International Journal of Modern Physics B
Vol. 33, No. 10 (2019) 1950089 (21 pages)
© World Scientific Publishing Company
DOI: 10.1142/S0217979219500899

Community detection in facebook activity networks and presenting a new multilayer label propagation algorithm for community detection

Fatemeh Alimadadi

Department of Computer Engineering and Information Technology,
Islamic Azad University Tehran North Branch, Tehran, Iran
f.alimadadi@gmail.com

Ehsan Khadangi*

Department of Computer Engineering, Shahed University, Tehran, Iran
School of Computer Science,
Institute for Research in Fundamental Sciences (IPM),
P. O. Box 19395-5746, Tehran, Iran
**khadangi@shahed.ac.ir; khadangi@gmail.com*

Alireza Bagheri

Department of Computer Engineering and Information Technology,
Amirkabir University of Technology, Tehran, Iran
Department of Computer Engineering and Information Technology,
Islamic Azad University Tehran North Branch, Tehran, Iran
ar_bagheri@aut.ac.ir

Received 1 November 2018
Revised 21 January 2019
Accepted 30 January 2019
Published 23 April 2019

The emergence of online social networks has revolutionized millions of web users' behavior so that their interactions with each other produce huge amounts of data on different activities. Community detection, herein, is one of the most important tasks. The very recent trend is to detect meaningful communities based on users' interactions or the activity network. However, in many of such studies, authors consider the basic network model while almost ignoring the model of the interactions in the multi-layer network. In this research, an experimental study is done to compare community detection in Facebook friendship network to that of activity network, considering different activities in Facebook OSN such as sharing. Then, a new community detection evaluation metric based on homophily is proposed. Eventually, a new method of community detection based on different activities in Facebook social network is presented. In this

*Corresponding author.

method, we generalized three familiar similarity methods, Jaccard, Common Neighbors and Adamic-Adar for multi-layered network model.

Keywords: Activity network; community detection; facebook; homophily; label propagation; multi-layer network; online social network; similarity metric.

PACS number: 64.60.aq

1. Introduction

Social networks have been studied fairly extensively over the last couple of decades in the general context of analyzing interactions between people in order to determine the important structural patterns in such interactions. The trends in recent years have been focused on online social networks enabled as an Internet application. Some examples of such networks are Facebook, LinkedIn and Myspace. These social networking services have been rapidly growing in popularity for they are no longer constrained by the geographical limitations of a conventional social network connecting people through face-to-face contact, or personal friendships.[1] The use of social networks like Facebook and Twitter during recent years shows that these networks are not only a means to spend time but also an evolution in human interactions. In other words, there is no doubt that online social networking websites have changed our ways of communication and caused us to spend plenty of time wandering over their web pages. Therefore, social network analysis might be of interest to different groups, including sociologists to analyze users' social behavior, computer engineers to design more efficient services and also security authorities to assess propagation channels of information and gossips,[2] decision makers in enterprises to use the underlying information in social interaction context to assist them for decision making in various contexts.[3]

One of the most important tasks when studying these networks is community detection. Community structure has a long history in social science, but it has become the focus especially in the fields of physics and computer science in recent years by Newman and Girvan's research.[4] In addition, by the emergence and massive success of social media, a new environment has been created for the community. The determination of these communities is useful in the context of a variety of applications in social network analysis, including customer segmentation, recommendations, link inference, vertex labeling, and influence analysis.[5]

1.1. *A Brief review on community detection*

There have been lots of researches regarding community detection such as Refs. 6–10 among which Fortunato has done an extensive research on community detection methods. However, they are mostly concerned with methodological foundations of community detection.[6] On the other hand, Papadopoulos *et al.* investigated community detection problem in the context of social media and conducted a comparative study on some popular community detection algorithms in terms of execution time, memory usage and attained community structure precision.[8]

Most of these algorithms model the social network as a graph or, in other words, the mono-dimensional network in which nodes represent users and edges between them represent friendship or following connections. However, in the real world, networks are heterogonous and multi-dimensional by nature and contain multiple connections between users. For instance, users in social networks share posts and participate in events besides making friendship relations. In other words, the interaction between users does not follow the declaration of friendship relationship in all cases. Therefore, detecting communities considering friendship networks not only is not enough, but also may not match the membership of users in reality. Chun and Moon named the network, in which nodes represent users and directed edges between them represent the activity from a user to another, "activity network".[11]

1.1.1. *Interaction based community detection*

As a result, discovering meaningful communities based on Individual's different interactions in a social network has attracted more attention in recent years which some of them described as below:

In Ref. 12, the author builds the idea of how to detect communities based on two observations:

- The interaction degree of pairs of users can be different,
- The interaction between mutual friends plays an important role in community detection problem.

They quantify these interactions as tie strength in two phases and then construct a probability graph in which the edge weight indicates the probability of two users being in the same community. At the end, they use hierarchical clustering to identify the communities.

In Ref. 13, the authors proposed an algorithm to detect local communities for the target user. In other words, the aim of this research is to find predefined communities by the use of the interaction between users whom they only consider texts named status in Facebook. Their method consists of two phases; at first, some initialized group is identified using the frequent item set approach and then based on similarity designed indices, small groups are merged until the number of groups reaches the predefined parameter.

The idea of the authors in Ref. 5 is to partition the edges based on the content and structure using matrix factorization.

In Ref. 14, a new community-detection algorithm based on label propagation algorithm (LPA) named SemPostLP is proposed which uses the edge weighting strategy and the semantics of the RDF description of social networks in order to reveal active and meaningful communities among users.

In Ref. 15, in order to find the same user across different OSNs , authors employ both user profiles and social network structure.

1.1.2. *Multilayer community detection*

Although taking users' interactions into account by merging all connections into one type and considering tie strength would be a more sophisticated analysis of the network, considering them as a single aggregated network may lead to a great extent of information loss regarding the heterogonous nature of the connections. Besides, as in the real-world problem, the number of communities is usually unknown, the local community detection such as the one proposed in Ref. 13 cannot help much.

