frontiers
in Big Data

Discovering Topic-Oriented Highly Interactive Online Communities

Swarna Das and Md Musfique Anwar*

Department of Computer Science and Engineering, Jahangirnagar University, Dhaka, Bangladesh

Community detection is an interesting field of online social networks. Most existing approaches either consider common attributes of social network users or rely on only social connections among the users. However, not enough attention is paid to the degree of interactions among the community members in the retrieved communities, resulting in less interactive community members. This inactivity will create problems for many businesses as they require highly interactive users to efficiently advertise their marketing information. In this paper, we propose a model to detect topic-oriented densely-connected communities in which community members have active interactions among each other. We conduct experiments on a real dataset to demonstrate the effectiveness of our proposed approach.

Keywords: online social network, interaction strength, active community, query cohesiveness, structure cohesiveness

1. INTRODUCTION

Nowadays, Online Social Networks (OSN) are widely used by a large part of the general population. Similar interests, choices, and hobbies tend to form a group of users in a social network known as online community. There have been many attempts to detect these online communities for the purpose of business, marketing, recommendations, biological research, etc. Often the mere use of connection links does not provide an effective group of users. As a result, these groups do not bring efficient results.

There are two types of network topology. One is global, where information of a whole network is captured and another is local, i.e., a network that works with the similar nodes (Tang et al., 2017). There have been many approaches to detect communities and serve various other fields with it (Fortunato and Hric, 2016). An approach to detecting communities is Affinity Propagation, where the network is divided and a multiobjective evolutionary algorithm is introduced (Shang et al., 2016). For the purpose of local community formation, dynamic membership function can be used (Luo et al., 2018). Fuzzy relations can be used for non-overlapping community detection. The nearest node with each node's greater centrality and fuzzy relations are combined for the desired result (Luo et al., 2017).

Recent research works consider social users' topical interests in OSNs, e.g., (Yang et al., 2013), in order to find meaningful communities. However, these methods did not focus on the topical interactions among the community members. Therefore, such communities contain many members who have very inactive topical interactions among them which perform poorly in viral marketing. In order to avoid the inactivity problem authors (Lim and Datta, 2016) have proposed an approach where interaction pattern and frequency are considered rather than only counting the following/follower links.

Our observation is that social users have different degrees of topical intimacy among them. In this work, we propose an approach to discover topic-oriented *highly interactive* communities in OSNs, where the members in the community should have a certain degree of topical interactions with each other related to a given query. We also emphasize that the members in the retrieved communities should actively interact with at least k other members within the community. Below, we summarize our contributions:

- We propose a methodology to discover highly interactive online communities where community members have a high degree of interactions with each other on similar topics;
- We quantify the topical interaction strengths among the users;
- We perform experiments on a real dataset to demonstrate the effectiveness of the proposed method.

2. RELATED WORK

Earlier methods for community detection are based on structural information of the social graph such as modularity (Clauset et al., 2004), edge betweenness (Newman and Park, 2003), and neighborhood concepts (Cohen, 2008). Some approaches also considered the textual content published by the users along with social connections to detect like-minded users. For example, SA-Cluster applied random-walk to measure the closeness of a node in an augmented attributed graph (Zhou et al., 2009). A Topic-Link LDA model (Liu et al., 2009) is proposed which considers both the linkage structure and similarities of the contents of edges to detect communities. A probabilistic generative model named as CESNA is proposed by Yang et al. (2013) and combines community memberships, node attributes, and the network topology to find the communities.

More recently, some approaches have focused on the interaction strength between the users in order to find active communities. Dev et al. (2014) considered the impact of interaction between users as well as the impact of the group behavior without considering topical attributes of the nodes. Lim and Datta (2016) proposed the Highly Interactive Community Detection (HICD) method, which constructs a weighted network using the frequency of direct interactions between users. Correa et al. (2012) proposed the *iTop* algorithm, which constructs a weighted graph based on user interactions and maximizes the local modularity to detect topic-oriented communities based on a set of seed users. However, all these methods ignored topic-wise users' inter-activeness. Our goal is to discover communities where users have high interactions with others with regard to the given query consisting a set of topics.

