Soc. Netw. Anal. Min. (2017) 7:22
https://doi.org/10.1007/s13278-017-0440-7

ORIGINAL ARTICLE

HellRank: a Hellinger-based centrality measure for bipartite social networks

Seyed Mohammad Taheri[1] · Hamidreza Mahyar[2] · Mohammad Firouzi[1] · Elahe Ghalebi K.[2] · Radu Grosu[2] · Ali Movaghar[1]

Received: 11 October 2016 / Revised: 3 May 2017 / Accepted: 6 May 2017 / Published online: 22 May 2017

Abstract Measuring centrality in a social network, especially in bipartite mode, poses many challenges, for example, the requirement of full knowledge of the network topology, and the lack of properly detecting top-*k behavioral representative users*. To overcome the above mentioned challenges, we propose HellRank, an accurate centrality measure for identifying central nodes in bipartite social networks. HellRank is based on the Hellinger distance between two nodes on the same side of a bipartite network. We theoretically analyze the impact of this distance on a bipartite network and find upper and lower bounds for it. The computation of the HellRank centrality measure can be distributed, by letting each node uses local information only on its immediate neighbors. Consequently, one does not need a central entity that has full knowledge of the network topological structure. We experimentally evaluate the performance of the HellRank measure in correlation with other centrality measures on real-world networks. The results show partial ranking similarity between the HellRank and the other conventional metrics according to the Kendall and Spearman rank correlation coefficient.

Keywords Bipartite social networks · Top-*k* central nodes · Hellinger distance · Recommender systems

✉ Hamidreza Mahyar
hmahyar@cps.tuwien.ac.at

Seyed Mohammad Taheri
mtaheri@ce.sharif.edu

Mohammad Firouzi
mfirouzi@ce.sharif.edu

Elahe Ghalebi K.
eghalebi@cps.tuwien.ac.at

Radu Grosu
radu.grosu@tuwien.ac.at

Ali Movaghar
movaghar@sharif.edu

[1] Department of Computer Engineering, Sharif University of Technology (SUT), Tehran, Iran

[2] Department of Computer Engineering, Vienna University of Technology (TU Wien), Vienna, Austria

1 Introduction

Social networking sites have become a very important social structure of our modern society with hundreds of millions of users nowadays. With the growth of information spread across various social networks, the question of " how to measure the *relative importance* of users in a social network?" has become increasingly challenging and interesting, as important users are more likely to be infected by, or to infect, a large number of users. Understanding users' behaviors when they connect to social networking sites creates opportunities for richer studies of social interactions. Also, finding a subset of users to statistically represent the original social network is a fundamental issue in social network analysis. This small subset of users (the behaviorally representative users) usually plays an important role in influencing the social dynamics on behavior and structure.

The centrality measures are widely used in social network analysis to quantify the *relative importance* of nodes within a network. The most central nodes are often the nodes that have more weight, both in terms of the number of interactions as well as the number of connections to other nodes (Silva et al. 2013). In social network analysis, such a centrality notion is used to identify influential users

(Mahyar 2015; Silva et al. 2013; Wei et al. 2013; Yustiawan et al. 2015), as the influence of a user is the ability to popularize a particular content in the social network. To this end, various centrality measures have been proposed over the years to rank the network nodes according to their topological and structural properties (Cha et al. 2010; Friedkin and Johnsen 2011; Zhao et al. 2014). These measures can be considered as several points of view with different computational complexity, ranging from low-cost measures (e.g., Degree centrality) to more costly measures (e.g., Betweenness and Closeness centralities) (Wehmuth and Ziviani 2013; Muruganantham et al. 2015). The authors of Stephenson and Zelen (1989) concluded that centrality may not be restricted to shortest paths. In general, the global topological structure of many networks is initially unknown. However, all these structural network metrics require full knowledge of the network topology (Wehmuth and Ziviani 2013; Mahyar et al. 2015a).

An interesting observation is that many real-world social networks have a bi-modal nature that allows the network to be modeled as a bipartite graph (see Fig. 1). In a bipartite network, there are two types of nodes and the links can only connect nodes of different types (Zhao et al. 2014). The *Social Recommender System* is one of the most important systems that can be modeled as a bipartite graph with users and items as the two types of nodes, respectively. In such systems, the centrality measures can have different interpretations from conventional centrality measures such as Betweenness, Closeness, Degree, and PageRank (Kitsak et al. 2010). The structural metrics, such as Betweenness and Closeness centrality, are known as the most common central nodes' identifier in one-mode networks, although in bipartite social networks they are not usually appropriate in identifying central users that are perfect representative for the bipartite network structure.

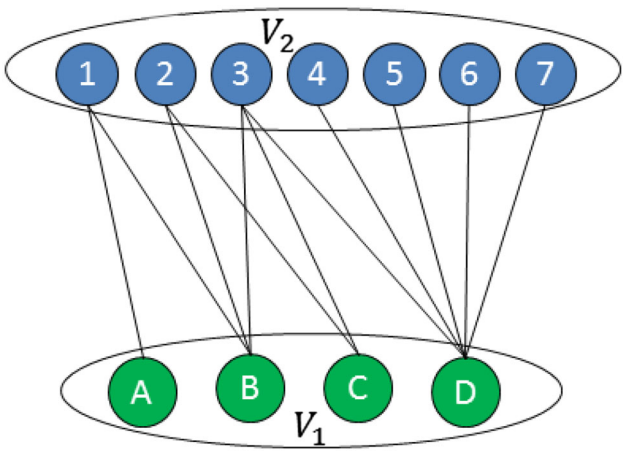

Fig. 1 Bipartite graph $G = (V_1, V_2, E)$ with two different nodes set $V_1 = \{A, B, C, D\}$, $V_2 = \{1, 2, 3, 4, 5, 6, 7\}$ and link set E that each link connects a node in V_1 to a node in V_2

For example, in a social recommender system (Mahyar et al. 2017; Taheri et al. 2017, that can be modeled by the network graph in Fig. 1, user $D \in V_1$ is associated with items that have too few connections and have been considered less often by other users; meanwhile user D is considered as the most central node based on these common centrality metrics, because it has more connections. However user $B \in V_1$ is much more a real representative than D in the network. In the real-world example of an online store, if one user buys a lot of goods, but these goods are low consumption, and another buys fewer goods, but these are widely, we treat the second user as being a synecdochic representative of all users of the store. This is quite different from a conventional centrality metric outcome.

Another interesting observation is that the common centrality measures are typically defined for non-bipartite networks. To use these measures in bipartite networks, different projection methods have been introduced to converting bipartite to monopartite networks (Zhou et al. 2007; Sawant and Pai 2013. In these methods, a bipartite network is projected by considering one of the two node sets, and if each pair of these nodes shares a neighbor in the network, two nodes will be connected in the projected one-mode network (Latapy et al. 2008; Liebig and Rao 2014. In the projected network of example that's shown in Fig. 1, user B is the most central node based on common monopartite centrality metrics, which seems that the more behavioral representative user is detected. One of the major challenges is that every link in a real network is formed independently, but this does not happen in the projected one-mode network. Because of lack of independency in the formation of links in the projected network, analysis of the metrics that use the random network (Bollobás 1984) as a basis for their approach, is difficult. Classic random networks are formed by assuming that links are being independent from each other (Opsahl 2013). The second challenge is that the projected bipartite network nodes tend to form *Cliques*. A clique is a fully connected subset of nodes that all of its members are neighbors. As a result, the metrics that are based on triangles (i.e., a clique on three nodes) in the network, can be inefficient (such as structural holes or clustering coefficient measures) (Lind et al. 2005; Opsahl 2013).

Despite the fact that the projected one-mode network is less informative than its corresponding bipartite representation, some of the measures for monopartite networks have been extended to bipartite mode (Kitsak and Krioukov 2011; Opsahl 2013). Moreover, because of requirement of full knowledge of network topology and lack of proper measure for detection of more behavioral representative users in bipartite social networks, the use of conventional centrality measures in the large-scale

networks (e.g., in recommender systems) is a challenging issue. In order to overcome the aforementioned challenges and retain the original information in bipartite networks, proposing an accurate centrality measure in such networks seems essential (Kettle 2012; Liebig and Rao 2014).

Motivated by these observations and taking into account users' importance indicators for detection of central nodes in social recommender systems, we introduce a new centrality measure, called HellRank. This measure identifies central nodes in bipartite social networks. HellRank is based on the *Hellinger distance* (Nikulin 2001), a type of *f-divergence* measure, that indicates structural similarity of each node to other network nodes. Hence, this distance-based measure is accurate for detecting the more behavioral representative nodes. We empirically show that nodes with high HellRank centrality measure have relatively high Degree, Betweenness and PageRank centrality measures in bipartite networks. In the proposed measure, despite of different objectives to identify central nodes, there is a partial correlation between HellRank and other common metrics.

The rest of the paper is organized as follows. In Sect. 2, we discuss related work on behavioral representative and influence identification mechanisms. We also discuss centrality measures for bipartite networks, and highlight the research gap between our objectives and previous works. In Sect. 3, we introduce our proposed measure to solve the problem of centrality in bipartite networks. Experimental results and discussions are presented in Sect. 4. We conclude our work and discuss the future works in Sect. 5.

