic imaging process. As noted in **Section 2.1**, our choice of using very thin tissue samples and a magnification not higher than ×40 is indeed motivated by the goal of minimizing the non-linear response of the optic microscopy, by keeping the optical depth small compared to the aperture of the image. For this reason, $\delta\rho$ in our brain samples strictly is a measure of the contrast of optical absorptions along the line of sight, which we assume to be a proxy for the density contrast for the sake of comparing with cosmological samples. We applied standard Fast Fourier Transform with periodic boundary conditions to compute the power spectra of cosmic web samples (as the domains are truly periodic), while in the case of the brain samples we used a standard zero-padding technique to embed the observed samples into a 2×2 larger and empty area, and applied apodisation at the interfaces between the empty area and the data, in order to minimize spurious edge effects, as commonly done in simulations [e.g., Ref. 23].

The resulting power spectra are shown in **Figure 2**. It shall be stressed that power spectra are free to be slid horizontally in the plot, in the sense that the reference scale L related to $k = 1$ is decided a-posteriori. In the following, after a preliminary comparison of spectra we adjusted the horizontal scale so that $k = 1$ corresponds to $L = 1.6$ mm in brain samples, and to $L = 100$ Mpc in the cosmic web. This corresponds to a scaling ratio of $1.875 \cdot 10^{27}$ between the two systems. The amplitude of spectra in the vertical direction, instead, is self-normalized to the total variance of $\delta\rho$ within each sample. As a consequence, the brain samples are differently normalized at $k = 1$, since when a lower magnification is used and larger spatial scales are sampled, ≤ 1.6 mm scales contribute proportionally less to the variance of $\delta\rho$ within the entire sample. In the first panel, we compare the spectra of a random cosmic web slice with random brain slices obtained with different magnifications. The comparison strikingly shows (in line with what suggested on **Figure 1**) that a remarkable similarity with spectra is obtained when comparing ≤ 1 mm scales in brain samples to ≤ 100 Mpc scales of the cosmic web. Most of the neuronal cells observed in our cerebellar samples are granule cells, with somata having a $\sim 5\,\mu$m diameter, while their dendrites have dendrites with a typical $\sim 13\,\mu m$ length. The axon length (although variable depending on the cortical areas) is on average in the range of several millimeters [e.g., Ref. 24]. Considering that the slices used for microscopic inspection most often are not parallel to the axis of axons, it is likely that fragments of these fibers around $\sim 1 - 2$ mm in length are visible in the slices. Therefore, the excess power of neural power spectra in this spatial regime reflects the abundance of structures with this typical size distribution.

On the other hand, the fluctuations measured on $\geq 1 - 2$ mm scales in brain samples present a steeper spectral shape than in cosmic web spectra. For this reason, in the remainder of the analysis we focused on datasets of 40× brain samples for a close comparison with cosmic web slices. In the second panel, we show $P(k)$ both for the dark matter and gas distribution of all slices, which are almost identical on large scales ($\geq$ Mpc) and more diffuse on smaller scales due to smoothing effect of gas pressure. As for the cosmic web spectra, we show the envelope containing all spectra of all 40× samples with shaded areas. We find a large agreement across nearly ~ 2 decades in spatial scales. The similarity between the cerebellum on $0.01 - 1.6$ mm scales and the dark matter distribution of the cosmic web on $1 - 10^2$ Mpc scales is remarkable. On smaller scales, the cortex sample displays significant more power than the cerebellum, owing to the distribution of small neurons in the granular layer described above, while the baryon distribution of the cosmic web has less power, due to the (well-known) effect of gas pressure in smoothing out the fluctuations of baryon gas density on small enough scale for hydrodynamical effects to be relevant. In all cases we measure broken power laws, unlike what is expected for (simpler) fractal distributions [e.g., Ref. 25]. This is in line with several works, which have shown that at small scales, $r \le 20$ Mpc the galaxy correlation function scales as $\propto r^{-1}$ (where r is the spatial scale in the 2-point correlation function) while on larger scales the density only weakly (logarithmically) depends on the system size [e.g., Refs. 26 and 27].

Lastly, we produced control power spectra for other randomly drawn samples of natural networks (sky clouds, tree branches, water turbulence, and magneto-hydrodynamic turbulence - all available at https://cosmosimfrazza.myfreesites.net/cosmic-web-and-brain-network-datasets), with the goal of double checking that our method is not biased to produce similarity between truly different physical systems. As shown by the gray lines in the right panel of **Figure 2**, such systems display a more regular power-law spectral behavior, clearly at variance with what found in the main networks analyzed in this work - even if in the latter case we did not perform a full analysis across the entire dynamical range of such systems, looking for the emergence of possible spectral features as in the case of the brain and the cosmic samples.

However, power spectra are blind to phase correlations in the continuous field, hence two morphologically different distributions can still produce similar spectra [28]. In the following section we will thus also rely on non-spectral methods to compare the different samples.

Network Analysis

Network science have proliferated into various physical disciplines, including neuroscience [e.g., Ref. 29–32] as well as cosmology [e.g., Refs. 33 and 34]. Complex network analysis can partially soften the problem of not having perfectly consistent density estimators across our samples, in the sense that defining the nodes of the various networks is less sensitive to the exact mapping details of the images. We focus here on two simple network parameters commonly used in graph theory and network analysis [e.g., Refs. 35 and 36]. The first is the degree centrality, C_d, which measures the degree of connectivity of a network within the localized area (determined by a maximum linking length, $l_{\rm link}$):

$$C_d(j) = \frac{k_j}{n-1} \qquad (1)$$

where k_j is the number of (undirected) connections to/from each j-node and n is the total number of nodes in the entire network. The second parameter is the clustering coefficient, C, which quantifies the existence of structure within the local vicinity of nodes, compared to a network of random points (i.e., the ratio of connected triangles of nodes to all possible triples in a given connected cluster). It is measured as

$$C(j) = \frac{2y_j}{k_j\left(k_j-1\right)}, \qquad (2)$$

in which y_j is the number of links between neighboring nodes of the j-node.

While sophisticated methods to identify nodes and filaments in the simulated cosmic network [e.g., Ref. 4] or in the neuronal network [e.g., Ref. 37] have been proposed, here we explore a simpler approach with the advantage of being readily applicable to both networks. The method is inspired by standard "halo finding" procedures in cosmology to identify the self-gravitating halos in the cosmic web [38]. In detail: 1) we marke the highest intensity peaks in all maps (i.e., pixels in the top 10% of the intensity distribution of each map); 2) we compute the enclosed average intensity of pixels within circles of increasing radius, until a low threshold value, Δ, is matched. The radius of the circle reaching the Δ value defines the radius of each node in the network (r_Δ); 3) all pixels at a distance $\le r_\Delta$ are assumed to belong to that node. In the case of the cosmic web we tailored the procedure so that $\Delta = 330\bar{\rho}$, while in the case of the brain networks we tailored the values of Δ so that the radius of nodes in the networks reasonably matches the size suggested by visual inspection.

We then built the adjacent matrix of nodes, M_{ij}, i.e., a matrix with rows/columns equal to the number of detected nodes, with value $M_{ij} = 1$ if the nodes are separated by a distance $\le l_{\rm link}$, or $M_{ij} = 0$ otherwise. The choice of $l_{\rm link}$ is arbitrary, but a full scan of network parameters as a function of $l_{\rm link}$ is beyond the goals of this first exploratory work. We thus focused on one specific choice for the linking length, motivated by the recent analysis of observed galaxies by de Regt et al. [33], who suggested $l_{\rm link} = 1.2$ Mpc as the reference "linking length" for matter halos in the cosmic web (i.e., $\sim L/100$ in **Figure 2**). Based on the similarity of power spectra after opportunely renormalizing the spatial scales presented in **Section 3.3**, we thus consistently rescaled the linking length in 40× brain samples to $l_{\rm link} = 16\,\mu$m. **Figure 3** gives close up view of the nodes and networks reconstructed for three slices of our dataset.

This method selects from $N \sim 3800 - 4700$ nodes in our cosmic web slices, with an average number of $\langle k\rangle \sim 3.8 - 4.1$ connections per node. For the cerebellum slices we measured $\langle k\rangle \sim 1.9 - 3.7$, while for the cortex we measured $\langle k\rangle \sim 4.6 - 5.4$ for the $N \sim 1800 - 2000$ identified nodes. On the other hand, the estimated average number of nodes for the simulated cosmic web is $\sim 40\%$ smaller of the results reported from real galaxy surveys by de Regt et al. [33], which is understood because of the much smaller thickness of our model slices (a factor ~ 4 thinner in comoving depth compared to observations).

Both statistics clearly show that the brain and cosmic web networks are very different from Erdös–Rényi random networks of the same size, which would instead predict for the two parameters $C_{\rm random} \approx \langle k\rangle/N$ ($\le 2\cdot 10^{-3}$ in our case) and $C_{\rm d,random} \approx C_{\rm random}(1 - C_{\rm random})/N$ ($\le 10^{-7} - 10^{-6}$ in our case), in the limit of large N [e.g., Ref. 39]. We can see that instead all

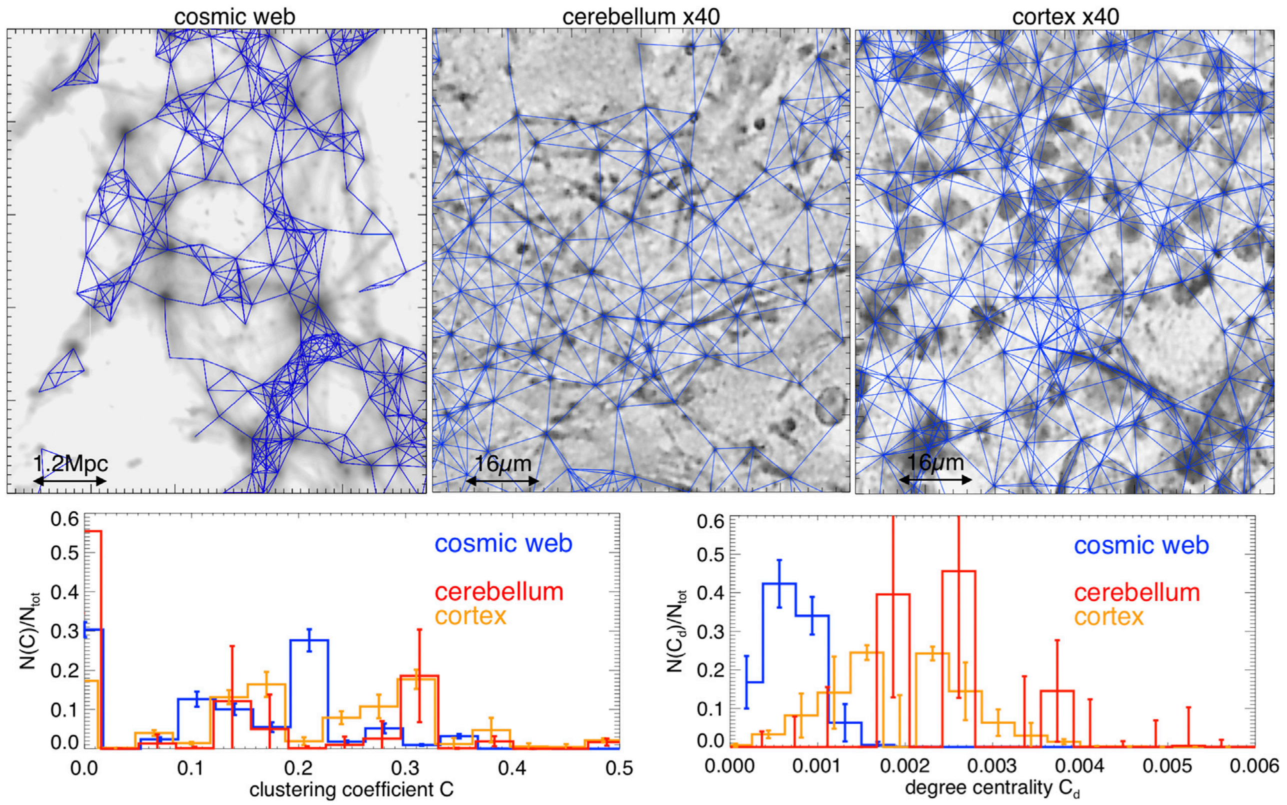

FIGURE 3 | Top panels: zoomed details of the reconstructed connections between nodes for three example of networks in our sample (blue lines, superimposed to the density contrast maps). Bottom panels: distributions of clustering coefficient and of degree centrality for all slices (the error bars give the scatter within each investigated sample of slices). The reconstructed connections do not take into account the long-range neural connections, and the clusters shown are purely spatial.

measured distributions of C measure a few different peaks in the $C \sim 0.1 - 0.4$ range, clearly indicating that all networks are highly correlated, i.e., their links tend to be highly clustered together. In the case of the cosmic web, similar sparse peaks were measured in real data by de Regt et al. [33], and are ascribed to galaxies in moderate ($C \sim 0.1$) or rich ($C \sim 0.3$) environments, like filaments or large clusters of galaxies. Only the residual part of the distributions, with $C \leq 10^{-2}$, marks instead regions of the network in which the connectivity is close to random (e.g., nodes in void regions). The networks also have values of degree centrality clearly much larger (by three to four orders of magnitude) than corresponding random networks. In the cosmic network, the distributions of C_d are approximately Poissonian and in line with the galaxy network studied by Ref. 33, even if the peaks of the distribution are at lower values than the brain samples. The latter is compatible with the enhanced presence of small neurons in the granular layer, already discussed above, which leads to the presence of more closely packed clusters of nodes.

