2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Attribute Driven Temporal Active Local Online Community Detection

Badhan Chandra Das
Dept. of Computer Sciencee and Engineering
Jahangirnagar University
Dhaka, Bangladesh
badhan0951@gmail.com

Md Musfique Anwar
Dept. of Computer Science and Engineering
Jahangirnagar University
Dhaka, Bangladesh
manwar@juniv.edu

Md. Al-Amin Bhuiyan
Dept. of Computer Engineering
King Faisal University
Al Hasa, Saudi Arabia
mbhuiyan@kfu.edu.sa

***Abstract*—This research investigates how the behavior of online social users' and topical activeness vary with time and how these parameters can be employed in order to improve the quality of the detected online local community. For a given input query, comprising a query node (user) and a set of attributes, this paper intends to find a densely connected community in which community members are temporally similar in terms of their activities related to the query attributes. To address the proposed problem, we develop a temporal activity biased weight model which assigns higher weight to users' recent activities and develop an algorithm to search effective community. The effectiveness of the proposed methodology is justified using three benchmark datasets.**

***Index Terms*—Online local Community, Temporal topical activeness, Query attributes.**

I. Introduction

A considerable amount of research has been devoted towards the problem of community detection in online social networks (OSNs). More recently, a related but different problem community search (aka local community) has been reported, where the main objective is to ascertain the best potential meaningful community that contains the query node(s) and query attributes [1]. Most of the existing investigations did not consider users' temporal behaviors towards the query attributes and also ignored an important aspect, namely the topical activeness of the community members. As a result, the resulting communities may have very low active users as well as may have users who might not show their inclination towards the query attributes in recent times.

This paper introduces a novel concept of user's *temporal activeness* which indicates user's topical degree of interest for a certain period of time with the notion that users' have different degrees of topical activeness which vary widely over time. For example, two users (A and B) are great fan of football (posting photos, messages related to football) are more likely to be grouped together than someone (C) who occasionally show her interest in football. However, all users (A, B, C) related to football would like to show their inclination during world cup football. The implication of these temporal activity-oriented communities outcomes significant quality enhancement in the detected communities. Our goal of this work is to search query oriented activity driven temporal active communities (ATAC).

IEEE/ACM ASONAM 2020, December 7-10, 2020
978-1-7281-1056-1/20/$31.00 © 2020 IEEE

The proposed approach is commenced on measuring the degree of activeness for each candidate community member with respect to the given query attributes to enhance the quality of the detected desired community. An active online local community is considered as a connected induced subgraph in which each node has a degree of at least k(k−core) which indicates the structure cohesiveness of the desired community. The main contributions of this research are summarized as follows:

- Propose a model that applies time-based forgetting factor to incorporate users' temporal activeness towards query attributes to determine communities of users who have similar temporal tendency.
- Develop a greedy algorithmic framework to search the desired query oriented temporal active community.
- Conduct extensive experiments to justify the efficacy of our proposed approach using benchmark datasets.

II. Related Works

Topical Community Search in Attributed Graphs. Yang et al. [1] proposed a prototype for community search over attributed graphs designed with k-cores for a given set of attributes and a single query node. Xin Huang et al. [7] presented a community search model for mining a community comprising multiple query nodes based on k-trusses. Yang et al. [4] performed spatial-aware community search method to determine groups of people involving homogeneous query attributes and are also geographically close to each other. However, none of these methods address the user's temporal degree of interest towards the given query attributes. As a result, these methods are not capable of determining the active communities for a given query.

III. Problem Statement

Before defining the problem statement, some relevant concepts are being introduced.

Attributed Graph: An attributed graph is expressed by *G = (U, E, $\mathcal{A}$)*, where *U* indicates set of social users (nodes), *E* denotes the social connections between users and $\mathcal{A} = \{a_1, ..., a_m\}$ represents the set of attributes associated with the users in U.

k-CORE: Given an integer *k ($k \geq 0$)*, the k-core of a graph G, denoted by C^k, is the maximal connected subgraph 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.