2 Related work

We organize the relevant studies on social influence analysis and the problem of important users in three different categories. First, in Sect. 2.1, we study existing work on behavioral representative users detection methods in social networks. Second, in Sect. 2.2, we review previous mechanisms for identifying influential users in social networks by considering the influence as a measure of the relative importance. Third, in Sect. 2.3, we focus in more details on centrality measures for bipartite networks.

2.1 Behavioral representative users detection

Unlike influence maximization, in which the goal is to find a set of nodes in a social network who can maximize the spread of influence (Chen et al. 2009; Kempe et al. 2003), the objective of behavioral representative users detection is to identify a few average users who can statistically represent the characteristics of all users (Landauer 1988). Another type of related work is social influence analysis. Anagnostopoulos et al. (2008) and Singla and Richardson (2008) proposed methods to qualitatively measure the existence of influence. Crandall et al. (2008) studied the correlation between social similarity and influence. Tang et al. (2009a) presented a method for measuring the strength of such influence. The problem of sampling representative users from social networks is also relevant to graph sampling (Leskovec and Faloutsos 2006; Maiya and Berger-Wolf 2011; Ugander et al. 2013). Zhu et al. (2007) introduced a novel ranking algorithm called GRASSHOPPER, which ranks items with an emphasis on diversity. Their algorithm is based on random walks in an absorbing Markov chain. Benevenuto et al. (2009) presented a comprehensive view of user behavior by characterizing the type, frequency, and sequence of activities users engage in and described representative user behaviors in online social networks based on clickstream data. Giroire et al. (2008) found significant diversity in end-host behavior across environments for many features, thus indicating that profiles computed for a user in one environment yield inaccurate representations of the same user in a different environment. Maia et al. (2008) proposed a methodology for characterizing and identifying user behaviors in online social networks.

However, most existing work focused on studying the network topology and ignored the topic information. Sun et al. (2013) aimed to find representative users from the information spreading perspective and Ahmed et al. (2014) studied the network sampling problem in the dynamic environment. Papagelis et al. (2013) presented a sampling-based algorithm to efficiently explore a user's ego network and to quickly approximate quantities of interest. Davoodi et al. (2012) focused on the use of the social structure of the user community, user profiles and previous behaviors, as an additional source of information in building recommender systems. Tang et al. (2015) presented a formal definition of the problem of sampling representative users from social network.

2.2 Identifying influential users

Goyal et al. (2010) studied how to infer social probabilities of influence by developing an algorithm to scan over the log of actions of social network users using real data. Bharathi et al. (2007), Tang et al. (2009b) focused on the influence maximization problem to model the social influence on large networks. TwitterRank, as an extension of PageRank metric, was proposed by Weng et al. (2010) to identify influential users in Twitter. Chen et al. (2013) used the Susceptible-Infected-Recovered (SIR) model to examine the spreading influence of the nodes ranked by different influence measures. Xu et al. (2012) identified influencers using joint influence powers through Influence

network. Zhu (2013) identified influencial users by using user trust networks. Li et al. (2014) proposed the weighted LeaderRank technique by replacing the standard random walk to a biased random walk. Sun et al. (2015) presented a novel analysis on the statistical simplex as a manifold with boundary and applied the proposed technique to social network analysis to rank a subset of influencer nodes. Tang and Yang (2012) proposed a new approach to incorporate users' reply relationship, conversation content and response immediacy to identify influential users of online health care community. Du et al. (2014) used multi-attribute and homophily characteristics in a new method to identify influential nodes in complex networks.

In the specific area of identifying influential users in bipartite networks, Beguerisse Díaz et al. (2010) presented a dynamical model for rewiring in bipartite networks and obtained time-dependent degree distributions. Liebig and Rao (2014) defined a bipartite clustering coefficient by taking differently structured clusters into account, that can find important nodes across communities. The concept of clustering coefficient will be discussed in further detail in the Sect. 2.2.1.

2.2.1 Clustering coefficient

This measure shows the nodes' tendency to form clusters and has attracted a lot of attention in both empirical and theoretical work. In many real-world networks, especially social networks, nodes are inclined to cluster in densely connected groups (Opsahl and Panzarasa 2009). Many measures have been proposed to examine this tendency. In particular, the global clustering coefficient provides an overall assessment of clustering in the network (Luce and Perry 1949), and the local clustering coefficient evaluates the clustering value of the immediate neighbors of a node in the network (DiChristina 2007).

The global clustering coefficient is the fraction of two paths (i.e., three nodes connected by two links) that are closed by the presence of a link between the first and the third node in the network. The local clustering coefficient is the fraction of the links among a node's interactions over the maximum possible number of links between them (DiChristina 2007; Opsahl 2013).

Due to structural differences, applying these general clustering coefficients directly to a bipartite network, is clearly not appropriate (Borgatti and Everett 1997). Thus the common metrics were extended or redefined, and different clustering measures were defined in these networks. In one of the most common clustering coefficients in bipartite networks, 4-period density is measured instead of triangles (Robins and Alexander 2004; Zhang et al. 2008). However, this measure could not consider the triple closure concept in the clustering as it actually consists of two nodes. This kind of measure can only be a measure of the level of support between two nodes rather than the clustering of a group of nodes. Accordingly, Latapy et al. (2008) defined the notion of clustering coefficient for pairs of nodes capturing correlations between neighborhoods in the bipartite case. Addtionally, Opsahl (2013) considered the factor, C^*, as the ratio of the number of closed 4-paths (τ^*_Δ) to the number of 4-paths (τ^*), as:

$$C^* = \frac{\text{closed } 4-\text{paths}}{4-\text{paths}} = \frac{\tau^*_\Delta}{\tau^*} \tag{1}$$

2.3 Centrality measures for bipartite networks

Various definitions for centrality have been proposed in which centrality of a node in a network is generally interpreted as the relative importance of that node (Freeman 1978; Chen et al. 2013). Centrality measures have attracted a lot of attentions as a tool to analyze various kinds of networks (e.g., social, information, and biological networks) (Faust 1997; Kang 2011). In this section, we consider a set of well-known centrality measures including Degree, Closeness, Betweenness, Eigenvector and PageRank, all of them redefined for bipartite networks. Given bipartite network $G = (V_1, V_2, E)$, where V_1 and V_2 are the two sides of network with $|V_1| = n_1$ and $|V_2| = n_2$. The link set E includes all links connecting nodes of V_1 to nodes of V_2. For the network in Fig. 1, n_1 and n_2 are equal to 4 and 7, respectively. Let Adj be the adjacency matrix of this network, as shown below:

$$Adj(G) = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \end{array} \tag{2}$$

2.3.1 Degree centrality

In one-mode graphs, degree centrality of node i, d_i, is equal to the number of connections of that node. In bipartite graphs, it indicates the number of node's connections to members on the other side. For easier comparison, degree centrality is normalized: the degree of each node is divided by the size of the other node set. Let d_i^* be the normalized degree centrality of node i. This is equal to (Borgatti and Everett 1997; Faust 1997):

$$d_i^* = \frac{d_i}{n_2},\ d_j^* = \frac{d_j}{n_1};\quad i \in V_1, j \in V_2 \tag{3}$$

As the network size becomes increasingly large, employing degree centrality is the best option (You et al. 2015). On the other hand, this centrality is based on a highly local view around each node. As a consequence, we need more

informative measures that can further distinguish among nodes with the same degree (Kang 2011).

2.3.2 Closeness centrality

The standard definition of Closeness c_i for node i in monopartite networks, refers to the sum of geodesic distances from node i to all $n-1$ other nodes in the network with n nodes (Sabidussi 1966). For bipartite networks, this measure can be calculated using the same approach, but the main difference is normalization. Let c_i^* be the normalized Closeness centrality of node $i \in V_1$. This is equal to (Borgatti and Everett 1997; Faust 1997):

$$c_i^* = \frac{n_2 + 2(n_1 - 1)}{c_i}; \quad i \in V_1 \tag{4}$$

$$c_j^* = \frac{n_1 + 2(n_2 - 1)}{c_j}; \quad j \in V_2 \tag{5}$$

For the bipartite network shown in Fig. 1, normalized Closeness centrality of the nodes A, B, C, and D are respectively equal to 0.35, 0.61, 0.52 and 0.68. It specifies that node D is the most central node which says that Closeness centrality cannot help us very much in the objective to find more behavioral representative nodes in bipartite social networks.

2.3.3 Betweenness centrality

Betweenness centrality of node i, b_i, refers to the fraction of shortest paths in the network that pass through node i (Freeman 1977). In bipartite networks, maximum possible Betweenness for each node is limited by relative size of two nodes sets, as introduced by Borgatti and Halgin (2011):

$$\begin{aligned} b_{\max}(V_1) = & \frac{1}{2}[n_2^2(s+1)^2 \\ & + n_2(s+1)(2t-s-1) - t(2s-t+3)] \end{aligned} \tag{6}$$

where $s = (n_1 - 1)\ div\ n_2$ and $t = (n_1 - 1)\ mod\ n_2$; and

$$\begin{aligned} b_{\max}(V_2) = & \frac{1}{2}[n_1^2(p+1)^2 \\ & + n_1(p+1)(2r-p-1) - r(2p-r+3)] \end{aligned} \tag{7}$$

where $p = (n_2 - 1)\ div\ n_1$ and $r = (n_2 - 1)\ mod\ n_1$.