We point out that in this study we analyzed only a fraction of the cortex, and not the whole Central Nervous System, whose architecture is obviously different. Actually, while proximity can accurately describe the cosmic web, neural webs are based on connections and therefore our analysis is not sensitive to long-range connectivity. But indeed long-range connectivity is known to be a crucial feature of neural webs. We defer the application of more complex network statistics [e.g., Ref. 40]) to future work.

DISCUSSION

We have presented a detailed comparison between the neuronal network and the cosmic web, two of the most fascinating and complex networks in Nature, with the goal of assessing the level of similarity between these two physical systems in an objective way.

We have also applied homogeneous statistical approaches to real lab samples of both the brain and the cerebellar cortex (**Section 2.1**), and to slices of the simulated distribution of dark matter and ordinary in the cosmic web (**Section 2.2**), and quantified their morphological and network properties using spectral analysis (**Section 3.3**) as well as network parameters from graph theory (**Section 3.4**). Within the range of simplifying assumptions we used to define both networks (e.g., based on the proximity of nodes identified from the continuous matter distribution rendered by different imaging techniques) our findings hint at the fact that similar network configurations can emerge from the interaction of entirely different physical processes, resulting in similar levels of complexity and self-organization, despite the dramatic disparity in spatial scales (i.e., $\sim 10^{27}$) of these two systems.

We are aware that this approach has several limitations. First, our comparison focused on density of matter. The selection of neurofilaments to outline the neuronal network was based on the fact that they are quite evenly expressed in the cytoplasmic compartment of the neurons. Our results should be further validated with different markers, as microfilaments or microtubules. Second, we assumed that the highest stain density is located at the level of neuronal Soma, which is an approximation, leading to a non-standard definition of nodes. Further studies are required to validate our results with functional neural network data and without losing anatomical-visual definition. Third, our study has been based on histological slices, which can obviously show only a tiny portion of the brain network itself. Moreover, while the cosmic web uses proximity to define its network, neural webs are based on connections that can be significantly long-range spatially, and which could not be properly assessed through our analysis due to technical limitation of the method. For the above limitations, we could not present a systematic and complete connectivity analysis of networks, as we focused on simple proximity and not on long-range connectivity. A key Frontier of this line of comparative research is the possibility of measuring the memory capacity of both networks, a task presently made challenging by the radically different approaches presently available to measure to monitor the flow of information within them. An interesting factoid well illustrates that possible similarities also exist in this respect. The total memory capacity of the human brain has been recently estimated using section electron microscopy to reconstruct the 3D distribution of dendritic spines and of their synapses, and finding 26 distinct synaptic strengths, which accounts to an average of ~ 4.7 bits of information per neuronal cell [41]. Extrapolated to the total average number of nodes in the neuronal network, this yields $\approx 2 \cdot 10^{16}$ bits, i.e., ~ 2.5 Petabytes as the memory capacity of the human brain. For the cosmic web, a radically different idea based on Information Theory can been used to quantify how much information is encoded by the 3-dimensional structure of the cosmic web [42, 43]. Through the computation of the "statistical complexity" that characterizes the dynamical evolution of simulated universes, it has been argued that $\sim 3.5 \cdot 10^{16}$ bits (i.e., ≈ 4.3 Petabytes of memory) are necessary to store the information of cosmic structure within the entire observable Universe (≈ 13.8 Gpc). Such close agreement may appear as a mere coincidence, considering that, given ambiguities in defining both networks, particularly the cosmic web, these numbers are known only approximately.

Together with the rest of the analysis presented in this work, such similarities are meant to motivate the development of more powerful and discriminating algorithms to pinpoint analogies and differences of these fascinating systems, almost at the conceivable extremes of spatial scales in the Universe.

DATA AVAILABILITY STATEMENT

All brain samples analysed in this work, as well as relevant samples of the simulated cosmic web and the reconstructed network connectivity are publicly accessible at this URL: https://cosmosimfrazza.myfreesites.net/cosmic-web-and-brain-network-datasets.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of the University Hospital of Modena. All subjects gave written informed consent in accordance with the Declaration of Helsinki.

AUTHOR CONTRIBUTIONS

Both authors contributed to the writing of the manuscript and to the interpretation of results. FV is responsible for the production of the cosmological simulations and for numerical methods adopted in the paper. AF is responsible for the extraction of the brain samples used in this work.

FUNDING

FV acknowledges financial support from the ERC Starting Grant "MAGCOW", no. 714196.

ACKNOWLEDGMENTS

We gratefully acknowledge Elena Zunarelli (Department of Anatomic Pathology, University Hospital Policlinico di Modena, Modena, Italy) for producing the slices used in **Figure 1**. The cosmological simulation of this work have been produced using the ENZO code (http://enzo-project.org), which is the product of a collaborative effort of scientists at many universities and national laboratories. The simulations/analysis of this work have been produced on Piz Daint supercomputer at CSCS-ETHZ (Lugano, Switzerland) and on Marconi at CINECA (Bologna, Italy) under projects s805 and INA17_C4A28 with (FV as Principal Investigator). FV acknowledges the usage of online storage tools kindly provided by the INAF Astronomical Archive (IA2) initiave (http://www.ia2.inaf.it). We gratefully acknowledge careful reading of this manuscript by D. Wittor.

REFERENCES

1. Sporns O. The human connectome: a complex network. *Ann N Y Acad Sci.* (2011) 1224:109–25. doi:10.1111/j.1749-6632.2010.05888.x
2. Battaglia D, Witt A, Wolf F, Geisel T. Dynamic effective connectivity of inter-areal brain circuits. *PLoS Comput Biol.* (2012) 8:1–20. doi:10.1371/journal.pcbi.1002438
3. Schneider P. *Extragalactic astronomy and cosmology: an introduction.* Berlin, Heidelberg: Springer (2015) doi:10.1007/978-3-642-54083-7
4. Libeskind NI, van de Weygaert R, Cautun M, Falck B, Tempel E, Abel T, et al. Tracing the cosmic web. *Mon Not R Astron Soc Lett.* (2018) 473:1195–217. doi:10.1093/mnras/stx1976
5. Doré O, Hirata C, Wang Y, Weinberg D, Eifler T, Foley RJ, et al. WFIRST: the essential cosmology space observatory for the coming decade. arXiv:1904.01174. arXiv e-prints (2019), https://arxiv.org/abs/1904.01174
6. Lima M, *Brain + Universe* (2009). Visual complexity. http://www.visualcomplexity.com/vc/blog/?p=234
7. Neyrinck M, Elul T, Silver M, Mallouh E, Aragón-Calvo M, Banducci S, et al. Exploring Connections Between Cosmos & Mind Through Six Interactive Art Installations in "As Above As Below". arXiv:2008.05942. arXiv e-prints (2020).
8. Hsu SM. [39] immunohistochemistry. M Wilchek EA Bayer, editors *Avidin-biotin technology. Methods in enzymology.* Vol. 184. Cambridge, MA: Academic Press (1990) p. 357–63. doi:https://doi.org/10.1016/0076-6879(90)84293-P
9. Bryan GL, Norman ML, O'Shea BW, Abel T, Wise JH, Turk MJ, et al. ENZO: an adaptive mesh refinement code for astrophysics. *Astrophys J.* (2014) 211:19. doi:10.1088/0067-0049/211/2/19
10. Vazza F, Brueggen M, Gheller C, Hackstein S, Wittor D, Hinz PM. Simulations of extragalactic magnetic fields and of their observables. *Class Quant Grav.* (2017) 34:234001. doi:10.1088/1361-6382/aa8e60
11. Condon JJ, Matthews AM. ΛCDM cosmology for astronomers. *Publ Astron Soc Pac.* (2018) 130:073001. doi:10.1088/1538-3873/aac1b2
12. Conselice CJ, Wilkinson A, Duncan K, Mortlock A. The evolution of galaxy number density at z < 8 and its implications. *Astrophys J.* (2016) 830:83. doi:10.3847/0004-637X/830/2/83
13. Azevedo FA, Carvalho LR, Grinberg LT, Farfel JM, Ferretti RE, Leite RE, et al. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *J Comp Neurol.* (2009) 513:532–41. doi:10.1002/cne.21974
14. Herculano-Houzel S. The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. *Proc Natl Acad Sci USA.* (2012) 109:10661–8. doi:10.1073/pnas.1201895109
15. Planck Collaboration, Ade PAR, Aghanim N, Arnaud M, Ashdown M, Aumont J, et al. Planck 2015 results. XIII. Cosmological parameters. *Astron Astrophys.* (2016) 594:63. doi:10.1051/0004-6361/201525830.
16. O'brien JS, Sampson EL. Lipid composition of the normal human brain: gray matter, white matter, and myelin. *J Lipid Res.* (1965) 6(4):537–44.
17. Biochemistry and the central nervous system. *A.M.A. Archives Neurol Psychiatry.* (1957) 77:56. doi:10.1001/archneurpsyc.1957.02330310066012
18. Pfeifer S, McCarthy IG, Stafford SG, Brown ST, Font AS, Kwan J, et al. The Bahamas project: effects of dynamical dark energy on large-scale structure. arXiv e-prints (2020) arXiv:2004.07670.
19. Maxwell W, Graham D. Loss of axonal microtubules and neurofilaments after stretch-injury to guinea pig optic nerve fibers. *J Neurotrauma.* (1997) 14:603–14. doi:10.1089/neu.1997.14.603
20. Jafari SS, Nielson M, Graham DI, Maxwell WL. Axonal cytoskeletal changes after nondisruptive axonal injury. ii. intermediate sized axons. *J Neurotrauma.* (1998) 15:955–66. doi:10.1089/neu.1998.15.955. PMID: 9840768
21. Fournier AJ, Hogan JD, Rajbhandari L, Shrestha S, Venkatesan A, Ramesh KT. Changes in neurofilament and microtubule distribution following focal axon compression. *PLoS One.* (2015) 10:1–21. doi:10.1371/journal.pone.0131617
22. Price R, Lasek R, Katz M. Microtubules have special physical associations with smooth endoplasmic reticula and mitochondria in axons. *Brain Res.* (1991) 540:209–16. doi:https://doi.org/10.1016/0006-8993(91)90509-T
23. Vazza F, Brunetti G, Gheller C, Brunino R, Brüggen M. Massive and refined. II. The statistical properties of turbulent motions in massive galaxy clusters with high spatial resolution. *Astron Astrophys.* (2011) 529:A17. doi:10.1051/0004-6361/201016015
24. Manto M, Gruol D, Schmahmann J, Koibuchi N, Rossi F. *Handbook of the cerebellum and cerebellar disorders.* Berlin, Heidelberg: Springer (2013) p. 1–24. doi:10.1007/978-94-007-1333-8
25. Sylos Labini F, Vasilyev NL, Baryshev YV. Power law correlations in galaxy distribution and finite volume effects from the sloan digital sky survey data release four. *Astron Astrophys.* (2007) 465:23–33. doi:10.1051/0004-6361:20065321
26. Sylos Labini F, Vasilyev NL, Baryshev YV. Large-scale fluctuations in the distribution of galaxies from the two-degree galaxy redshift survey. *Astron Astrophys.* (2009) 496:7–23. doi:10.1051/0004-6361:200810575
27. Sylos Labini F Inhomogeneities in the universe. *Class Quant Grav.* (2011) 28:164003. doi:10.1088/0264-9381/28/16/164003
28. Coles P. Phase correlations and topological measures of large-scale structure. *Data Anal.* (2009) 665:493–522. doi:10.1007/978-3-540-44767-2_15
29. Bassett DS, Bullmore E. Small-world brain networks. *Neuroscientist.* (2006) 12:512–23. doi:10.1177/1073858406293182. PMID: 17079517
30. Meng L, Xiang J. Brain network analysis and classification based on convolutional neural network. *Front Comput Neurosci.* (2018) 12:95. doi:10.3389/fncom.2018.00095
31. Joyce KE, Laurienti PJ, Burdette JH, Hayasaka S. A new measure of centrality for brain networks. *PLoS One.* (2010) 5:1–13. doi:10.1371/journal.pone.0012200
32. Sporns O, Honey CJ, Ktter R. Identification and classification of hubs in brain networks. *PLoS One.* (2007) 2:e1049. doi:10.1371/journal.pone.0001049
33. de Regt R, Apunevych S, von Ferber C, Holovatch Y, Novosyadlyj B. Network analysis of the COSMOS galaxy field. *Mon Not R Astron Soc Lett.* (2018) 477:4738–448. doi:10.1093/mnras/sty801
34. Tsizh M, Novosyadlyj B, Holovatch Y, Libeskind NI. Large-scale structures in the ΛCDM Universe: network analysis and machine learning. arXiv e-prints (2019) arXiv:1910.07868.
35. Hansen DL, Shneiderman B, Smith MA, Himelboim I. Social network analysis: measuring, mapping, and modeling collections of connections. In: DL Hansen, B Shneiderman, MA Smith, I Himelboim, editors *Analyzing social media networks with NodeXL.* 2nd ed. Chap. 3, Morgan Kaufmann (2020) p. 31–51. doi:https://doi.org/10.1016/B978-0-12-817756-3.00003-0
36. Golbeck J. Network structure and measures. In: J Golbeck, editor *Analyzing the social web.* Chap. 2. Boston: Morgan Kaufmann (2013) p. 25–44. doi:https://doi.org/10.1016/B978-0-12-405531-5.00003-1
37. Stanley M, Moussa M, Paolini B, Lyday R, Burdette J, Laurienti P. Defining nodes in complex brain networks. *Front Comput Neurosci.* (2013) 7:169. doi:10.3389/fncom.2013.00169
38. Knebe A, Knollmann SR, Muldrew SI, Pearce FR, Aragon-Calvo MA, Ascasibar Y, et al. Haloes gone MAD14: the halo-finder comparison project. *Mon Not Roy Astron Soc.* (2011) 415:2293–318. doi:10.1111/j.1365-2966.2011.18858.x
39. Albert R, Barabási AL. Statistical mechanics of complex networks. *Rev Mod Phys.* (2002) 74:47–97. doi:10.1103/RevModPhys.74.47
40. van den Heuvel MP, Sporns O. Rich-club organization of the human connectome. *J Neurosci.* (2011) 31:15775–86. doi:10.1523/JNEUROSCI.3539-11.2011
41. Bartol TM, Jr, Bromer C, Kinney J, Chirillo MA, Bourne JN, Harris KM, et al. Nanoconnectomic upper bound on the variability of synaptic plasticity. *Elife.* (2015) 4:e10778. doi:10.7554/eLife.10778
42. Vazza F On the complexity and the information content of cosmic structures. *Mon Not R Astron Soc: Lett.* (2017) 465:4942–55. doi:10.1093/mnras/stw3089
43. Vazza F How complex is the cosmic web?. *Mon Not R Astron Soc: Lett.* (2019) 491:5447–63. doi:10.1093/mnras/stz3317

