TÜBİTAK

Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2019) 27: 1228 – 1242
© TÜBİTAK
doi:10.3906/elk-1806-91

Understanding attribute and social circle correlation in social networks

Pranav NERURKAR[*], **Madhav CHANDANE**, **Sunil BHIRUD**
Department of Computer Engineering and IT, Veermata Jijabai Technological Institute, University of Mumbai,
Mumbai, India

Received: 12.06.2018 • **Accepted/Published Online:** 18.12.2018 • **Final Version:** 22.03.2019

Abstract: Social circles, groups, lists, etc. are functionalities that allow users of online social network (OSN) platforms to manually organize their social media contacts. However, this facility provided by OSNs has not received appreciation from users due to the tedious nature of the task of organizing the ones that are only contacted periodically. In view of the numerous benefits of this functionality, it may be advantageous to investigate measures that lead to enhancements in its efficacy by allowing for automatic creation of customized groups of users (social circles, groups, lists, etc). The field of study for this purpose, i.e. creating coarse-grained descriptions from data, consists of two families of techniques, community discovery and clustering. These approaches are infeasible for the purpose of automation of social circle creation as they fail on social networks. A reason for this failure could be lack of knowledge of the global structure of the social network or the sparsity that exists in data from social networking websites. As individuals do in real life, OSN clients dependably attempt to broaden their groups of contacts in order to fulfill different social demands. This means that 'homophily' would exist among OSN users and prove useful in the task of social circle detection. Based on this intuition, the current inquiry is focused on understanding 'homophily' and its role in the process of social circle formation. Extensive experiments are performed on egocentric networks (ego is user, alters are friends) extracted from prominent OSNs like Facebook, Twitter, and Google+. The results of these experiments are used to propose a unified framework: feature extraction for social circles discovery (FESC). FESC detects social circles by jointly modeling ego-net topology and attributes of alters. The performance of FESC is compared with standard benchmark frameworks using metrics like edit distance, modularity, and running time to highlight its efficacy.

Key words: Exploratory data analysis, social circles, social networks

1. Introduction

Social networks (graphs) have become a common vocabulary for representing complex systems across various domains. In this form of representation, systems are modeled as graphs such that the entities of these systems are represented as nodes and the relationships between these entities are edges. Online social networks (OSNs) are popular service platforms, such as Facebook, Google+, and Twitter, among others, which are commonly represented as graphs for purpose of efficient analysis [1]. These platforms give computerized socialization and are widely used. For example, Facebook's social network has been increasing in popularity and had recorded close to 200 million active client accounts by late 2018, with around 10 million messages being posted each hour and 46% of youthful clients logging in to their Facebook accounts as the first thing in their day. Behind this ubiquity lies a rich wellspring of data that could be legitimately coordinated and broken down for better

[*] Correspondence: panerurkar_p16@ce.vjti.ac.in

comprehension of the process of online socialization. This has raised the need for exploratory data analysis of such OSNs using social network analysis [2].

In exploratory data analysis of social networks, a key task is to detect social groupings of a particular user's friends. This information has to be estimated from the friendship network of that user, which is also known as an egocentric network (or ego-net). An ego-net contains an ego and alters. The alters are the friends of the ego (user). Both the ego and the alters are represented in the form of nodes in a graph. In such a graph, the social circles are subsets of the alters grouped manually by the user under certain measurements. The concept of grouping friends into subnetworks is provided in various popular social networking platforms such as Google+ (circles), Facebook (lists), and Twitter (groups) among others. Social circles have potential applications in the domains of content filtering, handling of information overload, recommendation systems, friendship predictions, group recommendations, inferring personal attributes, and human–computer interaction for controlling information boundaries between different groups of users [2–5]. It is argued that just as individuals do in real life, OSN clients attempt to broaden their groups of friends in order to fulfill different social demands [4].

Although it enhances user experience, the notion of social circles is not well received by users. This is due to the requirement of physically creating social circles and periodically reassigning member connections to appropriate social circles. As these connections change over time, the process of keeping these lists up-to-date becomes strenuous for the user [2, 6]. It is observed that for users that are less active than their peers, these lists no longer reflect the ground reality [7]. Hence, it remains an essential undertaking to outline techniques that will naturally and viably relegate users to respective social circles. Tracing this line of research, it is noticed that the current literature focuses on addressing detection of overlapping social circles or detection of social circles based on entity-annotated texts [2, 8–10]. Another line of literature focuses on attempting to enhance circle detection in a target ego-net by utilizing circle data from other ego-nets. However, a suitable framework for detecting social circles using both ego-net structural information and node attributes of alters is not available.

A related field is community detection, which involves identification of latent groups of entities in data. In network sciences such subgroups (modules) or communities are identified using network topology. Community detection algorithms aim to find such communities in undirected as well as directed graphs. Be that as it may, recognizing communities in a social network requires knowing the global system structure, effectively making this approach infeasible. Existing social network communities as identified by community discovery techniques are found to be finer or broader than actual social circles [6]. An additional challenge in the task of social circle identification is the high amount of missing values in social network data [3]. Standard clustering techniques ignore or misinterpret this incompleteness in data and hence results obtained by these are found to deviate significantly from the desired ones [2].