3. METHODOLOGY

First we formally formulate the problem of discovering highly interactive topical communities in OSNs. Then we give an overview of our proposed approach.

Attributed Social Graph: An attributed social graph is denoted as $G = (U, E, \mathcal{A})$, where U represents the set of social users (nodes), E indicates the set of links (edges) between the users,

and $\mathcal{A}=\{T_1, T_2, ..., T_m\}$ is the set of topics discussed by the social users in G.

In Twitter, users mention each other using "@." In order to construct a link (a, b) between users, @mention is used, i.e., $M_{a,b}$ denotes that user a has posted a tweet which contains @b.

k-Core: Given an integer k ($k \geq 0$), the k-core of a graph G, denoted by C^k, is the maximal connected sub graph of G, such that $\forall u \in C^k$, $deg_{C^k}(u) \geq k$, where $deg_{C^k}(u)$ refers to the degree of a node u in C^k. A k-core component H_j^k is considered as a community from a structural point of view.

Node Core Number: The core number of a social user u in a k-core induced sub graph from G indicates the maximum k for which u belongs to that k-core sub graph.

Topic: A topic contains a set of related words that represents the topic. For example, the politics topic has words like election, vote, democracy, political party, etc.

Activity: Any action performed by a social user is referred to as an activity. For example, posting a new tweet or retweeting an existing tweet is considered as an activity. In our work, we consider only those actions that are performed between any two social users. For example, a user u in Twitter replies to a tweet posted by user v. This activity is recorded as an activity tuple $\langle u, v, \psi_{uv} \rangle$, where ψ_{uv} indicates the set of attributes (topics) exchanged between u and v (Anwar et al., 2018).

Query: An input query $Q=\{T_1, T_2..., T_n\}$ contains a set of query topics.

Active Interaction Edge: If any two social users u and v in G have a certain number of direct interactions ($\gamma(\geq 1)$) between them related to Q, then we consider the interaction link between those two users as an active interaction edge (e_{uv}). Factor w_{uv} indicates their involvement in direct interactions compared with the most active pair of users in the network.

$$w_{uv} = \frac{|ACTS(u, v, \psi_{uv})|}{max_{x,y \in U^Q}|ACTS(x, y, \psi_{xy})|} \tag{1}$$

where $ACTS(u, v, \psi_{uv})$ indicates the number of direct interactions between u and v containing $\psi_{uv} \subseteq Q$.

Active User: The users of an active interaction edge e_{uv} are considered as active users. The set of all the active users for a given query Q is denoted as U^Q.

3.1. Problem Definition

Given a graph $G = (U, E, \mathcal{A})$, an input query Q and an integer k, we first find the set of active edges between the social users by measuring interaction strength $w_{uv}(w_{uv} \in [0, 1])$. Then an induced sub graph H_j^k is considered as an active interactive community if it satisfies the following criteria.

1. **Connectivity.** $H_j^k \subset G$ is connected;
2. **Structure cohesiveness.** $\forall u \in H_j^k$ has interaction degree of at least k;
3. **Active interaction.** $\forall e_{uv} \in H_j^k$, the interaction strength of e_{uv} is $w_{uv} \geq \theta$ and $\theta \in [0, 1]$ is a threshold.

Figure 1 shows a social graph G with the core number for each node, e.g., the three-core nodes are {A,B,C,I}. **Table 1** represents
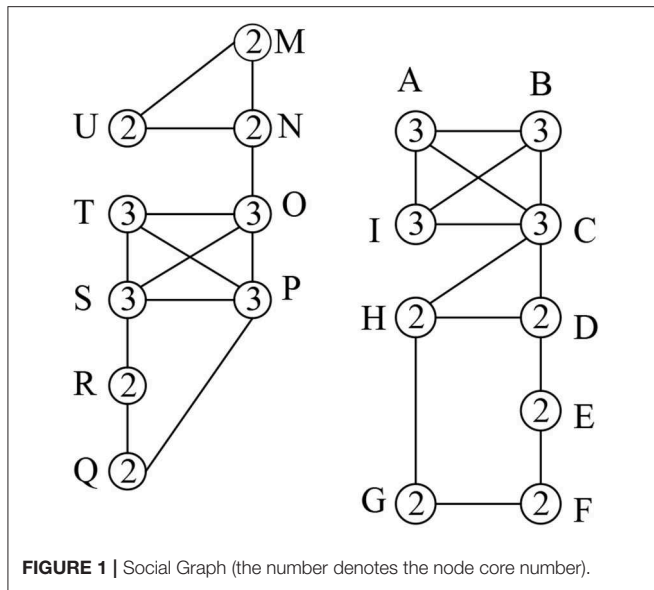
FIGURE 1 | Social Graph (the number denotes the node core number).