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Physics

ORIGINAL RESEARCH
published: 23 December 2020
doi: 10.3389/fphy.2020.584294

Analyzing the Motion of Symmetric Tops Without Recurring to Analytical Mechanics

Zsolt I. Lázár[1]*, *Antal Jakovác*[2] *and Péter Hantz*[3,4]

[1]*Faculty of Physics, Babes-Bolyai University, Cluj–Napoca, Romania,* [2]*Wigner Research Centre for Physics, Budapest, Hungary,* [3]*Centre for Ecological Research, Budapest, Hungary,* [4]*Fibervar LLC., Cluj–Napoca / Kolozsvár, Romania*

Characterizing the dynamics of heavy symmetric tops is essential in several fields of theoretical and applied physics. Accordingly, a series of approaches have been developed to describe their motion. In this paper, we present a derivation based on elementary geometric considerations carried out in the laboratory frame. Our framework enabled the simple derivation of the equation of motion for small nutations. The introduced formalism is also employed to determine the alteration of the dynamics of heavy, symmetric, spinning tops in a rotating force field, that is compared to the precession characteristics of a quantum magnetic dipole in rotating magnetic field.

Keywords: spinning top, precession, geometric interpretation, small nutations, rotating force fields

OPEN ACCESS

Edited by:
José Fernando Cariñena,
University of Zaragoza, Spain

Reviewed by:
Marin I. Marin,
Transilvania University of Brasov,
Romania
Haydar Uncu,
Adnan Menderes University, Turkey

****Correspondence:***
Zsolt I. Lázár
zsolt.lazar@ubbcluj.ro
Péter Hantz
hantz@general.elte.hu

Specialty section:
This article was submitted to
Mathematical and Statistical Physics,
a section of the journal
Frontiers in Physics

Received: *20 July 2020*
Accepted: *23 October 2020*
Published: *23 December 2020*

Citation:
Lázár ZI, Jakovác A and Hantz P (2020)
Analyzing the Motion of Symmetric
Tops Without Recurring to
Analytical Mechanics.
Front. Phys. 8:584294.
doi: 10.3389/fphy.2020.584294

INTRODUCTION

Mainstream methods for determining the equation of motion of heavy symmetric tops can be classified according to the theoretical approaches used, and the reference frames applied. The framework can employ the toolkits of the more elementary Newtonian, or those of the analytical mechanics. The coordinate systems used include mixed ones (like certain triplets of Euler-angles), or rotating frames attached to the body (like the principal axes used when solving the Euler equations). Euler angles offer a natural parametrization of the rigid body attitude simply revealing the first integrals (constants of motion) within the framework of the Lagrangian formalism.

First, we recapitulate the most well known solutions developed up to this time. The majority of them [1–9] use the Euler-angles (φ precession angle, ψ spinning angle, ϑ nutation angle) to deduce the Euler-Lagrange equations.

The Euler angles φ and ψ (**Figure 1**) are cyclic coordinates with corresponding conserved conjugate momenta. These are the vector projections of the total angular momentum to the vertical axis, $L_z \equiv p_\varphi$, and onto the symmetry axis, namely $L_n \equiv p_\psi$. Finally, the Euler-Lagrange equation for the nutation angle, ϑ, is a second-order differential equation reducible to a first order equation by applying the conservation of the energy, E. The equation of motion for ϑ is uniquely determined by the three conserved quantities, L_n, L_z and E and is analogous to that of a particle in an effective potential. The time evolution of the other two angles, namely φ and ψ can be obtained as the direct integration of expressions in ϑ. This approach confers all three degrees of freedom distinct roles and different dynamics. Nutation, however, stands out of the triplet since it does not have an associated conserved quantity and unidirectionally modulates the other two degrees of freedom.

A series of methods to solve the problem of spinning tops avoid using analytical mechanics. Wittenburg [10] uses the Newton-Euler equation expressed in a precessing coordinate system. The textbook of Morin [11] also presents an elementary deduction, using Euler angles and a mixed system, where the Newton-Euler equation is also transformed to the precessing frame. In Ref. [12] it

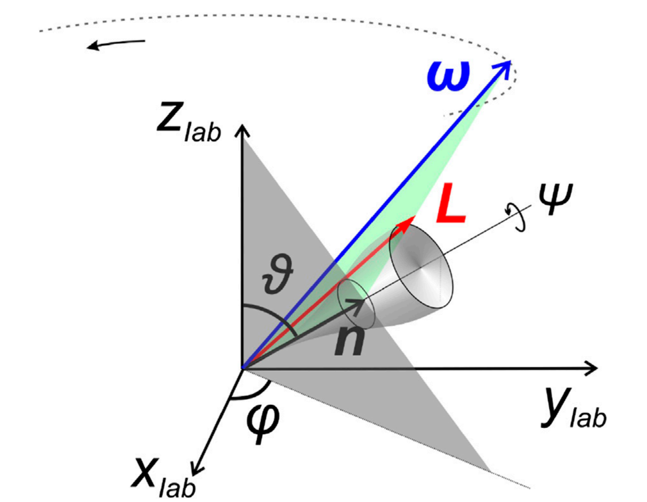

FIGURE 1 | Precessing and nutating heavy top. Angles φ (precession angle), ϑ (nutation angle) and ψ (spinning angle) parametrizing its attitude, the angular momentum, $\boldsymbol{L}$, and angular velocity, $\boldsymbol{\omega}$.

is shown that the three Euler-equations can be replaced by just as many conservation laws. Euler equations in rotating frame have also been applied to solve the problem [14].

As an alternative to the above more formal descriptions, pure precession has been intuitively explained by the so-called "square wheel model" where the spinning top is replaced by an ideal fluid flowing on a square-formed tube. This approach allows the explanation of the "hovering" of the top by forces acting on it, instead of the less intuitive conservation laws [15].

Here we present an alternative based on simple yet rigorous geometric considerations while employing only the elementary methods of Newtonian mechanics. The approach naturally leads to the separation of nutation from the other rotational degrees of freedom and makes possible the usage of a compact matrix formalism in the latter two dimensional subspace.

GEOMETRIC PRELIMINARIES

The spinning heavy top has two special directions that play an essential role in the relationships describing the dynamics of its vectorial quantities. One is the symmetry axis $\boldsymbol{n}$, while the other one is the direction of the gravitational field $\boldsymbol{z}$ (see **Figure 1**). These two unit vectors, spanning a plane, and the direction orthogonal to this plane, namely $\boldsymbol{e}_{\mathrm{nu}} \equiv \boldsymbol{n}\times\boldsymbol{z}/|\boldsymbol{n}\times\boldsymbol{z}|$, serve as a natural basis for investigating our three dimensional model. The spontaneous emergence of this basis is the reason behind the incontestable usefulness of Euler angles ψ, φ and ϑ for specifying the orientation of a spinning, symmetric rigid body. The rates of change of the these angles are denoted by $\dot{\psi} \equiv \omega_s, \dot{\varphi} \equiv \omega_p$ and $\dot{\vartheta} \equiv \omega_{\mathrm{nu}}$. Using the above basis, any vector $\boldsymbol{a}$ can be decomposed as

$$\boldsymbol{a} = a_s\boldsymbol{n} + a_p\boldsymbol{z} + a_{\mathrm{nu}}\boldsymbol{e}_{\mathrm{nu}} = \boldsymbol{a}_s + \boldsymbol{a}_p + \boldsymbol{a}_{\mathrm{nu}}, \tag{1}$$

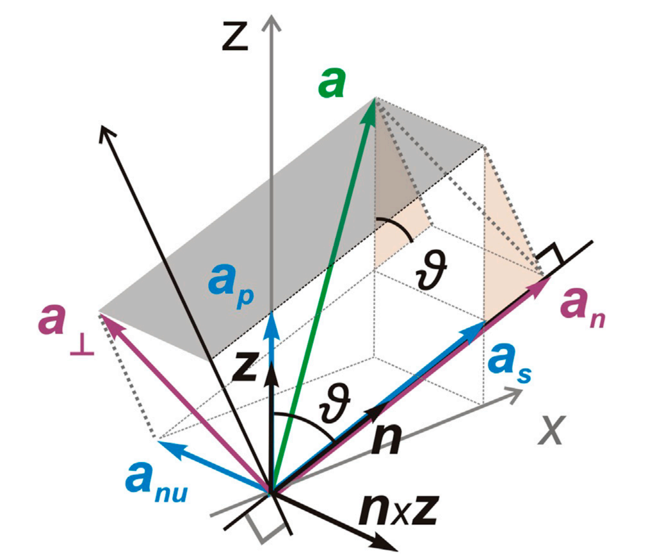

FIGURE 2 | Decomposition of a vector $\boldsymbol{a}$ in the non-orthogonal basis $\boldsymbol{n}$, $\boldsymbol{z}$, $\boldsymbol{n}\times\boldsymbol{z}$. For the definition of a_n, a_s and a_p and a_z, see **Eq. 1** and the corresponding text.

where the three terms are vector projections of $\boldsymbol{a}$ parallel to the respective basis vectors (see **Figure 2**). Since the chosen basis is not orthogonal, the scalar projections $a_n = \boldsymbol{a}\cdot\boldsymbol{n}$, $a_z = \boldsymbol{a}\cdot\boldsymbol{z}$ and $a_{\mathrm{nu}} = \boldsymbol{a}\cdot\boldsymbol{e}_{\mathrm{nu}}$ also claim a role in the description. Alternatively, one can project $\boldsymbol{a}$ to one of the basis vectors and to the corresponding orthogonal plane:

$$\begin{aligned}\boldsymbol{a} &= \boldsymbol{a}_n + \boldsymbol{a}_\perp,\\ \boldsymbol{a}_n &= \boldsymbol{n}a_n = (\boldsymbol{n}\circ\boldsymbol{n})\boldsymbol{a},\\ \boldsymbol{a}_\perp &= (\boldsymbol{a}-\boldsymbol{a}_n) = (\mathbb{I}-\boldsymbol{n}\circ\boldsymbol{n})\boldsymbol{a}.\end{aligned} \tag{2}$$

The dynamics of the top is such that these three directions are associated with qualitatively different phenomena (spin, precession and nutation). The nutation stands out of the trio as will become apparent also from this study. Therefore we shall introduce a formalism that manifestly separates the description into aspects confined to the rotating $(\boldsymbol{n},\boldsymbol{z})$ plane and aspects involving the direction perpendicular to it. Due to the linear connection between different decompositions and between kinematic and dynamic quantities such as angular velocity and angular momentum a matrix formalism will be useful.

Figure 2 reveals a number of geometric relations including

$$\begin{aligned}a_n &= a_s + \cos\vartheta a_p,\\ a_z &= a_p + \cos\vartheta a_s,\end{aligned} \tag{3}$$

that can be expressed compactly as

$$\boldsymbol{a}_{n,z} = \widehat{\boldsymbol{G}}\boldsymbol{a}_{s,p}, \tag{4}$$

where

$$\boldsymbol{a}_{n,z} = \begin{pmatrix} a_n \\ a_z \end{pmatrix}, \qquad \boldsymbol{a}_{s,p} = \begin{pmatrix} a_s \\ a_p \end{pmatrix},$$

$$\widehat{\boldsymbol{G}} \equiv \begin{pmatrix} 1 & u \\ u & 1 \end{pmatrix}, \qquad \widehat{\boldsymbol{G}}^{-1} = \frac{1}{s^2}\begin{pmatrix} 1 & -u \\ -u & 1 \end{pmatrix},$$

with

$$u = \cos\vartheta, \qquad s = \sqrt{1-u^2} = \sin\vartheta,$$

and

$$(\boldsymbol{a}_z - \boldsymbol{a}_n)^2 = \sin^2\vartheta\left(\boldsymbol{a}_s + \boldsymbol{a}_p\right)^2, \tag{5}$$

that will be applied *in Time Evolution of the Spin and Precession Angles*. The proof for **Eq. 5** is shown in *Proof of **Eq. (5)*** of the **Supplementary Material**. Note that the connections between a_n, a_z, a_s and a_p are solely determined by ϑ.

RELATIONSHIPS BETWEEN THE COMPONENTS OF *L* AND *ω*

The components of the angular momentum along the symmetry axis $\boldsymbol{n}$ and the orthogonal ones to this are referred to as $L_n = C\omega_n$, $L_\perp = A\omega_\perp$, where C and A are the corresponding principal moments of inertia.

Therefore,

$$\begin{aligned} \boldsymbol{L} = \boldsymbol{L}_n + \boldsymbol{L}_\perp &= [C\boldsymbol{n}\circ\boldsymbol{n} + A(\mathbb{I} - \boldsymbol{n}\circ\boldsymbol{n})]\boldsymbol{\omega} \\ &= A\boldsymbol{\omega} + (C-A)\boldsymbol{n}(\boldsymbol{n}\cdot\boldsymbol{\omega}) = A\boldsymbol{\omega} + (C-A)\omega_n\boldsymbol{n}. \end{aligned} \tag{6}$$

The above linear interdependence between **L**, **ω** and **n** reveals their coplanarity. Note that for asymmetric tops this property does not hold.