For the bipartite network shown in Fig. 1, normalized Betweenness centrality of the nodes A, B, C, and D are respectively equal to 0, 0.45, 0.71 and 0.71. It specifies that nodes C and D are the most central nodes which says that Betweenness centrality cannot help us much objective in finding more behavioral representative nodes in bipartite social networks.

2.3.4 Eigenvector and PageRank centrality

Another important centrality measure is Eigenvector centrality, which is defined as the principal eigenvector of adjacency matrix of the network. A node's score is proportional to the sum of the scores of its immediate neighbors. This measures exploits the idea that nodes with connections to high-score nodes are more central (Bonacich 1972). The eigenvector centrality of node i, e_i, is defined as follows (Faust 1997):

$$e_i = \lambda \sum a_{ij} e_j \tag{8}$$

where $Adj(G) = (a_{ij})_{i,j=1}^n$ denotes the adjacency matrix of the network with n nodes and λ is the principal eigenvalue of the adjacency matrix. Our interest regarding to the eigenvector centrality is particularly focused on the distributed computation of PageRank (You et al. 2015). The PageRank is a special case of eigenvector centrality when the adjacency matrix is suitably normalized to obtain a column stochastic matrix (You et al. 2015). HITS algorithm (Kleinberg 1999) is in the same spirit as PageRank. The difference is that unlike the PageRank algorithm, HITS only operates on a small subgraph from the web graph. HITS ranks the seed nodes according to their authority and hub weights.

The PageRank vector $R = (r_1, r_2, \ldots, r_n)^T$ is the solution of the following equation:

$$R = \frac{1-d}{n}.1 + dLR \tag{9}$$

where r_i is the PageRank of node i and n is the total number of nodes. d is a damping factor, set to around 0.85 and L is a modified adjacency matrix, such that $l_{i,j} = 0$ if and only if node j does not have a link to i and $\sum_{i=1}^n l_{i,j} = 1$, where $l_{i,j} = \frac{a_{i,j}}{d_j}$, and $d_j = \sum_{i=1}^n a_{i,j}$ is the out-degree of node j (Okamoto et al. 2008).

For the bipartite network shown in Fig. 1, normalized PageRank centrality of the nodes A, B, C and D are respectively equal to 0.05, 0.11, 0.08 and 0.21. It specifies that node D is the most central node which says that PageRank centrality cannot help us much objective to finding more behavioral representative nodes in bipartite social networks.

3 Proposed method

In this paper, we want to identify more behavioral representative nodes in bipartite social networks. To this end, we propose a new similarity-based centrality measure, called HellRank. Since the similarity measure is usually inverse of the distance metrics, we first choose a suitable distance

measure, namely Hellinger distance (Sect. 3.1). Then we apply this metric to bipartite networks. After that, we theoretically analyze the impact of the distance metric in the bipartite networks. Next, we generate a distance matrix on one side of the network. Finally, we compute the HellRank score of each node, accordingly to this matrix. As a result, the nodes with high HellRank centrality are more behavioral representative nodes in bipartite social networks.

3.1 Select a well-defined distance metric

When we want to choose a base metric, an important point is whether this measure is based on a well-defined mathematical metric. We want to introduce a similarity-based measure for each pair of nodes in the network. So, we choose a proper distance measure as base metric, because the similarity measures are in some sense the inverse of the distance metrics. A true distance metric must have several main characteristics. A metric with these characteristics on a space induces topological properties (like open and closed sets). It leads to the study of more abstract topological spaces. Hunter (2012) introduced the following definition for a distance metric.

Definition 1 A metric space is a set X that has a notion of the *distance function* $d(x, y)$ between every pair of points $x, y \in X$. A *well-defined distance metric* d on a set X is a function $d : X \times X \rightarrow \mathbb{R}$ such that for all $x, y, z \in X$, three properties hold:

1. *Positive Definiteness*: $d(x, y) \geq 0$ and $d(x, y) = 0$ if and only if $x = y$;
2. *Symmetry*: $d(x, y) = d(y, x)$;
3. *Triangle Inequality*: $d(x, y) \leq d(x, z) + d(z, y)$.

We define our distance function as the difference between probability distribution for each pair of nodes based on f-divergence function, which is defined by:

Definition 2 An *f-divergence* is a function $D_f(P||Q)$ that measures the difference between two probability distributions P and Q. For a convex function f with $f(1) = 0$, the *f-divergence* of Q from P is defined as (Csiszár and Shields 2004):

$$D_f(P||Q) = \int_{\Omega} f\left(\frac{dP}{dQ}\right) dQ \tag{10}$$

where Ω is a sample space, which is the set of all possible outcomes.

In this paper, we use one type of the *f-divergence* metric, called *Hellinger distance* (aka *Bhattacharyya distance*), that was introduced by *Ernst Hellinger* in 1909 (Nikulin 2001). In probability theory and information theory, *Kullback–Leibler divergence* Kullback and Leibler (1951) is a more common measure of difference between two probability distributions, however it does not satisfy both the symmetry and the triangle inequality conditions (Van der Vaart 2000). Thus, this measure is not intuitively appropriate to explain similarity in our problem. As a result, we choose Hellinger distance to quantify the similarity between two probability distributions (Van der Vaart 2000). For two discrete probability distributions $P = (p_1, \ldots, p_m)$ and $Q = (q_1, \ldots, q_m)$, in which m is length of the vectors, Hellinger distance is defined as:

$$D_H(P||Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{m} (\sqrt{p_i} - \sqrt{q_i})^2} \tag{11}$$

It is obviously related to the Euclidean norm of the difference of the square root of vectors, as:

$$D_H(P||Q) = \frac{1}{\sqrt{2}} \|\sqrt{P} - \sqrt{Q}\|_2 \tag{12}$$

3.2 Applying Hellinger distance in bipartite networks

In this section, we want to apply the Hellinger distance to a bipartite network for measuring the similarity of the nodes on one side of the network. Assume x is a node in a bipartite network in which its neighborhood is $N(x)$ and its degree is $deg(x) = |N(x)|$. Suppose that the greatest node degree of the network is Δ. Let l_i be the number of x's neighbors with degree of i. Suppose the vector $L_x = (l_1, \ldots, l_\Delta)$ be the non-normalized distribution of l_i for all adjacent neighbors of x. Now, we introduce the Hellinger distance between two nodes x and y on one side of the bipartite network as follows:

$$d(x, y) = \sqrt{2}\, D_H(L_x \| L_y) \tag{13}$$

The function $d(x, y)$ represents the difference between two probability distribution of L_x and L_y. To the best of our knowledge, this is the first work that introduces the Hellinger distance between each pair of nodes in a bipartite network, using degree distribution of neighbors of each node.

3.3 Theoretical analysis

In this section, we first express the Hellinger distance for all positive real vectors to show that applying this distance to bipartite networks still satisfies its metricity (lemma 1) according to Definition 1. Then, we find an upper and a lower bound for the Hellinger distance between two nodes of bipartite network.

Lemma 1 Hellinger distance for all positive real vectors is a well-defined distance metric function.

Proof Based on the true metric properties in Definition 1, for two probability distribution vectors P and Q, the following holds:

$$D_H(P\|Q) \geq 0 \tag{14}$$

$$D_H(P\|Q) = 0 \quad \Leftrightarrow \quad P = Q \tag{15}$$

$$D_H(P\|Q) = D_H(Q\|P) \tag{16}$$

If we have another probability distribution R similar to P and Q, then according to the triangle inequality in norm 2, we should have:

$$\frac{1}{\sqrt{2}}\|\sqrt{P} - \sqrt{Q}\|_2 \leq \frac{1}{\sqrt{2}}(\|\sqrt{P} - \sqrt{R}\|_2 + \|\sqrt{R} - \sqrt{Q}\|_2) \Rightarrow D_H(P\|Q) \leq D_H(P\|R) + D_H(R\|Q) \tag{17}$$

It shows that the triangle inequality in Hellinger distance for all positive real vectors is a well-defined distribution metric function. □

Using this distance measure, we have the ability to detect differences between local structures of nodes. In other words, this distance expresses similarity between the local structures of two nodes. If we normalize the vectors (i.e., sum of the elements equals to one), then differences between local structures of nodes may not be observed. For example, there does not exist any distance between node x with $deg(y) = 10$ that its neighbors' degree are 2 and node y with $deg(y) = 1$ that its neighbor's degree are 2. Therefore, our distance measure with vectors normalization is not proper for comparing two nodes.

Then, we claim that if the difference between two nodes' degree is greater (or smaller) than a certain value, the distance between these nodes cannot be less (or more) than a certain value. In other words, their local structures cannot be similar more (or less) than a certain value. In the following theorem, we find an upper and a lower bound for the Hellinger distance between two nodes on one side of a bipartite network using their degrees' difference.