The present inquiry centers around building a unified mathematical model for social circle recognition based on structural characteristics of the ego-net as well as attribute information of the alters. Thus, global network information is not required. The reason to prefer locally available information is the assumption that homophily exists in the network. Homophily is characterized as a person's inclination to interface with 'like-minded' people to amplify individual incentives [11]. An analysis of close to 110 ego-nets extracted from Facebook, 938 ego-nets extracted from Twitter, and 128 ego-nets extracted from Google+ identify the presence of homophily in social networks. Based on these experiments a joint framework is proposed for social circle identification that factors in the local network structure of the ego along with attribute information of the alters. Such an approach is demonstrated to be effective even if all the node information about the alters is not available. The organization of this paper is as follows: after the introduction of the paper in Section 1, Section 2 provides the review of concepts and techniques and Section 3 provides the mathematical model for the proposed strategy. The experimental results and conclusion are presented in Sections 4 and 5, respectively.

1.1. Preliminaries

1.1.1. Problem definition

$G(V, E)$ is a social network with E edges and V nodes. $V^a \in R^p$ is used to denote the p-dimensional vertex attribute vector of $a \in V$. $V^a = [v_1^a, ..., v_p^a]$ are the attributes of node a, where p is the number of node attributes. The adjacency matrix of social network $G(V, E)$ is denoted by A. If $(i, j) \in E$, $A_{ij} = 1$, denoting that the edge (i, j) is present; otherwise, $A_{ij} = 0$. Using A, V information, the aim is to find sets $S_1, ..., S_k$ such that $\forall a \in V; a \in \phi, S_1, ..., S_k$ and $S_i \cap S_j$ may or may not be ϕ.

1.1.2. Social network generative model: J-R model

Jackson et al. proposed a social network generative model where nodes of the social network are allowed to form links to other nodes using a hybrid strategy that encapsulates elements of the preferential attachment model and the Erdos–Rényi model. Thus, if there are preexisting m nodes in a network, then a newborn node links to $a * m$ of them chosen uniformly at random (chance-based links) and $(1 - a) * m$ using a neighborhood search strategy (choice-based links) and attaches to them. Thus, the distribution of the expected degree is:

$$dd_i(t)/dt = \frac{a * m}{t} + (1 - a) * \frac{d_i(t)}{2t}, d_i(i) = m, \tag{1}$$

where:

- $dd_i(t)/dt$ is change in degree if node i with time,
- $d_i(i)$ is degree of node i at time $= 0$,

$$d_i(t) = (m + 2am/(1 - a))(t/i)^{(1-a)/2} - 2am/(1 - a). \tag{2}$$

The frequency distribution is given by:

$$F(d) = 1 - [\frac{(m + 2am/(1 - a))}{d + 2am/(1 - a)}]^{2/(1-a)}. \tag{3}$$

The hyperparameter a (ratio of chance-based interactions to choice-based) of the model has to be estimated and can be found out by random parameter search or grid search. The value has to be selected that gives the best fit to the degree distribution. This model is thus effective in estimating the generative process of a social network. It is used in the proposed social circle detection technique for obtaining the characteristics of the degree distribution of the ego-nets.

2. Review of the literature

Clustering and community detection methods are abundantly available in literature. Clustering approaches have their own biases in identifying coarse-grained summaries of data. Each type of clustering algorithm has an objective function (measurement of similarity) and an optimizing criterion. Hence, there exist multiple algorithms for clustering due to the multiple and diverse interpretations of similarity. Standard clustering techniques tend to ignore the 'hidden' incompleteness of social network data and are found to be unsuitable for detection of social circles [2, 12]. Community detection techniques based on Newman's modularity as a quality measure aim to partition a graph into communities (modules) in which the edges mostly reside among nodes that belong to the same communities. These seek subgraphs with small cut sizes and are infeasible for use on ego-nets as they require knowledge of the global network structure [6, 13, 14].

Perozzi *et al.* argued that the existing approaches for social circle detection focus on unattributed or plain graphs and hence only utilize the structure of the social network. A review of existing social circle detection methods led the authors to speculate that a majority of these approaches only quantified the structural quality of a network and failed to utilize node attributes. The authors proposed a new measure to overcome this drawback of existing approaches. This measure, 'normality', was a quality function related to topology of the network as well as the attributes (side information) of the nodes [15]. McAuley et al. proposed the affiliation graph model (AGM), a new line of research for detecting communities by utilizing metadata associated with the entities. However, the AGM framework is unsuitable for social circle detection as it does not allow any individual node to have high participation strength to multiple separate communities concurrently. Finally, it has a large time complexity and computation cost. This makes it unsuitable for application on graphs with more than 1000 nodes [8, 9]. Yang et al. extended the framework of AGM to propose a bipartite affiliation graph (BAG). The BAG is denoted as $B(X,C,M)$, with X as the nodes, C as the attribute value, and M as the directed edge from X to C if node X has attribute value C. Detecting a set of communities $S = S_1, S_2, ..., S_k$ in $B(X,C,M)$ is done by using a variant of the nonnegative matrix factorization method [10]. The metadata or attribute information used by these techniques is the local optimal community membership of nodes. Thus, both AGM and BAG techniques are unsuitable for ego-nets with attributed data.