TABLE 1 | Interaction log.

(e_{ab})	T_1	T_2	(e_{ab})	T_1	T_2
(M,N)	6	18	(A,B)	20	7
(M,U)	3	14	(A,C)	18	6
(N,U)	5	15	(A,I)	14	6
(N,O)	6	8	(B,C)	10	13
(O,T)	20	13	(B,I)	13	8
(O,P)	19	7	(C,I)	15	9
(O,S)	11	6	(C,D)	7	12
(P,T)	14	8	(C,H)	6	17
(P,S)	18	9	(D,H)	12	9
(S,T)	16	9	(D,E)	7	18
(S,R)	12	4	(H,G)	5	8
(R,Q)	10	5	(E,F)	4	9
(P,Q)	20	9	(F,G)	6	20

the interaction frequencies among the users for topic T_1 and T_2. In **Table 2**, we show the interactive communities for a query $Q = \{T_1, T_2\}$. We get different community members for different values of Q, k, and θ. For example, when $Q=\{T_1\}$, $k = \{2\}$, and $\theta = \{0.4\}$, we get $H_1^2 = \{A,B,C,I\}$, $H_2^2 = \{O,P,Q,R,S,T\}$ while for the same values of Q and θ with an increase value of $k = \{3\}$, we get $H_1^2 = \{A,B,C,I\}$, $H_2^2 = \{O,P,S,T\}$. Again, for $Q=\{T_1, T_2\}$, $k = \{2\}$ and $\theta = \{0.5\}$, we get $H_1^2 = \{A,B,C,D,H,I\}$, $H_2^2 = \{O,P,Q,R,S,T\}$ and $H_3^2 = \{M,N,U\}$

3.2. Highly Interactive Community Detection Approach

In this work, we propose a method to detect highly interactive communities for a given a query Q in an online social attributed graph G. The desired communities from the graph G can be identified in the following three steps:

TABLE 2 | Community members for different values of Q, k, and θ.

Query	Community
$Q=\{T_1\}$, $k=\{2\}$, $\theta=\{0.4\}$	$H_1^2=\{A,B,C,I\}$, $H_2^2=\{O,P,Q,R,S,T\}$
$Q=\{T_1\}$, $k=\{3\}$, $\theta=\{0.4\}$	$H_1^3=\{A,B,C,I\}$, $H_2^3=\{O,P,S,T\}$
$Q=\{T_2\}$, $k=\{2\}$, $\theta=\{0.4\}$	$H_1^2=\{B,C,D,E,F,G,H,I\}$, $H_2^2=\{M,N,O,P,S,T,U\}$
$Q=\{T_2\}$, $k=\{2\}$, $\theta=\{0.5\}$	$H_1^2=\{C,D,H\}$, $H_2^2=\{M,N,U\}$
$Q=\{T_1,T_2\}$, $k=\{2\}$, $\theta=\{0.4\}$	$H_1^2=\{A,B,C,D,E,F,G,H,I\}$, $H_2^2=\{M,N,O,P,Q,R,S,T,U\}$
$Q=\{T_1,T_2\}$, $k=\{2\}$, $\theta=\{0.5\}$	$H_1^2=\{A,B,C,D,H,I\}$, $H_2^2=\{O,P,Q,R,S,T\}$, $H_3^2=\{M,N,U\}$
$Q=\{T_1,T_2\}$, $k=\{3\}$, $\theta=\{0.4\}$	$H_1^3=\{A,B,C,I\}$, $H_2^3=\{O,P,S,T\}$

Algorithm 1 Query Algorithm

Input: $G=(U,E)$, Q, k, θ
Output: set of active interactive communities $\Phi_Q=\{H_1^k, H_2^k, ..., H_n^k\}$