As it was already stated, the simple model of connections, i.e., simple graphs are not adequate and not always enough to effectively model the user's interactions containing multiple preferences, multifaceted behaviors and complex interactions,[16,17] so it is necessary to outgrow the simple network and explore more realistic frameworks.

Recently, there have been great attempts to explore networks with multiple types of connections, whereas in past decades such systems were only investigated in fields such as sociology. However, the effort to develop frameworks to study multilayer complex systems and to expand familiar tools and concepts is a new phenomenon. These new structures were introduced under the topics of "multilayer network", "multiplex network", "multi-dimensional network" and "multigraphs". Although all these structures have some differences, they all emphasize the multiple nature of users' interactions.[18]

In spite of the fact that identifying communities in the simple network is mature enough, the development of community detection in the multilayer network is in its early stages as well, hence, there are few methods to solve the community detection problem in the multilayer network which can be followed in below researches.

In Ref. 19; community detection in these networks are classified into two main groups, the first class consists of the approaches which are based on the existing monoplex community detection algorithm; In these approaches the idea is to modify the problem to one of the community detection algorithms in monoplex network which could be classified into two methods: (1) Using aggregation to transform the multiplex network to monoplex and employing monoplex community detection algorithms,[21] in Ref. 20 also the authors take analytical strategy to reduce the dimension of the multiplex network by dimension reduction methods. (2) Applying the monoplex community detection algorithm on each layer and then combining all partitions by the use of ensemble clustering approaches.[17] The second class, on the other hand, contains approaches extending existing algorithms to deal directly with multilayer networks.[19,22,23] In Ref. 18, the authors extend the seed-centric approach named 'Licod' which was designed for the monoplex network into multiplex network and call it mux-Licod. The underlying idea of these approaches is to find the specific nodes called seeds around which the communities can be discovered. In Ref. 23, the approach is similar to that of the popular Louvain method[24] the aim of which is to maximize the Modularity. This quality function was generalized to the case of multiplex network called 'Multislice Modularity'. In Ref. 21, the known Infomap community detection algorithm[25] is extended. It is based on the compression of the network flows that can identify modular flows, both within and across layers in nonaggregated multilayer networks, which results in overlapping communities.

1.2. *Data set*

In general, social networks share common features such as the ability to interact with others, like friendship-like relationship, but it does not mean that they are alike and created for the same purpose. Social networks like Facebook encourage users to share personal information while Twitter encourages users to share trends and memes with the world. Accordingly, analyzing each of these networks individually is an important matter.

Facebook, one of the most popular OSN, announced 500 million users and now in 2017, the number of active users has reached above 2.01 billion on June 30th. This shows the increasing development of social network usage.[26] On the other hand, Alexa has announced Facebook as the Third website after Google and YouTube in terms of visits. Based on these facts, this OSN was selected for this research.

The dataset collection Methodology used in this research is described in Ref. 27. We sampled the collected dataset and used it in current research. The sampled data include the friendship network and profile information of the Facebook users. Some attributes of the collected profile information are gender, city and age. In addition, for users' friends, the information about the number of activities including likes, comments, shares and posts were collected every one month over a period of 3 years from January 1, 2011 to December 31, 2013. Then, the activity network for each type was created based on the information gathered about activities.

1.3. *High level analysis of the collected data*

In this section, we present high level analysis; measurement results include average degree, clustering coefficient, average path length, size and number of connected components, degree distribution and reciprocity on the collected data that are borrowed from Khadangi's and Bagheri's previous research.[27]

The high level analysis of the activity networks like, comment, share, post and mixed, which is a mixture of the previous four, is presented in Table 1.

Table 1. High level characteristics of activity networks.

Network	Clustering coefficient	Avg. path length	# of components	Size of giant component
Mixed	0.16	5.4	15	0.99
Like	0.15	5.4	23	0.98
Comment	0.09	7	107	0.93
Post	0.039	6.2	198	0.75
Share	0.04	5	181	0.6

Scale-free networks are networks whose degree distribution follows the power-low and a few examples of the real world network claimed to be among these networks are social-networks, protein–protein interaction networks and collaboration networks. Facebook various activity networks follow the power-low distribution as well.

According to social ethics, when person A connects to person B based on an activity, B is supposed to respond to A conversely and suitably. Having presented this introduction, reciprocity represents the ratio of reciprocal links in a directed network. This, in the activity network means how probable is that if a person interacts with another, the other will interact the first person via the same activity. Obviously, if A interacts with B, B will conversely interact with A with more probability than random. Different interactions in Facebook have different natures and this difference and the pattern of users' behavior may cause different reciprocity. Table 2 presents the percentage of reciprocal links in various interaction networks. According to Table 2, the activity network is not highly reciprocated. In the best reciprocal mode, the Like network has 33% reciprocal links. Other networks also have little reciprocity so that post and/ share networks have 6% and 11% reciprocal links, respectively.

According to the results presented above, friendship and activity network are fundamentally different. Facebook users interact only with a few of their friends through different activities, including like, comment, post and share. Therefore, the activity network is much sparser than the friendship network, whereas links in the activity network show stronger relationships than those in the friendship network. The most important similarity between these two, however, is their degree distribution.

Moreover, the structural properties of activity and friendship networks are fundamentally different. The activity network has lower density, lower clustering coefficient, lower average degree and higher average path length. These differences cause major differences in analysis and the result of applying different processes on them. Hence, as it is mentioned in Refs. 27 and 28, activity network models the real world network better than friendship network as it is representative of users' interactions.

The contribution of this research is two-fold. First, we compare the community detection algorithm between the friendship and activity networks based on

Table 2. Networks reciprocity of different facebook interactions.

Network	Reciprocal links in binary network
Mixed	33%
Like	26%
Comment	6%
Post	11%
Share	36%

some of the community detection evaluation metric such as modularity and a new community detection evaluation measure proposed in this research which is based on homophily phenomena. Second, we extend the well-known LPA[29] to the case of multilayer network. Meanwhile, we introduce some similarity measures for the multilayer network and then we evaluate our algorithm using the data described in this section.

The remainder of this paper is structured as follows. In Sec. 2, we compare community detection in friendship and activity network and propose the homophily-based community detection evaluation measure. In Sec. 3, the new multi-layered community detection algorithm is proposed then in Sec. 4, the evaluation and experimental result of the proposed algorithm is described. Finally, in Sec. 5, results and future studies are stated.

2. Empirical Study of Community Detection in Friendship and Activity Networks

As we stated earlier, the network which considers interactions between users, i.e., activity network, is a more accurate way of modelling relationships between users in online social networks; which leads us to the following question in the context of community detection:

Is it more accurate to use activity network instead of friendship network to detect communities in Facebook OSN?

To answer this question, we compute communities based on some known algorithms including edge-betweenness,[4] walk trap,[30] Fast-Greedy,[31,32] Leading Eigenvector,[33] Info Map,[25] Label Propagation,[29] and Multilevel.[24] Then, we compare the computed communities based on two approaches. First, we evaluate the obtained communities using existing community detection evaluation measures such as Modularity and then we propose a new metric based on homophily to compare these communities.

2.1. *Comparison based on the existing community detection evaluation measures*

In this section, we take an approach almost similar to the one in Ref. 18. Therefore, we compute the evaluation metric value of the obtained partition of the friendship network with respect to the activity network and then compare that one to the evaluation metric value of the obtained partition of the activity network. According to this approach, we evaluated this statement that *using the activity network instead of the friendship network would lead us to detect better communities in Facebook OSN*.

In general, there are two criteria when thinking about how good a partition is. The first is the number of edges between the members of the partition, and the second is the number of edges between the members of the partition and the remainder of the network.

Table 3. Community detection metrics.[36]

Metric name	Details
Modularity	Tests a given division of a network against the random division
Internal Density	Density is defined by the number of edges (m_s) in subset S divided by the total number of possible edges between all nodes ($n_s(n_s - 1)/2$). The "2" is there to cancel out duplicated edges
Expansion	This measure of separability gives the average of the number of external connections (C_s) per node (n_s) in subset S has with graph G. It can be thought of as "External Degree"
Cut Ratio	This metric is a measure of separability and can be considered as "External Density". It is the fraction of external edges (C_s) of subset S out of the total number of possible edges in graph G
Conductance	This measures the ratio of edges inside the cluster to the number of edges leaving the cluster (captures surface area to volume)
Normalized Cut	This represents how well subset S is separated from graph G. It sums up the fraction of external edges over all edges in subset S (conductance) with the fraction of external edges over all noncommunity edges

As a result, the objective functions or metrics will be grouped into two classes. The first group which is referred to multi-criterion scores combines both criteria (the number of edges inside and the number of edges crossing) into a single objective function while the second group of objective functions employs only a single of the two criteria (e.g., volume of the cluster or the number of edges cut).[34,35]

The selected metrics in this research detailed in Table 3 are from Ref. 36 which are among multi-criterion scores except for the modularity. Equations based on which these metrics were calculated are as follows, in "Eqs. (1)–(6)" where $G(V, E)$ is considered as an undirected graph with $n = |V|$ nodes and $m = |E|$ edges.

$$\text{Conductance}: \ f(S) = \frac{c_s}{2m_s + c_s}. \tag{1}$$

$$\text{Expansion}: \ f(S) = \frac{c_s}{n_s}. \tag{2}$$

$$\text{Internal density}: \ f(S) = 1 - \frac{m_s}{\frac{n_s(n_s-1)}{2}}. \tag{3}$$

$$\text{Cut ratio}: \ f(S) = \frac{C_s}{n_s(n - n_s)}. \tag{4}$$

$$\text{Normalized Cut}: \ f(S) = \frac{c_s}{2m_s + c_s} + \frac{c_s}{2(m - m_s) + c_s}. \tag{5}$$

$$\text{Modularity}: \ f(S) = \frac{1}{4m(m_s - E(m_s))}. \tag{6}$$

$$F = \frac{f(S_1) + f(S_2) + f(S_3) + \cdots + f(S_n)}{n}, \qquad (7)$$

where n_s is the number of nodes in partition S, $n_s = |S|$; m_s is the number of edges in S, i.e., $m_s = |\{(u,v) : u \in S, v \in S\}|$; c_s is the number of edges on the boundary of S; $c_s = |\{(u,v) : u \in S, v \notin S\}|$ and $E(m_s)$ is the expected number of edges between the nodes in set S in a random graph with the same node degree sequence.[34]

After calculating each evaluation metric for each community, we compute the evaluation metric of the whole network based on "Eq. (7)".

In this research, we construct different graphs, including those of like, comment, post, share, and mixed as well as the aggregated graph of four basic graphs based on three applications Trust, Closeness and Spent Time which is based on the user's spent time on different activities.[27]

To measure the importance of activities for different applications, we initially implemented a Facebook application which has asked several Facebook users questions about their friends and according to the replies, we measured trust and closeness between 506 couples of users as well.

We also extracted the number of activities such as like, comment, post and share between pairs using this application. Based on the collected data and using linear regression, the importance of different activities was obtained for three applications. We make use of attribute weighting approaches to improve the result of linear regression which as presented in Table 4 results in weights below for different activities.

We also created different types of graphs, i.e., directed-weighted, undirected-weighted, directed-binary, undirected-binary, directed-unweighted, and undirected-unweighted for each of the activity networks. Then, we executed appropriate algorithms on each. Next, we calculated the evaluation metric value of the obtained partition of that specific activity network and compared it to the evaluation metric value of the obtained partition of the friendship network with respect to the activity network.

Table 5 showing six comparison cases is an example of these comparisons based on famous modularity evaluation metric in which Modularity of communities obtained from the activity network is more than that of friendship network. We did these comparisons for each activity network type and evaluated different metrics for them as it was mentioned earlier. The total number of comparisons was 880. In most of them (616 (70%)) our assumption was proved.

Table 4. The importance of different activity networks in different applications.

	Like	Comment	Share	Post
Trust	9	7	23	11
Closeness	8	16	10	16
Spent Time	1	4	4	10

Table 5. Comparison between modularity of communities obtained from the activity network and friendship network.

Algorithm	Friendship network	Activity network (Spend Time-Weighted and Undirected)
Walktrap	0.06883801	0.6261926
Fastgreedy	0.06590997	0.4978876
Leading.eigenvector	0.08069273	0.5018335
Infomap	0.06249074	0.1426414
Label_Propagation	0.06503851	0.4902367
multilevel	0.102058	0.6834549

Table 6. Percentage of cases which proved the assumption per metric.

Metric	Percentage of cases which proved the assumption
Modularity	96%
Conductance	64%
Normalized Cut	58%
Cut Ratio	79%
Expansion	75%
Internal Density	76%

Table 7. Percentage of cases which proved the assumption Per Graph Type.

Graph type	Percentage of cases which proved the assumption
Unweighted and Undirected	86%
Weighted and Undirected	71%
Weighted and Directed	67%
Unweighted and Directed	91%
Binary and undirected	96%
Binary and directed	71%

As it is demonstrated in Table 6, it is also found that almost all of the evaluation metrics prove our assumption and modularity is the best one for proving it. The order of proving the assumption based on each metric is modularity, Cut ratio, internal density, Expansion, Conductance, and Normalized Cut.

As it is shown in Table 7, among weighted graphs, the undirected graph is a better proof of the assumption and among unweighted graph types, the directed graph is a better proof of the assumption.

2.2. *Comparison based on homophily*

One of the most accepted and similar to reality definition of community is based on the fact that users tend to group with whom they are more similar to than with others.[17,37–39] In other words, users tend to create relationships or interact with people who are similar to them in, for example, certain profile attributes, such as gender. This phenomenon is named homophily, assortative mixing or mixing patterns in literature.[40,41]

Due to the fact that communities and relations are usually formed based on homophily, we used it as a metric to evaluate the community structures i.e., network partitions obtained algorithmically. First, we sampled the friendship and activity networks in such a way that there is no missing data among the considered attributes. Then, the Walktrap and edge-betweenness algorithms were executed as we wanted to run the algorithms on directed and weighted graphs. After that, we calculated homophily based on the desired attributes (age, city, gender) for each community structure.

The main idea of quantifying assortativity mixing is to compute the variation of numbers of links which connect nodes of the same or similar type from what would be expected if the type (attributes) were placed randomly.[38,42] Assortative mixing can be calculated for enumerative and scalar characteristics presented as follows.

Homophily for enumerative characteristics such as race and gender is characterized by a quantity e_{ij} which is defined to be a fraction of the edges in a network that connects a vertex of type i to the one of type j. On an undirected network, this quantity is symmetric in its indices $e_{ij} = e_{ji}$, although on the directed network it may be asymmetric. To quantify the level of assortativity, assortativity coefficient is defined as below

$$r = \frac{\sum_i e_{ii} - \sum_i a_i b_i}{1 - \sum_i a_i b_i}, \tag{8}$$

where

$$a_i = \sum_j e_{ij}, \tag{9}$$

$$b_j = \sum_i e_{ij}, \tag{10}$$

a_i and b_i are fractions of each type of an end of the edge that is connected to vertices of type i. It can be noticed that on undirected graph $(a_i = b_i)$.

This formula gives $r = 0$, where there is no assortative mixing, $r = 1$ where it is completely assortative and when the network is disassortative, it gives negative assortativity coefficient.[43]

$$r_{\min} = -\frac{\sum_i a_i b_i}{1 - \sum_i a_i b_i}. \tag{11}$$

Assortative mixing for scalar characteristics like age and weight can be calculated as follows:

$$r = \frac{\sum_{xy} xy(e_{xy} - a_x b_x)}{\sigma_a \sigma_b}, \tag{12}$$

where

$$a_x = \sum_y e_{xy}, \quad b_y = \sum_x e_{xy}, \tag{13}$$

where e_{xy} is a fraction of all edges in the network, that join vertices together with values x and y for the age or other scalar variables of interest. The values x

and y could be either discrete in nature (e.g., integers, such as age) or continuous (the exact age). a_x and b_y are fractions of edges that start and end at vertices with values x and y, respectivly.[40]

As it was stated earlier, the purpose of homophily-based metrics is to answer this question that whether the majority of the connected nodes in the network are similar in terms of attributes based on which homophily was computed or not. Since having the same value for one type of attribute would not fit the definition, i.e., the functions would return NA values and our desired communities are the communities which have just one attribute type, we replaced these values with 1.

We calculated homophily based on three attributes including gender, city, and age. These attributes were selected since the missing values for these were below 45%. Then different activity networks including Trust, Closeness, Spent Time, Mixed, Comment, Post, Share, Like and friendship network were constructed, additionally for each network. Different kinds of graphs including directed/weighted, undirected/weighted, directed/unweighted, undirected/unweighted, directed/binary and undirected/binary were constructed and as a result, we had 48 cases for each attribute obtained from different activity networks. Then, Edge-betweenness and Walktrap algorithm were executed on each network. After generating different sub-graph based on community membership of each node which was obtained from mentioned algorithm, homophily of each community was calculated with the use of two stated formulas "Eq. (8)" or "Eq. (12)". After that, average homophily was calculated based on formula (7), where n is the number of communities calculated by the algorithm and $f(S_1)$ is the homophily of subgraph S_1 (community 1).

In order to compare the accuracy of community detection in activity network and friendship network in Facebook OSN, average homophily based on each attribute of each mentioned activity network was compared to its counterparts in the friendship network. As presented in Table 8, for instance, in 79% of the 48 related cases, the average homophily based on gender of different activity networks' community structures is more than that of their counterparts in the friendship network.

In order to take all the attributes into account when comparing the community structure of the friendship to activity network and to give different importance to each, each of them is weighted by the attribute weightening approach, including Information Gain, Information Gain Ratio, Gini Index and Rule regarding commu-

Table 8. Percentage which average of homophily of community structures of activity is more than that of friendship network.

Network model	Percentage which average of homophily of community structures of activity is more than that of friendshipnetwork		
Algorithm	**Edge-betweenness**	**Walk-trap**	**Total average of both**
Gender	89.5833	68.75	79.17
City	52.08333	54.1667	53.13
Age	60.4167	64.5833	62.5

nity membership of friendship and mixed activity networks as the output label, we obtain weights for each attribute.

We use Rapidminer attribute weightening operators including Information Gain operator, Weight by Information Gain Ratio operator, Gini Index operator and Weight by Rule operator as tools to calculate these weights. These operators calculate the weight of attributes with respect to the community membership, by using the information gain, information gain ratio, computing the Gini index and constructing a single rule for each attribute.

Information gain would measure how important an attribute is by calculating entropy as follows where A denotes an attribute, S denotes the data set sample and Values (A) is the collection of all the values of attribute A.

$$\mathrm{Gain}(S,A) \equiv \mathrm{Entropy}(S) - \sum_{v\in\mathrm{Values}(A)} \frac{|S_v|}{|S|}\mathrm{Entropy}(S_v). \tag{14}$$

Information gain suffers from a problem when applied to attributes that can take on a large number of distinct values which is solved by the information gain ratio by introducing Split Information.

$$\text{Split information}(S,A) = -\sum_{i=1}^{c} \frac{|S_i|}{|S|}\log_2\frac{|S_i|}{|S|}. \tag{15}$$

$$\text{Gain Ratio}(S,A) \equiv \frac{\mathrm{Gain}(S,A)}{\text{Split information}(S,A)}. \tag{16}$$

Gini Index is a measure of impurity of a dataset. It is a measure of how often a randomly chosen element from a set of elements would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset. The probability of a correct labeling can be computed by summing the probability of choosing each item multiplied by the probability of correctly labeling it. In this setting, the probability of correctly labeling an item is equal to the probability of choosing that item. Therefore, Gini Index can be computed as

$$\text{gini index}(S,A) = 1 - \sum_{c\,\epsilon\,\mathrm{classes}} P(C)^2. \tag{17}$$

By calculating weights of each attributes using these four weightening approaches, we reach to this sequence of importance: city is the most important feature, then age and gender respectively.

Now, we can answer our question *Whether using activity network instead of friendship network help us to detect communities more accurately in Facebook OSN or not*? As it is explained above, we compare these two networks based on our new homophily-based and some known evaluation metrics and present the result in Tables 6–8 According to these results, the activity network will lead us to detect more accurate communities.

3. Multilayer Community Detection Algorithm

In this section, we propose a new community detection algorithm for multilayer networks. As it was stated before, community detection in the multilayer network is classified into two classes: approaches which are based on the existing monoplex community detection algorithms and approaches that extend the existing algorithms to deal directly with the multilayer network. Our proposed algorithm fits in the second category. In other words, it is an extension of the LPA. Figure 5 shows a simple view of the multilayer network with four layers.

There are lots of community detection algorithms that identify community structures. Many of them need prior information about the number and size of communities, what is not often predictable beforehand and is not applicable to online social networks most of the time. Hence, algorithms which are able to detect communities without the need for prior knowledge are essential. LPAs are among those using the network structure alone as their guide and do not need either prior information about communities or predefined objective functions.[32] The key idea behind LPA is to propagate the labels of nodes all over the network. The idea of flowing through the network is from Ref. 45 research and the idea of label flooding is that a single label can quickly become dominant in a community as it is difficult to cross to other regions. So they are expected to be trapped inside a densely connected group.

Since this algorithm was introduced in Ref. 29, it has been extended in various manners. In Refs. 44 and 45 authors extend this algorithm to detect overlapping community in different ways. In Ref. 46, the author extends this algorithm to detect communities in dynamic networks. In Ref. 47, the authors introduce new updating rule in LPA and in Ref. 48, the shortcomings of the primary LPA algorithm such as randomness, weak robustness is resolved by introducing CK-LPA based on the fact that the LPA is simple, effective, and a nearly linear time method, this algorithm was selected to generalize for the multilayer network.

In LPA, each node holds a label and is updated iteratively based on the majority label in its neighborhood and at the convergence level, disjoint communities are identified. In order to define this algorithm in the multilayer network, we should define the neighborhood concept in the next section.

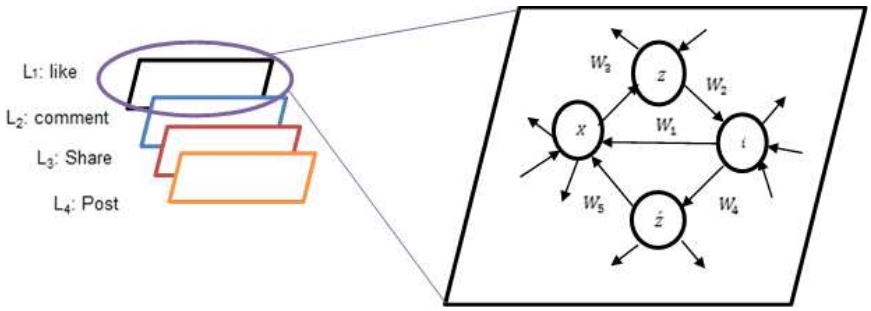

Fig. 1. (Color online) A simple view of the multilayer network with four layers.

3.1. *Neighborhood in multilayer networks*

There are a number of ways to generalize the notation of the neighborhood in the multilayer network. The simplest way to define the neighborhood in the multilayer network is an aggregate of its neighbor's sets in each layer.

- aggregate function: intersection function: i is the neighbor of j if i is connected to j in each layer[49];
- aggregate function: set union function: i is the neighbor of j if i is connected to j in at least one layer[49];
- trade-off between these two: i is the neighbor of j if i is connected to j in at least m layers. $(1 < m < a)$[50];
- trade-off based on node similarity: select all the neighbors of a node across all layers, which are most similar to the considered node.

$$\Gamma_{\delta\text{multiplex}(i)} = \{c \in \Gamma(i)^{\text{total}} : \text{sim}(c, i) \geq \sigma\}, \tag{18}$$

where $\Gamma(i)^{\text{total}}$ is the set of neighbors of node i across all layers and $\sigma \in [0, 1]$ is a similarity threshold, different similarity function can be used and an example could be the Jaccard similarity measure defined as follows[19]:

$$\text{sim}_{\text{jaccard}}(i, j) = \frac{\Gamma(i)^{\text{total}} \cap \Gamma(j)^{\text{total}}}{\Gamma(i)^{\text{total}} \cup \Gamma(j)^{\text{total}}}. \tag{19}$$

3.1.1. *Similarity metrics for multilayer network*

Taking into account the fourth way of finding nodes neighbors, and considering the fact that different levels of layers may have different levels of importance, we propose a new generalizations of the known similarity metrics Jaccard, Common Neighbors, and Adamic-Adar for multilayer networks. These measures are presented in Table 9. In this table, L_j denotes the weight of layer j. These weights represent the importance of different activities, for three applications.

$$N = \tau^j_{\text{out\_}i} \cap \tau^j_{\text{in\_}x}, \tag{20}$$

$$\acute{N} = \tau^j_{\text{in\_}i} \cap \tau^j_{\text{out\_}x}, \tag{21}$$

Table 9. Similarity measures for the multilayer network.

Metric	Definition
Jaccard (x,i)	$\frac{\sum_{j=1}^{k}\left(\frac{1}{2L_j}\left(\frac{\lvert N\rvert}{\lvert\tau^j_{\text{out\_}i}\cup\tau^j_{\text{in\_}x}\rvert}+\frac{\lvert\acute{N}\rvert}{\lvert\tau^j_{\text{in\_}i}\cup\tau^j_{\text{out\_}x}\rvert}\right)\right)}{\sum_{j=1}^{k} L_j}$
Common Neighbor (x,i)	$\frac{\sum_{j=1}^{k}\left(\frac{1}{2L_j}(\lvert N\rvert+\lvert\acute{N}\rvert)\right)}{\sum_{j=1}^{k} L_j}$
Adamic-Adar (x, i)	$\frac{\sum_{j=1}^{k}\left(\frac{1}{2L_j}\left(\sum_{Z\in N}\frac{1}{\log\lvert\tau^j_{Z}\rvert}+\sum_{\acute{Z}\in\acute{N}}\frac{1}{\log\lvert\tau^j_{\acute{Z}}\rvert}\right)\right)}{\sum_{j=1}^{k} L_j}$

where $\tau^{j}_{\text{out\_}i}$ and $\tau^{j}_{\text{in\_}x}$ denote out-neighbors and in-neighbors of node i in layer j, respectively, and as it is shown in Fig. 3, Ź and Z are common neighbors of node i and x where Z is X's out-neighbor and i's in-neighbor and Ź is X's in-neighbor and i's out-neighbor.

In this work, by similarity we mean the defined Jaccard similarity metric as it was the most popular metric used in previous studies.

3.1.2. *MNLPA: Label propagation algorithm for the multilayer network*

As we propose this algorithm for the Facebook activity network which is a weighted and directed network, we define the algorithm based on such network, but it is so straightforward to do it for the undirected network as well. In this case, one should do some small changes in similarity functions and consider each neighbors in general as opposed to in-neighbors in the algorithm.

Each node in this algorithm sends out its information (label) along outgoing links to its neighbors and receives information (label) along the incoming edges from neighbors, similar to the LabelRankT algorithm.[45] In other words, in order to identify each node's label, we identify its in-neighbors whom it receive information from. Neighbors in this algorithm are defined based on node similarity which was introduced in the previous section.

In a nutshell, MNLPA algorithm consists of following stages (see Fig. 2 for the pseudo-code):

(1) Each node is initiated with its own node id (i.e., unique label).

Algorithm: MNLPA: label propagation for Multilayer network algorithm

```
$C_x(0) = x$
$T = 1$
**For** $i = 1 : n$ **do**
    **For** $l = 1 : L$ **do**
        $Neighbors(i_l) = Node(i_l).getin\_neighbors()$
        $Y = union\_set\_of\_all\_neighbors\_in\_all\_layers(Neighbors(i_l))$
    **For** $j = 1 : m$ **do**
        $if\ Similarity(i, j) > \sigma\ then\ Nodes(j).addNeighbors(i)\ add\ j\ to\ i\text{'s}\ neighbors$
$while\ (T < \varphi\ and\ labelchanging = TRUE)$
    $Y = Randomized(Nodes)$
    **For** $X = 1 : n$ **do**
        $C_x(t) = f(V(C_{xi1}(t), \ldots, V(C_{xim}(t)V(C_{xi(m+1)}(t-1), \ldots, V(C_{xik}(t-1)))$
$Post\ Processing:\ CreateCommunities(Nodes)$
```

Fig. 2. Proposed MNLPA algorithm.

(2) In-neighbors of each node is calculated based on the following steps:
   (2.1) Union set of all in-neighbors of the considered node (i) is calculated (vector Y);
   (2.2) The similarity between the considered node (i) and each node in Y is calculated by one of the similarity functions proposed in the previous section;
   (2.3) The nodes which have similarity more than the defined threshold will be considered as node i's in-neighbors.

(3) The following steps are repeated until the stop criterion is satisfied:
   (3.1) Nodes are ordered randomly:
   (3.2) For each node (i), following steps are repeated:
      (3.2.1) Each similar in-neighbor of considered node (i) sends out its label to (i);
      (3.2.2) Node i receives the labels which have the max value, that is, the label which is popular and also more valuable.

(4) Finally, the post processing is done, that is, to create communities from the labels.

In this algorithm, f returns the label with the highest value among neighbors and ties are broken uniformly randomly and V calculates the value of each neighbor's label as follows where z denotes the neighbors' id of node x.

$$V(C_{xiz}(t)) = \sum_{j=1}^{L}(\text{edge weight}(z \to x) * (l_j)) \quad z \in (1, k). \tag{22}$$

Updating the label procedure in this algorithm is synchronous. In the label propagation process, at an instance, there exist some nodes in the network that have undergone iteration and the remaining nodes may not have undergone the iteration. In synchronous updating, a node at tth iteration updates its label based on its neighbors at iteration t as well as $t-1$, where $xi1 \ldots xim$ neighbors of x already has been updated in the current iteration while $xi(m+1) \ldots xik$ are the ones which have not been updated yet. The stop criterion in this algorithm is defined when the labels are not changed in the consecutive iteration, or simply, stops when the predefined maximum number of iterations φ is reached.

4. Experimental Results

In this section, we compare the proposed algorithm with layer aggregation approaches. In other words, we compare Multilayer LPA to the simple LPA. That is, comparison of the first class of multilayer community detection approach to the second one.

The direction is considered in the simple LPA, as opposed to the initial LPA which we run on the aggregate network, that is, comparison of the first class of multilayer community detection approach to the second one.

To the best of our knowledge, there do not exist any multiplex network with ground truth communities.[19] Thus, we cannot use supervised metrics like F_score, NMI (Normalized Mutual Information) and Omega Index to evaluate the community structures. As a result, we use our unsupervised metrics presented in Sec. 3.2. The dataset used in this section is described in Sec. 2. The proposed algorithm was executed on the multilayer activity network which contains post, like, comment, and share layers, each layer of which was weighed based on the application mentioned in Table 4, and the algorithm was run for the three applications trust, closeness and Spent Time. Then, for evaluation, the proposed homophily evaluation method was used for each community structure of each multilayer network; in other words, homophily for each community based on age, city and gender was calculated.

In order to consider the direction in the activity network, the initial LPA was extended and then executed on the three aggregated networks for each application, that is, trust, closeness and weighted mixed, then similar to evaluating the proposed algorithm, homophily was measured based on age, city and gender for each.

As it is shown in Tables 10–12, the MNLPA algorithm's resultant communities are more accurate than those of the initial algorithm which was run on the aggregate network. In other words, as the evaluation measure is based on homophily, we can strongly state that the resultant communities are similar nodes who get together and it truly fits the definition of the community as well.

Table 10. Evaluation based on homophily by age.

Network type	Evaluation score-based one homophily by age	
Network model	**Aggregated Network**	**Multi-layered Network**
Closeness	−0.03854627	0.1103173
Trust	−0.05579456	0.1221395
Spent Time	−0.1011532	0.1599364

Table 11. Evaluation based on homophily by city.

Network type	Evaluation score-based one homophily by age	
Network model	**Aggregated network**	**Multi-layered network**
Closeness	0.09840076	0.19686
Trust	−0.05579456	0.219795
Spent Time	0.08895607	0.1918399

Table 12. Evaluation based on homophily by gender.

Network type	Evaluation score-based one homophily by age	
Network model	**Aggregated network**	**Multi-layered network**
Closeness	0.8807012	0.8947729
Trust	0.8877749	0.9015552
Spent Time	0.891177	0.9210245

5. Conclusion and Future Works

Based on the comparison done in this paper between the results of community detection from the friendship network and activity network in Facebook OSN using existing community detection evaluation measures such as modularity and our proposed evaluation homophily-based measure, which is based on the definition of the community, the Activity network leads us to detect more accurate communities in Facebook OSN.

In this research, we proposed a new multi-layered community detection algorithm, which is a generalization of LPA for the multilayer network which is a well-fitted system for the activity network. The examination of our new multilayer community detection algorithm on real datasets using our proposed evaluation measures shows that our approach yields better results than the classical approach 'layer_aggregation methods'. Besides, in order to calculate the neighborhood for MNLPA, we generalized the known similarity measures, Jaccard, Adamic-Adar, and Common-neighbors for multilayer networks as well.

For future work, it is quite reasonable to do a comparison between similarity measures which were introduced in this paper as well as a comparison between the results of our new algorithm using these three similarity measures. In addition, other experiments should be done to include other multiplex algorithms. Other datasets may also be gathered in order to be able to do a thorough comparison between different approaches especially from different OSNs since as it is mentioned in Ref. 51 the platform may have impact the user activity. Besides heterogeneity, considering other main characteristics of online social networks, that is, their dynamicity is of great importance.

Acknowledgments

This research was in part supported by a grant from Institute for Research in Fundamental Sciences (IPM) (No. CS1397-4-75).

References

1. E. Khadangi and A. Bagheri, *Comput. Human Behavior* **73**, 64 (2017).
2. E. Khadangi and A. Bagheri, Comparing MLP, SVM and KNN for predicting trust between users in Facebook, in *Computer and Knowledge Engineering* (*ICCKE*) (Ferdowsi University of Mashhad, Iran, 2013), pp. 466–470.
3. J. Leng and P. Jiang, *IEEE Trans. Syst. Man Cybern. Syst.* **47**, 276 (2017).
4. M. E. Newman and M. Girvan, *Phys. Rev. E, Stat. Nonlinear Soft Matter Phys.* **69**, 026113 (2004).
5. G.-J. Qi, C. C. Aggarwal and T. Huang, Community detection with edge content in social media networks, in *Data Engineering (ICDE)* (IEEE, 2012), pp. 534–545.
6. S. Fortunato, *Phys. Rep.* **486**, 75 (2010).
7. B. Yang, D. Liu and J. Liu, Discovering communities from social networks: methodologies, in *Handbook of Social Network Technologies and Applications*, ed. B. Furht (Springer US, 2010), pp. 331–346.

8. S. Papadopoulos *et al.*, *Data Min. Knowl. Discov.* **24**, 515 (2012).
9. M. Coscia, F. Giannotti and D. Pedreschi, *Stat. Anal. Data Min.* **4**, 512 (2011).
10. M. Planti, Survey on social community detection, in *Social Media Retrieval Computer Communications and Networks*, eds. N. Ramzan, R. Van Zwol and J.-S. Lee, *et al.* (Springer London, 2013), pp. 65–85.
11. H. Chun and S. Moon, Comparison of online social relations in terms of volume vs. interaction: A case study of cyworld, in *IMC '08 Proc. the 8th ACM SIGCOMM Conf. Internet Measurement* (ACM, 2008), pp. 57–70.
12. H. Dev, M. E. Ali and T. Hashem, User interaction based community detection in online social networks, in *Database Systems for Advanced Applications*, eds. S. S. Bhowmick, E. C. Dyreson and S. C. Jensen, *et al.* (Springer International Publishing, 2014), pp. 296–310.
13. Y. Chen, C. Chuang and Y. Chiu, *J. Assoc. Inf. Sci. Technol.* **65**, 539 (2014).
14. S. Kianian, M. R. Khayyambashi and N. Movahhedinia, *J. Inf. Sci.* **13**, 1 (2015).
15. Jiangtao Ma *et al.*, *IEEE Access* **5**, 12031 (2017).
16. Santa Agreste *et al.*, *IEEE Trans. Syst. Man Cybern. Syst.* **45**, 559 (2015).
17. M. Berlingerio, F. Pinelli and F. Calabrese, *Data Min. Knowl. Discov.* **27**, 294 (2013).
18. M. Kivelä *et al.*, *J. Complex Netw.* **2**, 203 (2014).
19. M. Hmimida and R. Kanawati, *Netw. Heterogeneous Med.* **10**, 71 (2015).
20. András Vörös, A. Tom and B. Snijders, *Soc. Netw.* **49**, 93 (2017).
21. M. Berlingerio, M. Coscia and F. Giannotti, Finding and characterizing communities in multidimensional networks, in *Int. Conf. Advances in Social Networks Analysis and Mining* (IEEE, 2011), pp. 490–494.
22. M. De Domenico *et al.*, *Phys. Rev. X* **5**, 011027 (2015).
23. P. J. Mucha *et al.*, *Science* **328**, 876 (2010).
24. V. D. Blondel *et al.*, *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).
25. M. Rosvall, D. Axelsson and C. T. Bergstrom, *Eur. Phys. J. Spec. Top.* **178**, 13 (2009).
26. E. Khadangi and A. Bagheri, Empirical Analysis of Structural Properties, Macroscopic and Microscopic Evolution of Various Facebook Activity Networks. under review.
27. E. Khadangi, A. Bagheri and A. Zarean, *Quality & Quantity*, **52**, 249 (2018).
28. B. Viswanath *et al.*, On the evolution of user interaction in facebook, in *Proc. the 2nd ACM Workshop on Online Social Networks* (ACM, New York, USA, 2009), pp. 37–42.
29. U. N. Raghavan, R. Albert and S. Kumara, *Phys. Rev. E* **76**, 036106 (2007).
30. P. Pons and M. Latapy, *Comput. Inf. Sci. — ISCIS* **3733**, 284 (2005).
31. K. Wakita and T. Tsurumi, Finding community structure in mega-scale social networks, in *Proc. The 16th Int. Conf. World Wide Web* (ACM, 2007), pp. 1275–1276.
32. A. Clauset, M. E. J. Newman and C. Moore, *Phys. Rev. E* **70**, 066111 (2004).
33. M. E. J. Newman, *Phys. Rev. E* **74**, 036104 (2006).
34. J. Leskovec, K. J. Lang and M. Mahoney, Empirical comparison of algorithms for network community detection, in *Empirical Comparison of Algorithms for Network Community Detection* (ACM New York, New York, USA, 2010), pp. 631–640.
35. J. Yang and J. Leskovec, *Knowl. Inf. Syst.* **42**, 181 (2015).
36. Lab41, Market Survey: Community Detection, http://lab41.github.io/survey-community-detection.
37. K. Orman, Günce, Contribution to the interpretation of evolving communities in complex networks: Application to the study of social interactions, 2014.
38. A. Mislove *et al.*, You are who you know: Inferring user profiles in online social networks, in *Proc. The Third ACM Int. Conf. on Web Search and Data Mining — WSDM '10* (ACM, 2010), pp. 251–260.

39. P. Dev, *J. Public Econ. Theory* **18**, 268 (2016).
40. M. E. J. Newman and M. Girvan, Mixing patterns and community structure, in *Statistical Mechanics of Complex Networks*, eds. R. Pastor-Satorras, M. Rubi and A. Diaz-Guilera (Springer Berlin Heidelberg, 2003), pp. 66–87.
41. M. Mcpherson, L. Smith-lovin and J. M. Cook, *Ann. Rev. Soc.* **27**, 415 (2001).
42. K. Pelechrinis, *Soc. Netw. Anal. Min.* **4**, 188 (2014).
43. M. E. J. Newman, *Phys. Rev. E* **67**, 026126 (2003).
44. S. Gregory, *New J. Phys.* **12**, 103018 (2010).
45. J. Xie, B. K. Szymanski and X. Liu, SLPA: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process, in *2011 IEEE 11th Int. Conf., Data Structures and Algorithms; Physics and Society, Data Mining Workshops* (*ICDMW*) (Washington, DC, USA, 2011), pp. 344–349.
46. J. Xie, M. Chen and B. K. Szymanski, LabelRankT: Incremental community detection in dynamic networks via label propagation, in *Proc. the Workshop on Dynamic Networks Management and Mining* (ACM, New York, USA, 2013), pp. 25–32.
47. H. Lou, S. Li and Y. Zhao, *Phys. A, Stat. Mech. Appl.* **392**, 3095 (2013).
48. Z. Lin *et al.*, *Phys. A, Stat. Mech. Appl.* **416**, 386 (2014).
49. P. Bródka and P. Kazienko, Multilayered Social Networks, in *Encyclopedia of Social Network Analysis and Mining*, eds. R. Alhajj and J. Rokne (Springer, New York, 2014).
50. P. Kazienko, P. Brodka and K. Musial, Individual neighbourhood exploration in complex multi-layered social network, in *2010 IEEE/WIC/ACM Int. Conf. on Web Intelligence and Intelligent Agent Technology* (IEEE, 2010), pp. 5–8.
51. P. De Meo *et al.*, *ACM Trans. Intell. Syst. Technol.* **5**, 14:1 (2013).