Theorem 1 If we have two nodes x and y on one side of a bipartite network, such that $deg(x) = k_1$, $deg(y) = k_2$, and $k_1 \geq k_2$, then we have a lower bound for the distance between these nodes as:

$$d(x, y) \geq \sqrt{k_1} - \sqrt{k_2} \tag{18}$$

and an upper bound as:

$$d(x, y) \leq \sqrt{k_1 + k_2} \tag{19}$$

Proof To prove the theorem, we use the Lagrange multipliers. Suppose $L_x = (l_1, \ldots, l_\Delta)$ and $L_y = (h_1, \ldots, h_\Delta)$ are positive real distribution vectors of nodes x and y. Based on (13) we know $d(x, y) = \sqrt{2}\, D_H(L_x\|L_y)$, so one can minimize the distance between these nodes by solving $\min_{L_x, L_y} \sqrt{2}\, D_H(L_x\|L_y)$, which is equivalent to find the minimum square of their distance:

$$\min_{L_x, L_y} 2D_H^2(L_x\|L_y) = \min_{L_x, L_y} \sum_{i=1}^{\Delta} \left(\sqrt{l_i} - \sqrt{h_i}\right)^2$$

So, Lagrangian function can be defined as follows:

$$F(L_x, L_y, \lambda_1, \lambda_2) = \sum_{i=1}^{\Delta} \left(\sqrt{l_i} - \sqrt{h_i}\right)^2 + \lambda_1 \left(k_1 - \sum_{i=1}^{\Delta} l_i\right) + \lambda_2 \left(k_2 - \sum_{i=1}^{\Delta} h_i\right)$$

Then, we take the first derivative with respect to l_i:

$$\frac{\partial F}{\partial l_i} = 1 - \frac{\sqrt{h_i}}{\sqrt{l_i}} - \lambda_1 = 0 \Rightarrow h_i = l_i(1 - \lambda_1)^2$$

Due to $\sum_{i=1}^{\Delta} l_i = k_1$ and $\sum_{i=1}^{\Delta} h_i = k_2$, we have:

$$\sum_{i=1}^{\Delta} h_i = k_2 \rightarrow \sum_{i=1}^{\Delta} l_i(1 - \lambda_1)^2 = k_2 \rightarrow (1 - \lambda_1) = \pm\sqrt{\frac{k_2}{k_1}}$$

But in order to satisfy $\sqrt{h_i} = \sqrt{l_i}(1 - \lambda_1)$, the statement $1 - \lambda_1$ must be positive, thus:

$$h_i = l_i(1 - \lambda_1)^2 = l_i \frac{k_2}{k_1}$$

After derivation with respect to h_i, we also reach similar conclusion. If this equation is true, then equality statement for minimum function will occur, as:

$$\begin{aligned} \min_{L_x, L_y} 2D_H^2(L_x\|L_y) &= \sum_{i=1}^{\Delta} \left(\sqrt{l_i} - \sqrt{\frac{k_2}{k_1}}\sqrt{l_i}\right)^2 \\ &= \sum_{i=1}^{\Delta} l_i \left(1 - \sqrt{\frac{k_2}{k_1}}\right)^2 \\ &= (1 - \sqrt{\frac{k_2}{k_1}})^2 \sum_{i=1}^{\Delta} l_i \\ &= \left(1 - \sqrt{\frac{k_2}{k_1}}\right)^2 k_1 \quad \Rightarrow \\ \min_{L_x, L_y} \sqrt{2}\, D_H(L_x\|L_y) &= \sqrt{k_1}\left(1 - \sqrt{\frac{k_2}{k_1}}\right) \\ &= \sqrt{k_1} - \sqrt{k_2} \end{aligned}$$

So, the lower bound for distance of any pair of nodes on one side of the bipartite network could not be less than a certain value by increasing their degrees difference.

Now, we want to find an upper bound according to Equation (19). As we know, the following statement is true for any p_i, p_j and q_i, q_j:

$$\begin{aligned}&\left(\sqrt{p_i}-\sqrt{q_i}\right)^2+\left(\sqrt{p_j}-\sqrt{q_j}\right)^2\\&\quad\le\left(\sqrt{p_i+p_j}-0\right)^2+\left(\sqrt{q_i+q_j}-0\right)^2\\&\quad=p_i+p_j+q_i+q_j\end{aligned}$$

Suppose in our problem, $p_i = l_i$, $p_j = l_j$, and $q_i = h_i$, $q_j = h_j$, then this inequality holds for any two pairs of elements in L_x and L_y. Eventually we have:

$$d(x,y)\le\sqrt{\left(\sqrt{k_1}-0\right)^2+\left(\sqrt{k_2}-0\right)^2}=\sqrt{k_1+k_2}$$

We can conclude that it is not possible for any pair of nodes on one side of the bipartite network that their distance to be more than a certain value by increasing their degrees. □

As a result, we found the upper and the lower bounds for the Hellinger distance between two nodes on one side of the bipartite network using their degrees' difference.

3.3.1 An example with probabilistic view

In this example, we want to analyze the similarity among nodes based on Hellinger distance information in an artificial network. We examine how we can obtain required information for finding similar nodes to a specific node x as the expected value and variance of the Hellinger distance. Suppose that in a bipartite artificial network with $|V_1| = n_1$ nodes on one side and $|V_2| = n_2$ nodes on the other side, nodes in V_1 is connected to nodes in V_2 using Erdös–Rényi model $G(m, p)$. In other words, there is an link with probability p between two sides of the network. Distribution function L_x of node $x \in V_1$ can be expressed as a multinomial distribution form as:

$$\begin{aligned}P(l_1,\ldots,l_\Delta|deg(x)=k)&=P(L_x|deg(x)=k)\\&=\binom{k}{l_1,\ldots,l_\Delta}\prod P_i^{l_i}\end{aligned}\tag{20}$$

where $P_i=\binom{n_2-1}{i-1}p^{i-1}(1-p)^{n_2-i}$ is a binomial distribution probability $B(n_2,p)$ for x's neighbors that their degree is equal to i.

According to the central limit theorem (Johnson 2004), binomial distribution converges to a Poisson distribution $Pois(\lambda)$ with parameter $\lambda = (n_2 - 1)p$ and the assumption that $(n_2 - 1)p$ is fixed and n_2 increases. Therefore, average distribution of $P(L_x|deg(x) = k)$ will be $\mu = (kp_1, kp_2, \ldots, kp_\Delta)$. In addition, degree distribution in Erdös–Rényi model converges to Poisson distribution by increasing n_1 and n_2 ($\lambda = n_1 p$ for one side of network and $\lambda = n_2 p$ for another one).

The limit of average distribution of $P(L_x|deg(x) = k)$ by increasing Δ, approaches k times of a Poisson distribution. Thus, normalized L_x vector is a Poisson distribution with parameter $\lambda = (n_2 - 1)p$. To find a threshold for positioning similar and closer nodes to node x, we must obtain expectation and variance of the Hellinger distance between x and the other nodes in node set V_1. Before obtaining these values, we mention the following lemma to derive equal expression of Hellinger distance and difference between typical mean and geometric mean.

Lemma 2 Suppose two distribution probability vectors $P = (p_1, \ldots, p_m)$ and $Q = (q_1, \ldots, q_m)$ that P is k_1 times of a Poisson distribution probability vector $P_1 \sim \text{Poisson}(\lambda_1)$ and Q is k_2 times of a Poisson distribution probability vector $P_2 \sim \text{Poisson}(\lambda_2)$[1]. The square of Hellinger distance between P and Q is calculated by:

$$D_H^2(P\|Q)=\frac{k_1+k_2}{2}-\sqrt{k_1k_2}\left(1-e^{-\frac{1}{2}\left(\sqrt{\lambda_1}-\sqrt{\lambda_2}\right)^2}\right)\tag{21}$$

Proof The squared Hellinger distance between two Poisson distributions P_1 and P_2 with rate parameters λ_1 and λ_2 is (Torgersen 1991):

$$D_H^2(P_1\|P_2)=1-e^{-\frac{1}{2}\left(\sqrt{\lambda_1}-\sqrt{\lambda_2}\right)^2}\tag{22}$$

Therefore, the squared Hellinger distance for probability vectors P and Q, will be equal to $(\sum_{i=1}^m p_i = k_1, \sum_{i=1}^m q_i = k_2)$:

$$\begin{aligned}D_H^2(P\|Q)&=\frac{1}{2}\sum_{i=1}^m\left(\sqrt{p_i}-\sqrt{q_i}\right)^2\\&=\frac{1}{2}\sum_{i=1}^m\left(p_i+q_i-2\sqrt{p_iq_i}\right)\\&=\frac{k_1+k_2}{2}-\sqrt{k_1k_2}\left(1-e^{-\frac{1}{2}\left(\sqrt{\lambda_1}-\sqrt{\lambda_2}\right)^2}\right)\end{aligned}\tag{23}$$

□

However, in the special case of $\lambda_1 = \lambda_2$, we have:

$$D_H^2(P\|Q)=\frac{k_1+k_2}{2}-\sqrt{k_1k_2}\tag{24}$$

It means that the squared Hellinger distance is equal to difference between typical mean and geometric mean.

