IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 11, 2021, accepted June 23, 2021, date of publication June 29, 2021, date of current version July 8, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3093368

Attribute Driven Temporal Active Online Community Search

BADHAN CHANDRA DAS[1], MD. MUSFIQUE ANWAR[1], MD. AL-AMIN BHUIYAN[2], IQBAL H. SARKER[3], (Member, IEEE), SALEM A. ALYAMI[4], (Member, IEEE), AND MOHAMMAD ALI MONI[5,6]

[1]Department of Computer Science and Engineering, Jahangirnagar University, Savar Union, Dhaka 1342, Bangladesh
[2]Department of Computer Engineering, King Faisal University, Al Hofuf 400-31982, Saudi Arabia
[3]Department of Computer Science and Engineering, Chittagong University of Engineering and Technology, Chattogram 4349, Bangladesh
[4]Department of Mathematics and Statistics, Faculty of Science, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 13318, Saudi Arabia
[5]WHO Collaborating Centre on eHealth, UNSW Digital Health, Faculty of Medicine, University of New South Wales, Sydney, NSW 2052, Australia
[6]Healthy Ageing Theme, Garvan Institute of Medical Research, Darlinghurst, NSW 2010, Australia

Corresponding authors: Mohammad Ali Moni (m.moni@unsw.edu.au) and Badhan Chandra Das (badhan0951@gmail.com)

ABSTRACT Almost all of the existing approaches to determining online local community are typically deliberated like-minded users who have similar topical interests. However, such methodologies overlook the prospective temporality of users' interests as well as users' degree of topical activeness. As a result, the consequential communities might have extremely lower active users. This research investigates how online social users' behaviors and topical activeness vary over time and how these parameters can be employed in order to improve the quality of the detected local community. For a given input query, consisting a query node (user) and a set of attributes, this research intends to find 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 gives higher weight to users' recent activities and develop an algorithm to search an effective community. The effectiveness of the proposed methodology is justified using four benchmark datasets and compared with four other baseline methods. Experimental results demonstrate that our proposed framework yields better outcomes than the baseline methods for all four benchmark datasets.

INDEX TERMS Online local community, query attributes, temporal topical activeness.

I. INTRODUCTION

Information sharing and communication patterns of users on online social networks (OSNs) platforms can lead to the formation of online social groups or communities that consist of users with similar interests. Discovering meaningful communities in OSNs has recently occupied an overwhelming research interest owing to its diverse applications including online marketing, link prediction, information diffusion, friend/news recommendations etc. OSNs can be modeled as a graph, where social users are deliberated as nodes and the social connections between them are viewed as edges of the graph. One can then apply different graph clustering algorithms [1] to discover the communities in OSNs.

The associate editor coordinating the review of this manuscript and approving it for publication was Alberto Cano.

Nowadays, social networks are considered as attributed graphs due to the availability of rich attribute data associated with the nodes describing the properties of the nodes. A fair amount of topic-oriented methodologies have been proposed that consider the attributes of the users jointly with social connections to discover general communities based on the whole social graph [15], [34]. Another related but different problem is community search (aka local community) where the main objective is to ascertain the best potential meaningful community that contains the query node(s) and query attributes [2]. All these foregoing investigations did not consider users' temporal behaviour towards query attributes and also ignore 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. Hence, we are interested

in searching for communities in many application domains, where community members continuously pay active attention to the query attributes at a given time period. For example, Alex, a football lover, is a regular customer of a sports company. Then, the manager of the company can issue a query containing Alex as a query node and *football* as query attribute on a social network, with a hope to get an active community on *football* so that the people connected virtually with Alex, can be the company's advertising targets.

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) who are very fans of football sport (posts 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 in this work is to search query oriented activity driven temporal active communities (ATAC), where we show how the topical activeness of the users of OSNs can vary over time with different query attributes.

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 which indicates the structure cohesiveness of the desired community. The main contributions of this research are summarized as follows:

1) Propose a model that applies time-based forgetting factor to incorporate users' temporal activeness towards query attributes to determine communities of users who have a similar temporal tendency.
2) Modeling and evaluating users' degree of activeness towards different attributes of a given query;
3) Develop a greedy algorithmic framework to search the desired query oriented temporal active community.
4) Conduct extensive experiments to justify the efficacy of our proposed approach using benchmark data sets.

This paper is an extended version of a conference paper that appeared as [21]. The new contributions to this journal version are summarized as follows.

1) We introduce time-based forgetting factor to emphasize users' most recent activities. The idea behind time-based forgetting factor is that not all of a user's past activities are equally important and that the user's most recent activities can imply the most about his or her interests. Therefore, the changes in the user's interests over time can be compensated by providing more importance to the latest activities.
2) To speed up the calculation procedure, we select the set of active users who are within h hops away from the query node instead of finding the set of active users from the whole social graph. Consequently, we made changes in related equations.
3) We use a Twitter[1] (a very popular social networking platform) dataset. Registered users of Twitter can post information, comments and opinions on anything by the use of short lengthened (at most 140 characters) message known as *tweets*. Being a very casual social platform, tweets are short and often noisy. So, in order to improve the quality of data and the performance of the subsequent steps, we apply some pre-processing steps on the original tweets, for example, tweets are being normalized using a normalization lexicon.
4) Instead of considering hashtags (#), (which used in conference version) as the representation of the topics in the Twitter dataset, we apply BERTopic (topic modeling approach) in order to identify latent semantic topics from the processed tweets.
5) We use text clustering algorithm like K-means in both Twitter and DBLP datasets to cluster similar topics.
6) New experiments are conducted on other social networks like, Flicker.
7) We add two new measures, namely Community member frequency (CMF) and Community pair-wise Jaccard CPJ) to study the effectiveness of the proposed methodology. We also measure an average number of activities related to the given query in the detected communities.
8) We also compared our proposed approach with some baseline frameworks.

The rest of the paper is organized as follows. Section 2 includes the relevant works of this field. We introduce some relevant terms and the problem statement in section 3. The approach of attribute driven temporal local active online community is described in section 4. Section 5 covers the experimental evaluation. Section 6 summarizes this research and illustrates some real-life aspects of this research. Finally, we conclude the paper in section 7.

II. RELATED WORKS

A. COMMUNITY SEARCH IN NON-ATTRIBUTED GRAPHS

Various methods have been proposed [3], [4] in the literature that deliberated only the structural properties of the social graph in order to search a community for a given query node q. Modularity [5], edge betweenness [8] and neighborhood concepts [9] are some more examples of community structure have been proposed so far. Wang *et al.* proposed local expanding procedure based on structural centers, in order to uncover overlapping community structures effectively. They also introduced a structural centrality and a locating strategy to find structural centers in networks. [11]

[1] https://twitter.com/

All these approaches carried out non-attributed graphs and ignored the valuable information of the nodes resulting in poor cohesion in query attribute sets among the community members.

B. TOPICAL COMMUNITY SEARCH IN ATTRIBUTED GRAPHS

Fang *et al.* [2] 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.* presented a community search model for mining a community comprising multiple query nodes based on k-trusses [20]. Yang *et al.* performed spatial-aware community search method to determine groups of people involving homogeneous query attributes and are also geographically close to each other [14]. Gu *et al.* proposed a social community detection scheme for mobile social networks based on social-aware, including social attribute similarity, node interest similarity and node mobility in mobile social networks [16]. Souravlas *et al.* presented a threaded binary tree approach for community detection. To determine a user's membership to a community, the method tries to locate ''stronger'' paths (with higher similarity based on users' interest) between the user's node and this community [17]. Though these frameworks yield communities with similarity in attributes, they cannot determine whether the community members in the resulting communities are active or not with respect to the given query attributes.

C. USER INTERACTION BASED FRAMEWORKS

Lim *et al.* have proposed an approach where interaction pattern and frequency is considered rather than only counting the following/follower links [23]. A temporal active intimate community search method has been reported in which community members have active temporal interactions among them with respect to the given set of query nodes and query attributes [25]. Dev *et al.* introduced a user interaction based community detection method considering group behaviors of the users [26]. The proposed methods of this category focused on the user interactions and considered the attributes of the community members in order to find communities. They did not consider the users' degree of interest among the resulting communities with respect to the query attributes.

D. COMMUNITIES OVER DYNAMIC ENVIRONMENTS

Event and role analysis in the social network to conceive the dynamic change of network over time investigated by Faganan *et al.* [35]. For the purpose of local community formation, dynamic membership function can be used [36]. A distance dynamic model is proposed to detect community of variable size using dynamic membership degree by Meng *et al.* [37]. An incremental bottom-up community detection model is introduced by Liu *et al.* in the dynamic graphs [38]. The frameworks of this category give real-time communities, but, they overlook the community members' topical activeness towards given query attributes.