Using the notation introduced in *Geometric Preliminaries*, **Eq. 6** can be rewritten as

$$\boldsymbol{L}_{n,z} = C\widehat{\boldsymbol{D}}\boldsymbol{\omega}_{n,z},\ L_{\mathrm{nu}} = A\omega_{\mathrm{nu}}, \tag{7}$$

where

$$\widehat{\boldsymbol{D}} \equiv \begin{pmatrix} 1 & 0 \\ (1-\alpha)u & \alpha \end{pmatrix}, \quad \alpha = A/C.$$

READING CONSERVED QUANTITIES *L*ₙ, *ω*ₙ, AND *L*_z

Let us to consider the Newton-Euler equation

$$\dot{\boldsymbol{L}} = w\boldsymbol{z}\times\boldsymbol{n}, \tag{8}$$

where w is the magnitude of the torque of the homogeneous gravitational field pointing into the $-\boldsymbol{z}$ direction. Due to **Eq. 8** we have $\boldsymbol{n}\cdot\dot{\boldsymbol{L}} = 0$. All points of the top are engaged in a rotation defined by $\boldsymbol{\omega}$. This is also true for the symmetry axis $\boldsymbol{n}$, that is,

$$\dot{\boldsymbol{n}} = \boldsymbol{\omega}\times\boldsymbol{n}, \tag{9}$$

revealing that $\dot{\boldsymbol{n}}$ is orthogonal to the plane spanned by $\boldsymbol{\omega}$ and $\boldsymbol{n}$. Due to the co-planarity of **L**, **ω** *and* **n** *we have* $\dot{\boldsymbol{n}}\cdot\boldsymbol{L} = 0$. *Therefore*

$$\dot{L}_n = \frac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{L}\cdot\boldsymbol{n}) = 0,$$

thus L_n is conserved. **Equation 7** entails that ω_n is conserved as well.

A similar but more straightforward consideration yields $L_z = \boldsymbol{z}\cdot\boldsymbol{L} =$ const. as $\dot{\boldsymbol{z}} = 0$. Note that since no dissipation is present, the energy of the system is also conserved.

TIME EVOLUTION OF THE SPIN AND PRECESSION ANGLES

Due to the conservation of the angular momentum components L_n and L_z it is worth connecting them directly with the kinematically relevant spin and precession angular velocities. Combining **Eqs 4** and **7** results in

$$\boldsymbol{L}_{n,z} = C\widehat{\boldsymbol{T}}\boldsymbol{\omega}_{s,p}, \qquad \widehat{\boldsymbol{T}} = \widehat{\boldsymbol{D}}\widehat{\boldsymbol{G}} = \begin{pmatrix} 1 & u \\ u & \alpha s^2 + u^2 \end{pmatrix}.$$

This enables the expression of the two angular velocities as

$$\boldsymbol{\omega}_{s,p} = \frac{1}{C}\widehat{\boldsymbol{T}}^{-1}\boldsymbol{L}_{n,z}, \qquad \widehat{\boldsymbol{T}}^{-1} = \frac{1}{\alpha s^2}\begin{pmatrix} \alpha s^2 + u^2 & -u \\ -u & 1 \end{pmatrix}. \tag{10}$$

This formula has a pivotal importance: it connects the kinematic quantities of interest to the conserved dynamic quantities.

Note that ω_s and ω_p solely depend on conserved components of angular momenta and the time-dependent polar angle $\vartheta(t)$ therefore become themselves constants of motion if $\omega_{\mathrm{nu}} \equiv \dot{\vartheta} = 0$, phenomenon called pure precession.

TIME EVOLUTION OF THE NUTATION ANGLE

The above results were obtained without making explicit reference to the conservation of energy. For moving beyond pure precession and describing nutation we have to quantify the migration of the energy between kinetic and potential components during nutation.

By expressing the angular frequency $\boldsymbol{\omega}$ from the linear **Eq. 6** the rotational energy of the top can be written as

$$T \equiv \frac{1}{2}\boldsymbol{L}\cdot\boldsymbol{\omega} = \frac{1}{2A}\boldsymbol{L}^2 + \frac{1}{2}\left(1 - \frac{C}{A}\right)\omega_n L_n, \tag{11}$$

while the potential energy reads

$$V(\vartheta) = w\cos\vartheta. \tag{12}$$

Exploiting the orthogonality of $\boldsymbol{e}_{\mathrm{nu}}$ to the $(\boldsymbol{n},\boldsymbol{z})$ plane combined with property **Eq. 5** we get

$$\boldsymbol{L}^2 = \left(\boldsymbol{L}_s + \boldsymbol{L}_p\right)^2 + \boldsymbol{L}_{\mathrm{nu}}^2 = \frac{(\boldsymbol{L}_n - \boldsymbol{L}_z)^2}{\sin^2\vartheta} + \boldsymbol{L}_{\mathrm{nu}}^2. \tag{13}$$

The dynamics ruling the nutation angle can be regarded as a one-dimensional motion in an effective potential, motion completely determined by the conservation of the effective energy. Having L^2 and V obtained, enables us to provide the formula for these effective energies

$$E_{\mathrm{eff}} = T_{\mathrm{eff}}\left(\dot{\vartheta}\right) + V_{\mathrm{eff}}(\vartheta), \tag{14}$$

where the reuse of **Eqs 7**, **11** and **13** gives

$$T_{\mathrm{eff}}\left(\dot{\vartheta}\right) = \frac{1}{2A}\boldsymbol{L}_{\mathrm{nu}}^2 = \frac{A}{2}\dot{\vartheta}^2,$$
$$V_{\mathrm{eff}}(\vartheta) = \frac{1}{2A}\frac{(\boldsymbol{L}_n - \boldsymbol{L}_z)^2}{\sin^2\vartheta} + V(\vartheta),$$
$$E_{\mathrm{eff}} = E - \frac{1}{2}\left(\frac{1}{C} - \frac{1}{A}\right)L_n^2.$$

For convenience **Eq. 14** can be rewritten in terms of $u \equiv \cos\vartheta$ and $\dot{u} = -\sqrt{1-u^2}\dot{\vartheta}$ as

$$\frac{A}{2}\dot{u}^2 + U_{\mathrm{nu}}(u) = \epsilon, \tag{15}$$

where

$$U_{\mathrm{nu}}(u) = \nu u + \frac{\kappa}{2}u^2 - \gamma u^3.$$

Here the Greek letters denote the following

$$\nu = w - \frac{L_n L_z}{A}, \tag{16}$$
$$\kappa = 2E - L_n^2\left(\frac{1}{C} - \frac{1}{A}\right),$$
$$\gamma = w,$$
$$\epsilon = E - \frac{L_n^2}{2C} - \frac{L_z^2}{2A}.$$

The above relationships reveal that the equation of motion for the nutation angle, ϑ, can be solved decoupled from the other two angles, namely the φ precession and ψ spinning angle.

Full solution of the problem requires to resolve the time evolution of Euler angles. **Equation 15** rules $\vartheta(t)$, while $\psi(t)$ and $\varphi(t)$ can be determined by integrating ω_s, repectively ω_p in **Eq. 10**.

SMALL NUTATIONS

In order to describe small nutations, we consider the minimum point u_0 of the one-dimensional potential $U_{\mathrm{nu}}(u)$:

$$\frac{\mathrm{d}U_{\mathrm{nu}}}{\mathrm{d}u}(u_0) = \nu + \kappa u_0 - 3\gamma u_0^2 = 0, \tag{17}$$
$$\frac{\mathrm{d}^2 U_{\mathrm{nu}}}{\mathrm{d}u^2}(u_0) = \kappa - 6\gamma u_0 = A\Omega_{\mathrm{nu}}^2 > 0\,.$$

providing

$$u_0 = \frac{\kappa - \sqrt{\kappa^2 + 12\nu\gamma}}{6\gamma},$$
$$\Omega_{\mathrm{nu}} = \sqrt{\frac{\kappa^2 + 12\nu\gamma}{A}},$$

where Ω_{nu} represents the angular frequency of small, nearly harmonic oscillations in the polar angle during nutation.

We intend to investigate small deviations from pure precession. In the presence of nutation, the conservation of quantities such as ω_s, ω_p or $\boldsymbol{L}^2$ does not hold any more.

In the low amplitude oscillation limit, $\Delta u(t) = u(t) - u_0 = \delta\cdot\cos(\Omega_{\mathrm{nu}}t)$, $\delta \ll 1$ and $\dot{u}(t) = -\Omega_{\mathrm{nu}}\delta\sin(\Omega_{\mathrm{nu}}t) = -\Omega_{\mathrm{nu}}\sqrt{\delta^2 - [\Delta u(t)]^2}$ is an oscillation with the same frequency but $\pi/2$ phase delay.

Let us to denote generically by $f(u)$ the physical quantities modulated by the nutation angle. Since the temporal alteration of f can be written as $\Delta f[u(t)] \approx f'(u_0)\Delta u(t)$, the physical quantities modulated by u will oscillate with the same frequency. Moreover, f will oscillate with an amplitude $f'(u_0)\delta$ around the mean value, $f(u_0)$, that is the pure precession value at u_0.

The deviation of the angular frequency components can be obtained from **Eq. 10**

$$\Delta\boldsymbol{\omega}_{s,p} = \Delta u \left.\frac{d\widehat{\boldsymbol{T}}^{-1}}{du}\right|_{u_0}\frac{\boldsymbol{L}_{n,z}}{C},$$

where we made use of the conserved character of $\boldsymbol{L}_{n,z}$ and assume the time dependence of u as implicit. By definition the nutation component of the angular frequency is

$$\omega_{\mathrm{nu}} = \dot{\theta} = -\frac{\dot{u}}{\sqrt{1-u^2}}.$$

For any vectorial quantity, $\boldsymbol{A}$, with rate of change $\dot{\boldsymbol{A}}$ its nutation motion, $\Delta\dot{\boldsymbol{A}}$, can be described as that of a time dependent geometric vector viewed from the purely precessing reference frame rotating with $\omega_p(u_0)\boldsymbol{z}$. The transformation to the rotating (precessing) reference frame is given by

$$\Delta\dot{\boldsymbol{A}} = \dot{\boldsymbol{A}} - \omega_p(u_0)\boldsymbol{z}\times\boldsymbol{A}. \tag{18}$$

The nutation of the symmetry axis, $\dot{\boldsymbol{n}}$, can be captured by combining **Eq. 18** with **Eqs 1** and **9**. In the laboratory frame its rate of change can be written as

$$\dot{\boldsymbol{n}} = \left(\omega_p\boldsymbol{z} + \omega_{\mathrm{nu}}\boldsymbol{e}_{\mathrm{nu}}\right)\times\boldsymbol{n}.$$

According to **Eq. 18**

$$\dot{\boldsymbol{n}}_{\mathrm{nu}} = \left(\Delta\omega_p\boldsymbol{z} + \omega_{\mathrm{nu}}\boldsymbol{e}_{\mathrm{nu}}\right)\times\boldsymbol{n} = -s\Delta\omega_p\boldsymbol{e}_{\mathrm{nu}} + \omega_{\mathrm{nu}}\boldsymbol{e}_{\perp}.$$

The two orthogonal terms are proportional to Δu and $\dot{u}$, respectively, indicating a rotational movement about the (purely) precessing symmetry axis. From **Eqs 11** and **12** we can see that

$$\frac{1}{2A}\boldsymbol{L}^2 = E - \frac{1}{2}\left(1 - \frac{C}{A}\right)\omega_n L_n + wu. \tag{19}$$

Apart from u all other quantities are either parameters or constants of motion revealing that during small nutations the square of the total angular momentum oscillates with amplitude $2AW\delta$ and frequency Ω_{nu} about its pure precession value.

PURE PRECESSION

By setting $\dot{u} = 0$ **Eqs 15** and **17** take the form

$$\nu u + \frac{\kappa}{2}u^2 - \gamma u^3 = \epsilon,$$
$$\nu + \kappa u - 3\gamma u^2 = 0,$$

yielding

$$\kappa = 3\gamma u - \frac{\nu}{u}, \qquad 2\epsilon = u\left(\nu + \gamma u^2\right). \tag{20}$$

By eliminating the energy, E, from the definitions of κ and ϵ in **Eq. 16**, we have

$$\kappa - 2\epsilon = \frac{L_n^2 + L_z^2}{A}.$$

Combining the above with **Eq. 20** we get

$$L_n^2 + L_z^2 - 2L_nL_z\chi = \boldsymbol{L}_{n,z}^{\top}\hat{\boldsymbol{Q}}\,\boldsymbol{L}_{n,z} = -\frac{\alpha s^4}{u}Cw,$$

where $\chi = \frac{1}{2}\left(u + \frac{1}{u}\right)$, $s^2 = 1 - u^2$, and

$$\hat{\boldsymbol{Q}} = \begin{pmatrix} 1 & -\chi \\ -\chi & 1 \end{pmatrix}.$$

Therefore

$$\boldsymbol{\omega}_{s,p}^{\top}\hat{\boldsymbol{T}}^{\top}\hat{\boldsymbol{Q}}\hat{\boldsymbol{T}}\boldsymbol{\omega}_{s,p} = -\frac{\alpha s^4}{u}\frac{w}{C},$$

wherein

$$\hat{\boldsymbol{T}}^{\top}\hat{\boldsymbol{Q}}\hat{\boldsymbol{T}} = \frac{\alpha s^4}{u}\begin{pmatrix} 0 & -\frac{1}{2} \\ -\frac{1}{2} & u(\alpha - 1) \end{pmatrix},$$

resulting in

$$\omega_p^2\cos\vartheta(A - C) - \omega_p\omega_s C + w = 0,$$

the well-known relationship between precession and spin angular velocities for a given value of the nutation angle ϑ.