To calculate the second moment of distance between node $x \in V_1$ and any other nodes $z \in V_1$ in the same side of the bipartite network based on the lemma 2, we have:

[1] Vector P=$(p_0, p_1, \ldots)$ is a Poisson distribution probability vector such that the probability of the random variable with Poisson distribution being i is equal to p_i.

$$
\begin{aligned}
&E_{z\in V_1}\left[d^2(x,z)\right] \\
&\quad = E\left[2\ D_H^2(L_x\|L)\right] \\
&\quad = \sum_{i=1}^{\infty}\left(\frac{e^{-n_1p}(n_1p)^i}{(n_1p)!}(k+i-2\sqrt{ki})\right) \\
&\quad \simeq \sum_{i=1}^{n_2}\left(\frac{e^{-n_1p}(n_1p)^i}{(n_1p)!}(k+i-2\sqrt{ki})\right)
\end{aligned}
\tag{25}
$$

where $L = (L_z|z \in V_1)$ and the infinite can be approximated by n_2 elements. Similarly, for distance expectation we have:

$$
E\left[\sqrt{2}\,D_H(L_x\|L)\right] \simeq \sum_{i=1}^{n_2}\left(\frac{e^{-n_1p}(n_1p)^i}{(n_1p)!}\sqrt{(k+i-2\sqrt{ki})}\right) \tag{26}
$$

In addition, variance can also be obtained based on these calculated moments:

$$
Var_{z\in V_1}(d(x,z)) = E_{z\in V_1}\left[d^2(x,z)\right] - \left(E_{z\in V_1}[d(x,z)]\right)^2 \tag{27}
$$

Hence, using these parameters, the required threshold for finding similar nodes to a specific node x, can be achieved. If we want to extend our method to more complex and realistic networks, we can assume that distribution L_x is a multiple of Poisson distribution (or any other distribution) vector with parameter λ_x, in which λ_x can be extracted by either the information about structure of the network or appropriate maximum likelihood estimation for node x. Therefore, the threshold will be more realistic and consistent with the structure of the real-world networks.

3.3.2 Generalization to weighted bipartite networks

The introduced distance metric function can be extended to weighted networks. The generalized Hellinger distance between two nodes of the weighted bipartite network can be considered as:

$$
d(x,y) = \sqrt{2}D_H(W_x\|W_y) \tag{28}
$$

where $W_x = (w'_1,\ldots,w'_\Delta)$, $w'_i = \sum_{\substack{j\in N(x)\\ deg(j)=i}} w_j$, and w_j is the vector of weights on the links of the network.

3.4 Rank prediction via HellRank

In this Section, we propose a new Hellinger-based centrality measure, called HellRank, for the bipartite networks. Now, according to the Sect. 3.2, we find the Hellinger distances between any pair of nodes in each side of a bipartite network. Then we generate an $n_1 \times n_1$ distance matrix (n_1 is the number of nodes in one side of network). The Hellinger distance matrix of G shown in Fig. 1 is as follows:

$$
Hell\text{-}Matrix(G) = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array}
\begin{array}{c}
\begin{array}{cccc} A & B & C & D \end{array} \\
\begin{pmatrix} 0 & 0.42 & 0.54 & 1 \\ 0.42 & 0 & 0.12 & 0.86 \\ 0.54 & 0.12 & 0 & 0.82 \\ 1 & 0.86 & 0.82 & 1 \end{pmatrix}
\end{array}
$$

According to the well-defined metric features (in Sect. 3.1) and the ability of mapping to Euclidean space, we can cluster nodes based on their distances. It means that any pair of nodes in the matrix with a less distance can be placed in one cluster by specific neighborhood radius. By averaging inverse of elements for each row in the distance matrix, we get final similarity score (HellRank) for each node of the network, by:

$$
HellRank(x) = \frac{n_1}{\sum_{z\in V_1} d(x,z)} \tag{29}
$$

Let $HellRank^*(x)$ be the normalized HellRank of node x that is equal to:

$$
HellRank^*(x) = HellRank(x).\min_{z\in V_1}(HellRank(z))
$$

where ' . ' denotes the multiplication dot, and $\min_{z\in V_1}(HellRank(z))$ is the minimum possible HellRank for each node

A similarity measure is usually (in some sense) the inverse of a distance metric: they take on small values for dissimilar nodes and large values for similar nodes. The nodes in one side with higher similarity scores represent more behavioral representation of that side of the bipartite network. In other words, these nodes are more similar than others to that side of the network. HellRank actually indicates structural similarity for each node to other network nodes. For the network shown in Fig. 1, according to Hellinger distance matrix, normalized HellRank of nodes A, B, C, and D are respectively equal to 0.71, 1, 0.94, and 0.52. It is clear that among all of the mentioned centrality measures in Sect. 2.2, only HellRank considers node B as a more behavioral representative node. Hence, sorting the nodes based on their HellRank measures will have a better rank prediction for nodes of the network. The nodes with high HellRank is more similar to other nodes. In addition, we find nodes with less scores to identify very specific nodes which are probably very different from other nodes in the network. The nodes with less HellRank are very dissimilar to other nodes on that side of the bipartite network.

4 Experimental evaluation

In this section, we experimentally evaluate the performance of the proposed HellRank measure in correlation with other centrality measures on real-world

networks. After summarizing datasets and evaluation metrics used in the experiments, the rest of this section addresses this goal. Finally, we present a simple example of mapping the Hellinger distance matrix to the Euclidean space to show clustering nodes based on their distances.

4.1 Datasets

To examine a measure for detection of central nodes in a two-mode network, South Davis women (Davis et al. 2009), is one of the most common bipartite datasets. This network has a group of women and a series of events as two sides of the network. A woman linked to an event if she presents at that event. Another data set used in the experiments is OPSAHL-collaboration network (Newman 2001), which contains authorship links between authors and publications in the arXiv condensed matter Section (-cond-mat) with 16726 authors and 22015 articles. A link represents an authorship connecting an author and a paper.

4.2 Evaluation metrics

One of the most popular evaluation metrics for comparison of different node ranking measures is Kendall's rank correlation coefficient (τ). In fact, Kendall is nonparametric statistic that is used to measure statistical correlation between two random variables (Abdi 2007):

$$\tau = \frac{N_{<\text{concordant pairs}>} - N_{<\text{discordant pairs}>}}{\frac{1}{2}n(n-1)} \tag{30}$$

where $N_{<S>}$ is the size of set S.

Another way to evaluate ranking measures is binary vectors for detection of top-k central nodes. All of vector's elements are zero by default and only top-k nodes' values are equal to 1. To compare ranking vectors with the different metrics, we use Spearman's rank correlation coefficient (ρ) that is a nonparametric statistics to measure the correlation coefficient between two random variables (Lehamn et al. 2005):

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{31}$$

where x_i and y_i are ranked variables and $\bar{x}$ and $\bar{y}$ are mean of these variables.

4.3 Correlation between HellRank and the common measures

We implement our proposed HellRank measure using available tools in NetworkX (Hagberg et al. 2008). To compare our measure to the other common centrality measures such as Latapy Clustering Coefficient [Sect. 2.2.1], Bipartite Degree [Sect. 2.3.1], Bipartite Closeness [Sect. 2.3.2], Bipartite Betweenness [Sect. 2.3.3], and PageRank [Sect. 2.3.4], we perform the tests on Southern Davis Women dataset. In Fig. 2, we observe the obtained ratings of these metrics (normalized by maximum value) for 18 women in the Davis dataset. In general, approximate correlation can be seen between the proposed HellRank metric and the other conventional metrics in the women scores ranking. It shows that despite different objectives to identify the central users, there is a partial correlation between HellRank and the other metrics.

Figure 3 shows scatter plots of standard metrics versus our proposed metric, on the Davis bipartite network. Each point in the scatter plot corresponds to a women node in the network. Across all former metrics, there exist clear linear correlations between each two measures. More importantly, because of the possibility of distributed computation of HellRank over the nodes, this metric can also be used in billion-scale graphs, while many of the most common metrics such as Closeness or Betweenness are limited to small networks (Wehmuth and Ziviani 2013). We observe that high HellRank nodes have high bipartite Betweenness, bipartite Degree, and bipartite Closeness. This reflects that high HellRank nodes have higher chance to reach all nodes within short number of steps, due to its larger number of connections. In contrast with high HellRank nodes, low HellRank nodes have various Latapy CC and projected Degree values. This implies that the nodes which are hard to be differentiated by these measures can be easily separated by HellRank.

To have a more analysis of the correlations between measures, we use Kendall between ranking scores provided by different methods in Table 1 and Spearman's rank correlation coefficient between top $k = 5$ nodes in Table 2. These tables illustrate the correlation between each pair in bipartite centrality measures and again emphasizes this point that despite different objectives to identify the central users, there is a partial correlation between HellRank and other common metrics.