1: **for** each $(u,v) \in E$ **do**
2:     compute w_{uv}
3:     **if** $w_{uv} > \theta$ **then**
4:         $U^Q.add(u)$
5:         $U^Q.add(v)$
6: compute the induced graph G^Q on U^Q
7: compute the maximal k-core $C^k(G^Q)$ of G^Q
8: Output the set of active connected components Φ_Q from $C^k(G^Q)$

---

1. Identify the set of active users based on their direct interaction with each other for a given query Q.
2. Refine the original social graph G by filtering the inactive social users.
3. Apply k-core technique on the refined social graph in order to detect the desired online communities.

The first step of our approach is measuring the interaction frequencies among the users for a given query Q in social graph G to filter the weakly connected topology links. For this purpose, we consider users who have direct communication with others via retweets or mentions and consider an interaction link between two users irrespective of whether they have a topology link or not.

After establishing the newly active interaction edges and filtering the inactive topology links from the social graph G, we apply k-core on the refined social graph to find the connected components in which every node has degree of at-least k.

We develop an algorithmic framework to detect highly interactive communities for a given Q.

Algorithm overview. The algorithm, called Query Algorithm, has three steps. First, it computes the interaction strength w_{uv} of each edge e_{uv} for a given query Q in order to find the set of active users (line 1-5). Next, we compute the induced sub graph G^Q from U^Q (line 6). Finally, we identify the maximal k-core of $C^k(G^Q)$ from the induced graph G^Q to find the set of active connected components (i.e., desired connected communities) (line 6-7) Φ^Q from $C^k(G^Q)$ (line 7-8).

4. EXPERIMENT AND RESULT

We conduct our experiment on an academic coauthor (DBLP) dataset (Jie et al., 2008) and choose research papers that were published within 2005 to 2011. This revised dataset is a