Activity: Each user u_i performs actions (such as posting tweets in Twitter, publishing research papers in coauthor network) known as activities at different time points (t_j) which may contain set of attributes ψ_{u_i}. An activity tuple $\langle u_i, \psi_{u_i}, t_j \rangle$ is used to represent an action. An activity stream S is a continuous and temporal sequence of activities i.e. $S = \{s_1, s_2, ..., s_r, ..\}$, such that each object (s_i) corresponds to an activity tuple.

Query: An input query $Q = \{u_q, \mathcal{A}_q\}$ consisting a query node u_q and a set of query attributes/topics $\mathcal{A}_q = \{a_1, ..., a_n\}$.

Active User: A user u_i in G is deliberated as an *active* user if u_i has accomplished at least $\gamma(\geq 1)$ actions associated with the $\mathcal{A}_q$ of Q, i.e., $|\{\langle u_i, \psi_{u_i}, t_j \rangle\}| \geq \gamma$, where $\psi_{u_i} \in \mathcal{A}_q$. The set of all candidate active users (who are in within h hops (connection, co-authorship, or follower-following relationship) away from query node u_q) is denoted by U^Q.

Time-Based Forgetting Factor: The idea behind the time-based forgetting factor is that not all the user's past activities are equally important and that the user's most recent activities can imply the most about his or her interests. This research uses the logarithmic time-decay function expressed in Equation 1 to assign lower importance (denoted as μ) to older activities, since they are less probable of corresponding to the user's recent interests.

$$\mu_{\langle u_i, \psi_{u_i}, t_j \rangle} = \frac{1}{1 + log_b(age_{\langle u_i, \psi_{u_i}, t_j \rangle} + 1)} \quad (1)$$

The base of the logarithm in Equation 1, denoted by b, controls the speed of decay and $age_{\langle u_i, \psi_{u_i}, t_j \rangle}$ as the amount of time elapsed since it happened.

Activeness Score: The activeness score (denoted by σ) for each candidate community member $u_i \in U^Q$ is computed using Equations 4 and 5, respectively where $\psi_{u_i} \in \mathcal{A}_q$. This investigation deliberates two factors that are closely associated with the distinct activeness of a user u_i. The first factor $f_1(u_i, \psi_{u_i})$ specifies the probability that u_i performs an activity related to Q.

$$f_1(u_i, \psi_{u_i}) = \frac{\sum \mu_{\langle u_i, \psi_{u_i}, t_j \rangle} \times |ACTS(u_i, \psi_{u_i})|}{|ACTS(u_i, *)|} \quad (2)$$

where, $ACTS(u_i, \psi_{u_i})$ represents the set of activities comprising the set of attributes $\psi_{u_i} \subseteq \mathcal{A}_q$ performed by u_i and $ACTS(u_i, *)$ denotes the set of all the activities containing any attribute(s) performed by user u_i.

The second factor $f_2(u_i, \psi_{u_i})$ designates the participation of user u_i compared to the total number of activities related to Q performed by U^Q.

$$f_2(u_i, \psi_{u_i}) = \frac{\sum \mu_{\langle u_i, \psi_{u_i}, t_j \rangle} \times |ACTS(u_i, \psi_{u_i})|}{\sum_{u_z \in U^Q} |ACTS(u_z, \psi_{u_z})|} \quad (3)$$

Then, the activeness (denoted as σ) of u_i related to Q is

$$\lambda_{(u_i, \psi_{u_i})} = f_1(u_i, \psi_{u_i}) \times f_2(u_i, \psi_{u_i}) \quad (4)$$

$$\sigma_{(u_i, \psi_{u_i})} = \frac{\lambda_{(u_i, \psi_{u_i})}}{max_{u_z \in U^Q}\{\lambda_{(u_z, \psi_{u_z})}\}} \quad (5)$$

Problem Definition: Given an attributed graph *G = (U, E, $\mathcal{A}$)* with activity stream S, an input query $Q = \{u_q, \mathcal{A}_q\}$, two positive integers h and k, an attributed active local community $\mathcal{C}_q$ is an induced subgraph that meets the following constraints.