In the next experiment, we compare the top-k central users rankings produced by Latapy CC, PageRank, Bipartite, and projected one-mode Betweenness, Degree, Closeness, and HellRank with different values of k. We employ Spearman's rank correlation coefficient measurement to compute the ranking similarity between two top-k rankings. Figure 4 presents result of Spearman's rank correlation coefficient between the top-k rankings of HellRank and the other seven metrics, in terms of different values of k. As shown in the figure, the correlation values of top k nodes in all rankings, reach almost constant limit at a specific value of k. This certain amount of k is

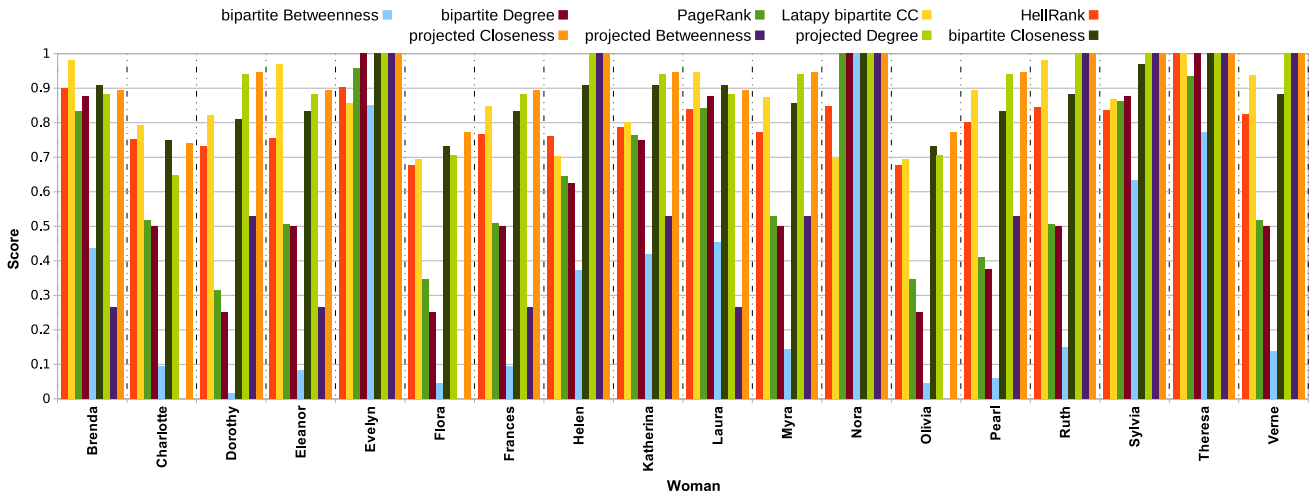

Fig. 2 Comparison between rankings for all women in the Davis dataset based on various centrality metrics

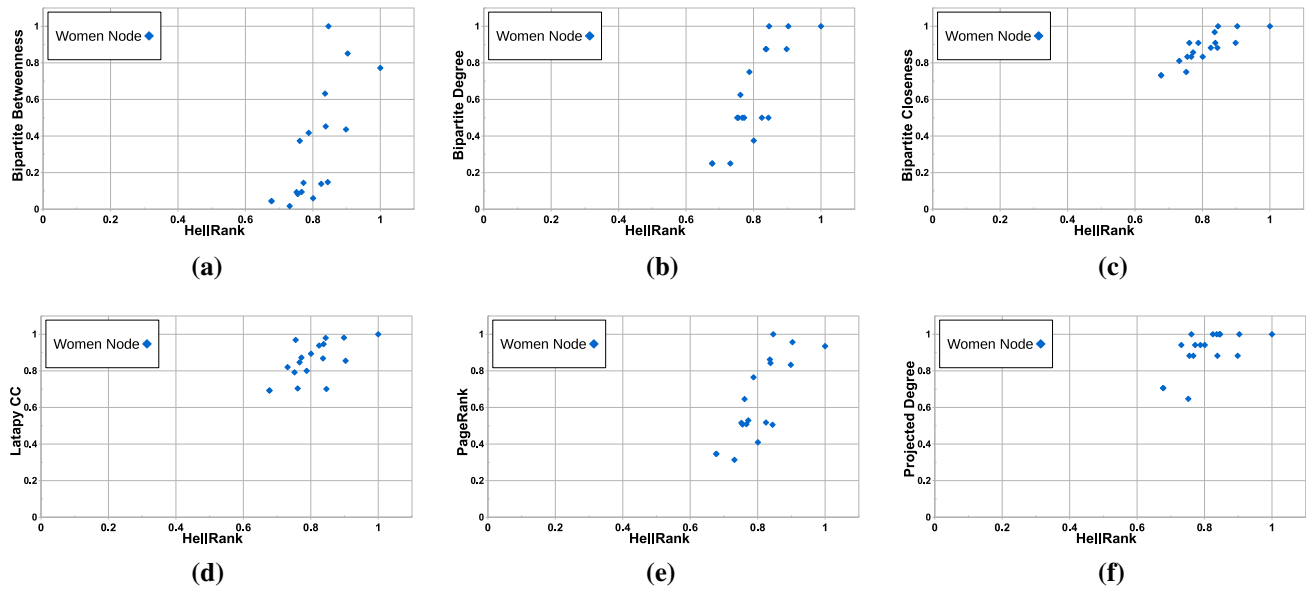

Fig. 3 The correlations between HellRank and the other standard centrality metrics on Davis (normalized by max value). **a** HellRank versus bipartite betweenness. **b** HellRank versus bipartite degree. **c** HellRank versus bipartite closeness. **d** HellRank versus Latapy CC. **e** HellRank versus PageRank. **f** HellRank versus projected Degree

Table 1 Comparison ratings results based on Kendall score in women nodes of Davis dataset [(2) means bipartite measures and (1) means projected one-mode measures]

Method	Latapy CC	Degree(2)	Betweenness(2)	Closeness(2)	PageRank	Degree(1)	Betweenness(1)	Closeness(1)
HellRank	0.51	0.7	0.67	0.74	0.59	0.51	0.53	0.51

approximately equal to 4 for all metrics. This means that the correlation does not increase at a certain threshold for $k = 4$ in the Davis dataset.

To evaluate the HellRank on a larger dataset, we repeated all the mentioned experiments for the arXiv cond-mat dataset. The scatter plots of standard metrics versus HellRank metric can be seen in Fig. 5. The results show that there exist almost linear correlations between the two measures in Bipartite Betweenness, Bipartite Degree and PageRank. In contrast to these metrics, HellRank has not correlation with other metrics such as Bipartite Closeness, Latapy Clustering Coefficient and

Table 2 Comparison top $k = 5$ important nodes based on Spearman's correlation in women nodes of Davis dataset [(2) means bipartite and (1) means projected one-mode]

Method	Latapy CC	Degree(2)	Betweenness(2)	Closeness(2)	PageRank	Degree(1)	Betweenness(1)	Closeness(1)
HellRank	0.44	0.72	0.44	0.72	0.44	0.44	0.44	0.44