network of 15,516 authors with 48,862 co-author relationships between these authors and contains 193,512 research papers. The co-author information in DBLP is considered as interaction between the authors. We extract the authors' details, publication year, and abstract from each research paper. We apply latent dirichlet allocation (LDA) topic modeling (Blei et al., 2003) on the abstracts of the research papers in order to find the research topics.
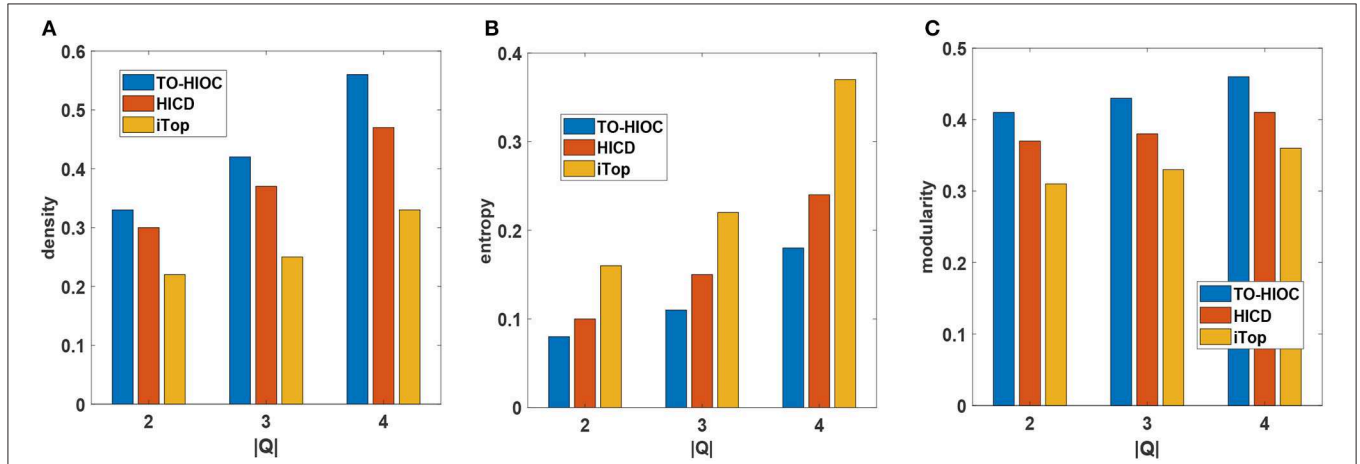
FIGURE 2 | Performance comparison on DBLP dataset **(A)** Density, **(B)** Entropy, **(C)** Modularity (in all cases, Q = {Semantic web, Data mining, Social network analysis}, k = {4}, θ = {0.5}, γ = {4}, the publications are chosen from the time period of 2005 to 2009).
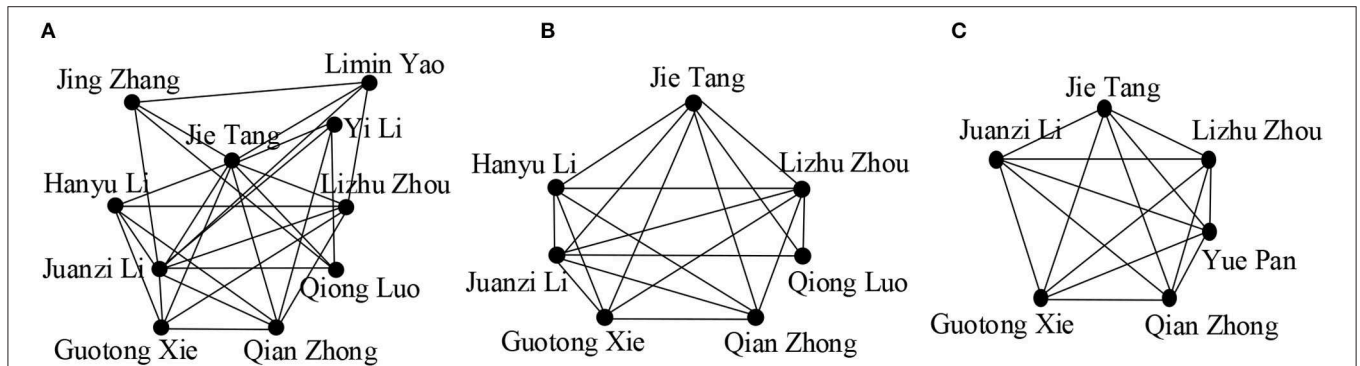
FIGURE 3 | (A) k = {3}; **(B)** k = {4}; **(C)** k = {5}; (in all cases, Q = {Semantic web}, θ = {0.5}, γ = {4}, the publications are chosen from the time period of 2005 to 2009).
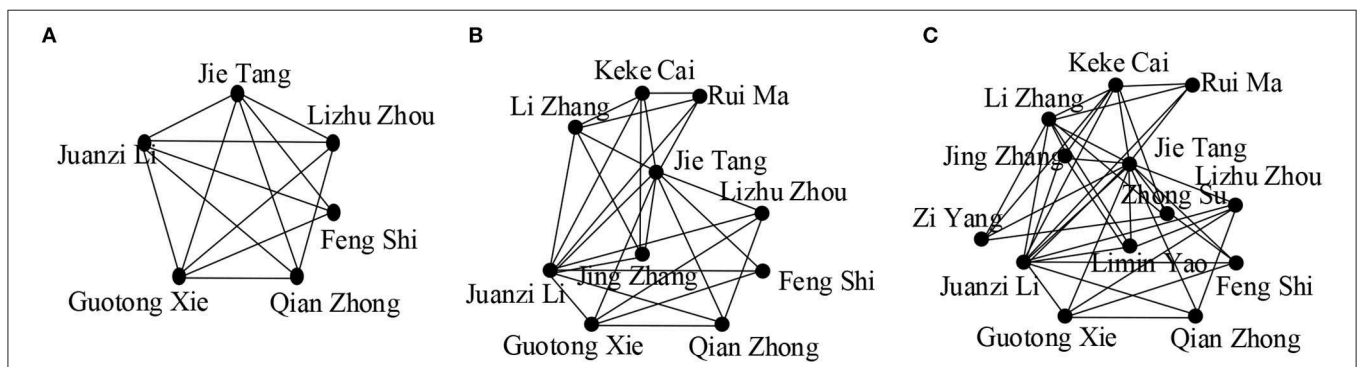
FIGURE 4 | (A) Q = {Semantic web}; **(B)** Q = {Semantic web, Data mining}; **(C)** Q = {Semantic web, Data mining, Social network analysis}; (in all cases, k = {4}, θ = {0.5}, γ = {4}, the publications are chosen from the time period of 2007 to 2011).