1) ***Connectivity.*** $\mathcal{C}_q \subset G$ is connected, $\mathcal{C}_q$ must include u_q;
2) ***Structure cohesiveness.*** $\forall u \in \mathcal{C}_q, deg_{\mathcal{C}_q}(u) \geq k$;
3) ***Query cohesiveness.*** $\forall u \in \mathcal{C}_q$, activeness score of an user u is $\sigma_{(u,Q)} \geq \theta_a$, where θ_a is the threshold ranged between 0 to 1.

IV. Active Online Local Community Detection Approach

A. Data Pre-processing for Topic Detection

User generated contents in OSNs are very short and often noisy. For example, Tweets are informally written and may contain grammatically incorrect text with misspellings and abbreviations (e.g., heellooo for hello, gr8 for great, netwrk or ntwork for network, and so on). So, in order to improve the quality of data and the performance of the subsequent steps, tweets are being normalized over the normalization lexicon proposed in [5].

B. Topic Detection from Social Data

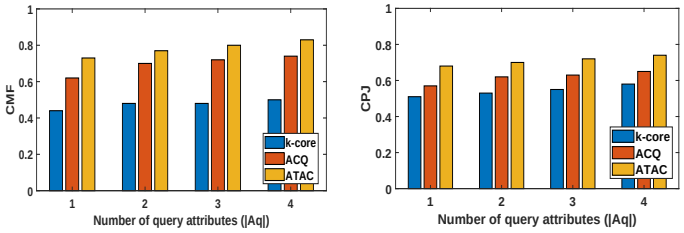
We apply topic modeling approach to detect the latent topics from user generated contents. Like social network, the academic co-author network consists of authors, research papers and co-author network, we apply Latent Dirichlet Allocation (LDA) model [6] to extract the topics from the abstracts of the papers. We choose Twitter-LDA (T-LDA) [8] for Twitter dataset as the tweets are very short in length.

C. Top-Down Algorithm

Algorithm 1 ATAC

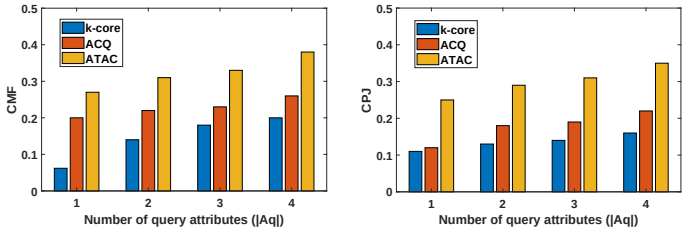
Require: $G = (U, E)$, $Q = \{u_q, \mathcal{A}_q\}$, θ_a, number k and h
Ensure: An online Active community $\mathcal{C}_q$ containing u_q

```
Find a set of nodes N_{u_q} who are within h hops away from the query node u_q
Compute the induced subgraph C_q on N_{u_q}, i.e. C_q = (N_{u_q}, E(N_{u_q})), where E(N_{u_q}) = (v_1, v_2) : v_1, v_2 ∈ N_{u_q}, (v_1, v_2) ∈ E
Maintain C_q as a k-core
for each u_i ∈ C_q do
    compute the activeness score σ_{(u_i, ψ_{u_i})}
    if σ_{(u_i, ψ_{u_i})} < θ_a then
        delete u_i and its incident edges from C_q
        Maintain C_q as a k-core
    end if
end for
Output the active connected k-core C_q
```



(a) CMF (T1) (b) CPJ (T1)

Fig. 1: Performance comparison on CRAWL dataset (in all cases, $k = 4$, $h = 3$, $\gamma = 150$, $\theta = 0.5$)



(a) CMF (b) CPJ

Fig. 2: Performance comparison on Flickr dataset (in all cases, $k = 4$, $h = 3$, $\gamma = 10$, $\theta = 0.5$)

Algorithm overview. The algorithm, known as ATAC, includes three-fold stages. First, it computes the induced subgraph $\mathcal{C}_q$ from the set of nodes $N(_{u_q})$ who are h hops away from u_q (line 1-2). Second, it discovers the k-core subgraph containing u_q from G (line 3). Third, it iteratively eliminates the inactive nodes i.e. whose activeness score $(\sigma_{(u_i,\psi_{u_i})})$ is less than a given threshold θ_a from $\mathcal{C}_q$, and preserves the remaining $\mathcal{C}_q$ as k-core, until no longer possible (line 4-8). Finally, it outputs the active attributed community.