PURE PRECESSION IN A ROTATING FORCE FIELD

Spins driven by rotating magnetic fields have been extensively studied due to their importance in resonance spectroscopy. Here we will study the effect of a horizontally rotating homogeneous field on a classical gyroscope. This force can be implemented, for example, by electrostatic interactions. In this case, the motion of a heavy spinning top without dissipation generally becomes erratic. Therefore we will limit our investigation to the situation when the precession is in synchrony with the driving field, meaning, that the rotating component of the field stays in the same vertical plane as the symmetry axis. In these special circumstances, the equations connecting kinematic and dynamic quantities such as **Eqs 10** and **11** are not affected by the particularities of the field. However, the conservation laws derived in Reading Conserved *Quantities L_n, !n, and L_Z* depend on the geometric relationship between the field and the symmetry axis of the top. If kept in the $(\boldsymbol{n}, \boldsymbol{z})$ plane the rotating field component will only change the magnitude of the torque in **Eq. 8** and not its direction. However, the potential energy in **Eq. 12** will modify as

$$\tilde{V}(\vartheta) = V(\vartheta) + b\sin\vartheta, \tag{21}$$

where b quantifies the effect of the horizontally rotating field component leading to the one dimensional effective potential

$$\tilde{U}_{\mathrm{nu}}(u) = U_{\mathrm{nu}}(u) + b\left(1 - u^2\right)^{3/2}. \tag{22}$$

Since the exhaustive investigation of the properties of the above function is beyond the scope of this paper we only remark that the main features of the dynamics are not affected by the additional term from above. For a simple yet quantitative conclusion we further confine our study to the limit of weak driving fields and view $\tilde{U}_{\mathrm{nu}}$ as the perturbation of U_{nu}. The stable solution of the perturbed nutation angle $\tilde{u}_0$ can be obtained from

$$\frac{\mathrm{d}\tilde{U}_{\mathrm{nu}}}{\mathrm{d}u}(\tilde{u}_0) = \frac{\mathrm{d}U_{\mathrm{nu}}}{\mathrm{d}u}(\tilde{u}_0) - 3b\tilde{u}_0\sqrt{1 - \tilde{u}_0^2} = 0. \tag{23}$$

The first order Taylor expansion around u_0 gives

$$\tilde{u}_0 - u_0 = b\frac{u_0\sqrt{1 - u_0^2}}{A\Omega_{\mathrm{nu}}^2}. \tag{24}$$

The above expansion procedure applied on the second derivative of $\tilde{U}_{\mathrm{nu}}$ yields

$$\frac{\mathrm{d}^2\tilde{U}_{\mathrm{nu}}}{\mathrm{d}u^2}(\tilde{u}_0) - \frac{\mathrm{d}^2U_{\mathrm{nu}}}{\mathrm{d}u^2}(u_0) = \mathcal{O}(b), \tag{25}$$

ensuring the stability of the perturbed solution. Note that the conclusions on the existence and stability of the stationary solution can be extended well beyond the perturbative range of the driving field component.

PRECESSING SPIN IN A ROTATING MAGNETIC FIELD

In a broader context precession is a term applicable to any axis with one of its points fixed and performing a circular motion along the surface of a cone. Outside the realm of inertial macroscopic motion [16] we encounter it in quantum mechanics of magnetic dipoles and it is the basis of nuclear magnetic resonance (NMR) [17] and ESR [18]. Atomic systems in strong fields obey dynamics where inertia has little or no role. However, the nature of the coupling between the angular momentum and the magnetic field produces a motion that is similar to the precession of a rigid body.

Let us consider a magnetic field that has a constant vertical and a rotating horizontal component, namely $\boldsymbol{B} = [b\sin(\omega t), b\cos(\omega t), B]$. Note that the horizontal component here rotates counterclockwise with respect to the third axis. The equation of motion for the quantum mechanical expectation value, $\boldsymbol{S}$, of the angular momentum coupled through the gyromagnetic factor γ to this field reads

$$\dot{\boldsymbol{S}} = \gamma \boldsymbol{S} \times \boldsymbol{B}. \tag{26}$$

Note that if no horizontal rotating component is present $\boldsymbol{S}$ precesses with the Larmor frequency $\omega_L = \gamma B$ (See ***Spin in Magnetic Field*** in the **Supplementary Material**). In this special case the attitude of $\boldsymbol{S}$ is arbitrary, i.e., determined by the initial condition. In the presence of dissipation the angle will relax to zero, i.e., parallel to the constant magnetic field.

In the general case, when the rotating component of the magnetic field is present, the stationary (particular) solution of **Eq. 26** will be a precession motion with the same ω frequency as the driving field and the angle φ enclosed with the vertical is

$$\cot\varphi = \frac{\omega_L - \omega}{\omega_l}, \quad \omega_l \equiv \gamma b. \tag{27}$$

Note that transients are disregarded. During this deduction the laboratory reference frame was used.

Though both refer to angular momenta, **Eqs 8** and **26** are far from being equivalent. The cross product in **Eq. 26** conserves the magnitude of the angular momentum. Therefore the magnitude oscillations described by **Eq. 19** are not present in the case of spins.

DISCUSSION

The dynamics of a heavy symmetric top is determined by the constants of motion L_n, L_z and E. An essential output of our approach is expressed in **Eq. 10**. This relationships represents the inversion in the $(\boldsymbol{n}, \boldsymbol{z})$ subspace of the linear **Eq. 6** such that the angular velocities are expressed in terms of L_n and L_z. The momentum $p_\vartheta = L_{\mathrm{nu}}$ associated with the third coordinate, ϑ, is not conserved. Nutation "remains alone" in a first order differential equation describing a one dimensional non-harmonic oscillator (see **Eq. 15**). This periodic conversion of the energy from potential to kinetic and back will modulate the spin and precession angular velocities through **Eq. 10**.

In the case of small nutations the only effective geometric parameter is the nutation angle ϑ characterizing the attitude of the top. The magnitude of the angular momentum harmonically oscillates around its value encountered in pure precession.

We also examine the case of the classical symmetric spinning rigid body and the quantum mechanical spin (without inertia) precessing in an external field having a rotating component. While the main features of the spin dynamics can be provided analytically, the case of a heavy spinning top driven by a rotating field seems to be more complex. For the case without dissipation, the dynamics of the system will be unpredictable, except the case when the precession frequency is in synchrony with the driving - the case discussed in first-order approximation in this paper.

Our paper employing matrix formalism combined with geometry provides another example that the problem of spinning top can be addressed by a multitude of approaches, each emphasizing a different facet of the phenomenon.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

ACKNOWLEDGMENTS

The authors are indebted to Gyula Dávid, László Forró, József Lázár, Eörs Szathmáry, András Málnási-Csizmadia, and Csaba András for their insightful comments. ZL is supported by the COFUND-FLAGERA II-CORTICITY grant.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphy.2020.584294/full#supplementary-material.

REFERENCES

1. Johns O. *Analytical mechanics for relativity and quantum mechanics.* New York, NY: Oxford University Press (2005) p. 656.
2. Lemos NA. *Analytical mechanics.* Cambridge, UK: Cambridge University Press (2018) p. 459.
3. Hand LN, Finch JD. *Analytical mechanics.* Cambridge, UK: Cambridge University Press (1998) p. 575.
4. Landau LD, Lifshitz EM. Mechanics. *Course of theoretical physics.*, Vol. 1. Oxford, UK: Pergamon Press (1969) p. 197.
5. Herbert Goldstein CP, Safko J. *Classical mechanics.* 3rd ed. London, UK: Pearson (2001) p. 20.
6. Lüdde CS, Dreizler RM. *Theoretical mechanics.* Berlin, Germany: Springer (2010) p. 402.
7. Ginsberg JH. *Engineering dynamics.* Cambridge, UK: Cambridge University Press (2008) p. 742.
8. Kibble TWB, Berkshire FH. *Classical mechanics.* London, UK: Imperial College Press (2004) p. 478.
9. Gregory RD. *Classical mechanics.* Cambridge, UK: Cambridge University Press (2006) p. 596.
10. Wittenburg J. *Dynamics of multibody systems.* Berlin, Germany: Springer (2008) p 223.
11. Morin D. *Introduction to classical mechanics.* Cambridge, UK: Cambridge University Press (2008) p. 719.
12. Romano A, Marasco A. *Classical mechanics with Mathematica®.* Cham, Switzerland: Springer International Publishing (2018) p. 506.
13. Berry MV, Shukla P. Slow manifold and hannay angle in the spinning top. *Eur J Phys.* (2010) 32:115–27. doi:10.1088/0143-0807/32/1/011

14. Bhattacharjee S. Rotating frame analysis of rigid body dynamics in space phasor variables. *Am J Phys.* (2013) 81:518–26. doi:10.1119/1.4803531
15. Hantz P, Lázár ZI. Precession intuitively explained. *Front Phys.* (2019) 7:5. doi:10.3389/fphy.2019.00005
16. Itu C, Öchsner A, Vlase S, Marin MI, Improved rigidity of composite circular plates through radial ribs. *Proc Inst Mech Eng L J Mater.* (2018) 233:1585–93. doi:10.1177/1464420718768049
17. Williamson MP. Drawing single NMR spins and understanding relaxation. *Nat Prod Commun.* (2019) 14:1–9. doi:10.1177/1934578x19849790
18. Bertrand P. *Electron paramagnetic resonance spectroscopy: fundamentals.* Cham, Switzerland: Springer (2020) p. 420.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

BRIEF RESEARCH REPORT
published: 15 January 2021
doi: 10.3389/fphy.2020.614596

Size of National Assemblies: The Classic Derivation of the Cube-Root Law is Conceptually Flawed

Giorgio Margaritondo *

Faculté des Sciences de Base, Ecole Polytechnique Fédérale de Lausanne, EPFL SB IPHYS LQM, Lausanne, Switzerland

For half a century, the analysis of the size of national assemblies was dominated by the famous cube-root relation with the population. However, a revisitation of that historical work with a physicist's approach reveals basic conceptual problems that fatally undermine its conclusions. Furthermore, the assembly size evaluation exceeds the accuracy of all power equations, which cannot be reliably used for political analysis.

Keywords: cube-root law, assembly size, members of parliament, representatives, optimal parliament, Taagepera

OPEN ACCESS

Edited by:
Haroldo V. Ribeiro,
State University of Maringá, Brazil

Reviewed by:
Matjaž Perc,
University of Maribor, Slovenia
Angel Akio Tateishi,
Federal Technological University of Paraná, Brazil
Eszter Bokanyi,
Corvinus University of Budapest, Hungary

****Correspondence:***
Giorgio Margaritondo
giorgio.margaritondo@epfl.ch

Specialty section:
This article was submitted to Social Physics, a section of the journal Frontiers in Physics

Received: *06 October 2020*
Accepted: *24 November 2020*
Published: *15 January 2021*

Citation:
Margaritondo G (2021) Size of National Assemblies: The Classic Derivation of the Cube-Root Law is Conceptually Flawed. Front. Phys. 8:614596. doi: 10.3389/fphy.2020.614596

INTRODUCTION

Could the "optimal" size for the national assembly of a country be evaluated with methods similar to physics research? This is a timely question: the debate about insufficient representation at the federal and state levels is raging in the USA. On the other side, there were recent initiatives to reduce the number of representatives in the national parliaments of many countries, including France, Hungary, Ireland, Japan, Mexico, the Netherlands, Portugal, Romania and the United Kingdom. And Italy just emerged from a referendum on this issue.

The classic reference is the 1972 work of Taagepera [1], who introduced the well-known cube-root formula to link A, the number of parliament members, and P_o, the population:

$$A = aP_o^{1/3}, \tag{1}$$

where a is a constant.

Alternate approaches were later presented [2–5]. In particular, Auriol and Gary-Bobo [2] derived a square-root law and then empirically obtained a 0.4 exponent from recent data for 100 countries. And the foundations of the cube-root law were criticized: notably, Jacobs and Otjes [6] questioned the cause-effect sequence that supposedly leads to it.

The relation between A and P_o size must be appreciated in a more general context [7–9]. Indeed, scaling power-law relations with the population were empirically and/or formally derived for other quantities like the number of election candidates [7, 8], patent production, personal income and the electrical cable length [9]. The general notion is that "*similarly to large-scale physical thermodynamic systems, large groups of interacting humans may exhibit universal statistical properties*" [7]. It is certainly not our scope here to challenge this notion, which is supported by a variety of facts and led to important contributions to the understanding of collective human phenomena. Our focused scope is to show that in the specific case of the cube-root relation for parliament members and population the derivation of Ref. [1] was flawed.

Furthermore, cube-root scaling laws have alternate mathematical explanations [10] with respect to [1] and were known as early as (at least) 1909 [11]. Therefore, our challenge of the classic derivation of Ref. [1] does not necessarily imply that the law itself is wrong.

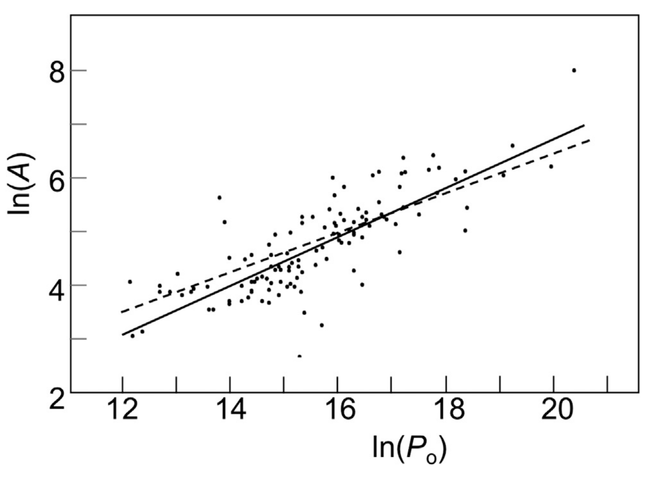

FIGURE. 1 | Log-log plot of the original data of Taagepera [1]. The solid line is the best fit given by **Eq. 3**, whereas the dashed line is the (forced) fit with a cube-root law, leading to **Eq. 4**.