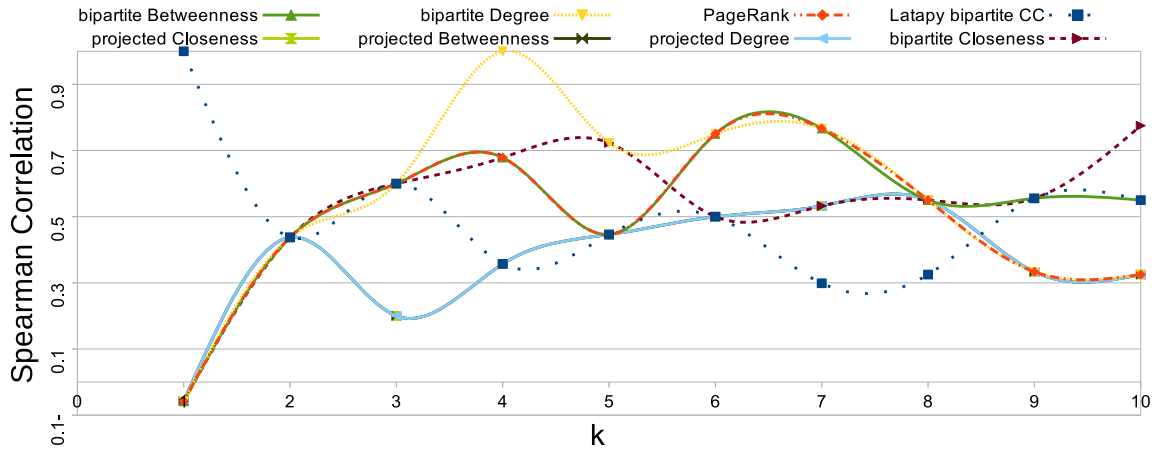

Fig. 4 Spearman's rank correlation with different values of k in Davis dataset

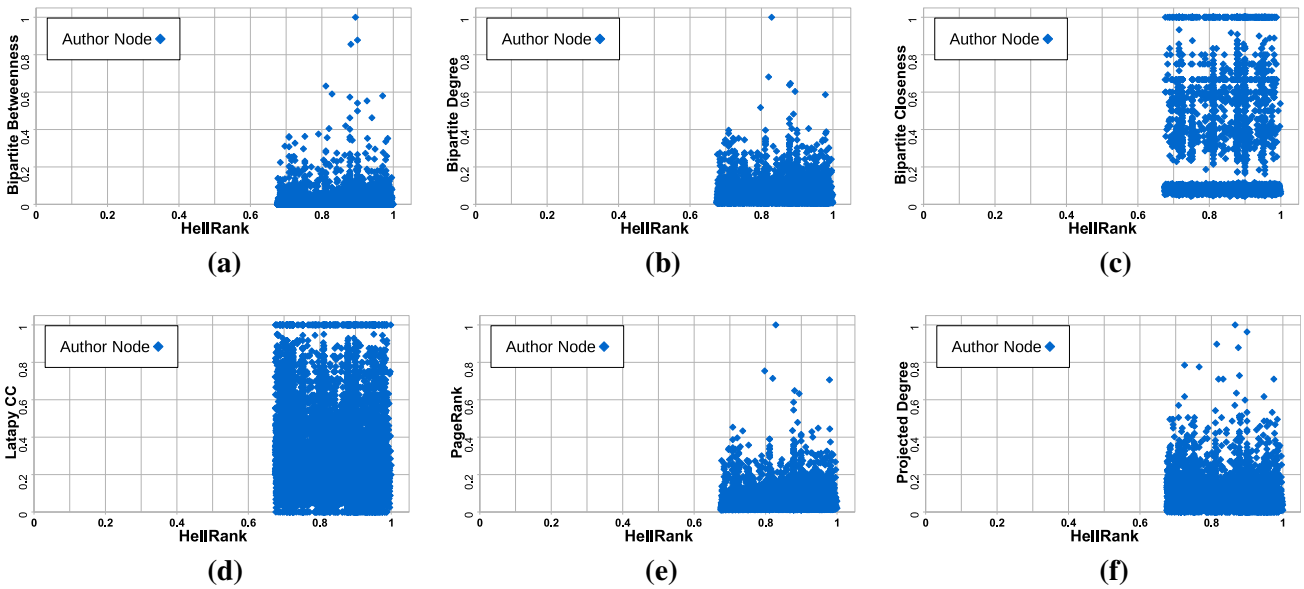

Fig. 5 The correlations between HellRank and the other standard centrality metrics on Opsahl (normalized by max value). **a** HellRank versus bipartite betweenness. **b** HellRank versus bipartite degree. **c** HellRank versus bipartite closeness. **d** HellRank versus Latapy CC. **e** HellRank versus PageRank. **f** HellRank versus projected Degree

Projected Degree. This implies that nodes that are hard to be differentiated by these metrics, can be separated easily by HellRank metric.

Moreover, Spearman's rank correlation with different values of k in arXiv cond-mat dataset can be seen in Fig. 6. We observe the correlation values from top k nodes in all rankings, with different values of k, reach almost constant limit at a specific value of k. This certain amount of k approximately equals to 1000 for all metrics except Bipartite Closeness and Latapy CC metrics. This means that the correlation does not increase at a certain threshold for $k = 1000$ in the arXiv cond-mat dataset.

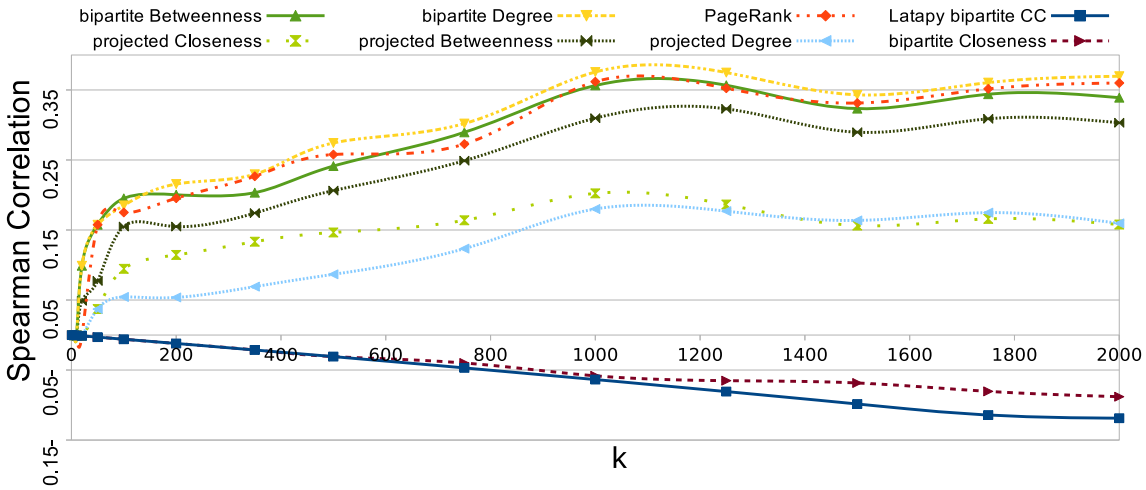

Fig. 6 Spearman's rank correlation with different values of k in arXiv cond-mat dataset

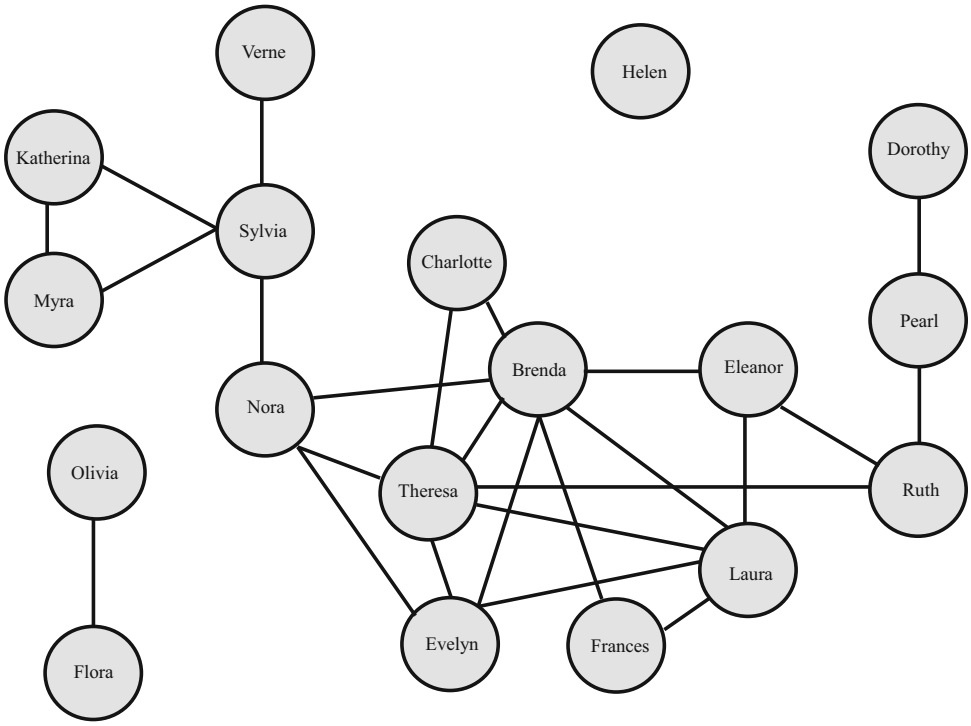

Fig. 7 Mapping Hellinger distance matrix to Euclidean Space. A tie indicates that the distance between two nodes is lesser than 0.50

4.4 Mapping the Hellinger distance matrix to the Euclidean space

Since we have a well-defined metric features and ability of mapping the Hellinger distance matrix to the Euclidean space, other experiment that can be done on this matrix, is clustering nodes based on their distance. This Hellinger distance matrix can then be treated as a valued adjacency matrix[2] and visualized using standard graph layout algorithms. Figure 7 shows the result of such an analysis on Davis dataset. This figure is a depiction of Hellinger Distance for each pair of individuals, such that a line connecting two individuals indicates that their Hellinger distance are less than 0.50. The diagram clearly shows the separation of Flora and Olivia, and the bridging position of Nora.

5 Conclusion and future work

In this paper, we proposed HellRank centrality measure for properly detection of more behavioral representative users in bipartite social networks. As opposed to previous work,

[2] In a valued adjacency matrix, the cell entries can be any nonnegative integer, indicating the strength or number of relations of a particular type or types (Koput 2010).

by using this metric we can avoid projection of bipartite networks into one-mode ones, which makes it possible to take much richer information from the two-mode networks. The computation of HellRank can be distributed by letting each node uses only local information on its immediate neighbors. To improve the accuracy of HellRank, we can extend the neighborhood around each node. The HellRank centrality measure is based on the Hellinger distance between two nodes of the bipartite network and we theoretically find the upper and the lower bounds for this distance.

We experimentally evaluated HellRank on the Southern Women Davis dataset and the results showed that Brenda, Evelyn, Nora, Ruth, and Theresa should be considered as important women. Our evaluation analyses depicted that the importance of a woman does not only depend on her Degree, Betweenness, and Closeness centralities. For instance, if Brenda with low degree centrality is removed from the network, the information would not easily spread among other women. As another observation, Dorothy, Olivia, and Flora have very low HellRank centralities. These results are consistent with the results presented in Bonacich (1978), Doreian (1979), and Everett and Borgatti (1993).