V. Experimental Evaluation

All experiments were performed on an Intel(R) Core(TM) i3-2350M 2.3 GHz Windows 10 PC with 4 GB RAM.

A. Data set

Twitter dataset. We conduct our experiment on a Twitter dataset named CRAWL [2]. We set the input query attributes as {*social media*, *politics*, *entertainment*, *sports*}.

Flickr dataset. For each user in Flickr dataset, we choose 30 most frequent tags of its associated photos as its attributes and choose {*nature*, *festival*, *architecture*, *portrait*} as input query

DBLP dataset. Our algorithm has been verified to an academic coauthor network dataset [3] which contains research papers that are published within 2000 to 2014. In DBLP dataset, the input query is set as {*data mining*, *natural language processing (NLP)*, *social network analysis (SNA)*, *machine learning*}.

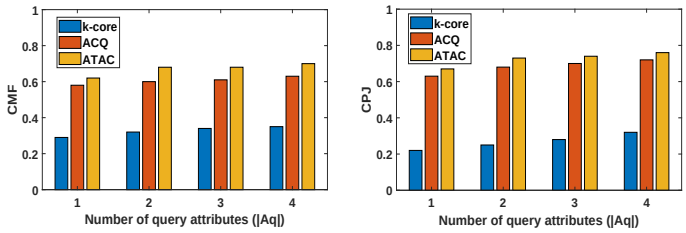
Table I shows the statistics of our experimental data.

Dataset	No. of Nodes	No. of Edges	No. of activities
CRAWL	9,468	1,474,510	6,211,653
Flickr	581,099	9,944,548	5,588,960
DBLP	15,516	48,862	193,512

TABLE I: Datasets

B. Comparison Methods

We compare our proposed ATAC algorithm with two other methods. We select ACQ method, proposed by Yang et al. [1], for community search over attributed graphs based on k-cores. The key distinction with our work is that ACQ doesn't consider users' topical activeness as well as ignore the prospective temporality of users' interests. Finally, we consider a baseline solution ($k-$core) which forms communities based on only k-core i.e. focusing only the structural cohesiveness.



(a) CMF (T1) (b) CPJ (T1)

Fig. 3: Performance comparison on DBLP dataset (in all cases, $k = 3$, $h = 3$, $\gamma = 3$, $\theta = 0.5$)

C. Community Quality Evaluation

We vary the length of query attributes $|\mathcal{A}_q|$ to $|\mathcal{A}_q|$= 1, 2, 3, 4 and use two measures of CMF, CPJ to assess the quality of the communities. Let us define them namely CMF and CPJ [1], for evaluating the attribute cohesiveness of the communities. Let $N(\mathcal{C}_q) = \{\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_\mathcal{L}\}$ be the set of $\mathcal{L}$ communities returned by an algorithm for a query node $u_q \in U$.

Community Member Frequency (CMF): Consider an attribute a of query attribute set $\mathcal{A}_q$. If a appears in most of the nodes (or members) of a community $\mathcal{C}_i$, then $\mathcal{C}_i$ can be considered to be greatly cohesive. The CMF measures the number of occurrences of query attributes in $\mathcal{C}_i$ to determine the degree of cohesiveness. Let $n_{i,p}$ be the number of nodes of $\mathcal{C}_i$ whose attribute sets contain the *p-th* attribute of $\mathcal{A}_q$. Then, $\frac{n_{i,p}}{|\mathcal{C}i|}$ is the relative occurrence frequency of this attribute in $\mathcal{C}_i$. The CMF is the average of this value in overall attributes in $\mathcal{A}_q$, and all communities in $N(\mathcal{C}_q)$:

$$CMF(N(\mathcal{C}_q)) = \frac{1}{\mathcal{L} \times |\mathcal{A}_q|} \sum_{i=1}^{\mathcal{L}} \sum_{p=1}^{|\mathcal{A}_q|} \frac{n_{i,p}}{|\mathcal{C}i|} \tag{6}$$