E. ACTIVE COMMUNITY SEARCH OVER ATTRIBUTED GRAPHS

Das *et al.* introduced a framework which detects attribute driven local active communities over attributed graph [21]. Anwar et. al. studied the problem of detecting attribute-driven active intimate community where they searched for densely-connected communities in which community members actively participate as well as have strong interaction with respect to the given query attributes, yet they counted all of their previous online actions equally, even if the actions performed a log ago. [24].

F. DEEP LEARNING-BASED FRAMEWORK FOR COMMUNITY DETECTION

Wu *et al.* proposed a combined model of Auto-encoder and Convectional Neural Network (AE-CNN) in order to detect communities on social networks [22], however, they did not consider neither topical interest nor topical activeness of the yielded communities from their framework. As a result, the connection between the users' might be strong, but, the falls short to detect topic oriented active communities, since the resulting communities contain the users who are not inclined to the given query attributes. Xin et. al. proposed a community detection approach in topologically incomplete networks (TIN), which are usually observed from real-world networks and where some edges are missing [27]. Dhilber et. al. proposed a deep neural network architecture for community detection multiple autoencoders have been incorporated and parameter sharing is applied [28]. The works under this class reported a different way to find communities over social graph, yet, as mentioned earlier, the frameworks of this class did not consider the users' topical activeness in the resulting communities, moreover, these methods fail to show how users' interest in the resulting communities varies time to time.

However, none of these methods addresses the user's 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. Moreover, these cited approaches did not contemplate how the users' interests for the given query attribute changes with time.

III. PROBLEM STATEMENT

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

A. 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 $\mathcal{A} = \{a_1, \ldots, a_m\}$ represents the set of attributes associated with the users in U.

B. INDUCED GRAPH

An induced subgraph H of a graph G is another graph, formed from a subset of the vertices of the graph G and all of the edges connecting pairs of vertices in that subset.

C. 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. Each node u in G has a core number which is the maximum k for which that u belongs in the k-core of G.

D. TOPIC

A topic is a collection of most representative words for that topic. For example, if sports is a topic, then the words of this topic will be like football, cricket, match, wicket, goal, run, foul, etc.

E. 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. $S = \{s_1, s_2, \ldots, s_r, ..\}$ such that each object (s_i) corresponds to an activity tuple.

F. QUERY

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

G. ACTIVE USER

An 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 away from query node u_q) is denoted by U^Q.

H. TIME-BASED FORGETTING FACTOR

The idea behind time-based forgetting factor is that all the' past activities of the user are not equally important and that the user's most recent activities can imply the most about his or her interests [6]. Therefore, the changes in the user's interests over time can be compensated by providing more importance (denoted as μ) to the latest activities. This research uses the logarithmic time-decay function expressed in Equation 1 to assign lower importance to older activities since they are less probable of corresponding to the user's recent interests.

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

The base of the logarithm in Equation 1 is 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.

I. 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 performs an activity related to Q.

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

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.

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

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

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

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

J. PROBLEM DEFINITION

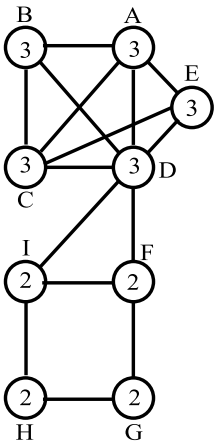
Given an attributed graph $G = (U, E, \mathcal{A})$ with activity stream S, an input query $Q = \{u_q, \mathcal{A}_q\}$, two positive integers h and k, an attributed active local community $\mathcal{C}_q$ is an induced subgraph that meets the following constraints.

1) **Connectivity.** $\mathcal{C}_q \subset G$ is connected, $\mathcal{C}_q$ must include u_q;
2) **Structure cohesiveness.** $\forall u \in \mathcal{C}_q$, $deg_{\mathcal{C}_q}(u_i) \geq k$;
3) **Query cohesiveness.** $\forall u \in \mathcal{C}_q$, activeness score of a user u is $\sigma_{(u_i, Q)} \geq \theta_a$ and $\theta_a \in [0, 1]$ is a threshold.

A social graph, G holding the core number for each node is illustrated in Figure 1(a). Different community members can be inferred for different query attributes and different values of k at different time intervals according to the users' actions log shown in Figure 1(b). For example, at time interval T_1, when $Q = \{D, a_0\}$ and $k = 2$, we get $\mathcal{C}_q = \{A, B, C, D\}$, while for same k with the addition of another attribute a_1, we see that new members $\{E, F, I\}$ are added with the existing $\mathcal{C}_q$ to $\mathcal{C}_q = \{A, B, C, D, E, F, I\}$ (Figure 1(c)). Again, at time interval T_3, when $Q = \{D, a_0\}$ and $k = 2$, we get $\mathcal{C}_q = \{A, C, D, E, F, G, H, I\}$, while for same Q with an increase value of $k = 3$, we get $\mathcal{C}_q = \{A, C, D, E\}$.

IV. ACTIVE ONLINE LOCAL COMMUNITY DETECTION APPROACH

The proposed framework includes three stages layout as presented in Figure 2 to search attribute driven temporal active community (ATAC). First, the pre-processing is performed to eliminate irrelevant data or noises from the activity stream S. Second, topic modeling method has been applied, in our case, we applied LDA/BERTopic (Topic Modelling Approach) to the filtered data to recognize the latent topics from S. Then similar topics are being clustered by employing k-means clustering algorithm [7]. Finally, an algorithm is developed and



(a) Social Graph (the number denotes node core number)

T1

	a_0	a_1
A	5	6
B	5	7
C	6	5
D	6	7
E	3	3
F	2	3
G	5	3
H	3	1
I	5	2

T2

	a_0	a_1
A	5	5
B	2	2
C	5	7
D	8	6
E	2	1
F	5	6
G	7	5
H	2	4
I	3	7

T3

	a_0	a_1
A	7	1
B	2	7
C	5	2
D	6	7
E	8	5
F	5	6
G	7	5
H	6	8
I	6	7

(b) Action log at different time intervals

Time Interval	Query	Community(C_q)
T1	Q={D, a_0}, k=2	{A, B, C, D}
	Q={D, a_0, a_1}, k=2	{A, B, C, D, E, F, I}
	Q={D, a_0}, k=3	{A, B, C, D}
	Q={D, a_0, a_1}, k=3	{A, B, C, D, E}
T2	Q={D, a_0}, k=2	{A, C, D}
	Q={D, a_1}, k=2	{A, C, D, F, I}
	Q={D, a_0, a_1}, k=2	{A, C, D, F, G, H, I}
T3	Q={D, a_0}, k=2	{A, C, D, E, F, G, H, I}
	Q={D, a_0}, k=3	{A, C, D, E}
	Q={D, a_1}, k=2	{D, F, G, H, I}

(c) Community members in C_q for different values of Q and k ($\gamma = 5$ in all cases)

FIGURE 1. **Active community search at different time intervals for different *Q* in a social graph *G*.**

applied to the processed activity streams to detect the desired community. The detailed description is given as follows.

A. DATA PRE-PROCESSING FOR TOPIC DETECTION

Here, we describe the first stage of our three-stage proposed framework, the data pre-processing. As Twitter is a very casual platform, people often post on Twitter in an informal way, like a short form of any text, or using numbers in words to make it short which is usually not allowed in standard writing conventions. We call these instances as noisy contents. Moreover, tweets are informally written and often contain grammatically incorrect sentence structures with misspellings and non-standard forms of words (e.g., toook for took, goooood for good, helloooo for hello), informal as well as misspelled abbreviations and short forms (e.g., tmrw for tomorrow, wknd for weekend, hlw for hello), phonetic substitutions (e.g., 4evafor forever, 2day for today, gr8 for great) etc. Oftentimes, the tweets also contain slang words. The amount of non-standard words in tweets results in significantly higher out-of-vocabulary (OOV) rates. So, in order to improve the quality of data and the performance of the subsequent steps, we need to clean the tweets using a linguistic procedure tweets using the normalization lexicon proposed in [18]. The remaining words are converted into a seed word (stemming word) for example: plays, playing, etc.-> play by using Lucene 4.9.0 Java API[1]. Apart from that, Twitter users often publish spam/noisy tweets which are irrelevant to their interest. Noisy tweets must be removed from the datasets so that we can get a more precise result. Omitting this step can lead to a false analysis which will obviously not be a good start for our experiment and for this it can occur significant rate of inconvenience on the accuracy rate of the desired result.

[1]https://lucene.apache.org

B. TOPIC DETECTION FROM SOCIAL DATA

We need to apply topic modeling approach to find the latent topics from the user generated contents.