METHOD

Taagepera's work [1] remains a milestone in many experts' view, is known by a broad public and is often used in political debates. For example, it was publicized by the media as "scientific" support for one of the sides in the recent Italian referendum [4]. We thus decided to directly look at its derivation from a physicist's prospective, and surprisingly found that the original work [1] is affected by four critical problems:

(1) The cube-root law was not derived from its data and the corresponding fit was arbitrarily forced.
(2) The theoretical steps that were used to derive **Eq. 1** incorrectly evaluated one of its key factors.
(3) The model assumed that each representative spends on the average equal times for communications inside and outside the parliament, an arbitrary hypothesis that has unrealistic consequences.
(4) No evaluation of the "optimal" size based on a power law, including the cube-root one, can reach a meaningful accuracy.

Concerning the first problem, the original article [1] did mention a power law more general than **Eq. 1**:

$$A = aP_o^n. \tag{2}$$

However, it surprisingly argued against using it to fit the data: "*The actual best fit of the data to an expression of the form A = aP*$_o$^{n} *. . . could be worked out, but this would be a dead end. . . It is more fruitful to look for a plausible theoretical model which would fit the observed general trend*". This argument is fundamentally flawed from a physicist's point of view: it considers only one hypothesis, renouncing *a priori* to demonstrate its superiority with respect to others.

RESULTS

We analyzed the consequences of the above argument by applying the same fitting procedure as Ref. [1] to the data of its Table 1, i.e., a least-square fit of the logarithms. Using **Eq. 2** instead of **Eq. 1**, i.e., an unrestricted fit (the solid line in **Figure 1**), we got:

$$A = 0.10\,P_o^{0.45 \pm 0.03} \tag{3}$$

The exponent $n = 0.45$ is actually closer to 0.5, the square-root law proposed by Auriol and Gary-Bobo [2], and to their empirical value 0.4.

If one forces the same data set to be fitted by a cube-root law, the result is:

$$A = 0.66\,P_o^{1/3}; \tag{4}$$

The corresponding fit (**Figure 1**, dashed line) is statistically inferior: the standard deviation, 250, is larger than for **Eq. 3**, 209.

To present the second and third of the problems affecting Ref. [1], we must consider the key steps in its derivation of the cube-root law. In a nutshell, the time spent in communications was considered as the essential factor in parliament effectiveness. And this time was linked to the number of communication channels.

Two kinds of channels were considered: first, those between each parliament member and his/her active constituency. The average number of such channels per member is:

$$C_C \approx kP_o/A, \tag{5}$$

Where kP_o is the fraction of the population that is politically involved.

The second type of communication channels connects different members of the assembly, to discuss and implement the measures identified by the first type of channels. While communicating between them, two assembly members share the same channel, and it was argued in [1] that the total number of channels is in this case:

$$C_A = (1/2)A(A-1), \tag{6}$$

Which, except for unrealistically small assemblies, can be approximated as:

$$C_A \approx A^2/2 \tag{7}$$

What is the relation between C_C and C_A? Ref. [1] simply assumed that for maximum effectiveness $C_C = C_A$, leading to:

$$A = (2k)^{1/3} P_o^{1/3} \tag{8}$$

That is, to the cube-root law of **Eq. 1**, with $a = (2k)^{1/3}$.

However, this logic frame is affected by two conceptual problems. First, **Eq. 5** applies to the channels between one member of the assembly and the corresponding constituency, whereas **Eqs. 6** and **7** give the number of inter-assembly channels for all members. For one member, instead of **Eqs. 6** and **7** one must use:

$$C_{\mathrm{A}} = (1/2)(A-1) \approx A/2 \tag{9}$$

Which, assuming again that $C_{\mathrm{C}} = C_{\mathrm{A}}$, leads to:

$$A = (2k)^{1/2} P_{\mathrm{o}}^{1/2} \tag{10}$$

Not a cube-root law but a square-root law [2].

To better understand why **Eq. 9** is correct and **Eqs. 6** and **7** are not, imagine that the inter-assembly "communication channels" are only used for speeches. A single assembly member shares with each speaker one channel, and the total number of his/her channels corresponds to the number of speakers, i.e., of representatives, and not to its square. This changes the cube-root law into a square-root law.

The other flaw in the above logic frame is that there is absolutely no evidence supporting its hypothesis that $C_{\mathrm{C}} = C_{\mathrm{A}}$. On the contrary, this assumption causes problems. In the original work of Ref. [1], it led to **Eq. 8**, and the corresponding forced best fit of **Eq. 4** would give $k \approx 14\%$, which hopefully is too low. And would become a catastrophic 0.3% with the unrestricted best fit of **Eq. 3**.

The balance between different types of communications can actually change from country to country and evolves with time. For example, modern communication instruments can reduce C_{C}. Symmetrically, effective negotiators can decrease C_{A}. Thus, assuming *a priori* that $C_{\mathrm{C}} = C_{\mathrm{A}}$ is arbitrary.

Supposing instead that $C_{\mathrm{C}}/C_{\mathrm{A}} = x$, **Eqs. 8** and **10** become:

$$A = (2k/\mathrm{x})^{1/3} \mathrm{P}_{\mathrm{o}}^{1/3} \tag{11}$$

$$A = (2k/\mathrm{x})^{1/2} \mathrm{P}_{\mathrm{o}}^{1/2} \tag{12}$$

In both cases, the multiplication factor is a combination of k and x, which cannot be disentangled from each other by best-fitting the data. One could perhaps estimate k from independent information like literacy, party membership and voter participation. But evaluating x is extremely difficult because of its multiple, competing and evolving causes and the lack of data.

DISCUSSION

The difficulties in evaluating x and k negatively impact the use of a power law to identify the "optimal" size of a national assembly. And other problems affect this approach.

Note that Ref. [1] tried to link the populations not to the "optimal" parliament sizes but to the real sizes, using data for countries of all kinds. Of these, many if not most were plagued by corruption, ineffective bureaucracy and/or authoritarian regimes. Thus, they could hardly lead to "optimal" values of A.

Hypothetically, one could try to extract an "optimal" value by using a subset of "good" countries, perhaps those with low indexes for corruption and bureaucratic ineffectiveness. However, not even filtering could solve the fourth problem affecting Ref. [1]: accuracy. In fact, any evaluation of A with a power law is very sensitive to the exponent. Taking the derivative of **Eq. 2** one obtains:

$$(\mathrm{d}A/\mathrm{d}n) = aP_{\mathrm{o}}^{\mathrm{n}} \ln(P_{\mathrm{o}}) = A \ln(P_{\mathrm{o}}) \ ,$$

$$\mathrm{d}A/A = 1\mathrm{n}(P_o)\mathrm{d}n$$

since P_{o} is large, an uncertainty dn, however small, is multiplied by a big factor $\ln(P_{\mathrm{o}})$ and produces a large relative uncertainty dA/A. For example, the dn uncertainty ±0.03 from **Eq. 3**, with a population of just ≈617,000, would bring dA/A to ≈40%, large enough to accommodate most political preferences.

In short, accurately evaluating the "optimal" size of a national assembly is illusory. And trying to inject additional factors besides the population cannot solve the above problems.

At most, this kind of approach can identify the countries that strongly deviate from the "average", as Ref. [2] did for France, the USA and Italy. However, without filtering the "average" is for a mix of "good" and "bad" countries, thus a deviation from it is not necessarily negative . . .and could even be positive!

In conclusion, we surprisingly found that the historical and very influential work of Taagepera [1] used a wrong equation to derive its famous cube-root law and arbitrarily assumed time equipartition between inter-assembly and assembly-constituency communications. An unrestricted best fit of the original data does not support the cube-root law and would favor instead a power law with an exponent larger than 1/3. These flaws fatally undermine the foundations of the cube-root law and disqualify - also for other reasons - its popular use to evaluate the "optimal" parliament size for a country.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: Reference 1 in the manuscript.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

Work supported in kind by the Ecole Polytechnique Fédérale de Lausanne (EPFL).

REFERENCES

1. Taagepera R. The size of national assemblies. *Soc Sci Res* (1972) 1:385–401. The data extracted from this article and used here can be found in https://sciencehistory.epfl.ch/physics-and-sociology/
2. Auriol E, Gary-Bobo RJ. On the optimal number of representatives. *Publ Choice* (2012) 153:419–445. doi:10.1007/s11127-011-9801-3
3. Auriol E, Gary-Bobo RJ. *The More the Merrier? Choosing the optimal number of representatives in modern democracies*. Available at: https://voxeu.org/optimal-number-representatives-democracy (2007).
4. De Sio L, Angelucci G. 945 sono troppi? 600 sono pochi? *Qual è il numero "ottimale" di parlamentari?*. Available at: https://cise.luiss.it/cise/2019/10/09/945-sono-troppi-600-sono-pochi-quale-e-il-numero-ottimale-di-parlamentari (2019).
5. Jacobs K, Otjes S. *Explaining reforms of assembly sizes*. Available at: https://ecpr.eu/Filestore/PaperProposal/3bc100be-56fe-4efc-8d8c-a1f0b85e7f24.pdf (2014)
6. Jacobs K, Otjes S. Explaining the size of assemblies. A longitudinal analysis of the design and reform of assembly sizes in democracies around the world. *Elect Stud* (2015) 40:280–292. doi:10.1016/j.electstud.2015.10.001
7. Mantovani MC, Ribeiro HV, Moro MV, Picoli S, jr, Mendes RS. Scaling laws and universality in the choice of election candidates. *European Phys. Letters* (2011) 96:48001. doi:10.1209/0295-5075/96/48001
8. Mantovani MC, Ribeiro HV, Lenzi EK, Picoli S, Mendes RS. Engagement in the electoral process: scaling laws and the role of political positions. *Phys Rev. E* (2013) 88:024802. doi:10.1103/PhysRevE.88.024802
9. Bettencourt LMA, Lobo HD, Kühnert C, West GB. Growth, innovation, scaling, and the pace of life in cities. *Proc Natl Acad Sci Unit States Am* (2007) 104: 7301–7306. doi:10.1073/pnas.0610172104
10. Kendall MG, Stuart A. The law of the cubic proportion in election results. *Br J Sociol* (1950) 1:183–196. doi:10.2307/588113
11. Gudgin G, Graham PJ (2012). *Seats, votes, and the spatial organisation of elections*. London, UK: ECPR Press, Chapter 3.

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Physics

PERSPECTIVE
published: 19 January 2021
doi: 10.3389/fphy.2020.584101

Possible Indications of Variations in the Directionality of Beta-Decay Products

Peter A. Sturrock[1], **Ephraim Fischbach**[2], **Oksana Piatibratova**[3] **and Felix Scholkmann**[4]*

[1]Center for Space Science and Astrophysics and Kavli Institute for Particle Astrophysics and Cosmology, Stanford University, Stanford, CA, United States, [2]Department of Physics and Astronomy, Purdue University, West Lafayette, IN, United States, [3]Geological Survey of Israel, Jerusalem, Israel, [4]Research Office for Complex Physical and Biological Systems, Zurich, Switzerland

Some experiments seem to yield strong evidence of variability of beta-decay rates, but other experiments may show little or no such evidence. Some recent experiments help clarify the situation. In particular, a certain oscillation appears in neutrino measurements made at the Super-Kamiokande Neutrino Observatory and in radon beta-decay measurements made at the Geological Survey of Israel, with identical frequency (9.43 years^{-1}), amplitude and phase, strengthening the case for an influence of neutrinos on beta decays. A review of current experimental information leads us to suggest that 1) beta-decay rates do not change, but 2) the angular distribution of decay products may be anisotropic, and 3) the angular distribution of decay products may be influenced by the ambient neutrino flux. It appears that experiments at standards laboratories tend to be insensitive to direction, and this may be the reason that they tend not to exhibit evidence of variability.

Keywords: radioactivity, radioactive decay, neutrinos, radon, anisotropy, solar interior

OPEN ACCESS

Edited by:
Nunzio Itaco,
University of Campania Luigi Vanvitelli,
Italy

Reviewed by:
Giuseppe Verde,
Ministry of Education, Universities and Research, Italy
Pasquale Migliozzi,
National Institute of Nuclear Physics of Naples, Italy

****Correspondence:***
Felix Scholkmann
Felix.Scholkmann@usz.ch

Specialty section:
This article was submitted to Nuclear Physics, a section of the journal Frontiers in Physics

Received: *16 July 2020*
Accepted: *17 November 2020*
Published: *19 January 2021*

Citation:
Sturrock PA, Fischbach E, Piatibratova O and Scholkmann F (2021) Possible Indications of Variations in the Directionality of Beta-Decay Products. Front. Phys. 8:584101. doi: 10.3389/fphy.2020.584101

INTRODUCTION, INCLUDING EARLY EVIDENCE FOR VARIABILITY

There has for some time been evidence that some beta decay processes exhibit some form of variability. Whether or not beta decays are intrinsically variable is significant for geologists who rely on radon measurements to probe the outer layers of the lithosphere. Whether or not the solar neutrino flux is variable is important not only to solar physicists, but also to physicists for whom solar-neutrino measurements yield a test of our comprehension of nuclear physics.