It is to be noted that the value of CMF($N(\mathcal{C}_q)$) ranges from 0 to 1. The larger its value, the more cohesive is a community.

Community pairwise Jaccard (CPJ): This is established on the similarity between the attribute sets of any pair of nodes of community $\mathcal{C}_i$. This research employs the Jaccard similarity, which is commonly used in the IR literature. Let $\mathcal{C}_{i,j}$ be the *j-th* node of $\mathcal{C}_i$. The CPJ is then the average similarity overall pairs of nodes of $\mathcal{C}_i$, and all communities of $n(\mathcal{C}_q)$:

$$CPJ(N(\mathcal{C}_q)) = \frac{1}{\mathcal{L}} \sum_{i=1}^{\mathcal{L}} \frac{1}{|\mathcal{C}_i|^2} \left[\sum_{j=1}^{|\mathcal{C}_i|} \sum_{k=1}^{|C_i|} \frac{|\mathcal{A}_q(\mathcal{C}_{i,j})| \cap |\mathcal{A}_q(\mathcal{C}_{i,k})|}{|\mathcal{A}_q(\mathcal{C}_{i,j})| \cup |\mathcal{A}_q(\mathcal{C}_{i,k})|} \right] \tag{7}$$

The value of CPJ($N(\mathcal{C}_q)$) ranges from 0 and 1. A higher value of CPJ($N(\mathcal{C}_q)$) indicates better cohesiveness.

Figure 1 shows the community quality evalution among the three methods on CRAWL dataset. ATAC always performs the best, in terms of CMF and CPJ (Figure 1(a), (b)). The reason is that each community member has to perform certain number (γ) of activities related to $\mathcal{A}_q$ to become an active user. As a result, most of the community members have to show their high degree of inclination towards multiple query topics. In case of ACQ, there are many low active community members who don't have interest in most of the query topics. So, the coverage of query topics within the communities are not that much better as in ATAC. On the other hand, the values of CMF and CPJ in k−core are very poor as it ignores users' association with the query topics while forming a community. As a result, most of the community members have no interest towards the given query topics.

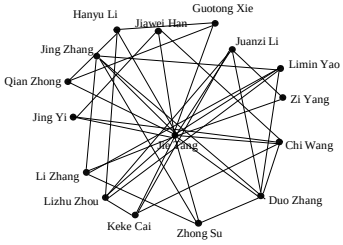
Figure 2 shows the results on effectiveness among the three methods on Flicker dataset. Similar to Twitter dataset, ATAC achieves better results in all cases due to the consideration of certain degree of topical activeness among the community members.

Figure 3 shows the community quality evolution among the three methods on DBLP datasets. We can see that ATAC outperforms the other two methods in all cases. In DBLP dataset, most of the users within a community have very similar research interests. As a result, the coverage of the query topics within a community are better than the other two datasets. Again, the number of activities (i.e. publishing research papers) in DBLP are very low compare with other datatsets. So, the performance of ACQ improves significantly.

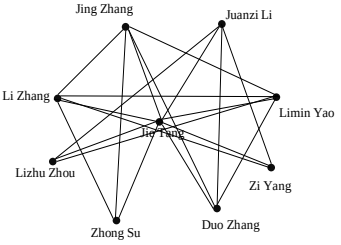
Larger values of $|\mathcal{A}_q|$ results more number of users having activeness in one or more query attributes as we see that the values of all the measures in every dataset are higher as $|\mathcal{A}_q|$ goes high for all the methods.