1) TOPIC MODELLING

Identifying latent topics from a bunch of text is called Topic Modeling. The early Topic Modelling Approach was proposed by Papadimitriou et. al. in 1998 [29]. It is an unsupervised machine learning technique of classification of documents, similar to clustering of numeric data, which finds some natural groups of items (topics). A document or a bunch of text can be a part of multiple topics. This is usually done by detecting patterns in a collection of documents called a *corpus*, and grouping the words used into topics. It can help in:

- Finding the hidden themes in the collection.
- Classifying the documents into the yielded themes.
- Using the classification to organize, summarize and search the documents.

We apply this Topic Modeling approach to detect the latent topics from user-generated contents. For a social network like academic coauthor network consisting of authors, research papers and coauthor network, we apply Latent Dirichlet Allocation (LDA) model [19] to extract the topics from the abstracts of the papers.

BERTopic [33] is a topic modeling approach which leverages BERT embeddings [32] and a class-based TF-IDF[2] to create dense clusters allows easily interpretable topics whilst keeping important words in the topic descriptions. We apply this BERTopic approach on our Twitter dataset (CRAWL and SNAP) to get the topics to which the tweets are corresponding to.

[2]https://towardsdatascience.com/creating-a-class-based-tf-idf-with-scikit-learn-caea7b15b858

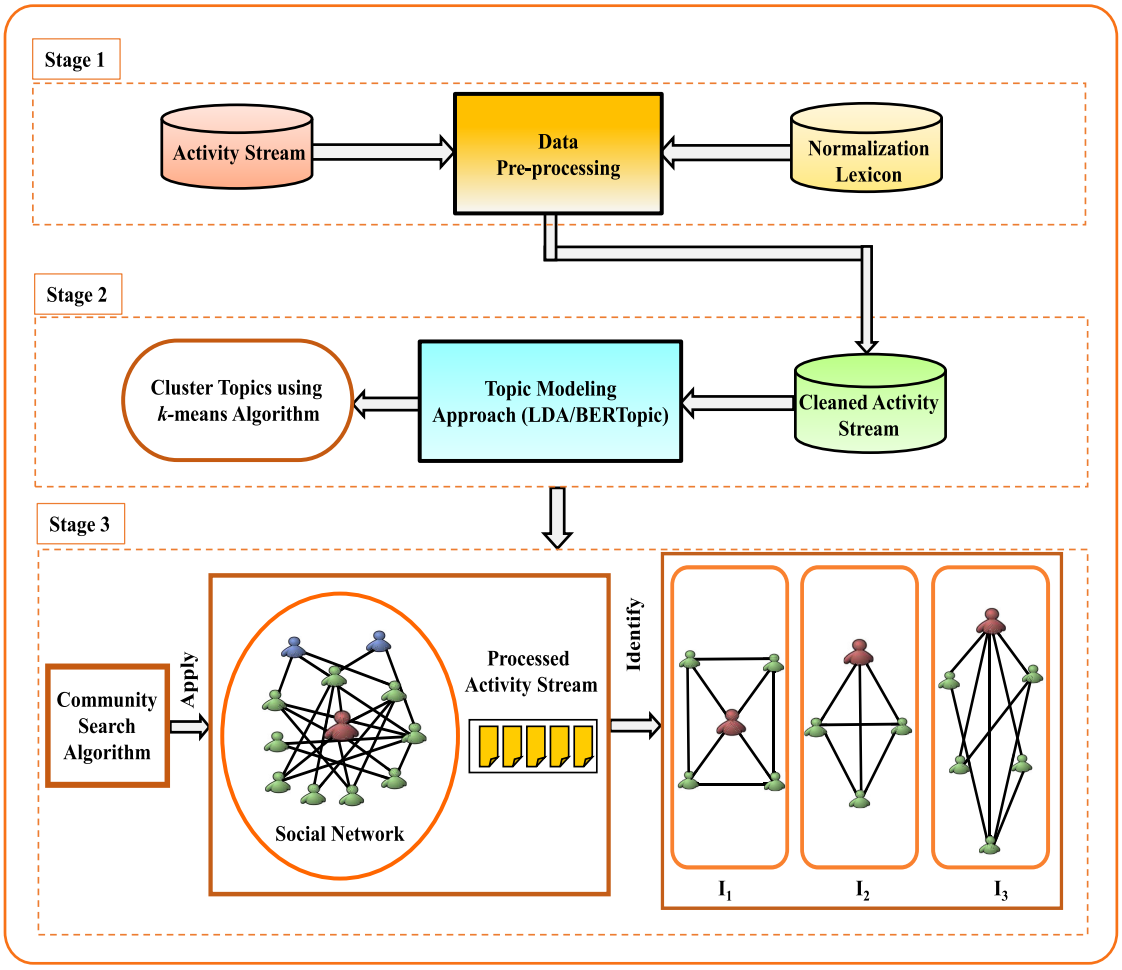


FIGURE 2. Overview of proposed methodology of activity driven temporal active community (ATAC).

Afterwards, *K*-means clustering algorithm [31] has been applied on the outputs yielded from the LDA/BERTopic in order to cluster similar topics to a topic category. The input of the *K*-means clustering algorithm is a set of documents. For DBLP dataset, we got the word-to-topic assignments from the LDA model and construct synthetic documents for each topic. Every synthetic document corresponds to a topic and consists of those words that are assigned to this topic. Table 1 shows the word to topic assignment yielded from the LDA model. Then we apply *K*-means clustering algorithm on these synthetic documents to get the clusters of similar topics. We used the *word2vec* [30] embedding in order to find the similarity between those synthetic documents.

The standard *K*-means algorithm works as follows. For a given pre-determined number of clusters $K = 4$ has been set and a set of data objects D, where $D = d1, d2, \ldots\ldots, d_n$, first of all, K data objects are randomly selected to initialize k clusters, each one works as the centroid of that cluster. The remaining objects are then assigned to a cluster represented by the nearest or most similar centroid. The centroid of a cluster in *K*-means algorithm is the average of all the objects in that cluster, i.e., the centroid value in each dimension is the arithmetic mean of that dimension over all the objects in the cluster.

Let D be a set of documents. Its centroid is defined as,

$$\vec{\omega_c} = \frac{1}{|D|} \sum_{\vec{\omega_d} \in D} \vec{\omega_d} \tag{6}$$

Here, $\vec{\omega_d}$ is are m-dimensional vector, where $\omega \in W$ in document $d \in D$, over the term set $W = \omega_1, \ldots, \omega_m$. Each dimension represents a term with its weight in the document. Equation 6 refers to the mean value of all the term vectors in the set. Next, new centroids are recomputed for each cluster and in turn, all documents are re-assigned based on the new centroids. This step iterates until a converged and fixed solution is reached, where all data objects remain in the same cluster after an update of centroids.

We can clearly see from the Table 1 that the words of Topic 1 and Topic 3 mostly relate to *machine learning*. Similarly, Topic 2 and Topic 5 correspond to *social media*, Topic 4 and Topic 6 indicate *natural language processing* and Topic 7 and Topic 8 indicate to the *data mining* respectively.

TABLE 1. Word to topic assignments yielded from LDA model.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
predict	online	detect	text	colleagu	semant	knowledg	data
machin	tag	train	noisi	friend	literatur	database	pattern
learn	user	label	clean	tweet	ontology	extract	context
supervis	influece	classif	word	node	keyword	retrive	mine
model	share	featur	dictionari	post	document	tupl	queri
unsupervis	profile	accuraci	textual	follow	language	inform	infer

TABLE 2. *K*-means clustering algorithm outcomes for different values of *k* for DBLP dataset.

No. of Clusters (k)	Clusters	Topics of corresponding clusters
K=5	Cluster 1	*[Topic 1, Topic 6]*
	Cluster 2	*[Topic 4, Topic 6]*
	Cluster 3	*[Topic 1, Topic 7]*
	Cluster 4	*[Topic 3, Topic 8]*
	Cluster 5	*[Topi 2, Topic 5]*
K=4	**Cluster 1**	***[Topic 2, Topic 5]***
	Cluster 2	***[Topic 4, Topic 6]***
	Cluster 3	***[Topic 1, Topic 3]***
	Cluster 4	***[Topic 7, Topic 8]***
K=3	Cluster 1	*[Topic 4, Topic 6, Topic 8]*
	Cluster 2	*[Topic 5, Topic 7, Topic 8]*
	Cluster 3	*[Topic 1, Topic 2, Topic 3]*

We find the best K for K-means clustering algorithm by comparing the inter-cluster similarity measurement. According to our observation, the lowest inter-cluster similarity score is produced for $k = 4$ for DBLP dataset. Table 2 shows the topics of corresponding clusters for different K values of K-means clustering algorithm. It is clearly seen from Table 2 that the clustering of similar topics has been performed best for the $k = 4$ in K-means clustering algorithm.