Alburger et al. [1] reported the results of their study at the Brookhaven National Laboratory (BNL) of the decay of ^{32}Si over the time period 1982–1986, using the long-lived nuclide ^{36}Cl as a calibration standard. Reviewing the ratio of the ^{32}Si count rate to the ^{36}Cl count rate, Alburger et al. noted "*small periodic annual deviations of the data points from an exponential decay curve* [that were] *of uncertain origin.*" One may note that the depths of modulation–of order 0.05%–for the two nuclides are similar, even though there is a wide difference in the decay half-lives (172 years for ^{32}Si, 300,000 years for ^{36}Cl).

Siegert et al. [2], at the Physikalisch-Technische Bundesanstalt (PTB), reported the results of a 20-year study of the beta decays of ^{152}Eu and ^{154}Eu, using ^{226}Ra as a standard. They noted annual oscillations in the measured decay rates of both ^{152}Eu and ^{226}Ra.

Falkenberg [3] claimed to find evidence of an annual oscillation in the beta decay rate of tritium, which he attributed to the annual variation of the Earth-Sun distance, suggesting a possible role of neutrinos.

Parkhomov [4] has found evidence of variability in beta decays but not in alpha decays.

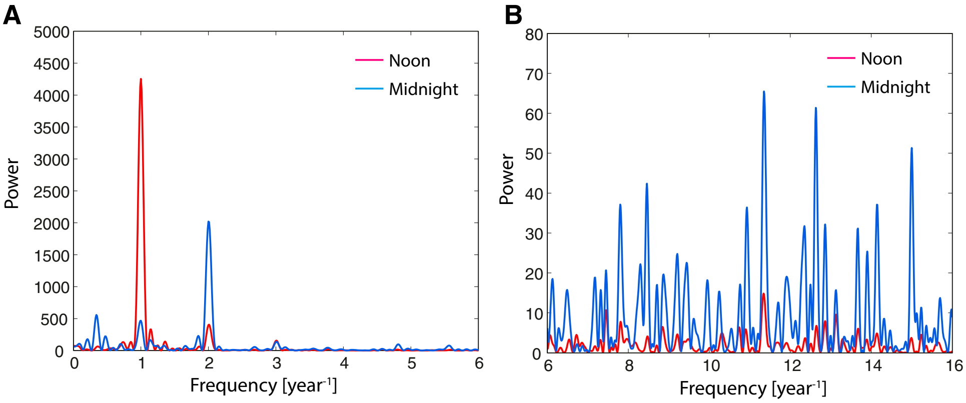

FIGURE 1 | (A) Power spectra formed from the 4-h band of measurements centered on noon (red) and midnight (blue) for the frequency band 0–6 years^{-1}. We see that the biggest daytime oscillation is at 1 year^{-1}; the biggest nighttime oscillation is at 2 years^{-1}. **(B)** Power spectra formed from the 4-h band of measurements centered on noon (red) and on midnight (blue) for the frequency band 6–16 years^{-1}. We see that there are strong oscillations in the expected rotational frequency band 10–14 years^{-1} (note especially the peaks at 11.35 and 12.64 years^{-1}) in the nighttime data, but comparatively small oscillations in the daytime data (Cf **Tables 3**, **4**).

FIschbach et al. [5], in their review of the field, presented an overview of (then) recent research dealing with "*the question of whether nuclear decay rates (or half-lives) are time-independent constants of nature, as opposed to being parameters which can be altered by an external perturbation.*" It was then not unreasonable to assume that variations in flux measurements should be interpreted as variations in decay rates. It was also not unreasonable to attribute an annual variation in these measurements to the annual variation in the Earth-Sun distance.

An overview of reported anomalies in decay rates can be found in the recently published work of McDuffie et al. [6].

In this article, we claim that more recent experiments yield conclusive evidence of variability. We point out that a reanalysis of the experimental results shows that an apparent conflict between experimenters who find evidence of variation and experimenters who do not find such evidence hinges on the conventional understanding of the role of neutrinos. We suggest that the conventional role may need to be revised, along lines suggested in the *Discussion* section.

EARLY EVIDENCE AGAINST VARIABILITY

Whether or not nuclear decay rates are constant or variable is clearly a question of interest to standards laboratories. As we noted in the *introduction, Including Early Evidence for Variability*, analysts at PTB reported apparent variations in decay measurements of ^{152}Eu and ^{226}Ra, but Nahle and Kossert [7] of PTB advanced reasons to discount the early results as evidence of variability. Kossert and Nahle [8] later claimed that measurements of ^{90}Sr/^{90}Y decays, in a specially designed experiment, gave no evidence of variability. However, a re-analysis of the Kossert-Nahle measurements [9] revealed evidence of variability.

Pommé, of the European Commission Joint Research Center, and his collaborators have published many articles discounting evidence of the variability of nuclear decay processes. An early article of this group [10] gives a summary of 67 measurements of the decay rates of several different nuclides, giving results from several different laboratories, covering several different decay mechanisms. Measurements were made by a wide variety of techniques, but most of the datasets were of limited length (less than 1,000 lines). The individual datasets were tested for annual oscillations, which were typically found to be a small fraction of a percent with phases that varied over a wide range. Pommé et al. concluded that "*the observed seasonal modulations could be attributed to instrumental instability*" [10]. For a recent publication and guide to earlier articles, see [11].

TABLE 1 | The annual oscillation and the leading two harmonics as derived from the GSI noon-centered measurements.

Frequency (year^{-1})	Power	Amplitude (%)	Phase of maximum
1	4,254	4.65	0.49
2	400	1.40	—
3	153	0.87	—

TABLE 2 | The annual oscillation and the leading two harmonics as derived from the midnight-centered measurements.

Frequency (year^{-1})	Power	Amplitude (%)	Phase of maximum
1	468	0.72	0.39
2	2020	1.48	—
3	134	0.38	—

RADON DECAY MEASUREMENTS ACQUIRED AT THE GEOLOGICAL SURVEY OF ISRAEL LABORATORY

The most extensive set of nuclear decay measurements is one that has been acquired at the Geological Survey of Israel (GSI)

TABLE 3 | Top 20 peaks in the power spectrum formed from GSI noon data in the frequency band 6–16 years^{-1}. Entries in bold comprise a triplet and two doublets with frequency separations close to 1 year^{-1}: 7.45, 8.45 and 9.45 year^{-1}; 11.35 and 12.35 year^{-1}; and 12.65 and 13.65 year^{-1}.

Frequency (year^{-1})	Power	Order
6.07	4.4	16
6.72	4.5	15
7.45	**10.7**	**2**
7.81	7.8	5
7.96	3.5	20
8.47	**4.1**	**17**
8.85	6.5	7
9.21	4.6	13
10.31	**5.0**	**11**
10.74	6.4	8
10.90	5.9	10
11.34	**14.9**	**1**
12.37	3.7	19
12.65	**6.8**	**6**
12.86	7.9	4
13.13	9.6	3
13.67	**6.0**	**9**
14.14	4.9	12
14.99	3.7	18
15.24	4.5	14

TABLE 4 | Top 20 peaks in the power spectrum formed from midnight data in the frequency band 6–16 years^{-1} Entries in bold comprise a triplet and two doublets with frequency separations close to 1 year^{-1}: 7.45, 8.45 and 9.45 year^{-1}; 11.35 and 12.35 year^{-1}; and 12.65 and 13.65 year^{-1}.

Frequency (year^{-1})	Power	Order
6.13	18.5	19
7.18	18.9	18
7.45	**20.7**	**15**
7.80	37.1	5
8.30	22.2	14
8.46	**42.4**	**4**
8.87	19.6	16
9.21	24.8	12
9.44	**22.6**	**13**
9.95	18.2	20
10.93	36.4	7
11.35	**65.5**	**1**
11.91	19.1	17
12.35	**31.7**	**9**
12.63	**61.4**	**2**
12.86	32.2	8
13.67	**31.1**	**10**
13.90	25.4	11
14.14	37.1	6
15.00	51.3	3

laboratory in Jerusalem. This experiment, in operation from day 86 of 2007 to day 312 of 2016, recorded, every 15 min, measurements of beta-related gamma rays, alpha radiation, and three environmental measurements (temperature, pressure and supply voltage), for a total of over 350,000 lines, each with seven entries [12].

Figures 1A,B show power spectra for the frequency ranges 0–6 year^{-1} and 6–16 year^{-1}, respectively. These figures show power spectra formed from gamma measurements acquired at local noon (shown in red) and at local midnight (shown in blue). The principal peaks in these power spectra are listed in **Tables 1–4**.

We find that the strongest oscillation is an annual oscillation, found primarily in the noon data, with a power of 4,254. (There is also a strong semiannual oscillation.) According to the standard expression

$$P = e^{-S} \tag{1}$$

for the probability of obtaining by chance a power of S or more at a selected frequency [13] from normally distributed random measurements, this value of the power corresponds to a false-alarm probability of less than 10^{-1700}. (This obviously rules out any environmental effect, such as has been suggested by Pommé [11].

We see from **Figure 1B** and **Table 4** that the two strongest oscillations in the frequency band 6–16 year^{-1} (which covers the frequency band expected for solar rotation) are found in the midnight data at 11.35 year^{-1} with $S = 65.5$ and at 12.63 year^{-1} with $S = 61.4$. The geometry of the experiment is such that the detector reveals signals traveling vertically upwards. Since these signals have originated in the Sun, they have traveled through the Earth, indicating that the radon beta-decay photons somehow have their origin in neutrinos.

However, the gamma detector records a stronger signal at noon (**Table 1**). By analogy with the solar influence (attributed to solar neutrinos) detected primarily at midnight, we are led to consider the possibility that the influence detected primarily at noon may be attributable to neutrinos traveling *toward* the Sun. These can only be cosmic neutrinos, which will be the topic of a later article.

EVIDENCE FOR ANISOTROPY

A significant variation of the basic GSI experiment was carried out at the GSI laboratory in late 2013 [14]. The setup comprised two cylinders at right angles to each other. One cylinder was oriented to be parallel to the Earth's rotation axis. The axis of the other cylinder was oriented to be in a vertical plane that contained the axis of the first cylinder, oriented to be normal to the axis of the first cylinder. If the measurements acquired by the two detectors appended to the two cylinders were subject to an isotropic influence (such as weather), the two detectors would have recorded identical measurements.

That did not happen. The apparent half-life of the radon source was found to be 0.861 ± 0.003 days in the pole direction and 2.308 ± 0.008 days in the orthogonal direction. The authors comment that *"the outcome is in conformity with observations on radon signals in confined conditions and their different manifestation at different directions."*

This experiment provides conclusive evidence that *whatever process influences the beta decay process is intrinsically anisotropic*. Any interpretation of beta-decay measurements must take this fact into account.

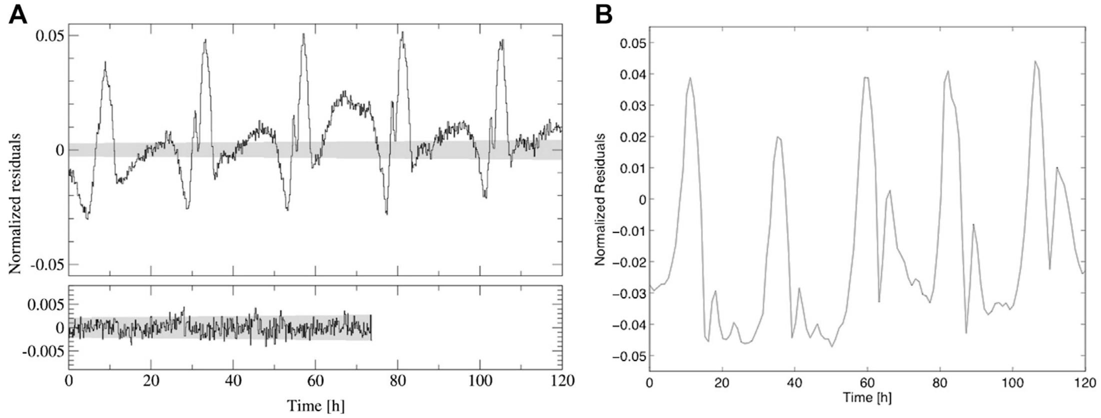

FIGURE 2 | (A) Normalized deviations from the exponential trend of the measured count rate for radon in air measured by Bellotti et al. [15]. The upper panel shows the data without polystyrene particles inside the glass sphere containing air charged with radon, the lower panel with these particles. Fluctuations within ±2σ are shown with shaded areas. Reprinted from Bellotti et al. [15] with permission from Elsevier. **(B)** A plot of a 120-h sample of the normalized hourly measurements of the gamma count rate registered by the GSI experiment (*Radon Decay Measurements Acquired at the Geological Survey of Israel Laboratory*), showing a diurnal variation very similar to that shown in **(A)**.

EVIDENCE FOR AN INFLUENCE OF THE ENVIRONMENT

Two experiments by Bellotti et al. [15] offer further information relevant to the mechanism of variability of nuclear decays.

The First Bellotti Experiment

A glass sphere (130 mm diameter) was connected through a pipe to a stainless steel cylinder containing 0.3 kg of rock rich in uranium. The radon from the radium decay fills the glass sphere which, after 5–6 days, was isolated from the radon source. Gamma rays from the radon progeny were detected by a 3” by 3” NaI crystal placed a few millimeters from the surface of the sphere. Both the detector and the glass sphere were enclosed in a 5 cm thick lead shield.

The normalized residual of the count rate, divided by the expected rate, is shown as a function of time in the upper panel of **Figures 2A**. The experimenters reported that “*instead of having a statistical distribution around zero, there is clearly a 24 h period.*” They found the same behavior if they took into account only the peaks due to ^{214}Pb and ^{214}Bi.