Network science (graph theory)-based approaches for social network analysis are increasingly being advocated for assessing the structure and functions of complex systems such as multilevel biological networks [16, 17], epidemiology networks [18, 19], social networks [20, 21], coauthorship networks [22], collective behavior [23], and political networks [24]. Researchers in these domains have traditionally relied on network-specific feature engineering using degree statistics or kernel functions from graph theory to analyze the complex systems in these domains that have been represented in graph format [25, 26]. This feature engineering made the network analysis process task-dependent. Hence, there is a need for investigation of frameworks for efficient task-independent feature learning. The frameworks that achieve this purpose are known in the literature as network representation learning or network embedding frameworks. The goal of such frameworks is to encode network structure into low-dimensional embeddings. The literature consists of several such NRL frameworks based on matrix factorization [27–36], the word2vec (skip-gram) model of Mikolov et al. [37–46], deep convolutional neural networks [47–51], the random walk and neural network unified framework [52], hyperbolic space embedding techniques [22, 53], latent-space models [54–59], and multidimensionality reduction [60, 61]. A key drawback of these frameworks is their inability to handle missing information.

Summarizing the literature review, it is found that only the work of Perozzi et al. utilized node attribute information along with the graph topology for social circle detection. The current inquiry aims to extend this research by focusing on the relationship that exists between an ego and its alters to detect social circles. This aspect is not captured in previous works that aimed to jointly model attributes and network structure to detect social circles. Social science research has observed homophily in human interactions, i.e. individuals choosing to connect to 'like-minded' individuals in order to obtain incentives. This offline behavior of people has been found to exist in human interactions on OSNs [4, 62, 63]. Thus, the assumption that homophily could play an important role in the online socialization process would be valid. Hence, understanding homophily could help us in identifying social circles in networks. The current inquiry is based on this intuition and aims to decode the social circles formation process in popular OSNs. A data-based inquiry is performed on ego-nets from Facebook, Google+, and Twitter to understand the presence of homophily in social networks. It proposes a mathematical model to obtain a representative set of features from an ego-net using homophily. Then it combines this information with the structural characteristics of the ego-net to detect social circles.

3. Feature extraction for social circles (FESC) discovery

This section provides the concepts that form the basis for the proposed FESC algorithm. The explanation of the algorithm and its stages are provided along with the block diagram in Figure 1.

3.1. FESC algorithm stage I: calculation of structural and behavioral characteristics of ego-nets

Adhesion or edge connectivity E for connected graph G is the minimum number of edges $\lambda(G)$ whose deletion from a graph G disconnects G.

Average path length $L = \sum_1^E (G) \frac{d(u,v)}{E(G)}$.

Degree distribution of graph $P(k) = \frac{n_k}{n}$ is fraction of nodes in the network with degree k, i.e. n_k where n is the graph order.

Verification of power laws $f(k) \propto k^{-\alpha}$ in networks related to eigenvectors distribution $x_1, x_2, ...x_{20}$, component distribution $C_1, C_2, ..., C_k$.

Assortativity measures the level of homophily of the graph:

$$r = \frac{\sum_{jk} jk(e_{jk} - q_j q_k)}{\sigma_q^2}, \tag{4}$$

where:

- q_k is number of edges leaving the node, other than the one that connects the pair j, k;
- σ_q is standard deviation of q;
- e_{jk} refers to the joint probability distribution of the remaining degrees of the two vertices.

Graph density (G_D) is the number of edges present in graph G among all possible edges in G. G_D for undirected and directed graphs is given by Eqs. (5) and (6), respectively.

$$\frac{2|E|}{|V|(|V|-1)} \tag{5}$$

$$\frac{|E|}{|V|(|V|-1)} \tag{6}$$

Reciprocity ρ is the probability of vertices in a directed network to be mutually linked.

$$\rho = \frac{\sum_{i \neq j (a_{ij} - \overline{a})(i \neq j (a_{ji} - \overline{a})}}{sum_{i \neq j (a_{ij} - \overline{a})^2}} \tag{7}$$

3.2. FESC algorithm stage II: comparison of degree distribution in the ego-net

```
Algorithm 1: Comparison of degree distribution.
    Result: $X^2$, P-value, df
    Divide degree distribution $P_k$ into 6 quantiles;
    Perform binning and obtain data in equal bins;
    For same graph order obtain degree distributions $P_K^1$ of J-R model;
    Perform Pearson's chi-square test and compare $P_k$ with $P_K^1$ Obtain $X^2$, P-value, degrees of
      freedom $df$
```

3.3. FESC algorithm stage III: calculate feature similarity value of alters

```
Algorithm 2: Calculate feature similarity value.
    Result: feature similarity scores
    select user social circles; while feature set is not empty do
        for each circle aggregate unique attribute values $f_c$;
        for each circle count total values $f_{total}$;
        Similarity value = $\frac{f_c}{f_{total}}$
    end
```

3.4. FESC algorithm and block diagram

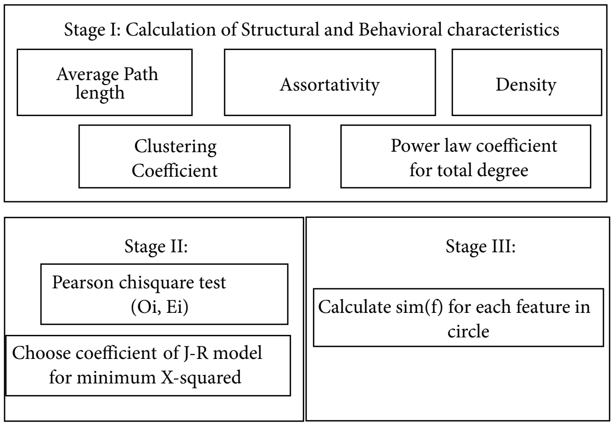
Stage I: Calculation of Structural and Behavioral characteristics
Average Path length
Assortativity
Density
Clustering Coefficient
Power law coefficient for total degree
Stage II:
Pearson chisquare test (Oi, Ei)
Choose coefficient of J-R model for minimum X-squared
Stage III:
Calculate sim(f) for each feature in circle

Figure 1. Block diagram of FESC algorithm shows the components of the three stages and individual components of each stage. The statistical features of the ego-nets are extracted from the social networks in Stage I. Stage II uses the J-R model to understand the characteristics of the degree distribution. The third stage provides a similarity value for each feature of the ego using the attribute information of the alters.