In Flickr, the photos are uploaded under a tag. These tags represent what the picture is about. For example, beach, island, mountains, forest altogether denote the topic of nature. Again, music, paintings, books represent the topic of arts and literature. We have considered these tags as our node attributes/topics in flick dataset in this research.

C. TOP-DOWN ALGORITHM

This subsection describes the development of an algorithmic framework to search the community for a given Q. This algorithm first finds a k-core connected subgraph containing the query node u_q, and then iteratively removes nodes with the smallest attribute score contribution.

1) ALGORITHM OVERVIEW

As mentioned before, the algorithm has activity driven temporal active communities, known as ATAC, includes three-fold stages. First, it pre-processes the social streams and then normalize the words to produce cleaned social streams (line 1-5). In the next stage, it first applies the topic modeling approach (LDA/BERTopic) on the processed social streams in order to get topic distribution for each topic discussed among the social users (line 6). Next, it prepares the list that consists the most related words from each topic and feed this list as an input to K-means algorithm (line 7-12). The algorithm then applies K-means algorithm to group similar topics (line 13). In the third and final stage of the proposed framework, the algorithm first computes the induced subgraph $\mathcal{C}_q$ from the set of nodes N_{u_q} who are within h hops away from the query node u_q (line 14-15). Next, it discovers the k-core subgraph containing u_q from $\mathcal{C}_q$ (line 16). Then,

Algorithm 1 ATAC

Input: $G = (U, E)$, $Q = \{u_q, \mathcal{A}_q\}$, activeness threshold θ_a, node core number k, h hops and activity stream S
Output: An online Active community $\mathcal{C}_q$ containing query node u_q at time point t_j

Stage 1:

1: **for** each activity tuple $\langle u_i, \psi_{u_i}, t_j \rangle \in S$ **do**
2: **for** each word $w_j \in \psi_{u_i}$ **do**
3: perform text normalisation on w_j
4: **end for**
5: **end for**

Stage 2:

6: Apply Topic Modeling approach (LDA/BERTopic) on the cleaned social stream S from Stage 1
7: $L \leftarrow Array()$
8: **for** each element $\phi_{T_m} \in$ word-topic distribution set Φ**do** {Each element ϕ_{T_m} in Φ contains most representative words for topic T_m }
9: **for** each word $w_j \in \phi_{T_m}$ **do**
10: $L.add(w_j)$
11: **end for**
12: **end for**
13: Apply K-means algorithm on word list L

Stage 3:

14: Find a set of nodes N_{u_q} who are within h hops away from the query node u_q at time point t_j
15: Compute the induced subgraph $\mathcal{C}_q$ on N_{u_q}, i.e. $\mathcal{C}_q = (N_{u_q}, E(N_{u_q}))$, where $E(N_{u_q}) = (v_1, v_2) : v_1, v_2 \in N_{u_q}, (v_1, v_2) \in E$
16: Maintain $\mathcal{C}_q$ as a k-core
17: **for** each $u_i \in \mathcal{C}_q$ **do**
18: compute the activeness score $\sigma_{(u_i, \psi_{u_i})}$, incorporating time-based forgetting factor $\mu_{\langle u_i, \psi_{u_i}, t_j \rangle}$
19: **if** $\sigma_{(u_i, \psi_{u_i})} < \theta_a$ **then**
20: DFS(u_i)
21: Maintain $\mathcal{C}_q$ as a k-core
22: **end if**
23: **end for**
24: Output the active connected k-core $\mathcal{C}_q$ at time point t_j.
25: **Procedure** DFS(u)
26: **for** each $v \in N(u, \mathcal{C}_q)$ **do** {$N(u, \mathcal{C}_q)$ is the list of u's neighborhood
27: remove edge (u, v) from $\mathcal{C}_q$
28: **if** $deg_{\mathcal{C}_q}(v) < k$ **then**
29: DFS(v)
30: **end if**
31: **end for**
32: remove node u from $\mathcal{C}_q$ $= 0$

it iteratively eliminates such nodes (we call them as inactive nodes) from $\mathcal{C}_q$, whose activeness score ($\sigma_{(u_i,\psi_{u_i})}$, calculated incorporating time-based forgetting factor as Equation 1) are less than a given threshold θ_a and preserves the remaining $\mathcal{C}_q$ as k-core, until no longer possible (line 17-23). Removal of an inactive node u requires to recursively delete all the nodes that violate the cohesiveness constraint using DFS procedure (lines 25-32). This is because the degree of u's neighbor nodes decrease by 1 due to the removal of the inactive node u. This may result into violation of the cohesiveness constraint by some of u's neighbors. Due to this, they cannot be included in the subsequent social groups, and thereby we need to delete them. Similarly, we also need to verify the neighbors at other hops (e.g., 2-hop, 3-hop, etc.) that they satisfy the cohesiveness constraint. Finally, the proposed algorithm `ATAC` provides the desired outputs as activity-driven temporal active community `(ATAC)` (line 24).

V. EXPERIMENTAL EVALUATION

This section assesses the performance of the algorithms on three real graph datasets. All the experiments have been performed on an Intel(R) Core(TM) i5-8265U 1.6 GHz - 1.8 GHz, Windows 10 PC with 12 GB RAM and 240 GB SSD.

A. DATA SET

1) TWITTER DATASET

Twitter is a free and popular micro-blogging service of social networking that enables registered users to broadcast short messages called tweets. Through different operating systems and devices, Twitter users may transmit tweets and follow tweets of other multiple users. These follow-following relationships is known as connections. Cellular phone text messages, online users, or posts on the website[3] can be used to post tweets and reply to them.

We conduct our experiment on a Twitter dataset named CRAWL [10]. In this dataset, we consider the user tweets from January, 2007 to December, 2012. We apply BERTopic model [33] to extract 100 latent topics in CRAWL and then cluster the similar topics into 20 categories using k-means algorithm. We set the input query attribute as {*social media*, *politics*, *entertainment*}.

We also conduct our experiment on another Twitter dataset named SNAP [39]. SNAP contains 467 million Twitter posts from 20 million users from June 11, 2009 to December 31, 2009. We choose 4,00,000 users and consider their tweets from June 11, 2009 to July 19, 2009. In the SNAP dataset, we set our input query attribute as {*iran election*, *michael jackson* and *social media*}(same topics as conference paper [21]).

2) FLICKR DATASET

Flickr[4] is an image and video hosting service, as well as serves an online community for professional and unseasoned photographers. It is one of the most popular platforms for amateur and professional photographers to host high-resolution photos and videos. The users of this photo-sharing platform do not need to create account to see photos and videos uploaded previously, but, having an account is mandatory to upload their own photo on the site. Registration of an account also enables users to create a profile page featuring images and videos posted by the user and also provides the option of adding another Flickr user as a contact, i.e. connection. They can get connected with each other according to their own choice. As mentioned earlier, there are different types of users in Flickr, for instance, people can upload photos of nature, portraits, festivals, landscape, architecture and so on. While uploading any content the user can categorize her one by adding one or more tags with it. It comes in both forms of the website, and mobile app for all platforms (android, windows and iOS).

Here, for each user in the Flickr dataset, we choose 30 most frequent tags of its associated photos as its attributes [2]. We then clustered the similar tags and choose {*nature*, *festival*, *architecture* } as input query attributes.

TABLE 3. Datasets.

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
SNAP	400,000	5,357,560	573,832

3) DBLP DATASET

DBLP[5] is an on-line reference for bibliographic information on major **computer science publications.** It has evolved from an early small experimental webserver to a popular open-data service for the whole computer science community. DBLP originally stood for database systems and logic programming. As an acronym, it has been taken to stand for the Digital Bibliography and Library Project. It grew from a small collection of HTML files, started at the University of Trier, Germany, in 1993, and became an organization hosting a database and a huge academic coauthor network.

Our algorithm has been verified to an academic coauthor network dataset [13] which includes detail information of each research paper and also the collaboration network among the authors. This research selects the research papers that are published within 2000 and 2014. The LDA topic modeling [19] approach were employed to extract 30 research topics from the abstracts of the papers and then cluster similar topics into 10 categories. In DBLP dataset, the input query attribute is set as {*data mining*, *natural language processing (NLP)*, *social network analysis (SNA)* }.

Table 3 shows the statistics of our experimental data.