Comparison Methods. We compare our Algorithm 1 (Query Algorithm), denoted here as TO-HIOC, with two other existing methods: HICD method (Lim and Datta, 2016) and iTop algorithm (Correa et al., 2012).

Evaluation Measures. We vary the length of the Q to $|Q| = 2, 3, 4$ and use three measures of density, entropy and modularity to evaluate the quality of the detected online communities discovered by different methods. The definition of density, entropy, and modularity are as follows.

$$density(\{H_j^k\}_{j=1}^n) = \sum_{j}^{n} \frac{|\{(u,v)|u, v \in H_j^k, (u,v) \in E\}|}{|E|} \quad (2)$$

where n denotes the total number of detected communities. Density measures the compactness of the communities in structure.

$$entropy(\{H_j^k\}_{j=1}^n) = \sum_{j}^{n} \frac{|U(H_j^k)|}{|U|} entropy(H_j^k), \text{ where}$$

$$entropy(H_j^k) = -\sum_{i=1}^{n} p_{ij} log_2 p_{ij} \quad (3)$$

and p_{ij} is the percentage of members in a community $\mathcal{G}_j$ who are active on the query topic T_i. $entropy(\{\mathcal{G}_j\}_{j=1}^n)$ measures the weighted entropy considering all the query topics over all the communities. Entropy indicates the randomness of the topics which are covered in the communities.

$$modularity(\{H_j^k\}_{j=1}^n) = \frac{1}{2m} \sum_{ij} [A_{ij} - \frac{d_i d_j}{2m}] \delta(s_i, s_j) \quad (4)$$

Here, m denotes the number of edges corresponding to an adjacency matrix A[1], d_i denotes the degree corresponding to node n_i, s_i denotes the community membership of node n_i and $\delta(s_i, s_j) = 1$ if $s_i = s_j$.

Generally, a good interactive community should have high density, high modularity, and low entropy.

Figure 2A shows the density comparison between all the methods on the DBLP dataset. We set $k = 4$ as there are usually many small-sized research groups existing in DBLP. We see that TO-HIOC achieves better performance compared to the other two methods because it considers query-oriented active interactions among the community members. The HICD method fails to achieve better density values as it requires interaction between users (authors) to the celebrities (i.e., very high profile researchers in DBLP), which is not very common. The iTop method ignores the interactions between the non-seed users, resulting in poor performance. We also observe that all the methods achieve better density values for higher values of $|Q|$. The reason is that the number of interactive connections of the

---

[1] An adjacency matrix of a network is represented by A, where $A_{uv} = 0$ means there is no edge (no interaction) between nodes u and v and $A_{uv} = 1$ means there is an edge between the two.

users increases as $|Q|$ increases, which results in large and more densely connected communities.

Figure 2B shows the entropy comparison between the three methods. TO-HIOC achieves better performance in the aspect of the entropy as it considers the topical relevance (with regard to the query topics) during the interactions between the authors while forming a community. On the other hand, HICD achieves higher entropy value because not all the connected authors in a community have interests or active interactions in the common research topics. iTop also achieves a higher entropy value due to the lack of active topical interactions between the seed users and their followers. We see in **Figure 2C** that our proposed method TO-HIOC outperforms HICD and iTop in modularity comparison.

We examined a community in a co-author dataset which includes Jie Tang, who is one of the leading researchers in the data mining area, to see the differences in the community members for different values of k and $Q =$ {semantic web, topic mining, social network analysis}.

We observe the effect of value k in **Figures 3A–C**. By varying the values of k, we get communities of different sizes. We see that the community size decreases for higher values of k as the cohesiveness constraint becomes more strict, resulting in the exclusion of some active community members, for example "Yi Li," "Jing Zhang," "Limin Yao" leave the group. We also see that more researchers joined the community when the length of Q is increased as higher values of $|Q|$ covered more interactive researchers (**Figures 4A–C**).