4. Experimental study

The procedure highlighted in Section 3 was applied on datasets from [64] of OSNs such as Facebook, Twitter, and Google+. The list of features extracted at the end of the analysis for each OSN showed the criteria of users in each OSN to form social circles.

4.1. Datasets

The Google+ (Gp-net) dataset has 128 ego-nets with features of both alters and egos. Each ego-net has different features related to gender, education, work, interests, etc. The Facebook (Fb-net) dataset has 110 ego-nets with

```
Algorithm 3: Feature extraction for social circles discovery.
    Result: Set of features $f \in f_1, f_2, .., f_n$
    Calculate structural characteristics such as diameter, average path length, density, degree
      distribution of graphs;
    Calculate behavioral characteristics such as reciprocity, transitivity, assortativity;
    Calculate features of circles such as number of members, circles per user;
    Compare degree distribution with Jackson–Rogers model to understand behavior;
    if Degree distribution Jackson–Rogers model then
        Calculate feature similarity value;
        Select features with lowest similarity values;
    else
        conclude no evidence of homophily;
    end
```

each ego-net having 57 features, such as name, education, work, and locale. All features are anonymized to hide the identity of the users. The Twitter (Twt-net) dataset has 938 ego-nets with each having different features. The features of the Twitter dataset are based on regular tweets by the users. In each dataset, ground truth communities of groups, circles, and lists are provided by the users for the ego-nets.

4.2. Results and discussion

Table 1, Table 2, and Table 3 give the structural and behavioral characteristics of the datasets. Socially generated networks tend to have the average distance between a pair of nodes on the order of the log of the number of nodes. The geodesic of such networks is also on the order of the log of the number of nodes. Both these features are evident from Table 1, Table 2, and Table 3 below. Transitivity in the below networks is larger than in networks where links are generated by an independent random process. Thus, homophily could be present in the networks and there is a need to compare the degree distribution of these networks with the J-R model.

Table 1. Calculated structural and behavioral characteristics of Fb-net.

Characteristics	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Order of graph	5.0	80.0	134.5	133.5	188.0	247.0
Size of graph	5	698	1828	2397	3549	17930
Degree distribution of total-degree	0.0028	0.0066	0.0101	0.0143	0.0169	0.2
Average path length	1.219	2.110	2.438	2.537	2.895	6.403
Assortativity	-0.6124	-0.1601	-0.0825	-0.0727	0.0068	0.6252
Density	0.0212	0.07387	0.1092	0.1417	0.1750	0.7115
Transitivity	0.0000	0.3892	0.5238	0.5219	0.6581	1.0000
Power law coefficient for total degree	-0.9725	0.1608	0.3293	0.3409	0.5129	1.4020

In Stage II of the algorithm, the degree distributions of the OSNs are compared with the J-R model to estimate the role of chance (network-based interactions) p_s and choice (random interactions) p_r among the actors (participants) in the OSNs.

Table 2. Calculated structural and behavioral characteristics of Twt-net.

Characteristics	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Order of graph	13.0	120.2	216.0	251.7	338.8	782.0
Size of graph	26.0	788.2	1926.0	3639.0	4409.0	26020.0
Degree distribution of total-degree	0.0013	0.003	0.0046	0.0073	0.0083	0.0769
Average path length	1.667	1.858	1.910	1.893	1.937	1.978
Assortativity	-0.5792	-0.1814	-0.1356	-0.1418	-0.0923	0.0972
Density	0.0223	0.0630	0.0902	0.1066	0.1418	0.3333
Transitivity	0.1477	0.3349	0.4354	0.4279	0.5113	0.7257
Power law coefficient for total degree	-0.1308	0.4249	0.5907	0.5616	0.7088	0.9589

Table 3. Calculated structural and behavioral characteristics of Gp-net.

Characteristics	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Order of graph	8.0	475.8	1424.0	1923.0	3374.0	4938.0
Size of graph	7	13110	71870	226300	310000	1615000
Degree distribution of total-degree	0.0003	0.0005	0.0012	0.0038	0.0025	0.1250
Average path length	1.000	2.166	2.477	2.466	2.688	3.942
Assortativity	-0.4958	-0.2275	-0.1552	-0.1625	-0.1006	0.1663
Density	0.0048	0.0285	0.0455	0.0607	0.0806	0.2496
Transitivity	0.0000	0.2519	0.3511	0.3738	0.4693	1.0000
Power law coefficient for total degree	-0.1003	0.4658	0.6573	0.6235	0.8016	1.2220

Table 4. Chance and choice in social networks.