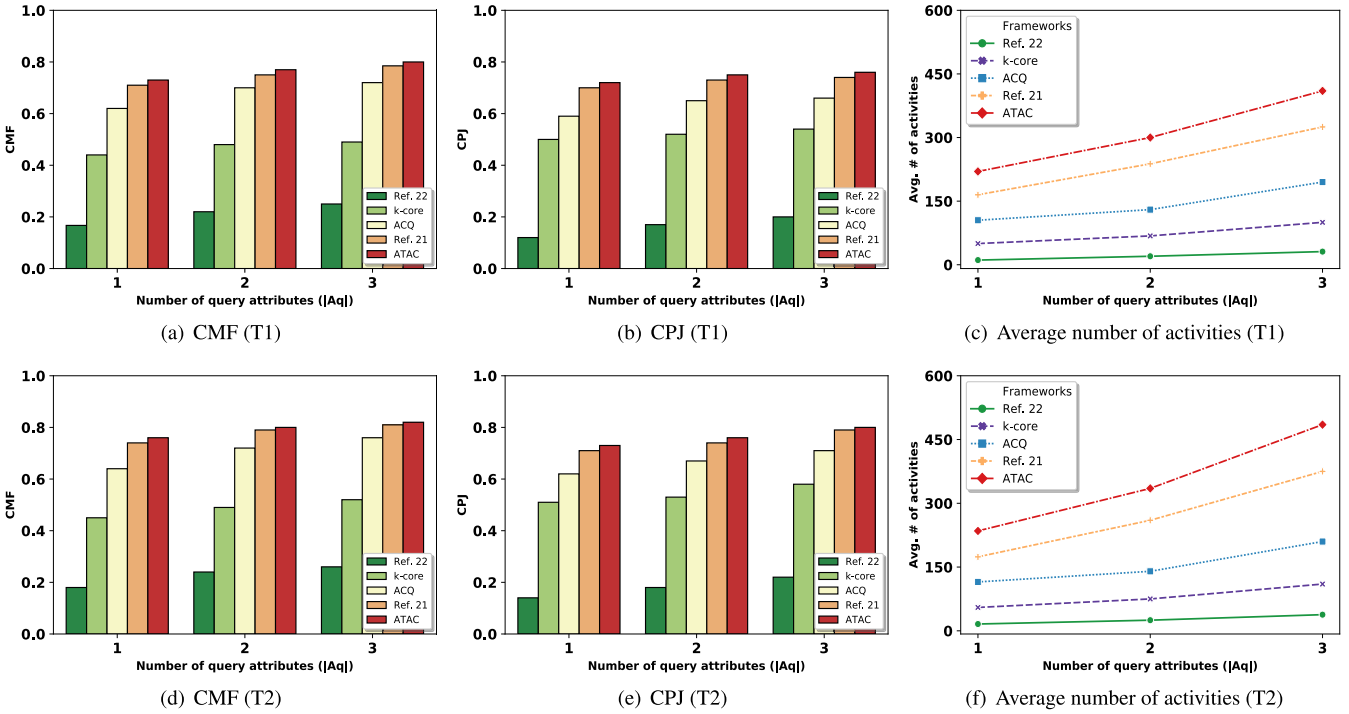
B. COMPARISON METHODS

We compare our proposed `ATAC` algorithm with four other methods. We select `ACQ` method, proposed by Fang *et al.* [2],

[3]https://twitter.com/

[4]https://www.flickr.com/explore

[5]https://dblp.org/



(a) CMF (T1) (b) CPJ (T1) (c) Average number of activities (T1)

(d) CMF (T2) (e) CPJ (T2) (f) Average number of activities (T2)

FIGURE 3. Performance comparison on CRAWL dataset in time interval T1 (01/01/2009 to 31/12/2010) and T2 (01/01/2011 to 31/12/2012) (in all cases, $k = 4$, $h = 3$, $\gamma = 150$, $\theta = 0.5$).

for community search over attributed graphs based on k-cores. The key distinction with our work is ACQ doesn't consider users' topical activeness as well as ignore the prospective temporality of users' interests. Again, ACQ doesn't take into account similar topics. Furthermore, we compare our method (*ATAC*) with the framework proposed by Das *et al.* [21]. Finally, we consider a baseline solution ($k-$core) which forms communities based on only k-core i.e. focusing only the structural cohesiveness, and another framework proposed by Wu *et al.* [22], where they considered no topics, instead they took only connections in order to form communities.

C. EVALUATION METRICS

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

1) COMMUNITY MEMBER FREQUENCY (CMF)

This is inspired by the classical document frequency measure. 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 $N(\mathcal{C}_q)$:

$$CMF(N(\mathcal{C}_q)) = \frac{1}{\mathcal{L} \times |\mathcal{A}_q|} \sum_{i=1}^{\mathcal{L}} \sum_{p=1}^{|\mathcal{A}_q|} \frac{n_{i,p}}{|\mathcal{C}i|} \tag{7}$$

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.

2) COMMUNITY PAIRWISE JACCARD (CPJ)

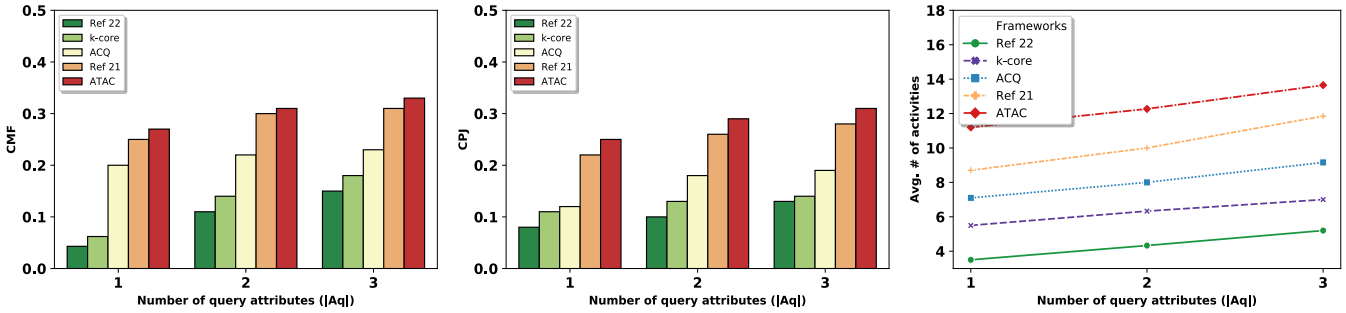
This is established on the similarity between the attribute sets of any pair of nodes of community $\mathcal{C}_i$. This research employs the Jaccard similarity, which is commonly used in the IR literature. Let $\mathcal{C}_{i,j}$ be the j-*th* node of $\mathcal{C}_i$. The CPJ is then the average similarity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

$$CPJ(N(\mathcal{C}_q)) = \frac{1}{\mathcal{L}} \sum_{i=1}^{\mathcal{L}} \frac{1}{|\mathcal{C}_i|^2} \left[\sum_{j=1}^{|\mathcal{C}_i|} \sum_{k=1}^{|\mathcal{C}_i|} \frac{|\mathcal{A}_q(\mathcal{C}_{i,j})| \cap |\mathcal{A}_q(\mathcal{C}_{i,k})|}{|\mathcal{A}_q(\mathcal{C}_{i,j})| \cup |\mathcal{A}_q(\mathcal{C}_{i,k})|} \right] \tag{8}$$

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.