5. CONCLUSION

In this paper, a topic-oriented highly interactive community detection approach is proposed. This method detects global communities where users have active interaction with each other on common topics. We observed that users have different degrees of interactions for different topics. As future work, we will consider the temporal factor to measure the recency behavior of users' interactions.

DATA AVAILABILITY

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

AUTHOR CONTRIBUTIONS

For this research paper, SD conducted the experiments (100%), designed the algorithm (75%), and wrote the paper (70%). MA designed the algorithm and experiments (25%), revised the paper (30%) as well as provided helpful insights and contribution in

REFERENCES

Anwar, M. M., Liu, C., and Li, J. (2018). "Uncovering attribute-driven active intimate communities," in *Australasian Database Conference* (Gold Coast, QLD: Springer), 109–122.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, 993–1022.

Clauset, A., Newman, M. E., and Moore, C. (2004). Finding community structure in very large networks. *Phys. Rev. E* 70:066111. doi: 10.1103/PhysRevE.70.066111

Cohen, J. (2008). Trusses: Cohesive subgraphs for social network analysis. *Natl. Secur. Agency Tech. Rep.* 16, 3–1. doi: 10.1.1.505.7006

Correa, D., Sureka, A., and Pundir, M. (2012). "itop: interaction based topic centric community discovery on twitter," in *Proceedings of the 5th Ph. D. Workshop on Information and Knowledge* (Maui, HI: ACM), 51–58.

Dev, H., Ali, M. E., and Hashem, T. (2014). "User interaction based community detection in online social networks," in *International Conference on Database Systems for Advanced Applications* (Bali: Springer), 296–310.

Fortunato, S., and Hric, D. (2016). Community detection in networks: a user guide. *Phys. Rep.* 659, 1–44. doi: 10.1016/j.physrep.2016.09.002

Jie, T., Jing, Z., Limin, Y., Juanzi, L., Li, Z., and Zhong, S. (2008). "Arnetminer: extraction and mining of academic social networks," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Beijing: ACM), 990–998.

Lim, K. H., and Datta, A. (2016). An interaction-based approach to detecting highly interactive twitter communities using tweeting links. *Web Intell.* 14, 1–15. doi: 10.3233/WEB-160328

Liu, Y., Niculescu-Mizil, A., and Gryc, W. (2009). "Topic-link lda: joint models of topic and author community," In *Proceedings of the 26th Annual International Conference on Machine Learning* (Montreal, QC: ACM) , 665–672.

Luo, W., Yan, Z., Bu, C., and Zhang, D. (2017). Community detection by fuzzy relations. *IEEE Trans. Emerg. Top. Comput.* 1–14. doi: 10.1109/TETC.2017.2751101

Luo, W., Zhang, D., Jiang, H., Ni, L., and Hu, Y. (2018). Local community detection with the dynamic membership function. *IEEE Trans. Fuzzy Syst.* 26, 3136–3150. doi: 10.1109/TFUZZ.2018.28 12148

Newman, M. E., and Park, J. (2003). Why social networks are different from other types of networks. *Phys. Rev. E* 68:036122. doi: 10.1103/PhysRevE.68. 036122

Shang, R., Luo, S., Zhang, W., Stolkin, R., and Jiao, L. (2016). A multiobjective evolutionary algorithm to find community structures based on affinity propagation. *Phys. A Stat. Mech. Appl.* 453, 203–227. doi: 10.1016/j.physa.2016.02.020

Tang, X., Xu, T., Feng, X., Yang, G., Wang, J., Li, Q., et al. (2017). Learning community structures: global and local perspectives. *Neurocomputing* 239, 249–256. doi: 10.1016/j.neucom.201 7.02.026

Yang, J., McAuley, J., and Leskovec, J. (2013). "Community detection in networks with node attributes," in *2013 IEEE 13th International Conference on Data Mining* (Dallas, TX: IEEE), 1151–1156.

Zhou, Y., Cheng, H., and Yu, J. X. (2009). Graph clustering based on structural/attribute similarities. *Proc. VLDB Endow.* 2, 718–729. doi: 10.14778/1687627.1687709