Sr. no.	Dataset	p_r	p_s
1	Fb-Net	0.1	0.9
2	Twt-Net	0.3	0.7
3	Gp-Net	0.2	0.8

Analysis of Google+, Facebook, and Twitter reveals a higher link creation between actors obtained through network-based meetings than chance-based meetings as given in Table 4. The intuitive reason behind a higher ratio of network-based links in these socially generated networks is due to the friendship recommendations given by these websites. These recommendations are usually obtained by selecting potential candidates from an actor's local neighborhood. Thus, a typical actor would obtain a relatively higher ratio of friends through the network-based search process than chance meetings with strangers. This agrees with the intuition that actors in social networking websites would form links with other actors known to them rather than complete strangers. Such networks tend to have higher clustering coefficients (transitivity). In Stage III of the FESC, it is clear that Twitter users create more groups for friends. The feature similarity index was used and features given in Table 5 were found to be influential.

Table 5. Descriptive information about social circles.

Sr. no.	Dataset	Friends per group	Groups created by user
1	Fb-Net	27	8
2	Twt-Net	26	13
3	Gp-Net	33	3

Table 6. Significant features in social circles.

Sr. no.	Dataset	Significant features	Less significant features
1	Fb-Net	Religion, political affiliation, work location, work description, educational institute, batch in educational institute, locale, gender	Birth date, hometown, work location, employer name, graduation degree
2	Twt-Net	Celebrity, music	Gender, literature
3	Gp-Net	Workplace, college, hobbies	Gender, hometown, school, work description

4.3. Social circle detection

The FESC algorithm is compared with the benchmark network embedding techniques of DeepWalk [37], LINE [27], and MF and community discovery techniques of Walktrap, Fastgreedy, and Infomap. Performance measures for evaluating the results are given below.

4.3.1. Performance measures

Modularity $Q \in -1, 1$ is a quality function for measuring the strength of division of a network into communities:

$$Q = \frac{1}{2m} \sum_{i,j}^{n} [A_{i,j} - \frac{k_i k_j}{2m}] \delta(C_i C_j), \tag{8}$$

where:

- A_{ij} is edge weight between nodes i and j;
- k_i and k_j are degrees of nodes i and j in the case of unweighted graphs;
- m is the sum of all edge weights in graphs;
- c_i, c_j are communities of nodes;
- δ is 1 if an edge exists between c_i, c_j and 0 otherwise.

Edit distance calculates the minimum number of edit operations needed for transformation of the predicted social circles into actual social circles. Converting the predicted social circles into actual social circles requires readjustment of the users in them. Each of the following operations for readjustment costs one edit:

- Add user to an existing circle.
- Remove user from a circle.

- Create a circle with one user.
- Delete a circle with one user.

4.3.2. Results and discussion on social circle detection

Community discovery techniques are classified into two families: modularity maximization and flow-based. Walktrap is a agglomerative, hierarchical clustering algorithm that allows finding community structures at different scales. Fastgreedy has a running time of O($mHlogn$) on a graph with m = $| E |$ edges, n = $| V |$ vertices, and h = height of the dendrogram. Both these techniques are modularity maximization methods and aim to find modules in the network that are densely connected regions. In contrast, for Infomap, the best partition is the one that yields minimum description for the random walk. Figure 2 gives these results. This flow-based method provides different partitions than the methods based on structural features of the network like modularity. The communities uncovered by these class of families are in stark variation, especially when the underlying network has directed links.

Network embedding techniques such as DeepWalk and MF preserve only first-order proximity, whereas LINE preserves both first- and second-order proximity of the nodes. However, these methods prove ineffective due to the large number of parameters $\sim O(V^2)$, nonsharing of parameters, and inability to handle attribute information of nodes. FESC integrates attribute and structural information in detection of social circles of an ego. The local structure of an ego-net as well as the attributes of the alters are factored for a unified framework of social circle detection. Additional information being factored in the framework gives it an advantage over traditional methods that consider only attribute information (clustering) or only structure (community discovery or node embedding).

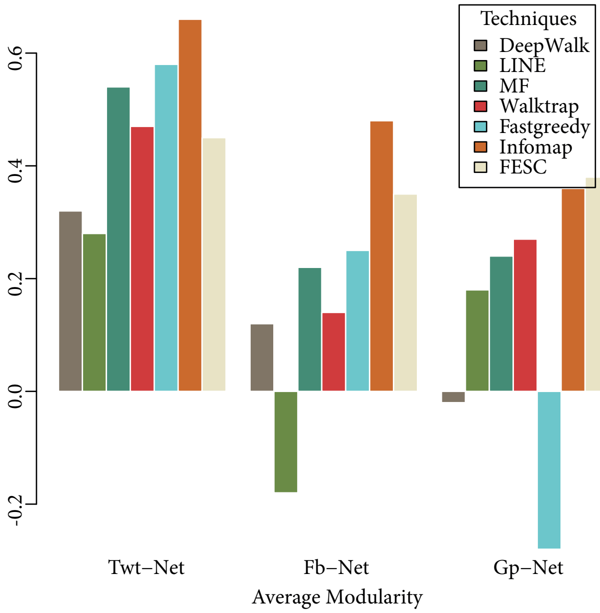

Figure 2. Results of average modularity metric in comparison of FESC with baseline frameworks. Modularity optimizing algorithms such as Walktrap and Fastgreedy identified community structure with high modularity. Infomap technique based on minimizing the expected description length of a random walkers trajectory outperformed other techniques. Network embedding techniques with embedding dimensions set to 5 resulted in low modularity partitions. Performance of FESC is comparative to community discovery techniques.