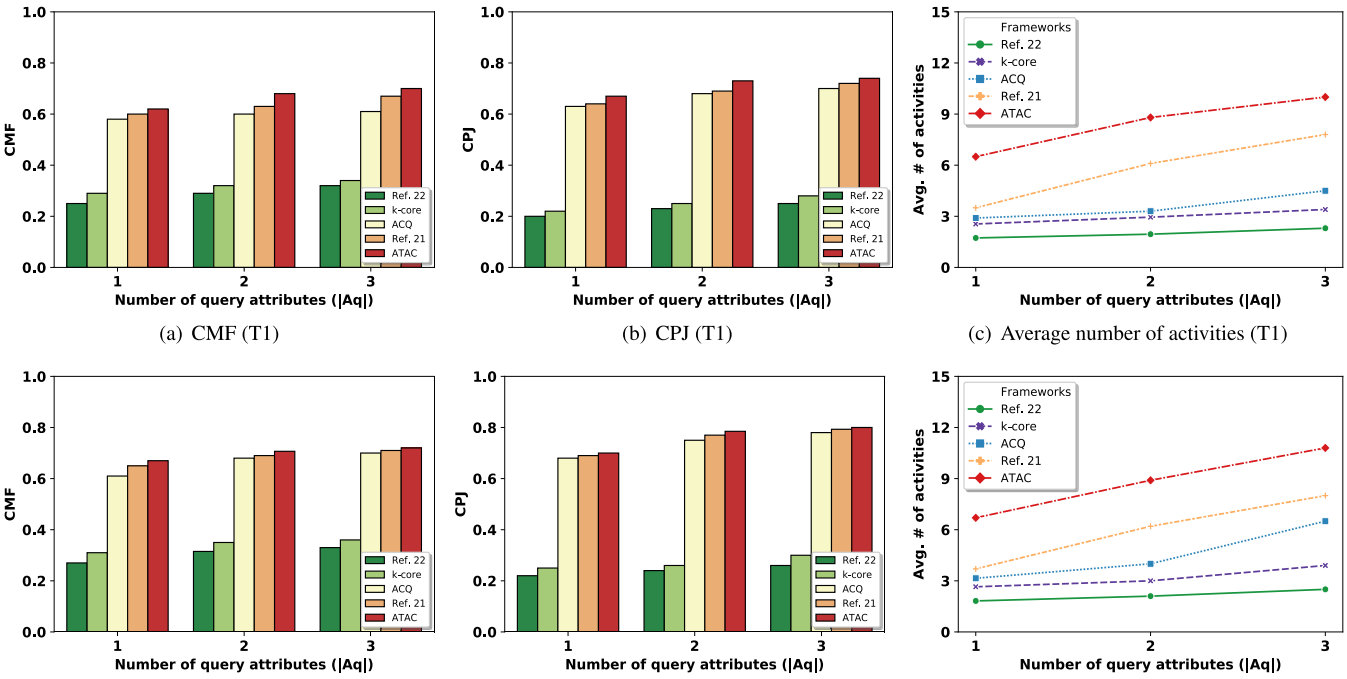
D. COMMUNITY QUALITY EVALUATION

Figure 3 shows the community quality evolution among the five methods on CRAWL dataset for two-time intervals.



(a) CMF (b) CPJ (c) Average number of activities

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



(a) CMF (T1) (b) CPJ (T1) (c) Average number of activities (T1)

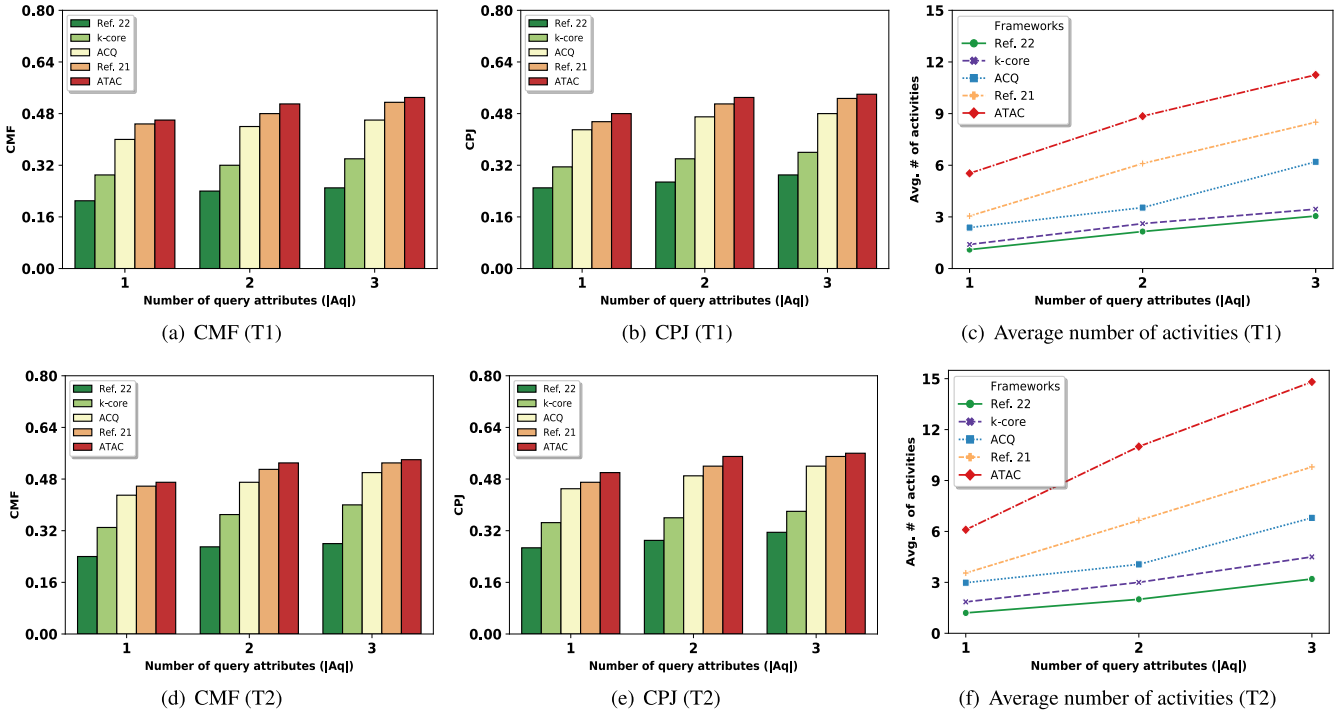
(d) CMF (T2) (e) CPJ (T2) (f) Average number of activities (T2)

FIGURE 5. Performance comparison on DBLP dataset in time interval T1 (2005-2007) and T2 (2008-2010) (in all cases, $k = 3$, $h = 3$, $\gamma = 3$, $\theta = 0.5$).

ATAC always performs the best, in terms of CMF and CPJ (Figure 3(a), (b), (d) and (e)). The reason is that each community member has to perform a 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 the framework of Ref. [21], they did not consider time based forgetting factor, so all of the users' all activities were equally counted. Consequently, there were some users who did not perform any online actions on a respective query attribute in recent times, but long ago. In the case of ACQ, there are many low active community members who don't have an interest in most of the query topics. So, the coverage of query topics within the communities is not that much better as in ATAC. For the same reason, the average number of activities are also low in ACQ comparing with ATAC. On the other hand, the values of CMF, CPJ and an average number of activities in k−core and Ref. [22] 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 in the given query topics.

For the same reasons, the result of ATAC outperforms the other four methods in SNAP dataset for both time intervals. The results of three evaluation measures are shown in Figure. 6

Figure 4 shows the results on effectiveness among the five methods on Flicker dataset. Similar to the Twitter dataset, ATAC achieves better results in all cases due to the consideration of the certain degree of topical activeness among the community members. This dataset doesn't contain the temporal information associated with users' activities and so we consider the entire time period as one-time interval.



(a) CMF (T1) (b) CPJ (T1) (c) Average number of activities (T1)

(d) CMF (T2) (e) CPJ (T2) (f) Average number of activities (T2)

FIGURE 6. Performance comparison on SNAP dataset in time interval T1 (11.06.2009 to 30.06.2009) and T2 (01.07.2009 to 19.07.2009) (for Ref. 22 $k = 4$ (k-means) and in all other cases, $k = 4$ (k-core), $h = 3$, $\gamma = 5$, $\theta = 0.5$).

Figure 5 shows the community quality evolution among the five methods on DBLP dataset for two-time intervals. We see that ATAC outperforms the other four 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 is better than the other three datasets. Again, the number of activities (i.e. publishing research papers) in DBLP are very low comparing with other datasets. So, we see that the performance of ACQ, Ref. [21] and ATAC improves significantly.

Larger values of $|\mathcal{A}_q|$ result 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.