The Second Bellotti Experiment

Bellotti and his colleagues speculated that the diurnal modulation evident in their first experiment could be attributed to the displacement of the radioactive nuclei inside the gas volume, together with a variation of the detector efficiency. The experimenters set out to evaluate that hypothesis by filling the sphere with polystyrene particles (diameter: 0.7–0.9 mm), so that the radon atoms were confined to the interstitial space between the polystyrene particles. Measurements made with that configuration exhibited no modulation, as shown in the lower panel of **Figure 2A**. The authors concluded that the displacement of radioactive atoms was the cause of the diurnal modulation.

As a further check of that hypothesis, the experimenters added a second NaI detector diametrically opposite to the first detector. They found that the variation with time of the difference in count rates of the two detectors was very similar to the difference in the temperatures at the locations of the two detectors, and inferred that the diurnal variation of the count rate evident in their first experiment was attributable to the diurnal variation of the location of the radioactive nuclei inside the gas volume.

However, we show in **Figure 2B** a short section (120 h) of gamma measurements recorded by the GSI experiment. We see that this experiment exhibits a diurnal oscillation very similar to that recorded by the Bellotti experiment (**Figures 2A**)–similar in both amplitude and structure. This strong similarity suggests that both experiments are responding to a similar or identical influence.

As a further test, Bellotti et al. modified the experiment to “immobilize” radon and its progeny, allowing at the same time for a rather high radon concentration. To achieve these goals, the experimenters diffused radon into olive oil which has a much higher viscosity than air and which permits a radon concentration 29 times higher than in air. To minimize the background and its fluctuations, the experiment was carried out underground at the Gran Sasso National Laboratory (LNGS). The olive oil, charged with radon, was contained in a copper tube, 10 cm diameter, with wall thickness 2 mm. The detector was again a 3” by 3” NaI detector (but its relationship to the tube has not been specified). The shielding was provided by at least 15 cm of lead and the laboratory temperature was kept between 12°C and 13°C.

Measurements were made for four intervals of lengths ranging from 1,185 to 1,462 h. Their analysis of these four intervals gave no evidence of variability: the relative half-life variation was 7×10^{-6}, one order of magnitude less than the statistical error. The experimenters concluded that their final result was a very precise value for the ^{222}Rn half life of $3.82146\ (16)_{\mathrm{stat}}(4)_{\mathrm{syst}}$ d. The experimenters remark that using radon diffused in olive oil removed the large fluctuations (presumably the diurnal oscillations) in the count rate that were a feature of the first experiment.

TABLE 5 | The frequency, amplitude and phase of the (nominally) 9.43 years^{-1} oscillation, as it occurs in Super-Kamiokande data and GSI data.

	Super-Kamiokande neutrino measurements	GSI radon-decay measurements
Frequency	9.43 ± 0.04 years^{-1}	9.44 ± 0.03 years^{-1}
Amplitude	6.8 ± 1.7%	7.0 ± 1.0%
Phase	124 ± 15°	124 ± 9°

AN OSCILLATION EVIDENT IN BOTH SUPER-KAMIOKANDE AND GEOLOGICAL SURVEY OF ISRAEL MEASUREMENTS

The Super-Kamiokande (SK) Observatory, which has been in operation for 35 years (since 1985, with one unfortunate lapse), began data-taking in 1996 and released 5 years of data in 2003. There have been a number of analyses of that dataset. One from the SK Consortium [16] claimed to establish that their dataset yields no evidence of variability. However, concerning the two most detailed analyses of that dataset, that of Sturrock and Scargle [17] revealed evidence of a significant oscillation at 9.43 year^{-1}, and that of Ranucci et al. [18] contains the following conclusion: "*multiple peaks significance assessment and alias prediction delineate a … complex picture in which a line at 9.42 cycles/year … emerges in the spectrum with an individual significance which cannot exclude the constant rate hypothesis, but accompanied by other indicators that do not fully endorse such a conclusion.*"

We have carried out analyses of the SK and GSI datasets using an extension of the Lomb-Scargle procedure that yields amplitude and phase as well as power. The result is shown in **Table 5** and **Figure 3**.We see that *there is remarkable agreement not only in frequency but also in amplitude and phase*.

We see from **Table 3** that not only is the 9.43 year^{-1} oscillation evident in GSI data, but we also find two annual sidebands (effectively at 8.43 and 7.43 year^{-1}). Such sidebands are attributable to oblique rotation, i.e., to rotation about an axis that is not normal to the ecliptic [19].

DISCUSSION

The results shown in the previous section concerning the GSI measurements present a strong case that some nuclear decay processes are in some sense variable. **Tables 1**, **2** and **Figures 1A** present evidence that beta-decay measurements are influenced by the varying Earth-Sun configuration. **Figure 1B** and **Tables 3**, **4** strongly suggest that these measurements are influenced by solar rotation. This inference is reinforced by the finding that some of the rotational oscillations are accompanied by sidebands with displacements of 1 year^{-1}. Such sidebands are comprehensible if the rotation axis departs significantly from the normal to the ecliptic [19] (and therefore differs significantly from the rotation axis inferred from optical observations). Different oscillations presumably correspond to different regions of the solar interior, the triplet at effectively 7.43, 8.43, and 9.43 year^{-1}, possibly corresponding to the solar core.

We saw in *Evidence for an Influence of the Environment* that a completely different experiment (the first Bellotti experiment) exhibits a diurnal oscillation very similar to that found in the midnight data of the GSI experiment. The agreement is one not only of shape but also of magnitude. Since the same pattern is found in two completely different experiments, it can hardly be attributed to any environmental process. One must suspect that there is a physical explanation for this relationship.

We saw in *An Oscillation Evident in Both Superkamiokande and GSI Measurements* that an oscillation (at 9.43 years^{-1}) is evident in both Superkamiokande solar neutrino measurements and GSI radon-decay measurements. Remarkably, *the agreement is not simply one of frequency but also one of amplitude and phase*. It is difficult to avoid the conclusion that both sets of measurements have a common cause. The simplest such interpretation is that *neutrinos somehow influence beta decays*.

The second Bellotti experiment, discussed in *Evidence for an Influence of the Environment*, shows that *evidence for variability is suppressed if the radiation is isotropized by scattering*. An apparent implication is that (as previously suggested [20]) *variability involves the directional characteristics of measurement*s, not simply time dependence. Indeed, it is possible that time variation of

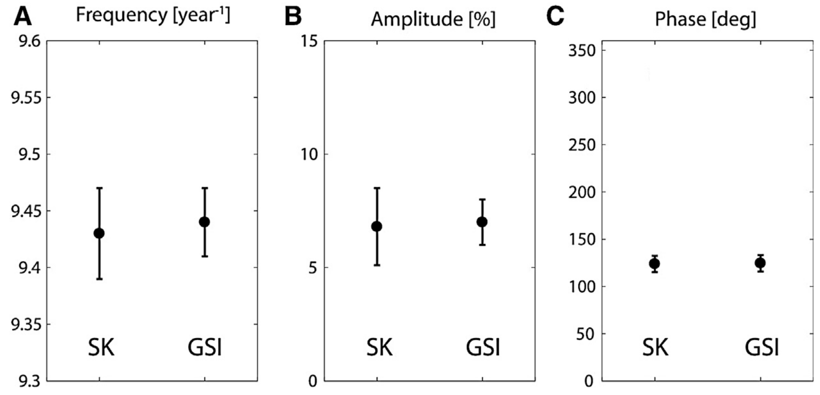

FIGURE 3 | Frequency, amplitude and phase of the 9.43 years^{-1} oscillation as it appears in Super-Kamiokande (SK) measurements and in GSI measurements.

measurements may actually be due to an anisotropy of what is being measured. *Measurements made by an experiment that isotropizes radiation would then yield no evidence of variability.*

Hence there may be no conflict between evidence of variability reviewed in this article and the null findings of many standards experiments, in which the target nucleus may be part of a molecule in a chemical form that is dissolved in a "cocktail" contained in a vial that may or may not be transparent, since such a design would tend to isotropise radiation from the nucleus under investigation.

The process by which neutrinos might influence beta decays is (if real) currently unknown. That such a process may exist would seem surprising, in view of the known very weak interaction of neutrinos with other particles. In seeking to reconcile these two apparently contradictory properties of neutrinos, one may consider as a possible analogy the interaction of electrons and ions in an electron-ion plasma. In that situation, there are two types of interaction: one is the direct short-range particle-particle interaction (typically negligible); the other is a long-range *collective* process by which large numbers of charged particles can interact [21]. One may therefore consider the possibility that the influence of neutrinos on nuclear processes may be a *collective* process, not a particle-particle process. This would require a mechanism for a long-range coupling of neutrinos, which may require the mediation of a boson to play the same role as the electromagnetic field in an electron-ion plasma. The examination of this hypothesis may require a new suite of experiments, including a search for evidence of spatial correlation that might be expected of a collective process, but mght not be expected of a non-collective process.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

Conceptualization, software, visualization, original draft preparation: PS; review and editing: EF, OP, FS.

ACKNOWLEDGMENTS

We thank the Super-Kamiokande consortium, the Geological Survey of Israel, and Karsten Kossert and Ohle Nahle of the Physikalisch Technische Bundesanstalt, for making their data available for analysis. We also thank Jeffrey Scargle and Guenther Walther for their valuable guidance concerning time-series analysis.

REFERENCES

1. Alburger DE, Harbottle G, Norton EF. 'Half-life of ^{32}Si'. *Earth Planet. Sci Lett* (1986) 78(2-3):168-76. doi:10.1016/0012-821x(86)90058-0
2. Siegert H, Schrader H, Schoetzig U. Half-life measurements of europium radionuclides and the long-term stability of detectors. *Appl Radiat Isot* (1998) 49(9-11):1397–401. doi:10.1016/s0969-8043(97)10082-3
3. Falkenberg ED. Radioactive decay caused by neutrinos? *Apeiron* (2001) 8(No. 2):32–45.
4. Parkhomov AG. Bursts of count rate of beta-radioactive sources during long-term measurements. *Int J Pure Appl Phys* (2005) 1:119–28.
5. Fischbach E., Buncher JB, Gruenwald JT, Jenkins JH, Krause DE, Mattes JJ, et al. Time-dependent nuclear decay parameters: new evidence for new forces?. *Space Sci Rev* (2009) 145(3-4):285–335. doi:10.1007/s11214-009-9518-5
6. McDuffie MH, Graham P, Eppele JL, Gruenwald JT, Javorsek II D, Krause DE, Fischbach E. Anomalies in Radioactive Decay Rates: A Bibliography of Measurements and Theory. https://arxiv.org/abs/2012.00153
7. Nahle O, Kossert K. Comment on "Comparative study of beta-decay data for eight nuclides measured at the Physikalisch-Technische Bundesanstalt" [Astropart. Phys. 59 (2014) 47-58]. *Astropart Phy* (2015) 66:8–10. doi:10.1016/j.astropartphys.2014.11.005
8. Kossert K, Nahle OJ. Disproof of solar influence on the decay rates of 90Sr/90Y. *Astropart Phy* (2015) 69:18–23. doi:10.1016/j.astropartphys.2015.03.003
9. Sturrock PA, Steinitz G, Fischbach E, Parkhomov A, Scargle JD. Analysis of beta-decay data acquired at the Physikalisch-Technische Bundesanstalt: evidence of a solar influence. *Astropart Phy* (2016) 84:8–14. doi:10.1016/j.astropartphys.2016.07.005
10. Pommé S, Stroh H, Paepen J, Van Ammel R, Marouli M, Altzitzoglou T, et al. Evidence against solar influence on nuclear decay constants. *Phy Letters B* (2016) 761:281–86. doi:10.1016/j.physletb.2016.08.038
11. Pommé S. Solar influence on radon decay rates: irradiance or neutrinos? *Eur. Phys. JC.* (2019) 79:73. doi:10.1140/epjc/s10052-019-6597-7
12. Steinitz G, Kotlarsky P, Piatibratova O. Radon signals in geological (natural) geogas and in a simultaneous enhanced confined mode simulation experiment. *Proc Math Phys Eng Sci* (2018) 474:2216. doi:10.1098/rspa.2017.0787
13. Scargle JD. Studies in astronomical time series analysis. II - Statistical aspects of spectral analysis of unevenly spaced data. *ApJ* (1982) 263:835–53. doi:10.1086/160554
14. Steinitz G, Kotlarsky P, Piatibratova O. Observations of the relationship between directionality and decay rate of radon in a confined experiment. *Eur Phys J Spec Top* (2015) 224(4):731–40. doi:10.1140/epjst/e2015-02403-2
15. Bellotti E, Broggini C, Di Carlo G, Laubenstein M, Menegazzo R. Precise measurement of the ^{222}Rn half-life: a probe to monitor the stability of radioactivity. *Phy Lett B* (2015) 743:526–30. doi:10.1016/j.physletb.2015.03.021
16. Yoo J, Ashie Y, Fukuda S, et al. Search for periodic modulations of the solar neutrino flux in Super-Kamiokande-I. *Phys Rev D* (2003) 68:092002
17. Sturrock PA, Scargle JD. Comparative analysis of Super-Kamiokande and SNO Solar-Neutrino Data and the photospheric magnetic field. *Sol Phys* (2006) 239(1-2):1–27. doi:10.1007/s11207-006-0143-0
18. Ranucci G. Likelihood scan of the Super-Kamiokande I time series data. *Phys Rev D* (2006) 73:103003. doi:10.1103/physrevd.73.103003
19. Sturrock PA, Bai T. Search for evidence of a clock related to the solar 154 day complex of periodicities. *ApJ* (1992) 397:337–46. doi:10.1086/171789
20. Sturrock PA, Steinitz G, Fischbach E. Analysis of gamma radiation from a radon source: II. Indications of influences of both solar and cosmic neutrinos. *Astropart Phys* (2018) 100:1–12.
21. Sturrock PA. *Plasma Physics: an introduction o the theory of astrophysical, geophysical and laboratory plasmas*. Cambridge, UK: Cambridge University Press, (1994).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership