Computers in Human Behavior 73 (2017) 64–79

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Presenting novel application-based centrality measures for finding important users based on their activities and social behavior

Ehsan Khadangi, Alireza Bagheri*

Computer Engineering and Information Technology Department, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Hafez Ave, 424, Iran

ARTICLE INFO

Article history:
Received 8 September 2016
Received in revised form
28 November 2016
Accepted 4 March 2017

Keywords:
Social network analysis
Social behavior
Centrality measure
Social media marketing
Activity network

ABSTRACT

There are more important relationships based on users' behavior and the done activities than those of friendship in online social networks. Study of social behavior of users in these networks has many applications. Analyzing online social networks' activity graphs, as a better representation of users' social behavior, may open new perspectives for real applications such as finding important users. Although detecting these influential nodes based on their friendship relationships is studied a lot, finding important nodes using users' behavior and activates has not attracted much attention. In this work, we study users' importance in various Facebook activity networks including like, comment, post, share, and mixed, then compare gained rankings with those of the friendship network and conclude that users influence analysis in activity networks represents very different results. Afterwards, we propose new centrality measures that can present different rankings suitable for different applications, further to have the potential for simultaneous consideration of various activities in a multilayer network. Experimental results highlights the benefits of using the presented methods. To the best of our knowledge, our methods are the first and only proposed centrality measures that can present different rankings for various applications based on users' social behavior.

1. Introduction

Online social networks develop different mechanisms for users' activity and interaction. Facebook as the most famous and popular social networking website is not an exception to this fact. Facebook activities could be classified into the two categories latent and active. Among Facebook active activities the users' behavior on which is visible to other users are posting new contents, taking comment or like on a post or comment, sharing content, tagging photos, joining groups, and using Facebook applications. In addition, Facebook supports latent activities such as chatting, sending instant messages, browsing profiles, and online games. Since different users use some Facebook activities more according to their personality, psychological, and sociological properties, their behavior pattern is not the same as each other. Therefore, it is necessary to present new network analysis methods so as to consider several activities simultaneously besides the friendship network in order to analyze the users' behavior.

Although many works have been presented on user ranking in social networks, ranking Facebook users based on their different activities and interactions requires more examination. Particularly, the vacancy of a method which can measure the influence of nodes for different applications further to the simultaneous consideration of users' activities and interactions is perceived. The main goal of this study is to fill this vacancy. For this purpose, we initially support this idea that different people's importance is different from different viewpoints. Then based on the simultaneous analysis of Facebook users' different behaviors and activities in a multilayer model, an application-based method has been presented which measures users' importance based on the considered application. A literature review of detecting influential nodes and activity network analysis is presented as follows.

1.1. Empirical research

Among the society's people, there are always some who have great power in affecting and guiding different people's thoughts, interests, and beliefs due to personal, scientific, and psychological properties and their social position. Based on the 80/20 rule, also most opinions in sociology, these people are few (Cha, Haddadi,

* Corresponding author.
E-mail addresses: khadangi@gmail.com (E. Khadangi), ar_bagheri@aut.ac.ir (A. Bagheri).

http://dx.doi.org/10.1016/j.chb.2017.03.014
0747-5632/

Benevenuto, & Gummadi, 2010) and are named innovators (Rogers, 2010), salesmen, connectors, hubs, mavens (Point, 2002), or etc. Finding these people is used for influencing other people (Xu, Li, & Song, 2012), marketing and advertising (Li, Lee, & Lien, 2012; Xu et al., 2012), and etc. The theory of "diffusion of innovation" which was presented by Rogers states that only 2.5 percent of the society's people are much brave, risky, and interested in trying new products and thoughts. These people whom Rogers called "innovators" are potential initial adapters for the social network marketing process (Rogers, 2010). By convincing such people, a product, idea, or thought could be spread in social networks extraordinarily.

One of the most important methods of measuring impact in social networks is using centrality measures. Although these measures are applicable in different fields such as gaining the most important roads in a road network (Tsiotas & Polyzos, 2013), finding the important papers in a citation network (Cheang, Chu, Li, & Lim, 2014), and discovering author impact in coauthorship networks (Yan & Ding, 2011; Ding, Yan, Frazho, & Caverlee, 2009; Ding, 2011), one of their most important applications is in the measuring the impact and importance of social networks' users for different applications (Kang, Molinaro, Kraus, Shavitt, & Subrahmanian, 2012; Mochalova & Nanopoulos, 2013; Kiss & Bichler, 2008; Kermani, Badiee, Aliahmadi, Ghazanfari, & Kalantari, 2016).

Node centrality measures calculate the importance of users based on their position in the network. Some works suggest new node centrality measures which have better results (Takes & Kosters, 2011; Chen, Lü, Shang, Zhang, & Zhou, 2012; Alahakoon, Tripathi, Kourtellis, Simha, & Iamnitchi, 2011; Campiteli, Holanda, Soares, Soles, & Kinouchi, 2013); some have focused on presentation of algorithms for increasing the speed of calculating centrality measures especially betweenness and PageRank (Bader, Kintali, Madduri, & Mihail, 2007; McSherry, 2005); some others dealt with finding important nodes for different applications such as targeted social media marketing and advertising (Li et al., 2012; Xu et al., 2012); and some others are involved with the empirical study of the presented measures in different networks particularly online social networks (Cheang et al., 2014; Heidemann, Klier, & Probst, 2010; Hu, Wang, & Lee, 2010; Valente, Coronges, Lakon, & Costenbader, 2008). Furthermore, different generalizations of classic centrality measures have also been presented for weighted networks, including different versions of degree, closeness, betweenness, eigenvector, and PageRank (Opsahl, Agneessens, & Skvoretz, 2010; Yan, Zhai, & Fan, 2013; Xing & Ghorbani, 2004).

The concept of activity network as the network which models the actual interactions between users was first proposed by Chun et al. (Chun et al., 2008). After that, few efforts were made for identifying important users by applying centrality measures to the activity network. For example, Heidemann et al. identified important Facebook users by applying weighted PageRank to Facebook wall-post activity network (Heidemann et al., 2010). Corbellini et al. used centrality measures for suggesting the software engineering group leader. They initially presented an application named Paynal for the software developers' coordination. The software used social network analysis methods in order to analyze interactions among users and achieve high-level knowledge on the development team's members. For example, they used centrality measures in order to reach a node which can manage the team (Corbellini, Schiaffino, & Godoy, 2012).

Although a few works have been done in the field of measuring influence based on the users' activities, to the best of our knowledge, most of them have dealt with only one activity and measure the influence based on that particular activity. In addition, almost in all the works in the field of analyzing users' influence, influence is defined as a non-flexible and rigid idea whereas a person may be important by one application, and unimportant by another one. In this paper, a method has been presented which makes it possible to measure application-based importance of Facebook users by analyzing their activities and behavior simultaneously.

The rest of the paper is organized as follows. In Section 2, we present a literature survey of previous works. Section 3 describes material and methods. In Section 4, we study centrality measures in friendship and various activity networks. Then, we propose an application-based multilayer PageRank in Section 5. Section 6 presents the results of our experiments. Finally, we conclude the paper in Section 7.

2. Material and methods

The methodology of our research consists of five parts. First data collection using BFS network sampling technique, second study of important nodes in different activity networks, third calculating the importance of different activities, forth presenting application-based centrality measures, and ultimately Comparing presented methods with the most popular centrality measures.

Study of important nodes comprises of two parts: first calculating the importance of various activities, then comparing influential users in different activity networks. We also used three measures to compare presented methods with other centrality measures: capability to nodes' differentiation, standard deviation & dynamic range, and robustness over time. These parts are presented in Fig. 1.

In this section, we first present a high level analysis of the used dataset. Then the pre-processing of the collected data is also presented.

2.1. Description of the dataset

We used the collected dataset of Facebook users' activity and friendship networks in our experiments. The dataset include the friendship network of 36204 Facebook users. For friend users, the information about the number of activities including like, comment, post, and share as well as the number of exchanged words in their comments are collected every one month over a period of 3 years from 1 January 2011 to 1 January 2014 (Khadangi, Bagheri, & Shahmohammadi, 2016; Shahmohammadi, Khadangi, & Bagheri, 2016; Khadangi, Bagheri & Zarean, 2017). In our experiments, we usually used a smaller dataset which is a subset of the original one including the information of 8079 Facebook users.

The average degree of the friendship graph is 138. In addition, friendship network shows assortative mixing by degree. The average path length and 90-percentile effective diameter of friendship network are also 4.23 and 5 respectively. In addition, the degree distribution of the friendship follows power-law distribution.

The high level characteristics of different activity networks have also been presented in Table 1. It should be noted that mixed network is a mixture of like, comment, post, and share networks.

High clustering coefficient and low average path length confirm the small-world nature of various Facebook activity networks. In addition, the degree distribution of different activity networks follow power-law or semi-power-law.

2.2. Calculating the importance of different activities

Different activities have different importance for various applications. Accordingly, the presented methods use the importance of different activities for calculating nodes' centrality for different applications.

To calculate the importance of activities, we initially

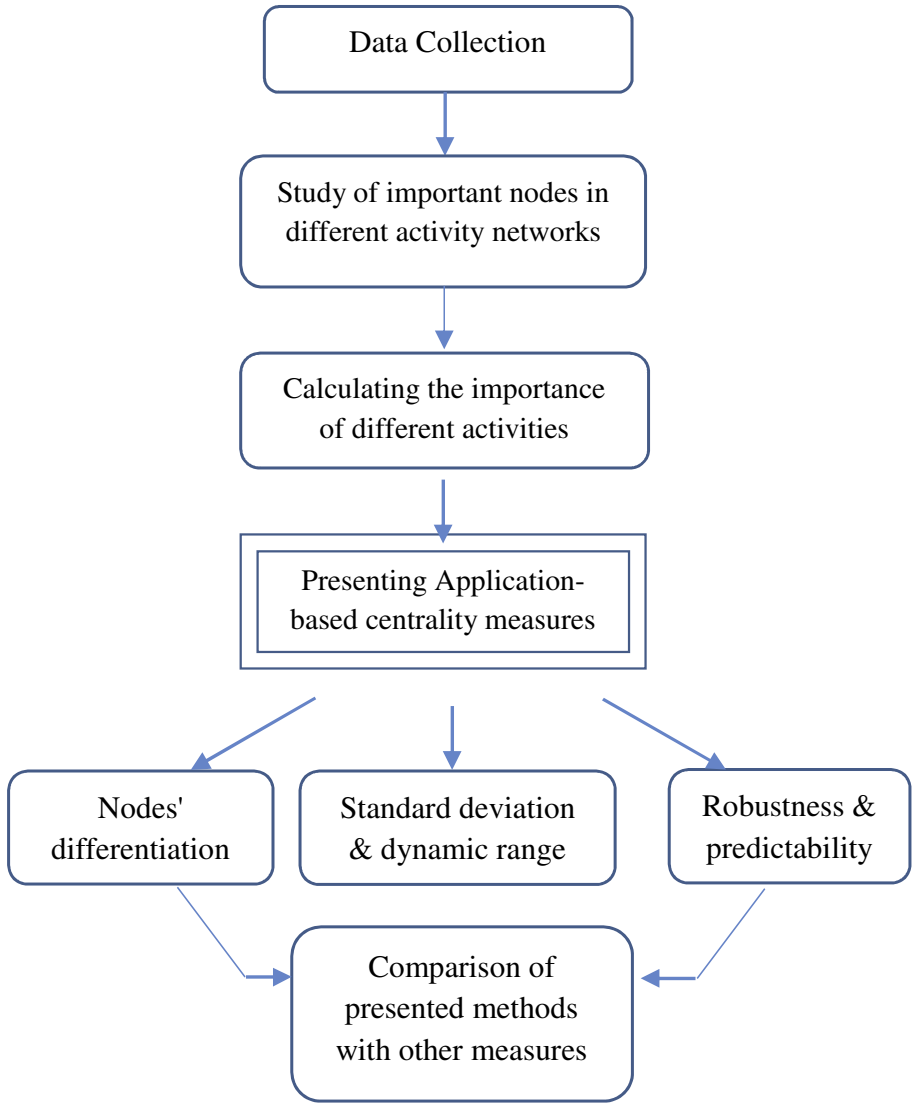

Fig. 1. An overall view of various parts of the research.

Table 1
Different structural properties of Like, Comment, Post, Share, and the mixture of other four activity networks.

	Clustering coefficient	Ave. Path Length	# of connected 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
share	0.04	5	181	0.75
post	0.039	6.2	198	0.6

implemented a Facebook application which has asked questions from different Facebook users about their friends. Then, according to the replies, we calculated the trust and closeness between 506 couples of users. In addition, we collected the number of like, comment, post, and share activities between pairs(Khadangi, Zarean, Bagheri & Iafarabadi, 2013). For calculating trust and closeness users answered four questions about their friends listed as follows. A in these questions represents a friend.

1. How much do you trust the information presented by A?
2. How close to A are your political, social and religious thoughts?
3. How strong is your friendship strength with A?
4. If A wants you to do something for him, how hard would you try for it?

After calculating trust and closeness based on the collected data and using linear regression, the importance of different activities is obtained for the two applications. At this stage, it should be noticed that further to users' interactions, other pieces of information such as gender, age, interests, and education may impact on trust and closeness. It is also possible that the relationship between trust and users' interactions is not absolutely linear. Therefore we used attribute weighting methods to improve the result of linear regression; the weights of different activities thereof were calculated as below.

1. Trust : $W_{like} = 9,\ W_{comment} = 7,\ W_{share} = 23\ \ and\ \ W_{post} = 11$
2. Closeness: Closeness : $W_{like} = 8,\ W_{comment} = 16,\ W_{share} = 10$ $and\ W_{post} = 16$

As seen, the importance of different activities is different for

various applications.

We also used another method to calculate activities' importance according to the time application. The importance of activities is calculated based on the time spent on different activities. For calculating them, by analyzing the collected data, the average number of words in each comment is calculated 10.7. In addition, upon some users' posts which lacked any pictures, the average number of words in each post is gained 22.5, according to which and the fact that some posts included pictures it can be estimated that each post takes approximately time 2.5 times of each comment. Therefore, the weights of different activities according to the time application was calculated as follows:

3. Closeness:
   Closeness : $W_{like} = 1,\ W_{comment} = 4,\ W_{share} = 4\ and\ W_{post} = 10$

3. Empirical study of centrality measures in activity networks

In this section, we study various centrality measures in friendship and activity networks to see if the influential nodes in friendship network are the same as the influential nodes in the activity network. In addition, the impact of the type of the activity link connecting nodes on the rank of different nodes is studied so as to suggest that the type of the links connecting nodes impacts considerably on the importance ranking of nodes so that a user may be important from the viewpoint of one activity, but not that important from the viewpoint of another activity.

3.1. Centrality measures in activity networks

The study of the activity network and examination of nodes' ranking based on different centrality measures can illustrate the differences between activity and friendship networks. It can also present researchers with better solutions for prominent users' identification. The influential nodes gained by applying different centralities to the activity and friendship networks have been presented in Table 2.

As seen in Table 2, a node may be influential in a network, but not such in another. In other words, considering Facebook users from different viewpoints, different rankings may be gained for them. Therefore, it is preferable to generalize centrality measures in a way that they can measure nodes' influence for different viewpoints. In addition, it is seen that different centrality measures present different nodes' rankings. Rankings in friendship network

Table 2
10 top-ranked nodes using different centrality measures In-degree, Out-degree, Closeness, Betweenness, Eigenvalue, Authority, and PageRank in different activity networks.

MixIn	MixOut	MixClos	MixBet	MixEig	MixPR	LikeIn	LikeOut	LikeClos	LikeBet	LikeEig	LikePR
112	7951	1	127	5952	112	112	7951	1	127	5952	112
231	112	112	1	6021	231	231	112	94	1	6021	231
24	7861	94	112	245	2	24	3903	112	112	468	2939
34	77	47	226	468	74	74	24	47	226	245	136
74	3903	231	74	94	2939	2939	32	135	7951	2939	74
2939	3330	135	231	2939	136	94	77	40	74	6011	2
94	24	40	24	6011	34	135	3330	32	24	94	24
135	32	32	94	5972	77	136	3328	231	94	5972	5969
ComIn	**ComOut**	**ComClos**	**ComBet**	**ComEig**	**ComPR**	**PostIn**	**PostOut**	**PostClos**	**PostBet**	**PostEig**	**PostPR**
112	245	47	1	6021	94	112	7861	245	112	112	3209
2939	47	112	127	245	231	245	112	112	245	245	7538
47	112	245	112	5952	2939	136	6801	5958	77	94	34
34	7951	231	47	47	112	5958	77	94	1394	77	112
24	24	2370	5458	5990	47	34	2	77	5447	44	3132
74	35	1	231	94	56	77	59	44	127	39	3136
231	2	44	5475	4592	5969	6021	6753	34	34	6021	2
6021	7074	94	7077	34	24	94	32	6112	6753	5958	6928
ShareIn	**ShareOut**	**ShareClos**	**ShareBet**	**ShareEig**	**SharePR**	**FGDeg**	**FGClose**	**FGBet**	**FGEig**	**...**	**...**
231	2	231	231	2	231	127	30	1	127	...	...
4596	118	5972	5972	118	2939	6817	1	127	6750	...	...
136	4824	47	94	129	74	6750	22	74	7867	...	...
4932	2092	1	47	136	3848	7867	71	112	7341	...	...
47	2421	94	1	1	5940	6809	41	136	6784	...	...
2939	6564	4	4596	132	2890	7270	5	135	7868	...	...
5972	5510	3197	4	34	1084	6784	64	70	6817	...	...
4824	6801	4596	4824	4	1085	7868	63	76	7849	...	...

Table 3
Spearman's rank correlation coefficient of different centrality measures between different five activity networks.

In	mix	like	com	post	Sha
mixed	1				
Like	0.98	1			
Comment	0.92	0.92	1		
Post	0.19	0.19	0.29	1	
Share	0.88	0.87	0.77	0.05	1

Out	mix	like	com	post	Sha
mixed	1				
Like	0.98	1			
Comment	0.72	0.67	1		
Post	0.27	0.23	0.27	1	
Share	0.37	0.33	0.25	0.05	1

PageRank	mix	like	com	post	Sha	wei.
mixed	1					
Like	0.95	1				
Comment	0.83	0.78	1			
Post	0.29	0.26	0.31	1		
Share	0.59	0.57	0.43	0.02	1	

Weighted PageRank	mix	like	com	post	Sha
mixed	1				
Like	0.96	1			
Comment	0.86	0.80	1		
Post	0.27	0.22	0.31	1	
Share	0.59	0.56	0.41	0.01	1

Hub	mix	like	com	post	Sha
mixed	1				
Like	0.98	1			
Comment	0.80	0.76	1		
Post	0	0	0	1	
Share	0.55	0.51	0.43	-0.01	1

Authority	mix	like	com	post	Sha	wei.
mixed	1					
Like	0.99	1				
Comment	0.95	0.95	1			
Post	0	0	0	1		
Share	0.91	0.90	0.87	0	1	

Closeness	mix	like	com	post	Sha
mixed	1				
Like	0.98	1			
Comment	0.51	0.48	1		
Post	0.26	0.25	0.16	1	
Share	0.40	0.37	0.27	0.18	1

Betweenness	mix	like	com	post	Sha
mixed	1				
Like	0.98	1			
Comment	0.87	0.87	1		
Post	0.45	0.45	0.52	1	
Share	0.33	0.33	0.26	0.09	1

Eigenvector	mix	Like	com	post	Sha
mixed	1				
Like	0.99	1			
Comment	0.92	0.91	1		
Post	-0.07	-0.07	-0.06	1	
Share	0.75	0.74	0.70	-0.05	1

are also considerably different from activity networks, which highlights the importance of calculating node importance by applying centrality measures to the activity network instead of the friendship network. The rankings are also different in different activity networks using a similar centrality measure.

3.2. Inter-network correlation of centrality measures

To study the differences of nodes' importance in various networks, the inter-network correlation of different centrality measures are presented in Table 3. We used Spearman's rank correlation coefficient for calculating correlations between different centrality measures. The Spearman's rank correlation coefficient can be calculated as follows:

$$R_{xy} = 1 - \left[\frac{6 \sum_i d_i^2}{n(n^2 - 1)} \right]$$

Where n is the number of items, and $d_i = x_i - y_i \cdot x_i$. is the rank of node i with respect to the first variable and y_i is the rank of it with respect to the second variable.

Table 3 shows that the correlation between nodes' ranking in various activity networks is sometimes low. This observation indicates again that users have different importance based on various activity networks.

3.3. Centrality measures over time

In this section, we study the dynamics of centrality measures and the importance change of nodes in different networks over time. Table 4 presents five top-ranked nodes based on applying different centrality measures to different snapshots of the mixed network. It can be seen that closeness changes little over time and influential nodes obtained using PageRank interchange their positions at a low rate. This change is average in other measures.

Fig. 2 presents the correlation between different centrality measures in different activity networks over time.

Significant facts can be realized from Fig. 2, whereas they were not available by the study of centrality in a static snapshot. Firstly, the correlation between hub and authority centralities is initially very low, but it increases with an almost monotone trend over time. It seems due to the sparseness of the initial snapshots and the low number of activities, some of them were good hubs and some others were good authorities and the number of nodes which are not important based on two measures in the first snapshots. However, users who were good hubs received incoming activities according to their many activities and gradually became rather good authorities. The same event has also happened inversely to some extent. However, in the post network, the change of hub and authority correlation does not follow a definite pattern. On the other hand, Indegree-outdegree correlation has an ascending trend similar to the authority-hub correlation. This value is almost zero at first and increases over time to reach approximately 0.5. This is probably due to a similar situation as described above. However, Indegree-Outdegree correlation is not high in different snapshots. Moreover, correlation between authority and indegree and hub and outdegree has been high since the beginning and has not changed considerably over time. The same phenomenon is almost true for the correlation between PageRank and authority.

3.4. Differences between friendship and activity networks based on the gained influential nodes

Since activity network models users' behavior better than friendship network, can distinguish between strong and weak links, and it is also closer to users' social behavior in real world, finding influential users in the activity network will give better and more real results. Accordingly, we can compare influential users in the two networks in order to study the goodness of centrality in the friendship network. For this, the number of common top-ranked nodes in friendship network (based on degree centrality) and mixed and weighted mixed activity networks (based on degree centrality and PageRank) is shown in Table 5.

It can be seen in Table 5 that ranking based on the number of friends is not even close the most similar measure in activity networks i.e. in-degree at all, that only 25% of 1000 top-ranked nodes using in-degree centrality of the activity network have been detected by the degree centrality in the friendship network. Because we know that the activity network represents users' behavior better, we can infer that friendship degree is usually not an appropriate method for finding influential nodes in real

Table 4
Five top-ranked nodes of different snapshots of mixed activity network obtained using In-degree, Closeness, PageRank, Betweenness, Hub & Authority.

one PR	two PR	three PR	four PR	five PR	six PR
112	112	112	112	112	112
5225	4500	92	2	231	231
135	4485	77	94	94	2
80	80	67	34	34	2939
83	83	80	92	2	34
one Auth	**two Auth**	**three Auth**	**four Auth**	**five Auth**	**six Auth**
94	112	2939	2939	2939	2939
77	34	6021	34	231	231
44	6021	245	6021	6021	34
34	5990	5990	94	34	6021
5969	2939	34	5958	94	5958
one in	**two in**	**three in**	**four in**	**five in**	**six in**
77	112	112	112	112	112
44	44	77	94	231	231
75	77	44	34	2939	24
34	244	244	2939	94	34
43	34	2939	44	34	2939

one Bet	two bet	three bet	four Bet	five Bet	six Bet
77	112	112	112	127	127
94	77	77	5447	112	1
43	244	244	127	1	112
44	94	35	5445	231	226
245	44	245	7782	94	231
one Hub	**two Hub**	**three Hub**	**four Hub**	**five Hub**	**six Hub**
245	245	5952	5952	5952	245
5972	5952	6021	6021	245	5952
5952	468	4599	245	6021	6021
32	6021	468	4599	468	468
468	5972	245	520	520	6083
one close	**two close**	**three close**	**four close**	**five close**	**six close**
94	245	245	94	112	1
77	94	77	112	231	94
245	77	94	245	94	112
44	244	112	244	1	47
43	112	244	77	47	231

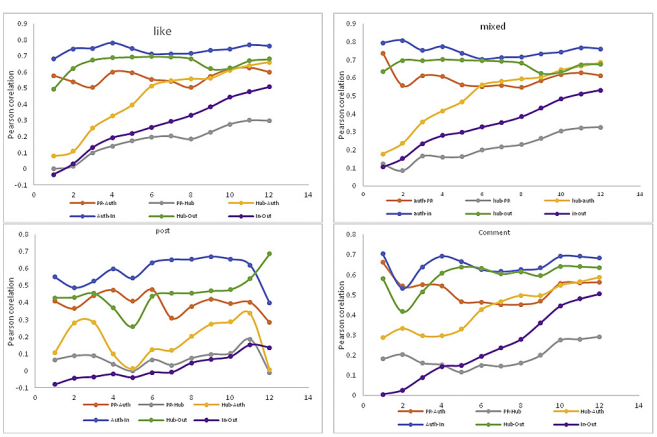

Fig. 2. Spearman's rank correlation coefficient between different centrality measures (In-degree, Out-degree, PageRank, Hub & Authority) applied on various activity networks over time.

Table 5
The number of common top-ranked nodes in friendship network (by applying In-degree centrality), mixed and weighted mixed activity networks (by applying In-degree, Out-degree, and PageRank).

	Number of the common top-ranked nodes							
Friendship Degree	10	20	50	100	500	1000	2000	4000
Mixed In	2	4	15	30	173	306	704	1723
WMixed In	1	3	16	26	131	252	552	1388
WMixed Out	0	0	3	7	69	154	408	1206
Mixed PR	0	0	0	0	7	23	99	208
Weighted PR	0	0	0	0	10	25	113	206

Table 6
The number of common top-ranked nodes in friendship network (by applying Betweenness) and important nodes in mixed network (by applying PageRank, Hub & Authority).

	Number of the common top-ranked nodes							
FGBet	10	20	50	100	500	1000	2000	4000
Mixed Bet	4	7	19	45	224	442	935	1933
Mixed PR	1	2	6	19	161	355	747	1572
WMixed PR	1	2	7	23	144	304	653	1414
Mixed Auth	0	1	1	3	67	169	442	1228
Mixed Hub	0	0	0	1	35	94	282	822

applications. By comparing 4000 top-ranked nodes in friendship network using Betweenness and important nodes gained by PageRank in mixed and weighted mixed activity networks, the results in Table 6 were gained.

It is interesting that although the betweenness is much more powerful and reliable than the degree centrality, it still cannot detect many of the influential nodes gained from the activity network. We believe that this fact is due to the weakness of the friendship network to model the users' interactions in real-world rather than the weakness of betweenness. Therefore, we suggest that the activity network which is closer to users' social behaviors is used to calculate users' influence, particularly for applications such as finding appropriate initial adopters for social media marketing.

4. Application-based centrality measure

As already mentioned, the study of users' importance on the friendship network does not provide us with appropriate result; therefore, it is better to apply centrality measures to activity networks rather than to the friendship network. In addition, different

activity networks have different structural properties and using each of them can present a different ranking of nodes. Therefore, we require a method which besides the capability to apply to networks individually be capable of considering different networks simultaneously in measuring the importance of nodes. It can also be claimed that almost all centrality measures have been designed regardless of the application although the node importance may be different in terms of different viewpoints and applications. For instance, the importance of a node is different in terms of the amount of activities, trust, closeness, spam, fame and time spent applications. In addition, in order to present a suitable centrality measure we should pay attention to the nature and structural properties of the activity network, especially the reciprocity of the network. Table 7 shows the reciprocity of activity networks.

It is seen in Table 7 that reciprocity is low in different networks. This fact is seen particularly in post and share networks. This shows that Facebook activities are intrinsically directed concepts. Therefore, the proposed centrality measure should consider the direction of links. In addition, the centrality measure should be capable of generalizing to weighed networks because the study of the weighted and unweighted activity networks presents different results.

Accordingly, we presented two types of directed centrality measures. The first type is directed centrality measures proposed for fame and spam applications. The second type is a PageRank based centrality measure for measuring the importance of nodes for different applications. On the one hand, PageRank is a suitable measure for directed networks; on the other hand, it can simply be generalized for applying to weighted networks. Since the two types must consider all activities simultaneously, it should be examined how it is possible to mix different activity networks with each other suitably. As each activity has its own importance in terms of the considered application, the calculated activities' importance in Section 3.2 will be the input for the two types of centrality measures, as explained below.

4.1. Centrality measures for fame and spam applications

Some users of Facebook have gained fame outside the network and use Facebook for communicating with general people. Among these are actors and actresses, politicians, sportsmen and sportswomen, singers, and famous scientists. Ranking users based on their fame may be very useful for different applications. What we exactly mean by "famous user" in this paper is the person who is also famous outside the social network. Therefore, the high value of different centrality measures does not equal the user's fame, as popular centrality measure shows the user's importance only in the social network. Most famous people have many incoming activities due to their fame and prestige; meanwhile, they do not have enough time to answer these input activities. Therefore, the famous node in the activity network can simply be defined as follows: a famous node is the node which holds many input activities versus its output; for example, a famous node's posts and contents are liked and shared by many people. In this paper, the famous node is defined based on the collected information of four activities like, comment, post, and share. The fame centrality can be calculated as follows.

$$Fame(i) = \frac{\sum_{a\in\{like,comment,post,share\}}[(\max(inweight_a(i) - outweight_a(i), 0))*W_a*purity_a(i)]}{\sum_{a\in\{like,comment,post,share\}} W_a} \quad (1)$$

where

$$inweight_a(i) = \sum_{j\in in-neighbors\{i\}} w_{ji}^a$$

and

$$outweight_a(i) = \sum_{j\in out-neighbors\{i\}} w_{ij}^a$$

W_a also shows the importance of the activity network $a\in\{like, comment, post, share\}$. This value may be different for various activities. For example, share and post activities are more important than comment and like in fame application. $purity_a(i)$ shows the purity of user i's activities. The philosophy of including the concept of purity in definition of fame is that famous nodes have high purity in terms of the weight of input and output links, since they have many input links and less output links. In other words, the subtraction of the input and output weight is not the only way to show one node's fame, but the purity in terms of the weight of input and output links at different layers of the activity network is also very important. There are two ways Gini Index and Entropy for calculating the amount of purity. These two measures for the like layer of the activity network can be calculated as below.

$$\begin{aligned} purity_{like}(i) &= 1 - gini_{like}(i) \\ &= \left(\frac{inweight_{like}(i)}{weight_{like}(i)}\right)^2 + \left(\frac{outweight_{like}(i)}{weight_{like}(i)}\right)^2 \end{aligned} \quad (2)$$

Or

$$\begin{aligned} purity_{like}(i) &= 1 - entropy_{like}(i) \\ &= 1 + \frac{inweight_{like}(i)}{weight_{like}(i)} log_2^{\frac{inweight_{like}(i)}{weight_{like}(i)}} \\ &\quad + \frac{outweight_{like}(i)}{weight_{like}(i)} log_2^{\frac{outweight_{like}(i)}{weight_{like}(i)}} \end{aligned} \quad (3)$$

where

$$weight_{like}(i) = \sum_{j\in in-neighbors^{like}\{i\}} w_{ji}^{like} + \sum_{j\in out-neighbors^{like}\{i\}} w_{ij}^{like}$$

In these equations, $inweight_{like}(i)$ is the sum of the weight of

Table 7
Reciprocity of various activity networks.

Network	Reciprocity of unweighted network	Reciprocity of weighted network
Like	0.33	0.26
Comment	0.26	0.3
Post	0.06	0.02
Share	0.11	0.2
Mixed	0.36	0.3

input links to the node i at the activity layer like. $weight_{like}(i)$ also equals the number of input and output likes of the node i. Similarly, purity for other activities can be defined.

In addition to the fame application, detecting spam users is also an important issue (Wang, 2010). Spam detection is not possible without using the activity network and the information of users' interactions. This is because the difference between spammers and others is by how they act. Based on this, the spam user or the user whose behavioral pattern is spam-like is the person with many outputs but few inputs. In other words, most spam users have many outgoing activities in the social network whereas they do not have many incoming activities. Equation (4) shows the spam centrality measure which calculates the amount of being spam. The more the spam centrality of a node is, the higher the spamming probability of the node will be.

$$Spam(i) = \frac{\sum_{a\in\{like,comment,post,share\}}[\max(outweight_a(i) - inweight_a(i), 0)]*W_a*purity_a(i)}{\sum_{a\in\{like,comment,post,share\}}W_a} \tag{4}$$

$inweight_a(i)$, $outweight_a(i)$ and $purity_a(i)$ can be calculated as mentioned above. It should be noted that we can normalize fame and spam centralities by dividing them by the maximum possible value of the two centralities. However, since we use these two measures for comparing users in a special activity network, the normalization is not essential.

4.2. Application-based PageRank for multilayer activity network

We should have involved the calculated importance of the activities for the considered application while calculating PageRank of nodes. In this regard, we initially modeled the activity network as a four-layer network where each layer represents one of Facebook activities, including like, comment, post, and share. Other layers may also be considered, such as number of mutual groups, events, and interests. For calculating the importance of nodes in a multi-layer network for a specific application, we should merely use the generalized PageRank equation below with regard to the weight of different activities for the considered application. In this equation C_{MLPR} shows the PageRank in the multi-layer network.

$$C_{MLPR}(i) = \frac{1}{\sum_{\alpha=1}^{L}W_\alpha}\sum_{\alpha=1}^{L}\left(W_\alpha\cdot\left(d_\alpha\cdot\sum_j A_{ji}^{\alpha}\frac{C_{MLPR}(j)}{\sum_k A_{jk}^{\alpha}} + \frac{(1-d_\alpha)}{n_\alpha}\right)\right) \tag{5}$$

where W_α is the weight of the layer or activity α which is different according to the application. A_{jk}^{α} is the weight of the link $j\rightarrow k$ in layer α. And L is the number of layers of the activity network, which is 4 in this paper. In addition, n_α and d_α are the number of nodes and the damping factor of layer α respectively. The different damping factor for different layers can be due to the nature of the considered layer, and also behavior of users based on the considered activities.

In most situations, it is better that the damping factor and the number of nodes be the same for all layers. In these cases, the equation of application-based PageRank is shortened as below.

$$C_{MLPR}(i) = \frac{d}{\sum_{\alpha=1}^{L}W_\alpha}\sum_{\alpha=1}^{L}\left(W_\alpha\cdot\sum_j A_{ji}^{\alpha}\frac{C_{MLPR}(j)}{\sum_k A_{jk}^{\alpha}}\right) + \frac{(1-d)}{n} \tag{6}$$

The equations presented so far are about the loop-less activity network. However, some of the collected activities are not interactive, but they are activities which users do on their own profile. More accurately speaking, due to the data collection process, the data of activities which a user does on his own timeline do not refer to any other user. We call such activities intra-actions and are modeled as self-loops in the activity graph. It is suggested that intra-actions be considered while calculating the importance of nodes since their consideration may impact on ranking the nodes considerably. Having considered intra-actions, the PageRank equation is reformed as below.

$$C_{MLPR}(i) = \frac{d}{\sum_{\alpha=1}^{L}W_\alpha}\sum_{\alpha=1}^{L}\left(W_\alpha\cdot\left(\sum_j\frac{A_{ji}^{\alpha}}{\gamma A_{jj}^{\alpha}+\sum_k A_{jk}^{\alpha}}\cdot C_{MLPR}(j) + \frac{\gamma A_{ii}^{\alpha}}{\gamma A_{ii}^{\alpha}+\sum_k A_{ik}^{\alpha}}\cdot C_{MLPR}(i)\right)\right) + \frac{(1-d)}{n}, \tag{7}$$

where A_{ii}^{α} is the weight of intra-actions belonging to user i in layer α. In addition, γ is a constant, often between 0 and 1, which shows the importance of intra-actions versus interactions. The more this value is, the importance of intra-actions is and when it is equal to one, the importance of intra-actions and interactions is the same. The importance of intra-actions is usually lower than interactions so often $0 \le \gamma \le 1$. This equation means that a node is important only if important people or many people have linked to it. However, the number of intra-actions is also influential in the overall importance. Obviously if there is no intra-action in the network, Equation (2) is gained.

Each of PageRank equations depending on the values of parameters lead to different results of rankings. In order to calculate the values of PageRank according to the equation above, different methods such as spectral graph theory and random-walk simulation may be used. Intuitively, simple PageRank can be obtained by the probability that a random surfer visits a node in a graph. Herein we used a two-stage random walk with restart on the weighted multilayer activity network. In the first stage, we jump to a node chosen uniformly at random with probability (1-d), otherwise select a layer according to the importance of the activity. Then, the random walk stay at the current node in the selected layer with the probability $W_{ii}^{\alpha}/w_{ii}^{\alpha}+\sum_j W_{ij}^{\alpha}$ and move to one of the nodes to which i has linked with a probability in proportion of the weight of output links. It should be noticed that we used $\gamma = 1$ in our algorithm. Algorithm 1 shows the pseudo-code of this algorithm. With a simple and effective approach, this algorithm can measure users'

Algorithm1. Application-based MultiLayer PageRank

```
Input: gLike,gComment,gPost,gShare,initial,maxIter,DF,vnum, activity weights
p=[likeWeight, commentWeight, postWeight, shareWeight]
for i=1 to vnum do
   Visits[i]<-ε
end
v<-initial node
i<-1

 while(not converge and i<maxIter) do
   rand<-a number uniformly at random between 0 and 1
   if(rand>DF) then
     W<-select a uniformly random node
   else
   begin
     while(degree^Act(v)=0) do
       Act<-select activity i with probability p[i]
     N<-neighbors^Act(v)
     foreach(vertex n ∈ N ∪ {v})
       prob[v]<-w_vn/(Σ_{n∈N} w_vn + w_vv)
     W<-select a vertex from N with probabilities of prob
   end
   V<-W
   Visits[v]<-Visits[v]+1
   i<-i+1
 end
 for all v ∈ V do
   PageRank[v]=Visits[v]/i
end
Normal PageRank:
LikeWeight=1, CommentWeight=1, PostWeight=1, ShareWeight=1
Trust application:
LikeWeight=9, CommentWeight=7, PostWeight=23, ShareWeight=11
Time application:
LikeWeight=1, CommentWeight=4, PostWeight=4, ShareWeight=10
Closeness application:
LikeWeight=8, CommentWeight=16, PostWeight=10, ShareWeight=16
```

influence for different applications. It only requires that we measure the weight of different layers of the network for the considered applications and give them to the algorithm as input. We named this algorithm A-BMLPR standing for Application-Based Multilayer PageRank.

A-BMLPR takes the four activity graphs of like, comment, post, and share as well as their importance as input. The number of input graphs which in fact are different layers of the multilayer network may easily change. In addition, "initial" shows the initial node and "DF" shows the damping factor which is between 0 and 1. Finally, the influence of nodes is measured between 0 and 1 which is according to the number of nodes' visits based on the two-stage random walk. It is quite obvious that if the network does not have any self-loops, $W_{vv} = 0$. Therefore, the algorithm is easily applicable to the graph which does not have any self-loops.

We can also calculate application-based PageRank using spectral graph theory. According to spectral graph theory, PageRank is the principal eigenvector of the stochastic transition matrix. Therefore, we only need to define the stochastic transition matrix based on the corresponding application. If we have a multilayer network without self-loop, the transition matrix for application-based random walk can be defined as follows:

$$M_{ij} = \frac{d}{W_{total}} \sum_{a \in activity} W_a \left(\frac{A_{ij}^{a}}{k_{out}^{a}(i)} \right) + \frac{(1-d)}{n} \tag{8}$$

where

$$W_{total} = \sum_{a \in activity} W_a$$

And And $= \{like, comment, post, share\}$, d is damping factor and A_{ji}^{a} is the weight of the link $j \rightarrow i$ in layer a.

Theorem 1. *For all loop-less activity networks, matrix M which models the application-based random walk process is a stochastic matrix.*

Proof. A matrix is stochastic if all its entries are non-negative and the sum of elements of each row equals 1. Since A_{ji}^{a} , $k_{out}^{a}(j)$, W_a, d, and $(1-d)$ are non-negative, all elements of M are non-negative. In addition, by summing i's row elements of M and using Equation (9) we have:

$$\sum_{j=1}^{n} M_{ij} = \sum_{j=1}^{n}\left[\frac{d}{W_{total}}\sum_{a\in activity} W_a\left(\frac{A_{ij}^{a}}{k_{out}^{a}(i)}\right)+\frac{(1-d)}{n}\right]$$
$$= (1-d)+\frac{d}{W_{total}}\sum_{j=1}^{n}\sum_{a\in activity} W_a\left(\frac{A_{ij}^{a}}{k_{out}^{a}(i)}\right)$$
$$= (1-d)+\frac{d}{W_{total}}\sum_{a\in activity} W_a\sum_{j=1}^{n}\left(\frac{A_{ij}^{a}}{k_{out}^{a}(i)}\right)$$
$$= (1-d)+\frac{d}{W_{total}}\sum_{a\in activity} W_a\frac{\sum_{j=1}^{n}A_{ij}^{a}}{k_{out}^{a}(i)}$$

Since we have

$$\frac{\sum_{j=1}^{n}A_{ij}^{a}}{k_{out}^{a}(i)} = 1$$

We have

$$\sum_{j=1}^{n} M_{ij} = (1-d)+\frac{d}{W_{total}}\sum_{a\in activity} W_a = 1$$

Since the sum of elements of each row equals 1, M is a stochastic matrix.

On the other hand, if our network contains self-loops, the application-based random walk can be modeled as follows:

$$\forall j\neq i\ :\ M'_{ij} = \frac{d}{W_{total}}\sum_{a\in activity} W_a\left(\frac{A_{ij}^{a}}{k_{out}^{a}(i)+\gamma A_{ii}^{a}}\right)+\frac{(1-d)}{n}, M'_{ii}$$
$$= \frac{d}{W_{total}}\sum_{a\in activity} W_a\left(\frac{\gamma A_{ii}^{a}}{k_{out}^{a}(i)+\gamma A_{ii}^{a}}\right)+\frac{(1-d)}{n} \tag{9}$$

where γ is a constant between 0 and 1 which shows the importance of intra-actions versus interactions.

Theorem 2. *For all activity networks which contains self-loops, matrix M′ presented in Equation (11) is a stochastic matrix.*

Proof. Similar to the proof of Theorem 1 all of the elements of *M*' are non-negative. In addition we have

$$\sum_{j=1}^{n} M'_{ij} = \sum_{j\neq i}\left[\frac{d}{W_{total}}\sum_{a\in activity} W_a\left(\frac{A_{ij}^{a}}{k_{out}^{a}(i)+\gamma A_{ii}^{a}}\right)+\frac{(1-d)}{n}\right]$$
$$+\frac{d}{W_{total}}\sum_{a\in activity} W_a\left(\frac{\gamma A_{ii}^{a}}{k_{out}^{a}(i)+\gamma A_{ii}^{a}}\right)+\frac{(1-d)}{n}$$
$$= \frac{d}{W_{total}}\sum_{a\in activity} W_a\sum_{j\neq i}\left(\frac{A_{ij}^{a}}{k_{out}^{a}(i)+\gamma A_{ii}^{a}}\right)$$
$$+\left(\frac{(n-1)(1-d)}{n}\right)$$
$$+\frac{d}{W_{total}}\sum_{a\in activity} W_a\left(\frac{\gamma A_{ii}^{a}}{k_{out}^{a}(i)+\gamma A_{ii}^{a}}\right)+\frac{1-d}{n}$$

Since we have

$$\sum_{j\neq i}\left(\frac{A_{ij}^{a}}{k_{out}^{a}(i)+\gamma A_{ii}^{a}}\right) = \frac{k_{out}^{a}(i)}{k_{out}^{a}(i)+\gamma A_{ii}^{a}}$$

Therefore

$$\sum_{j=1}^{n} M'_{ij} = (1-d)+\frac{d}{W_{total}}\left(\sum_{a\in activity} W_a\cdot\left(\frac{k_{out}^{a}(i)}{k_{out}^{a}(i)+\gamma A_{ii}^{a}}\right)\right.$$
$$\left.+\sum_{a\in activity} W_a\cdot\left(\frac{\gamma A_{ii}^{a}}{k_{out}^{a}(i)+\gamma A_{ii}^{a}}\right)\right)$$

We also have:

$$\sum_{a\in activity} W_a = W_{total}$$

Finally

$$\sum_{j=1}^{n} M'_{ij} = (1-d)+d\left(\frac{k_{out}^{a}(i)+\gamma A_{ii}^{a}}{k_{out}^{a}(i)+\gamma A_{ii}^{a}}\right) = (1-d)+d = 1$$

Therefore, M' is a stochastic matrix. Ultimately, the principal eigenvectors of *M* and *M*' represent the application-based PageRank of multilayer networks.

Now we can reach the application-based PageRank of the multilayer Facebook activity network by calculating the principal eigenvector of the presented application-based stochastic matrices. It should be noted that W_a is the importance of activity *a* that calculated based on the method presented in the previous section. Both "A-BMLPR algorithm" and "spectral-based method" have gained approximately similar results. Therefore, we named the two methods as A-BMLPR.

5. Experimental results

To study A-BMLPR, three types of experiments were done. In the first type, we intend to compare A-BMLPR for different applications.

Table 8
50 top-ranked nodes gained using A-BMLPR in different applications.

normal				trust		closeness		time		PR in simple activity graph	
1st	231	26th	56	2939	59	2	3068	112	56	112	468
2nd	2939	27th	2890	231	15	7951	135	2	35	94	75
3rd	112	28th	3206	112	6379	112	6564	7951	826	2	6911
4th	5969	29th	4635	5969	3193	5479	6011	5479	3903	135	4830
5th	6021	30th	6379	6021	3206	6021	826	77	804	34	4932
6th	74	31st	3193	2	3504	7861	44	6021	135	136	7937
7th	2	32nd	24	74	5952	32	804	136	4	74	66
8th	77	33rd	6357	77	4943	74	64	7861	7687	231	2354
9th	94	34th	5972	94	4949	77	4	74	244	2939	39
10th	47	35th	96	44	66	47	6013	32	6013	77	1
11st	135	36th	826	47	4401	24	59	231	839	24	40
12nd	2106	37th	244	2106	4715	118	67	55	4406	67	2911
13rd	44	38th	5685	135	56	96	245	3909	6011	70	244
14th	245	39th	6606	3197	65	136	6362	6801	33	6021	99
15th	3197	40th	3554	5990	244	6801	4557	47	4451	35	7936
16th	34	41st	4406	55	2890	94	3903	245	5972	5969	65
17th	55	42nd	194	136	4406	56	1	6564	5046	2353	3554
18th	136	43rd	4943	67	2353	2939	4824	96	5952	47	59
19th	5990	44th	4643	6013	24	55	7687	24	127	92	3
20th	67	45th	6011	6911	826	35	3197	118	6362	55	68
21st	59	46th	4401	245	6593	231	5549	48	3197	56	5958
22nd	6013	47th	15	4189	1462	6593	5972	44	5549	64	217
23rd	6911	48th	3221	4592	40	3909	346	6593	3078	44	5952
24th	4592	49th	3554	4635	6606	48	5046	94	4635	43	3107
25th	6362	50th	2382	34	465	3078	4949	2939	4824	4596	96

Table 9
The number of common top-ranked nodes between A-BMLPR and simple PageRank.

	10-top	20-top	30-top	50-top	100-top	200-top	300-top
normal	7	15	19	24	49	93	144
trust	6	13	17	26	51	95	141
closeness	4	12	15	21	37	77	114
time	5	10	15	19	39	83	117

Table 10
The number of common top-ranked nodes between trust-based PageRank and the result of applying PageRank on different activity networks and the betweenness in friendship network.

	10 top	20 top	30 top	50 top	100 top	200 top	300 top
like	3	4	4	8	18	39	60
comment	3	3	6	13	29	58	84
post	1	4	5	9	24	54	74
share	2	3	7	12	28	51	72
friendship	1	2	4	8	22	48	59

Table 11
The number of common top-ranked nodes between nodes gained by using trust-based PageRank and nodes gained by applying different centrality measures on mixed network.

	10-top	20-top	30-top	50-top	100-top	200-top	300-top
in-trust	5	13	17	26	42	80	133
out-trust	2	7	8	12	21	53	88
bet-trust	5	10	11	16	32	70	111
auth-trust	5	11	15	18	30	56	90
hub-trust	1	5	7	13	22	46	70

The objective herein is to examine if the application has a considerable impact on ranking or not. More completely, we want to demonstrate in the first type of our experiments that the presentation of different rankings for different applications is among the advantages of our proposed method. In the second type of experiments, we intend to compare A-BMLPR with other centrality measures including indegree, outdegree, betweenness, closeness, hub and authority, and etc. The goal of the third-type experiments is the study of the performance of A-BMLPR based on the dynamic

range, distribution and robustness over time.

5.1. A-BMLPR capability to nodes' differentiation based on different applications

In order to evaluate the potential of A-BMLPR for differentiating nodes based on the considered application, we measured Facebook users' influence for different applications by applying A-BMLPR to the Facebook activity network. Table 8 shows the gained top-ranked nodes. It is seen therein that by applying application-based PageRank to the multilayer activity network, more than a half of 50 top-ranked nodes are different from that gained by applying simple PageRank to mixed network. It is also seen that users' rankings for different applications are different. The most influential node is different by all applications. In addition, the further we go from the beginning of the top-ranked nodes list, the more different the nodes' rankings become. Therefore, application-based PageRank can differentiate between users acceptably according to different applications.

5.2. Comparison of A-BMLPR and other centrality measures

The number of common top-ranked nodes between application-based PageRank and simple PageRank is also demonstrated in Table 9.

It can be realized from this table that in most cases more than 50% of top-ranked nodes are different for different applications from the result of applying simple PageRank to mixed network. Therefore, it can be inferred that applying PageRank to mixed network cannot provide a suitable result for different applications and it is better to use A-BMLPR. It should be noted that when we compare results of A-BMLPR with PageRank in separate activity networks, the differences are much more. Table 10 compares trust-based PageRank with PageRank in separate activity networks and betweenness in the friendship network.

As seen in Table 10, the number of common top-ranked nodes between A-BMLPR and PageRank in different activity networks is small. In addition, the difference of the gained results with that of applying betweenness on the friendship network is less than separate activity networks. Therefore, if we consider A-BMLPR as a benchmark, finding influential users in activity networks, especially in mixed network, will give more real results than applying centrality measures to the friendship network. Table 11 also shows the number of common top-ranked nodes between trust-based PageRank and different centralities in mixed network. The reason of selecting the trust application among different applications in order to assess the proposed method is that trust is a relatively important application and also is more tangible versus other applications. According to Table 11, we realize that the difference between the results of application-based PageRank with that of applying different centrality measures in mixed network is very much. If we examine separate activity networks the difference will be much more.

5.3. Standard deviation and dynamics range of A-BMLPR

The seen differences between A-BMLPR and simple PageRank in different activity networks shows that A-BMLPR presents a different rankings from simple PageRank. The main reason of this difference is that this method has been specialized for a special application. Nonetheless, this difference does not inform very much about goodness of the A-BMLPR. In order to evaluate the performance of the proposed method, we compare its standard deviation and dynamic range with other centrality measures. The result has been shown in Table 12.

The dynamic range is herein defined as $max(centrality)/min(centrality)$ which accordingly the more this value is, it is better. We can see in Table 12 that the standard deviation of application-based PageRank is more than all centrality measures but the hub centrality. In addition, regardless of two exceptions, the dynamic range of the proposed method is more than other centrality measures. This shows that the presented method has a high potential for differentiation; therefore, it can detect influential nodes sharply and accurately. In order to study this fact further, we have shown the distribution of application-based PageRank values in Fig. 3.

As seen in Fig. 3, the distribution of A-BMLPR centrality is approximately linear in log-log scale; therefore, the proposed centrality follows the power-law distribution. So, there are many nodes with low centrality and the closer we go to higher centralities, the number of nodes decreases. Following the power-law distribution is another support for the statement that multilayer PageRank can detect influential nodes sharply.

5.4. Robustness of A-BMLPR over time

Another feature of a suitable centrality measure is robustness. The method the ranking of which is less variable by the network changes or errors such as edge deletion, node deletion, edge addition and node addition is better (Borgatti, Carley, & Krackhardt, 2006; Davidsen & Ortiz-Arroyo, 2012). For evaluating A-BMLPR performance, we study its robustness over time. Centrality of nodes by using a robust centrality measure changes over time, but nodes' ranking change less and thus its ranking result at a particular time can be a good estimation for the future. In order to study this issue, we initially made 10 different snapshots of the mixed network and also the multilayer activity network by three-month time windows. We named snapshots of mixed network as $\{MAN_1, MAN_2, \ldots, MAN_{10}\}$ and those of the multilayer network as $\{MLAN_1, MLAN_2, \ldots, MLAN_{10}\}$. Then, we measured the number of common top-ranked nodes between 9 snapshots with MAN_{10}. In addition, having applied trust-based multilayer PageRank to different snapshots of the multilayer network, we measured the number of common top-ranked nodes between them and $MLAN_{10}$. Fig. 4 shows the obtained results.

We can see in Fig. 4 that the number of common nodes of different snapshots with the last one increases over time and follows a predictable pattern. This fact is quite reasonable. Because

Table 12
Standard deviation and dynamic range of different centrality measures.

Measure	Dynamic range	Standard deviation	Measure	Dynamic range	Standard deviation
In	1586	0.024	PageRank	5591	0.04
Out	420	0.04	Hub	9608	0.063
Closeness	40	0.0003	Authority	8899	0.03
Betweenness	6304	0.014	Eigenvector	9010	0.042
Normal MLPR	12567	0.047	Trust-based PR	11902	0.05
Time-based PR	12488	0.048	Closeness-based PR	7371	0.063

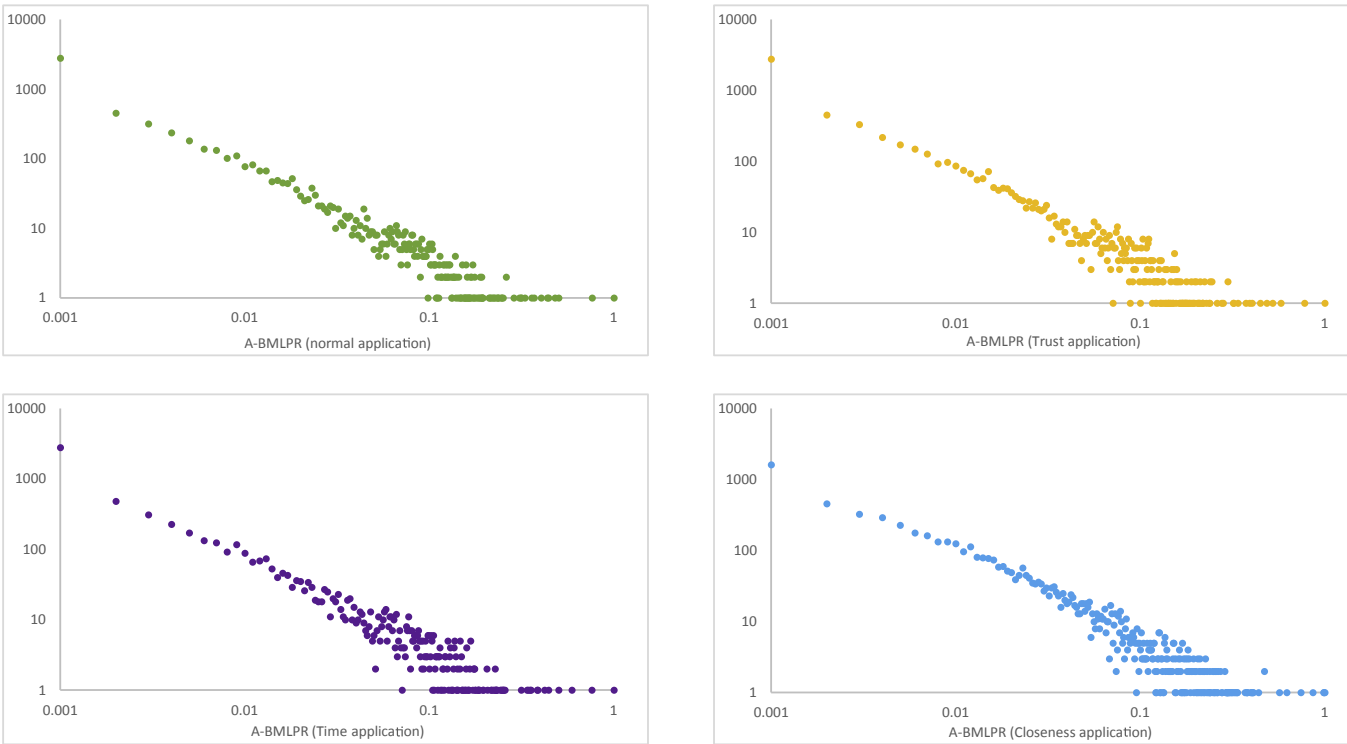

Fig. 3. A-BMLPR distribution in different applications.

networks which are at a less time distance from each other present rankings similar to each other and the more these distances are, the rankings will be more different. In addition, the diagram of trust-based PageRank for 10 and 100 top-ranked nodes is over all other diagrams and shows that this method has a more robustness and predictability. It also shows that 10 top-ranked nodes using application-based PageRank at a particular time are suitable estimation for 10-top ranked nodes in the future. The robustness of A-BMLPR in two other cases is also higher than most of the other centrality measures. Therefore, A-BMLPR predicts influential nodes in the future better than other centrality measures. Another point is that most of the influential nodes detected at a time by application-based PageRank are influential at other times.

Table 13 compares the results of A-BMLPR with those of simple PageRank. In all presented cases in the table, application-based PageRank has a higher or equal robustness versus simple PageRank. For instance, all 10-top ranked nodes found in the 7th snapshot are the ones which application-based PageRank found at the 10th snapshot, but in simple PageRank the number of common top-ranked nodes is 7 in 7th and 8th snapshots.

From another viewpoint, a method according to which nodes' influence does not change considerably at two successive snapshots is robust. Therefore, the more the number of common top k-ranked nodes by applying centrality measures to two successive snapshots is, the centrality is better. Based on the experiments done, on the average 82% of top 10-ranked nodes and 86% of top 100-ranked nodes using A-BMLPR between two successive snapshots are the same. This relatively high value at three-month intervals again supports the robustness of the proposed centrality measure.

6. Discussion

According to the experiments performed, finding important users in activity networks gives us more similarity to the real-world results. We also realized that detecting influential users based on different activity networks presents different rankings. This means that a user can be important based on one activity, but not such by another. Accordingly, we proposed Application-based centrality measures that can detect different influential users for different applications. A-BMLPR as the better method can appropriately differentiate between important users. Its standard deviation and dynamic range are more than those of in-degree, out-degree, closeness, Betweenness, PageRank, Authority, and eigenvector centralities. By studying nodes' ranking over time, we realized that the presented method is more robust over time. In addition to these good results, our methods can be considered the first centrality measures that can give different rankings for different applications.

There are several limitations to the current study. First, the data set used for analysis (8097 Facebook users) is not so big and not representative of Facebook users at large. Second, Facebook does not represent all the social relationships between people. In other words, Facebook users act in other social networks as well where the type of their acts and behavioral pattern may be different. In addition, social networks' users have different social relationships in the real world, which could not be analyzed due to inaccessibility of the information. Third, among visible activities (such as liking a post, posting and sharing new contents, commenting on a post, joining in groups, and tagging photos) and different hidden activities (such as profile browsing, social games, and chatting), four visible activities were studied. Although the experiments done did not consider all activities, the proposed methods works independently of social network activities. In other words, although analyzing the activities like, comment, post, and share have been carried out, any number of other activities could be input for calculating the influence of users in case of their information being available.

Based on the mentioned limitation, it is suggested that a bigger

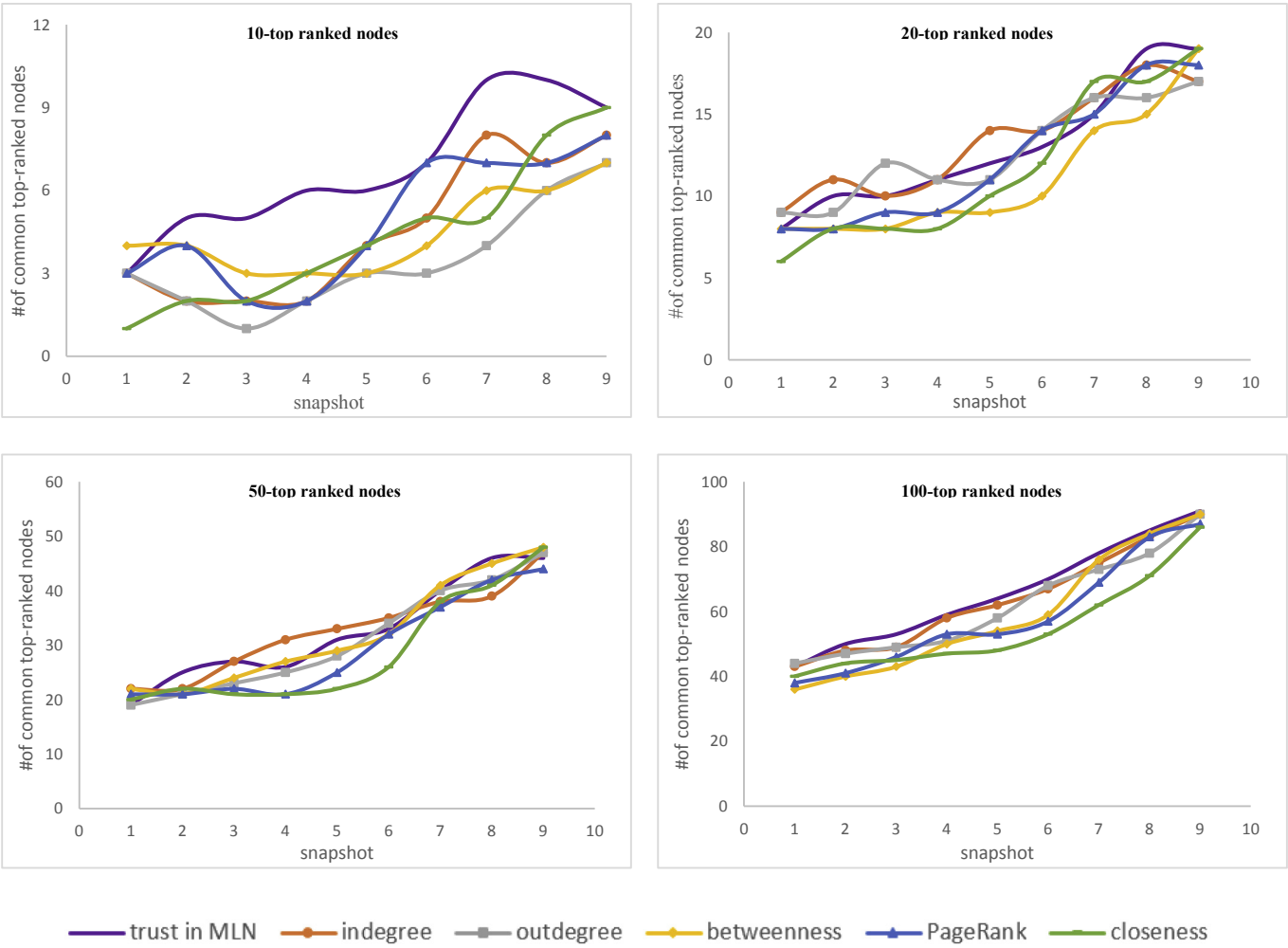

Fig. 4. Robustness of different centrality measures besides trust-based PageRank over time.

Table 13
Robustness of A-BMLPR versus simple PageRank.

	10 top-ranked		20 top-ranked		50 top-ranked		100 top-ranked	
	application-based PR	simple PR	application-based PR	simple PR	application-based PR	simple PR	application-based PR	simple PR
first snapshot	3	3	8	8	19	21	43	38
second snapshot	5	4	10	8	25	21	50	41
third snapshot	5	2	10	9	27	22	53	46
forth snapshot	6	2	11	9	26	21	59	53
fifth snapshot	6	4	12	11	31	25	64	53
sixth snapshot	7	7	13	14	33	32	70	57
seventh snapshot	10	7	15	15	40	37	78	69
eighth snapshot	10	7	19	18	46	42	85	83
ninth snapshot	10	9	19	18	46	44	93	90

data set be gained and the influential users therein be analyzed. For the future works, it is suggested that other Facebook activities, also the activities of users in different social networks and in the real world be input into the analyses and a comprehensive model be presented for measuring the influence of people in social networks. A-BMLPR may also be generalized to other applications such as information diffusion. In addition, since the basis of A-BMLPR is PageRank, presentation of an application-based ranking for other centrality measures such as hub and authority can reasonably be included among the future works.

7. Conclusion

According to this work, nodes' influence based on applying different centrality measures to the activity network is very different from that in the friendship network. The findings herein imply that we should use the activity network rather than the friendship network for measuring nodes' importance, particularly for applications such as social media marketing and advertising. Accordingly, we proposed two new types of centrality measures that not only can consider different activities in the multilayer

activity network simultaneously, but can also present different rankings suitable for different applications. The first type of the presented methods can measure nodes' importance for fame and spam applications. The second and more important type is Application-based Multi-layer PageRank (A-BMLPR). The results of the experiments showed that A-BMLPR has a higher capability for node-importance differentiation versus other centralities. This method is more robust over time and thus has more predictability, highly suggested for when we are considering a particular application of measuring nodes' importance.

Appendix A. Theoretical foundation

Graph: graph G is denoted as $G = (V,\ E)$ where $V = \{v_1, v_2, \ldots v_n\}$ is a set of nodes and $E = \{e_1, e_2, \ldots, e_m\}$ is a set of links. In undirected graphs, $e_i = \{u, v\}$ shows the links of the graph while in the directed one, $ei = (u, v)$.

Centrality measures: Node centrality measures calculate the importance of users based on their position in the social network. Some centrality measures such as betweenness, closeness, degree, and eigenvector are better to apply to undirected networks, and some others such as PageRank, Hub, Authority, Indegree, and Outdegree are better for directed networks.

Degree centrality: The simplest method to measure a node centrality is its degree, which in directed networks is calculated as in-degree and out-degree. In activity networks the in-degree and out-degree can be interpreted as the amount of nodes popularity and their activity level respectively. Generalizing degree in the weighted networks is called node strength, which is calculated as below (Opsahl et al., 2010):

$$C_{str}(i) = \sum_{j}^{N} W_{ij}$$

Closeness centrality: Closeness centrality sums the inverse geodesic-paths distance from a node to all other nodes in the network, and can be calculated as follows:

$$C_{close}(i) = \frac{1}{n-1} \sum_{j \neq i} \frac{1}{d_{ij}}$$

where d_{ij} is the distance of the shortest path between i and j (Newman, 2010).

Betweenness centrality: Betweenness centrality considers the number of the shortest paths which pass through the considered node as the importance of the node. It can describe the influence of a node in the entire network since it is known as a measure of gatekeeping, controlling the flow of information and brokering (Costenbader & Valente, 2003). Betweenness of node i may be calculated as below.

$$C_{Bet}(i) = \frac{1}{n^2} \sum_{\forall s,t \in V} \frac{n^i_{s,t}}{g_{s,t}}$$

where $n^i_{s,t}$ is the number of shortest paths from source s to target t which pass through node i and $g_{s,t}$ is the number of all shortest paths from s to t.

We may also generalize Betweenness and closeness to weighted networks by calculating shortest paths as below.

$$d^{w\alpha}(s,t) = \min\left(\frac{1}{(w_{sh})^{\alpha}} + \ldots + \frac{1}{(w_{ht})^{\alpha}}\right)$$

Where alpha is a positive tuning parameter (Costenbader & Valente, 2003).

Eigenvector centrality: Eigenvector centrality is a recursive notion of importance which proposes that a node is important if it has important neighbors. This measure is defined as the principle eigenvector of the adjacency matrix defining the network.

PageRank: PageRank is as one of the most important methods of measuring importance in directed networks, and was first suggested by Google and followed by some algorithms based on it afterwards. PageRank is used in the directed networks such as the web network and proposes that a node is important if it is linked to by many important nodes with a high PageRank and is calculated recursively as below:

$$C_{PR}(i) = d \cdot \sum_{j \to i} \frac{C_{PR}(j)}{k_j^{out}} + \frac{(1-d)}{n}$$

where d is a damping factor and Google has suggested to use d = 0.85.

Power-law distribution: Degree distribution is an important structural characteristic of a social network. In scale-free networks, as networks which follow the power-law degree distribution, the probability of nodes with degree k follows:

$$P(k) \propto k^{-\gamma} \quad \gamma > 1$$

where γ is a constant parameter of the distribution and called the power-law exponent (Clauset, Shalizi, & Newman, 2009).

Small-world network: In small-world networks the average distance is small compared to the size of the network. In other words, it increases logarithmically with the number of nodes.

Clustering coefficient: Clustering coefficient is a measure of the existence of local clustering within a network. It measures the average amount of how well connected nodes' neighbors with each other.

Reciprocity: Reciprocity represents the ratio of reciprocal links in a directed network, a number which in the activity network means if a person interacts with another, how probable is for the other to respond via the same activity. Obviously, if A has interacted with B, B's response would be more probable than random.

Activity: Different social networking platforms such as Facebook, Twitter, Linkedin, and Google + have presented different mechanisms for users' activity. Some of the activities are visible according to the website's policy and the user's privacy settings and some others are hidden. In this paper, we name any individual activity from the person as well as one-way and two-way interactions in the social network activity. Most Facebook activities could be considered directed, such as like, comment, post, share, and tagging. Some of them are mostly interactive, such as chatting, and some others are individual acts, particular to each user, such as status update. This paper is based on the four activities like, comment, post, and share.

Post: Facebook has implemented the post activity, so the users could present new contents such as photos, videos, and musics to their friends, or place contents on the profile of each other.

Comment: users can put comments on the posts of other users.

Like: if a Facebook user is interested in posts or comments of another user, he or she shows this interest by the like activity. However, there is no content-scoring mechanism or dislike in Facebook.

Share: users can share the contents presented by other users with their friends, with some possible changes.

Activity graph: This is a graph which models the users' activities and interactions. In this graph, the nodes show Facebook users and the links show their interactions. Depending on the application, the activity graph could be modeled weighted/unweighted and

directed/undirected.

References

Alahakoon, T., Tripathi, R., Kourtellis, N., Simha, R., & Iamnitchi, A. (2011). K-path centrality: A new centrality measure in social networks. In *Proceedings of the 4th workshop on social network systems* (p. 1).

Bader, D. A., Kintali, S., Madduri, K., & Mihail, M. (2007). Approximating betweenness centrality. In *Algorithms and models for the web-graph* (pp. 124–137). Springer.

Borgatti, S. P., Carley, K. M., & Krackhardt, D. (2006). On the robustness of centrality measures under conditions of imperfect data. *Social Networks, 28*, 124–136.

Campiteli, M. G., Holanda, A. J., Soares, L. D., Soles, P. R., & Kinouchi, O. (2013). Lobby index as a network centrality measure. *Physica A: Statistical Mechanics and its Applications, 392*, 5511–5515.

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, P. K. (2010). Measuring User Influence in Twitter: The Million Follower Fallacy. *ICWSM, 10*, 30.

Cheang, B., Chu, S. K. W., Li, C., & Lim, A. (2014). A multidimensional approach to evaluating management journals: Refining PageRank via the differentiation of citation types and identifying the roles that management journals play. *Journal of the Association for Information Science and Technology, 65*, 2581–2591.

Chen, D., Lü, L., Shang, M.-S., Zhang, Y.-C., & Zhou, T. (2012). Identifying influential nodes in complex networks. *Physica A: Statistical Mechanics and its Applications, 391*, 1777–1787.

Chun, H., Kwak, H., Eom, Y.-H., Ahn, Y.-Y., Moon, S., & Jeong, H. (2008). Comparison of online social relations in volume vs interaction: A case study of cyworld. In *Proceedings of the 8th ACM SIGCOMM conference on Internet measurement* (pp. 57–70).

Clauset, A., Shalizi, C. R., & Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM Review, 51*, 661–703.

Corbellini, A., Schiaffino, S., & Godoy, D. (2012). Intelligent analysis of user interactions in a collaborative software engineering context. In *Advances in new technologies, interactive interfaces and communicability* (pp. 114–123). Springer.

Costenbader, E., & Valente, T. W. (2003). The stability of centrality measures when networks are sampled. *Social Networks, 25*, 283–307.

Davidsen, S. A., & Ortiz-Arroyo, D. (2012). Centrality robustness and link prediction in complex social networks. In *Computational social networks* (pp. 197–224). Springer.

Ding, Y. (2011). Applying weighted PageRank to author citation networks. *Journal of the American Society for Information Science and Technology, 62*, 236–245.

Ding, Y., Yan, E., Frazho, A., & Caverlee, J. (2009). PageRank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology, 60*, 2229–2243.

Heidemann, J., Klier, M., & Probst, F. (2010). *Identifying key users in online social networks: A PageRank based approach*.

Hu, J., Wang, B., & Lee, D. (2010). Evaluating node importance with multi-criteria. In *Proceedings of the 2010 IEEE/ACM Int'l conference on green computing and communications & Int'l conference on cyber, physical and social computing* (pp. 792–797).

Kang, C., Molinaro, C., Kraus, S., Shavitt, Y., & Subrahmanian, V. (2012). Diffusion centrality in social networks. In *Proceedings of the 2012 international conference on advances in social networks analysis and mining (ASONAM 2012)* (pp. 558–564).

Kermani, M. A. M. A., Badiee, A., Aliahmadi, A., Ghazanfari, M., & Kalantari, H. (2016). Introducing a procedure for developing a novel centrality measure (Sociability Centrality) for social networks using TOPSIS method and genetic algorithm. *Computers in Human Behavior, 56*, 295–305.

Khadangi, E., Bagheri, A., & Shahmohammadi, A. (2016). Biased sampling from facebook multilayer activity network using learning automata. *Applied Intelligence*, 1–21.

Khadangi, E., Bagheri, A., & Zarean, A. (2017). Empirical analysis of structural properties, macroscopic and microscopic evolution of various facebook activity networks. *Quality & Quantity*. http://dx.doi.org/10.1007/s11135-016-0465-4.

Khadangi, E., Zarean, A., Bagheri, A., & Iafarabadi, A. B. (2013). Measuring relationship strength in online social networks based on users' activities and profile information. In *Computer and knowledge engineering (ICCKE), 2013 3th international econference on* (pp. 461–465).

Kiss, C., & Bichler, M. (2008). Identification of influencers—measuring influence in customer networks. *Decision Support Systems, 46*, 233–253.

Li, Y.-M., Lee, Y.-L., & Lien, N.-J. (2012). Online social advertising via influential endorsers. *International Journal of Electronic Commerce, 16*, 119–154.

McSherry, F. (2005). A uniform approach to accelerated PageRank computation. In *Proceedings of the 14th international conference on world wide web* (pp. 575–582).

Mochalova, A., & Nanopoulos, A. (2013). On the role of centrality in information diffusion in social networks. In *ECIS* (p. 101).

Newman, M. (2010). *Networks: An introduction*. Oxford University Press.

Opsahl, T., Agneessens, F., & Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks, 32*, 245–251.

Point, T. (2002). *How little things can make a big difference*. Malcolm Gladwell, Little, Brown.

Rogers, E. M. (2010). *Diffusion of innovations*. Simon and Schuster.

Shahmohammadi, A., Khadangi, E., & Bagheri, A. (2016). Presenting new collaborative link prediction methods for activity recommendation in Facebook. *Neurocomputing, 210*, 217–226.

Takes, F. W., & Kosters, W. A. (2011). Identifying prominent actors in online social networks using biased random walks. In *Proceedings of the 23rd benelux conference on artificial intelligence (BNAIC)* (pp. 215–222).

Tsiotas, D., & Polyzos, S. (2013). Introducing a new centrality measure from the transportation network analysis in Greece. *Annals of Operations Research*, 1–25.

Valente, T. W., Coronges, K., Lakon, C., & Costenbader, E. (2008). How correlated are network centrality measures? *Connections (Toronto, Ont.), 28*, 16.

Wang, A. H. (2010). Don't follow me: Spam detection in twitter. In *Security and cryptography (SECRYPT), proceedings of the 2010 international conference on* (pp. 1–10).

Xing, W., & Ghorbani, A. (2004). Weighted pagerank algorithm. In *Communication networks and services research, 2004. Proceedings. Second annual conference on* (pp. 305–314).

Xu, K., Li, J., & Song, Y. (2012). Identifying valuable customers on social networking sites for profit maximization. *Expert Systems With Applications, 39*, 13009–13018.

Yan, E., & Ding, Y. (2011). Discovering author impact: A PageRank perspective. *Information Processing & Management, 47*, 125–134.

Yan, X., Zhai, L., & Fan, W. (2013). C-index: A weighted network node centrality measure for collaboration competence. *Journal of Informetrics, 7*, 223–239.