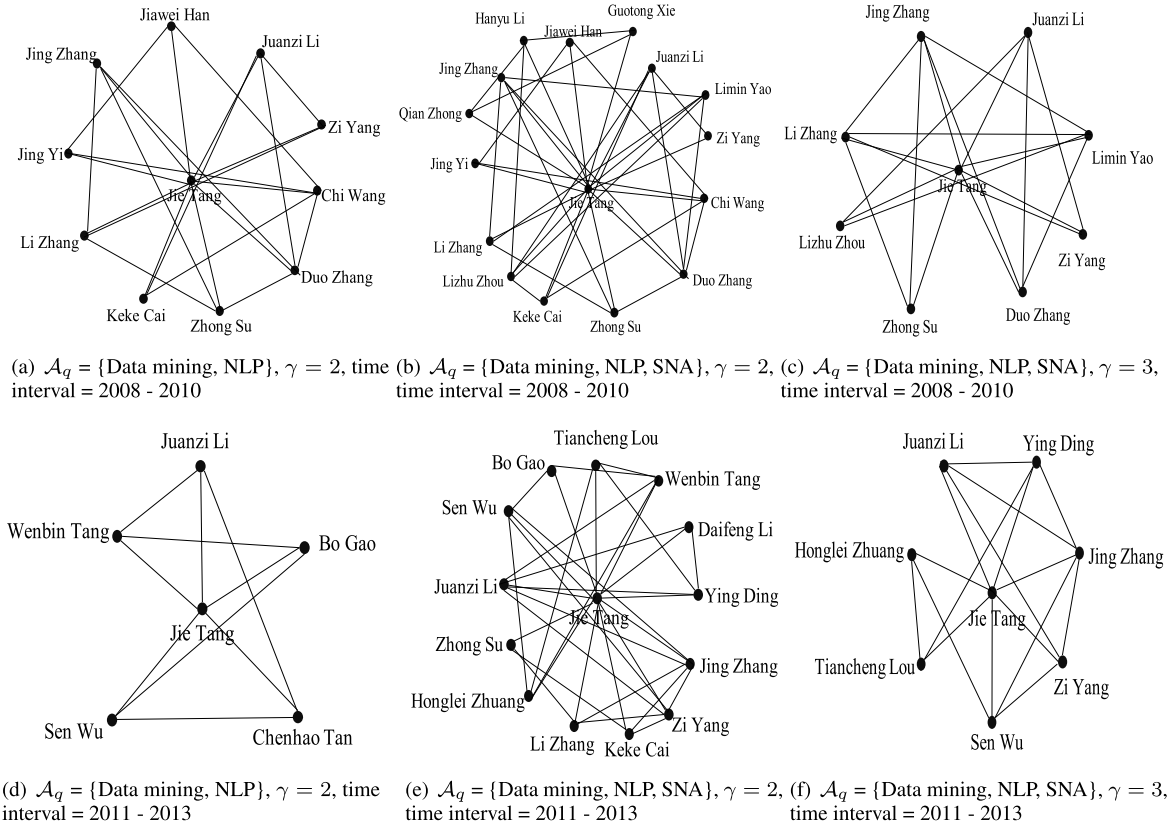
E. A CASE STUDY

This research investigated a local community in DBLP dataset which includes Jie Tang (as query node), who is one of the prominent researchers in the data mining area, to observe the distinctions in the community members for various values of γ and $\mathcal{A}_q$ depicted in 7. Here, we considered non-overlapping time intervals.

Figure 7(a) demonstrates the yielded community for the query Q as {*data mining*, *NLP*} and $\gamma = 2$ in 2008-2010 time interval. After adding one more attribute {*SNA*} in the same query in same time interval and $\gamma = 2$, the community size gets increased as 7(b). Then we show the community layout after setting the query Q as {*data mining*, *NLP*, *SNA*}, $\gamma = 3$ and time interval as 2008-2010 and show that some members of the community got deduced from the previous one after making $\gamma = 3$ 7(c). Now, shifting the time interval to 2011-2013, for the same consecutive query settings, our proposed framework produces the output as 7(d) for input query {*data mining*, *NLP*} and γ=2. Then we included one more topic *SNA*in the query set keeping the γ same, we got the output community as 7(e), Finally, in 7(f), we have illustrate the layout of the community after setting the query as {*data mining*, *NLP*, *SNA*} and $\gamma = 3$. For all six cases we kept the value of k, h and θ as 3, 3 and 0.5 respectively. It is easily noticed that as we increase the number of topics i.e., query attributes the size of the community gets bigger. Our observation is that the value of k, γ and θ can balance the trade-off between activeness and cohesiveness of a community.

VI. DISCUSSION

This research investigates the issue towards how the *topical activeness* of the members of a certain community varies over different time intervals and over different query attributes. In the real world, time is considered as the most significant factor that impacts on user activities [12]. We proposed a time-based forgetting factor in order to discount the weight of the users' past activities since those actions do not exhibit the current topic of interest to them. Equations have been formulated to calculate the degree of activeness of the users against different attributes of a given query. In this research, we have conducted an extensive experiment on four benchmark datasets including two Twitter datasets named CRAWL and SNAP, and other two are DBLP and Flickr. First of all, the prepossessing step has been performed in order to



(a) $\mathcal{A}_q$ = {Data mining, NLP}, $\gamma = 2$, time interval = 2008 - 2010 (b) $\mathcal{A}_q$ = {Data mining, NLP, SNA}, $\gamma = 2$, time interval = 2008 - 2010 (c) $\mathcal{A}_q$ = {Data mining, NLP, SNA}, $\gamma = 3$, time interval = 2008 - 2010

(d) $\mathcal{A}_q$ = {Data mining, NLP}, $\gamma = 2$, time interval = 2011 - 2013 (e) $\mathcal{A}_q$ = {Data mining, NLP, SNA}, $\gamma = 2$, time interval = 2011 - 2013 (f) $\mathcal{A}_q$ = {Data mining, NLP, SNA}, $\gamma = 3$, time interval = 2011 - 2013

FIGURE 7. **Case study: results for different $\mathcal{A}_q$ and γ in DBLP dataset (in all cases, u_q = "Jie Tang", $k = 3$, $h = 3$, $\theta = 0.5$).**

remove noisy and irrelevant parts of the activity stream S. Then, the topic modeling approach (LDA/BERTopic) to find the latent topic from the bunch of text, e.g. the topics from the abstracts of research papers. After that, we employed k-means clustering algorithm [7] to cluster the similar topics. A greedy algorithm `ATAC` has been introduced to find the attribute driven temporal local active community with respect the given query Q, containing the query node u_q over k-core connected subgraph, within h hops.

At the evaluation stage, we justify the efficiency of our proposed framework with three evaluation metrics, **Community Member Frequency (CMF), Community pairwise Jaccard (CPJ), and Average Number of Activities.** The comparison has also shown with four previously proposed methods: Wu *et al.* [22], $k-$`core`, `ACQ` [2], and Das *et al.* [21] The performance of our proposed framework `ATAC` is better than all other methods for all four datasets, as we considered topical activeness with time based forgetting factor of the community members with respect to the query attributes ($|A_q|$). It can be seen clearly that the number of community members varies when we make some changes in the query attributes. (Figure: 7 (a), (b)). It has also demonstrated that the members will also change if we set the query in a different time interval keeping all other attributes same, e.g. (Figure: 7 (a), (d)). Then talking about the efficacy of our proposed method, `ATAC` outperforms the two baseline frameworks for two different non-overlapping time intervals in all the cases and for all three datasets.

If we set a large value of k for $k-$`core`, then the number of community members will decrease depending on the dataset.

There is a wide range of important implications of this framework in real-life. Some of them are as follows.

- Diverse applications of viral marketing and information diffusion.
- Anticipating the number of members of a group of target customer according to the history of purchasing any specific product.
- Predicting the role of any community towards any particular upcoming events analyzing their group behaviour in likewise past events.

VII. CONCLUSION

Through this research, we analyzed the problem of active local community search in the attributed social graph. It has been observed how the users' topical activeness vary on OSNs over time and different query attributes. We have empirically shown that the users' individual activeness vary widely in different time intervals and over different attributes. This research outlined an activeness score function for the candidate community members in order to put more weights on the users' recent online actions and developed methods to search the query oriented active community i.e. `ATAC`. The

effectiveness of the proposed method has been demonstrated over extensive experiments on three real benchmark datasets. The efficacy of the ATAC has been illustrated through three evaluation metrics **(CMF, CPJ and an average number of activities)** for all the datasets in two different time intervals. It can be clearly noticed that our proposed framework ATAC outperforms the other two baseline methods in all the cases. Afterwards, We have also shown a case study over the **DBLP** dataset keeping the query node *Jie Tang*, for different query attributes over two non-overlapping time intervals to exhibit our claim graphically. At the last section of the paper, we discuss different real-life implications and aspects of this research. In future, our plan is to work on dynamic network and user interactions in order to detect real-time attribute driven local active community. We are also planning about to apply k-Truss and k-clique to extract subgraph from large networks.

REFERENCES

[1] C. C. Aggarwal and W. H. ArnetMiner, *Managing and Mining Graph Data*. Boston, MA, USA: Springer, 2010, pp. 13–68.
[2] Y. Fang, R. Cheng, S. Luo, and J. Hu, "Effective community search for large attributed graphs," *Proc. VLDB Endowment*, vol. 9, no. 12, pp. 1233–1244, Aug. 2016.
[3] N. Barbieri, F. Bonchi, E. Galimberti, and F. Gullo, "Efficient and effective community search," *Data Mining Knowl. Discovery*, vol. 29, no. 5, pp. 1406–1433, Sep. 2015.
[4] X. Huang, L. V. Lakshmanan, J. X. Yu, and H. Cheng, "Approximate closest community search in networks," *Proc. VLDB Endowment*, vol. 9, no. 4, pp. 267–287, 2015.
[5] P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti, "Generalized Louvain method for community detection in large networks," in *Proc. Int. Conf. Intell. Syst. Design Appl.*, 2011, pp. 88–93.
[6] I. H. Sarker, A. Colman, and J. Han, "RecencyMiner: Mining recency-based personalized behavior from contextual smartphone data," *J. Big Data*, vol. 6, no. 1, pp. 1–21, Dec. 2019.
[7] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *Social Netw. Comput. Sci.*, vol. 2, no. 3, pp. 1–21, May 2021.
[8] M. E. J. Newman and J. Park, "Why social networks are different from other types of networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 68, no. 3, Sep. 2003, Art. no. 036122.
[9] J. Cohen, "Trusses: Cohesive subgraphs for social network analysis," *J. Nat. Secur. Agency Tech. Rep.*, vol. 16, no. 3, pp. 3–29, 2008.
[10] P. Bogdanov, M. Busch, J. Moehlis, A. K. Singh, and B. K. Szymanski, "The social media genome: Modeling individual topic-specific behavior in social media," in *Proc. ASONAM*, Aug. 2013, pp. 236–242.
[11] X. Wang, G. Liu, and J. Li, "Overlapping community detection based on structural centrality in complex networks," *IEEE Access*, vol. 5, pp. 25258–25269, 2017.
[12] I. H. Sarker, "Context-aware rule learning from smartphone data: Survey, challenges and future directions," *J. Big Data*, vol. 6, no. 1, pp. 1–25, Dec. 2019.
[13] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "ArnetMiner: Extraction and mining of academic social networks," in *Proc. KDD*, 2008, pp. 990–998.
[14] Y. Fang, R. Cheng, X. Li, S. Luo, and J. Hu, "Effective community search over large spatial graphs," *Proc. VLDB Endowment*, vol. 10, no. 6, pp. 709–720, Feb. 2017.
[15] N. Natarajan, P. Sen, and V. Chaoji, "Community detection in content-sharing social networks," in *Proc. ASONAM*, 2013, pp. 82–89.
[16] K. Gu, D. Liu, and K. Wang, "Social community detection scheme based on social-aware in mobile social networks," *IEEE Access*, vol. 7, pp. 173407–173418, 2019.
[17] S. Souravlas, A. Sifaleras, and S. Katsavounis, "A parallel algorithm for community detection in social networks, based on path analysis and threaded binary trees," *IEEE Access*, vol. 7, pp. 20499–20519, 2019.
[18] B. Han, P. Cook, and T. Baldwin, "Lexical normalization for social media text," *ACM Trans. Intell. Syst. Technol.*, vol. 4, no. 1, pp. 1–27, Jan. 2013.
[19] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.
[20] X. Huang and L. V. S. Lakshmanan, "Attribute-driven community search," *Proc. VLDB Endowment*, vol. 10, no. 9, pp. 949–960, May 2017.
[21] B. C. Das, M. S. Ahmed, and M. M. Anwar, "Query-oriented active community search," in *Proc. Int. Joint Conf. Comput. Intell.*, 2020, pp. 495–505.
[22] L. Wu, Q. Zhang, C.-H. Chen, K. Guo, and D. Wang, "Deep learning techniques for community detection in social networks," *IEEE Access*, vol. 8, pp. 96016–96026, 2020.
[23] K. H. Lim and A. Datta, "An interaction-based approach to detecting highly interactive Twitter communities using tweeting links," *Book Web Intell.*, vol. 14, no. 1, pp. 1–15, 2016.
[24] M. M. Anwar, C. Liu, and J. Li, "Uncovering attribute-driven active intimate communities," in *Proc. ADC*, 2018, pp. 109–122.
[25] M. M. Anwar, "Query-oriented temporal active intimate community search," in *Proc. Australas. Database Conf.*, 2020, pp. 206–215.
[26] H. Dev, M. E. Ali, and T. Hashem, "User interaction based community detection in online social networks," in *Proc. Int. Conf. Database Syst. Adv. Appl.*, 2014, pp. 296–310.
[27] X. Xin, C. Wang, X. Ying, and B. Wang, "Deep community detection in topologically incomplete networks," *Phys. A, Stat. Mech. Appl.*, vol. 469, pp. 342–352, Mar. 2017.
[28] M. Dhilber and S. D. Bhavani, "Community detection in social networks using deep learning," in *Proc. Int. Conf. Distrib. Comput. Internet Technol.*, 2020, pp. 241–250.
[29] C. H. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala, "Latent semantic indexing: A probabilistic analysis," in *Proc. 17th ACM SIGACT-SIGMOD-SIGART Symp. Princ. Database Syst.*, 1998, pp. 159–168.
[30] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. 1st Int. Conf. Learn. Represent. (ICLR)*, Scottsdale, AZ, USA, 2013, pp. 1–13.
[31] X. Jin and J. Han, "K-means clustering," in *Encyclopedia of Machine Learning*. Boston, MA, USA: Springer, 2011.
[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805
[33] M. Grootendorst. (2021). *BERTopic: Leveraging BERT and c-TF-IDF to Create Easily Interpretable Topics*. Accessed: May 24, 2021. [Online]. Available: https://github.com/MaartenGr/BERTopic
[34] Y. Zhou, H. Cheng, and J. X. Yu, "Graph clustering based on structural/attribute similarities," *Proc. VLDB Endowment*, vol. 2, no. 1, pp. 718–729, Aug. 2009.
[35] J. Fagnan, R. Rabbany, M. Takaffoli, E. Verbeek, and O. R. Zaïane, "Community dynamics: Event and role analysis in social network analysis," in *Advanced Data Mining and Applications*. Cham, Switzerland: Springer, 2014, pp. 85–97.
[36] W. Luo, D. Zhang, H. Jiang, L. Ni, and Y. Hu, "Local community detection with the dynamic membership function," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 5, pp. 3136–3150, Oct. 2018.
[37] T. Meng, L. Cai, T. He, L. Chen, Z. Deng, W. Ding, and Z. Cao, "A modified distance dynamics model for improvement of community detection," *IEEE Access*, vol. 6, pp. 63934–63947, 2018.
[38] W. Liu, T. Suzumura, L. Chen, and G. Hu, "A generalized incremental bottom-up community detection framework for highly dynamic graphs," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2017, pp. 3342–3351.
[39] J. Yang and J. Leskovec, "Patterns of temporal variation in online media," in *Proc. 4th ACM Int. Conf. Web Search Data Mining (WSDM)*, 2011, pp. 177–186.

BADHAN CHANDRA DAS received the bachelor's and master's degrees in computer science and engineering from Jahangirnagar University, Bangladesh, under the supervision of Dr. Md. Musfique Anwar and Dr. Md. Al-Amin Bhuiyan. His research interests include data mining, social network analysis, and natural language processing.

MD. MUSFIQUE ANWAR received the Ph.D. degree from the Swinburne University of Technology, Australia, in 2018. He is currently an Associate Professor at Jahangirnagar University, Bangladesh. His research interests include data mining, social network analysis, natural language processing, and software engineering.

MD. AL-AMIN BHUIYAN received the Ph.D. degree from Osaka City University, Japan, in 2001. From 2001 to 2003, he worked as a COE Postdoctoral Researcher with the Intelligent Systems Research Division, National Institute of Informatics, Japan. He was involved as a Research Associate at the University of Hull, U.K. He is currently an Associate Professor at King Faisal University, Saudi Arabia. His research interests include image processing, pattern recognition, artificial intelligence, neural networks, and robotic vision.

IQBAL H. SARKER (Member, IEEE) received the Ph.D. degree from the Department of Computer Science and Software Engineering, Swinburne University of Technology, Melbourne, Australia, in 2018. He is currently working as a Faculty Member of the Department of Computer Science and Engineering, Chittagong University of Engineering and Technology. His research interests include data science, machine learning, AI-driven computing, NLP, cybersecurity analytics, behavioral analytics, the IoT-smart city technologies, and healthcare analytics. He has published a number of peer-reviewed journals and conferences in top venues. He is one of the research founders of the International AIQT Foundation, Switzerland, and a member of ACM.

SALEM A. ALYAMI (Member, IEEE) received the Ph.D. degree in bio-statistics from Monash University, Australia, in 2017. He has been working as an Assistant Professor with the School of Mathematics and Statistics, IMAMU, Riyadh, Saudi Arabia, since 2017, contributing/leading several grants in biostatistics projects. He has recently been appointed as the Dean of the Deanship of Scientific Research at IMAMU. His research interests include Bayesian networks, neural networks, Bayesian statistics, MCMC methods, and applications of statistics in biology and medicine.

MOHAMMAD ALI MONI received the Ph.D. degree in clinical bioinformatics and machine learning from the University of Cambridge. He is a Research Fellow and a Conjoint Lecturer at the University of New South Wales, Australia. His research interests include encompass artificial intelligence, machine learning, data science, medical image processing, epidemiology, public health, and clinical bioinformatics.