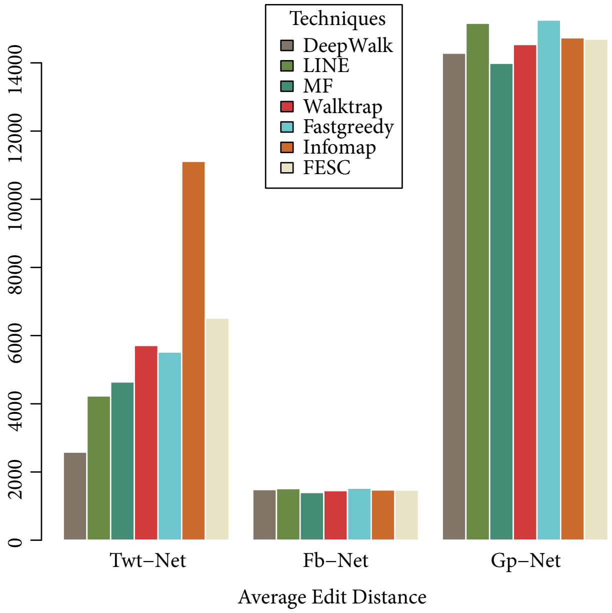

Figure 3. Results of average edit distance metric in comparison of FESC with baseline frameworks. Modules discovered by community detection frameworks do not correspond to social circle communities and hence the value of the edit distance metric was high for modularity maximizing techniques such as Walktrap and Fastgreedy. Network embedding techniques fared relatively better than community discovery techniques. FESC performed better than modularity maximization and random walk-based techniques of community discovery.

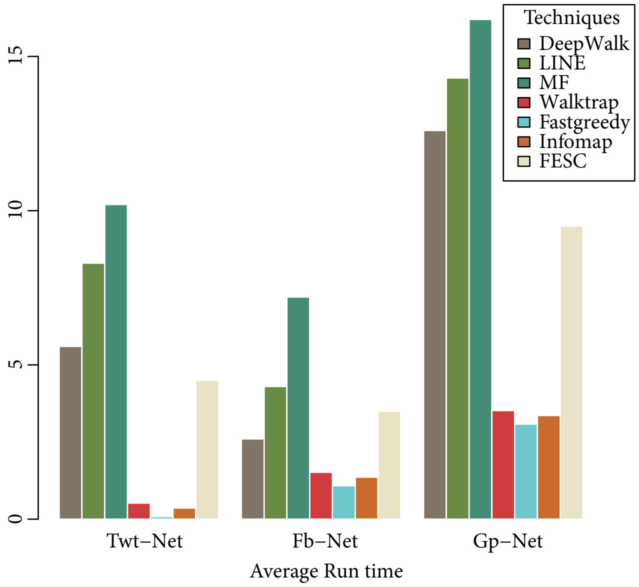

Figure 4. Results of average running time of FESC compared with baseline frameworks. Running time of network embedding techniques is high due to nonsharing of parameters. FESC and community discovery techniques had a time complexity of $\sim O(V+E)$, resulting in lower running time.

5. Conclusion

Groupings of friends are common facilities offered by OSNs. However, this facility was not well received by users. This is due to the requirement of physically creating social circles and allocating member connections to appropriate lists manually. The responsibility of keeping the social circles up-to-date lies with the users, making this task tedious. However, social circles have effective applications in multiple downstream social network analytic tasks such as link prediction, friend recommendation, node classification, and content recommendation. Hence, the current investigation focuses on improvement of the performance of social circle detection mechanisms by utilization of attribute information of alters as well as the structural characteristics of ego-nets (local network).

Existing techniques for exploratory data analysis such community discovery or clustering do not recognize the importance of homophily in creation of new connections by users. Also, other state-of-the-art frameworks either require knowledge of the global structure of the network or complete information of the entities in the social network. As these criteria are not fulfilled in real-world systems, there is a need for a framework that can perform the task but at the same time utilize local network information as well as handle incomplete data. The FESC framework achieves this by jointly modeling structure and attribute information of ego-nets to identify social circles. The extensive experiments as shown in Figure 3 and Figure 4 show this. These experiments performed on ego-nets of popular OSNs are used to highlight the efficacy of this approach. Performance measures reveal that the performance of FESC can match state-of-the-art community discovery and network embedding techniques.

Graph convolutional neural networks (GCNNs) are increasingly being used to generate node embeddings based on aggregating information from local neighborhoods. A modified FESC framework can be investigated in future work, which consists of a GCNN as a intermediate stage. Theoretically deep learning frameworks learn nonlinear relations in the data, which can be useful in social circle detection.

References

[1] Chandane M, Bhirud S, Nerurkar P, Shirke A. A novel heuristic for evolutionary clustering. Procedia Comput Sci 2018; 125: 780–789.

[2] Liu H, Lin Y, Sangaiah AK, Zhang S, Li X. A privacy-preserving friend recommendation scheme in online social networks. Sustain Cities Soc 2018; 38: 275-285.

[3] Li H, Jiang Q, Gao S, Ma X, Ma J. Armor: A trust-based privacy-preserving framework for decentralized friend recommendation in online social networks. Future Gener Comp Sy 2018; 79: 82–94.

[4] Akoglu L, Perozzi B. Discovering communities and anomalies in attributed graphs: Interactive visual exploration and summarization. ACM T Knowl Discov D 2018; 12: 24-40.