D. A Case Study

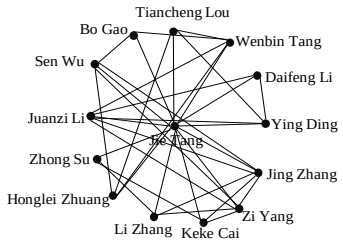
This research investigated a local community in DBLP dataset which includes Jie Tang (as query node), who is one of the prominent researchers in data mining area, to observe the distinctions in the community members for various values of γ and $\mathcal{A}_q$. Here, we considered non-overlapping time intervals. Figure 4(a) and 4(d) illustrate the community members when the attributes of the query Q is {*data mining*, *NLP*, *SNA*} for the time period of 2008 to 2010 and 2011 to 2013, respectively. It is observed that the number of community members decreases, when the higher degree of activeness (γ) of the researchers for a given Q is considered, (Figure 4(b) and Figure 4(d)). Our observation is that the value of k, γ and θ can balance the trade-off between activeness and cohesiveness of a community.



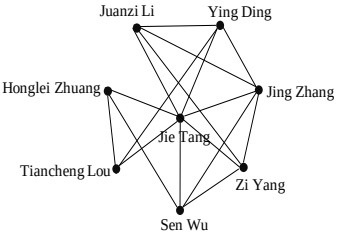
(a) $\mathcal{A}_q$ = {Data mining, NLP, SNA}, $\gamma = 2$, time interval = 2008 - 2010



(b) $\mathcal{A}_q$ = {Data mining, NLP, SNA}, $\gamma = 3$, time interval = 2008 - 2010



(c) $\mathcal{A}_q$ = {Data mining, NLP, SNA}, $\gamma = 2$, time interval = 2011 - 2013



(d) $\mathcal{A}_q$ = {Data mining, NLP, SNA}, $\gamma = 3$, time interval = 2011 - 2013

Fig. 4: Case study: results for different $\mathcal{A}_q$ and γ in DBLP dataset (in all cases, u_q = "Jie Tang", $k = 3$, $h = 3$, $\theta = 0.5$)

VI. Conclusion

Through this research, we analyzed the problem of active local community search in attributed social graph. It has been observed that the users' individual activeness vary widely for different attributes. This research outlined an activeness score function for the candidate community members and developed methods to search the query oriented active community. The effectiveness of the proposed method has been demonstrated over extensive experiments on real datasets.

References

[1] Fang, Y., Cheng, R., Luo, S., Hu, J.: *Effective Community Search for Large Attributed Graphs.*, VLDB, 1233–1244, 2016.

[2] P. Bogdanov, M. Busch, J. Moehli, A. K. Singh and B. K. Szymanski. The Social Media Genome: Modeling Individual Topic-Specific Behavior in Social Media. In ASONAM, pp. 236–242, 2013.

[3] Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: ArnetMiner: Extraction and Mining of Academic Social Networks. In: KDD, pp.990–998 (2008)

[4] Fang, Y., Cheng, R., Li, X., Luo, S., Hu, J.: *Effective community search over large spatial graphs.*, VLDB, 709–720, 2017.

[5] Han, B., Cook, P., Baldwin, T.: Lexical normalization for social media text. In: Journal ACM Transactions on Intelligent Systems and Technology (TIST), Volume 4, Issue 1 (2013)

[6] Blei, D. M., Ng, A. Y., Jordan, M. I.: Latent dirichlet allocation. In: Journal of Machine Learning Research, 3:993–1022 (2003)

[7] Huang, X., Lakshmanan, L. VS.: Attribute-driven community search. In: VLDB, pp. 949–960 (2017)

[8] Zhao, W. X., Jiang, J. , Weng, J., He, J., Lim, E., Yan, H., Li, X.: Comparing twitter and traditional media using topic models, In: ECIR, pp. 338–349 (2011)

[9] Zhou, Y., Cheng, H., Yu, J. X.: Graph clustering based on structural/attribute similarities. In: VLDB, pp. 718–729 (2009)