[5] Sood K, Cui L, Pham VV, Yu S. Privacy issues in social networks and analysis: a comprehensive survey. IET Netw 2017; 7: 74-84.

[6] Zhao G, Mei T, Qian X, Feng H. Personalized recommendation combining user interest and social circle. IEEE T Knowl Data En 2014; 26: 1763–1777.

[7] Kosinski M, Stillwell D, Mo F, Zhou J. Usage patterns and social circles on Facebook among elderly people with diverse personality traits. Educ Gerontol 2018; 44: 265–275.

[8] Leskovec J, Yang J, McAuley J. Community detection in networks with node attributes. In: 13th International Conference on Data Mining; 7–10 December 2013; Dallas, TX, USA. pp. 1151–1156.

[9] Leskovec J, Yang J. Overlapping community detection at scale: a nonnegative matrix factorization approach. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining; 4–8 February 2013; Rome, Italy. New York, NY, USA: ACM. pp. 587–596.

[10] Alaa AM, Van der Schaar M, Ahuja K. A micro-foundation of social capital in evolving social networks. IEEE T Netw Sci Engg 2018; 5: 14–31.

[11] Tang J, Dong Y, Chawla NV. User modeling on demographic attributes in big mobile social networks. ACM T Inform Syst 2017; 35: 35-67.

[12] Chandane M, Bhirud S, Nerurkar P, Shirke A. Empirical analysis of data clustering algorithms. Procedia Comput Sci 2018; 125: 770–779.

[13] Li X, Luo B, Huan J, Lan C, Yang Y. Learning social circles in ego-networks based on multi-view network structure. IEEE T Knowl Data Eng 2017; 29: 1681-1694.

[14] Zhang C, Zimmermann R, Hong R, Zhang L. Flickr circles: aesthetic tendency discovery by multi-view regularized topic modeling. IEEE T Multimedia 2016; 18: 1555–1567.

[15] Chen Z, Xu M, Mei T, Lu D, Sang J. Who are your "real" friends: analyzing and distinguishing between offline and online friendships from social multimedia data. IEEE T Multimedia 2017; 19: 1299–1313.

[16] Gosak M, Markovič R, Dolenšek J, Rupnik MS, Marhl M, Stožer A, Perc M. Network science of biological systems at different scales: a review. Phys Life Rev 2018; 24: 118–135.

[17] Gosak M, Markovič R, Dolenšek J, Rupnik MS, Marhl M, Stožer A, Perc M. Loosening the shackles of scientific disciplines with network science: reply to comments on network science of biological systems at different scales: a review. Phys Life Rev 2017; 24: 162-167.

[18] Jalili M, Perc M. Information cascades in complex networks. J Compl Netw 2017; 5: 665–693.

[19] Wang Z, Yamir M, Stefano B, Perc M. Vaccination and epidemics in networked populations—-an introduction. Chaos Soliton Fract 2017; 103: 177-183.

[20] Jalili M, Orouskhani Y, Asgari M, Alipourfard N, Perc M. Link prediction in multiplex online social networks. R Soc Open Sci 2017; 4: 1-11.

[21] Martinčić-Ipšić S, Močibob E, Perc M. Link prediction on Twitter. PLoS One 2017; 12: e0181079.

[22] Nickel M, Kiela D. Poincare embeddings for learning hierarchical representations. Adv Neur In 2017; 31: 6338–6347.

[23] Perc M, Jordan JJ, Rand DG, Wang Z, Boccaletti S, Szolnoki A. Statistical physics of human cooperation. Phys Rep 2017; 687: 1-51.

[24] Ribeiro HV, Alves LG, Martins AF, Lenzi EK, Perc M. The dynamical structure of political corruption networks. arXiv preprint. arXiv: 1801.01869.

[25] Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. arXiv preprint. arXiv: 1802.00543.

[26] Zitnik M, Leskovec J. Predicting multicellular function through multi-layer tissue networks. Bioinformatics 2017, 33: 190–198.

[27] Tang J, Qu M, Wang M, Zhang M, Yan J, Mei Q. Line: Large-scale information network embedding. In: Proceedings of the 24th International Conference on World Wide Web; 18–22 May 2015; Florence, Italy. New York, NY, USA: International World Wide Web Conferences Steering Committee. pp. 1067–1077.

[28] Huang X, Li J, Hu X. Label informed attributed network embedding. In: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining; 6–10 February 2017; Cambridge, UK. New York, NY, USA: ACM. pp. 731–739.

[29] Huang X, Li J, Hu X. Accelerated attributed network embedding. In: Proceedings of the 2017 SIAM International Conference on Data Mining; 16–22 May 2017; Notre Dame, IN, USA: ACM. pp. 633–641.

[30] Liao L, He X, Zhang H, Chua TS. Attributed social network embedding. arXiv preprint. arXiv:1705.04969.

[31] Bandyopadhyay S, Kara H, Kannan A, Murty MN. Fscnmf: Fusing structure and content via non-negative matrix factorization for embedding information networks. arXiv preprint. arXiv: 1804.05313.

[32] Tsitsulin A, Mottin D, Karras P, Muller E. Verse: Versatile graph embeddings from similarity measures. In: Proceedings of the 2018 World Wide Web Conference on World Wide Web; 23–27 April 2018; Lyon, France. Geneva, Switzerland: International World Wide Web Conferences Steering Committee. pp. 539–548.

[33] Ou M, Cui P, Pei J, Zhang Z, Zhu W. Asymmetric transitivity preserving graph embedding. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 13–17 August 2016; San Francisco, CA, USA. New York, NY, USA: ACM. pp. 1105–1114.

[34] Rozemberczki B, Davies R, Sarkar R, Sutton C. Gemsec: Graph embedding with self clustering. arXiv preprint. arXiv: 1802.03997.

[35] Rozemberczki B, Sarkar R. Fast sequence based embedding with diffusion graphs. In: International Conference on Complex Networks; 11–13 December 2018; France. Cambridge, UK: Springer. pp. 99-107.

[36] Yang Z, Cohen WW, Salakhutdinov R. Revisiting semi-supervised learning with graph embeddings. arXiv preprint. arXiv: 1603.08861.

[37] Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 24–27 August 2014; Washington, DC, USA. New York, NY, USA: ACM. pp. 701–710.

[38] Grover A, Leskovec J. node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 13–17 August 2016; San Francisco, CA, USA. New York, NY, USA: ACM. pp. 855-864.

[39] Sheikh N, Kefato Z, Montresor A. gat2vec: representation learning for attributed graphs. Computing 2018; 9: 1-23.

[40] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. Adv Neur In 2013; 23: 3111–3119.

[41] Cao S, Lu W, Xu Q. Grarep: Learning graph representations with global structural information. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management; 19–23 October 2015; Melbourne, Australia. New York, NY, USA: ACM. pp. 891–900.

[42] Liu Q, Li Z, Lui J, Cheng J. Powerwalk: Scalable personalized pagerank via random walks with vertex centric decomposition. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management; 24–26 October 2016; Indianapolis, IN, USA. New York, NY, USA: ACM. pp. 195–204.

[43] Pandhre S, Mittal H, Gupta M, Balasubramanian VN. Stwalk: learning trajectory representations in temporal graphs. In: Proceedings of the ACM India Joint International Conference on Data Science and Management of Data; 11–13 January 2018; Goa, India. New York, NY, USA: ACM. pp. 210–219.

[44] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint. arXiv: 1301.3781.

[45] Tran PV. Learning to make predictions on graphs with autoencoders. arXiv preprint. arXiv: 1802.08352.

[46] Wang Z, Ye X, Wang C, Wu Y, Wang C, Liang K. Rsdne: Exploring relaxed similarity and dissimilarity from completely-imbalanced labels for network embedding. Network 2018; 11: 475-482.

[47] Zhang M, Cui Z, Neumann M, Chen Y. An end-to-end deep learning architecture for graph classification. In: Proceedings of AAAI Conference on Artificial Inteligence; 2–7 February 2018; New Orleans, LA, USA: AAAI. pp. 531–538.

[48] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint. arXiv:1609.02907.

[49] Chen J, Ma T, Xiao C. Fastgcn: fast learning with graph convolutional networks via importance sampling. arXiv preprint. arXiv: 1801.10247.

[50] Donnat C, Zitnik M, Hallac D, Leskovec J. Spectral graph wavelets for structural role similarity in networks. arXiv preprint. arXiv: 1710.10321.

[51] Wu L, Fisch A, Chopra S, Adams K, Bordes A, Weston J. Starspace: Embed all the things! arXiv preprint. arXiv: 1709.03856.

[52] Perozzi B, Kulkarni V, Chen H, Skiena S. Don't walk, skip!: online learning of multi-scale network embeddings. In: Proceedings of the 2017 ACM International Conference on Advances in Social Networks Analysis and Mining; 1–3 August 2017; Sydney, Australia. New York, NY, USA: ACM. pp. 258–265.

[53] Desa C, Re C, Gu A, Sala F. Representation tradeoffs for hyperbolic embeddings. arXiv preprint. arXiv: 1804.03329.

[54] Goodreau SM. Advances in exponential random graph (p*) models applied to a large social network. Soc Networks 2007; 31: 231–248.

[55] Hoff PD, Raftery AE, Handcock MS. Latent space approaches to social network analysis. J Am Stat Assoc 2002; 64: 1090–1098.

[56] Snijders TAB. Longitudinal methods of network analysis. Enc Com Sys Sci 2009; 24: 5998–6013.

[57] Hoff PD. Dyadic data analysis with amen. arXiv preprint. arXiv: 1506.08237.

[58] Denny M. Social Network Analysis. Amherst, MA, USA: Academic Press, 2014.

[59] Denny M. Intermediate Social Network Theory. Amherst, MA, USA: Academic Press, 2015.

[60] Balasubramanian M, Schwartz EL. The isomap algorithm and topological stability. Science 2002; 295: 7.

[61] Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science 2000; 290: 2323–2326.

[62] Leskovec J, Yang J, McAuley J. Detecting cohesive and 2-mode communities indirected and undirected networks. In: Proceedings of the 7th ACM International Conference on Web Search and Data Mining; 24–28 February 2014; New York, NY, USA: ACM. pp. 323–332.

[63] Singh M, Murukannaiah P. Platys social: Relating shared places and private social circles. IEEE Internet Comput 2012; 16: 53–59.

[64] Robardet C, Boulicaut JF, Prado A, Plantevit M. Mining graph topological patterns: finding covariations among vertex descriptors. IEEE T Knowl Data En 2013; 25: 2090–2104.