As a future work, more meta data information can be taken into account besides the links in a bipartite network. Moreover, we can consider a bipartite network as a weighted graph (Mahyar et al. 2013, 2015b) in which the links are not merely binary entities, either present or not, but have associated a given weight that record their strength relative to one another. As one of the other possible future works, we can consider alpha-divergence (Cichocki and Amari 2010) as a generalization of squared Hellinger distance. Furthermore, as HellRank measure is proper for detection of more behavioral representative users in bipartite social network, we can use this measure in recommender systems (Taheri et al. 2017). In addition, we can detect top k central nodes in a network with indirect measurements and without full knowledge the network topological structure, using compressive sensing theory (Mahyar 2015; Mahyar et al. 2013, 2013, 2015a, b).

Acknowledgements Open access funding provided by TU Wien (TUW).

References

Abdi H (2007) The Kendall rank correlation coefficient. Encyclopedia of measurement and statistics. pp. 508–510. https://www.utdallas.edu/~herve/Abdi-KendallCorrelation2007-pretty.pdf

Ahmed NK, Neville J, Kompella R (2014) Network sampling: from static to streaming graphs. ACM Trans Knowl Discovery Data (TKDD) 8(2):7

Anagnostopoulos A, Kumar R, Mahdian M (2008) Influence and correlation in social networks. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 7–15. ACM

Beguerisse Díaz M, Porter MA, Onnela JP (2010) Competition for popularity in bipartite networks. Chaos 20(4): 043101. doi:10.1063/1.3475411. http://www.ncbi.nlm.nih.gov/pubmed/21198071

Benevenuto F, Rodrigues T, Cha M, Almeida V (2009) Characterizing user behavior in online social networks. In: Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference, pp. 49–62. ACM

Bharathi S, Kempe D, Salek M (2007) Competitive influence maximization in social networks. International Workshop on Web and Internet Economics. Springer Berlin Heidelberg, pp 306–311

Bollobás B (1984) The evolution of random graphs. Trans. Am. Math. Soc. 286(1):257–257. doi:10.1090/S0002-9947-1984-0756039-5. http://www.citeulike.org/group/3509/article/4012374

Bonacich P (1972) Factoring and weighting approaches to status scores and clique identification. J Math Soc 2(1):113–120. doi:10.1080/0022250X.1972.9989806

Borgatti SP, Everett MG (1997) Network analysis of 2-mode data. Soc Netw 19(3):243–269. doi:10.1016/s0378-8733(96)00301-2. URL<Go to ISI>://WOS:A1997XE81300004

Borgatti SP, Halgin DS (2011) Analyzing affiliation networks. Sage Handb Soc Netw Anal 1941:417–433

Cha M, Haddadi H, Benevenuto F, Gummadi PK (2010) Measuring user influence in Twitter: the million follower fallacy. Icwsm 10(10–17):30

Chen DB, Gao H, Lü L, Zhou T (2013) Identifying influential nodes in large-scale directed networks: the role of clustering. PLoS ONE 8(10):e77,455. doi:10.1371/journal.pone.0077455

Chen W, Wang Y, Yang S (2009) Efficient influence maximization in social networks. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 199–208. ACM

Cichocki A, Amari Si (2010) Families of alpha-beta-and gamma-divergences: flexible and robust measures of similarities. Entropy 12(6):1532–1568

Crandall D, Cosley D, Huttenlocher D, Kleinberg J, Suri S (2008) Feedback effects between similarity and social influence in online communities. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 160–168. ACM

Csiszár I, Shields PC (2004) Information theory and statistics: a tutorial. Found Trends® Commun Inf Theory 1(4):417–528. doi:10.1561/0100000004

Davis A, Gardner BB, Gardner MR (2009) Deep south: a social anthropological study of caste and class. University of South Carolina Press

Davoodi E, Afsharchi M, Kianmehr K (2012) A social network-based approach to expert recommendation system. In: International

conference on hybrid artificial intelligence systems, pp. 91–102. Springer
DiChristina M (2007) Small world. Sci Am sp 17(3):1–1. doi:10.1038/scientificamerican0907-1sp
Du Y, Gao C, Hu Y, Mahadevan S, Deng Y (2014) A new method of identifying influential nodes in complex networks based on TOPSIS. Phys A Stat Mech Appl 399:57–69. doi:10.1016/j.physa.2013.12.031
Faust K (1997) Centrality in affiliation networks. Soc Netw 19(2):157–191. doi:10.1016/S0378-8733(96)00300-0
Freeman LC (1977) A set of measures of centrality based on betweenness. Sociometry 40(1):35–41. JSTOR, http://www.jstor.org/stable/3033543
Freeman LC (1978) Centrality in social networks conceptual clarification. Soc Netw 1(3):215–239
Friedkin NE, Johnsen EC (2011) Social influence network theory: a sociological examination of small group dynamics, vol 33. Cambridge University Press
Giroire F, Chandrashekar J, Iannaccone G, Papagiannaki K, Schooler EM, Taft N (2008) The cubicle vs. the coffee shop: behavioral modes in enterprise end-users. In: International conference on passive and active network measurement, pp. 202–211. Springer
Goyal A, Bonchi F, Lakshmanan LV (2010) Learning influence probabilities in social networks. In: Proceedings of the third ACM international conference on Web search and data mining - WSDM '10 p 241. doi:10.1145/1718487.1718518
Hagberg AA, Schult DA, Swart PJ (2008) Exploring network structure, dynamics, and function using NetworkX. In: Proceedings of the 7th python in science ... 836, 11–15
Hunter JK (2012) An introduction to real analysis. University of California at Davis, California
Johnson O (2004) Information theory and the central limit theorem, vol 8. World Scientific, Imperial College Press, London
Kang U (2011) Centralities in large networks : algorithms and observations. In: SIAM international conference on data pp 119–130
Kempe D, Kleinberg J, Tardos É (2033) Maximizing the spread of influence through a social network. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp 137–146. ACM
Kettle D (2012) Street art. Strad 123(1467):56–59. doi:10.3386/w19846. http://www.nber.org/papers/w19846
Kitsak M, Gallos LK, Havlin S, Liljeros F, Muchnik L, Stanley HE, Makse Ha (2010) Identifying influential spreaders in complex networks. Nat Phys 6(11):36. doi:10.1038/nphys1746. http://arxiv.org/abs/1001.5285
Kitsak, M., Krioukov, D.: Hidden variables in bipartite networks. Phys Rev E Stat Nonlinear Soft Matter Phys 84(2):026114 (2011). doi:10.1103/PhysRevE.84.026114
Kleinberg JM (1999) Authoritative sources in a hyperlinked environment. J ACM 46(5):604–632. doi:10.1145/324133.324140
Koput KW (2010) Social capital: an introduction to managing networks. Edward Elgar Publishing, Cheltenham
Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22(1):79–86
Landauer TK (1988) Research methods in human-computer interaction. In: Helander M (ed) Handbook of human-computer interaction. Elsevier, New York, pp 905–928
Latapy M, Magnien C, Vecchio ND (2008) Basic notions for the analysis of large two-mode networks. Soc Netw 30(1):31–48. doi:10.1016/j.socnet.2007.04.006
Lehamn A, O'Rourke N, Stepanski L (2005) JMP For Basic Univariate And Multivariate Statistics. A step-by-step guide. SAS Institute, Cary, p 481
Leskovec J, Faloutsos C (2006) Sampling from large graphs. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 631–636. ACM
Li Q, Zhou T, Lü L, Chen D (2014) Identifying influential spreaders by weighted LeaderRank. Phys A Stat Mech Appl 404:47–55. doi:10.1016/j.physa.2014.02.041
Liebig J, Rao A (2015) Identifying influential nodes in bipartite networks using the clustering coefficient. In:Proceedings—10th International conference on signal-image technology and internet-based systems, SITIS 2014 pp 323–330
Lind PG, González MC, Herrmann HJ (2005) Cycles and clustering in bipartite networks. Phys Rev E 72(5)
Luce RD, Perry AD (1949) A method of matrix analysis of group structure. Psychometrika 14(2):95–116. doi:10.1007/BF02289146
Mahyar H (2015) Detection of top-K central nodes in social networks: a compressive sensing approach. In: IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), Paris, France, pp 902–909
Mahyar H, Ghalebi K. E, Morshedi SM, Khalili S, Grosu R, Movaghar A (2017) Centrality-based group formation in group recommender systems. In: Proceedings of the 26th international conference on world wide web companion (WWW), Perth, Australia, pp 1343–1351. doi:10.1145/3041021.3051153
Mahyar H, Rabiee HR, Hashemifar ZS (2013) UCS-NT: an unbiased compressive sensing framework for network tomography. In: IEEE international conference on acoustics, speech, and signal processing (ICASSP), Vancouver, Canada, pp 4534–4538
Mahyar H, Rabiee HR, Hashemifar ZS, Siyari P (2013) UCS-WN: an unbiased compressive sensing framework for weighted networks. in: conference on information sciences and systems (CISS), Baltimore, USA, pp 1–6
Mahyar H, Rabiee HR, Movaghar A, Ghalebi K E, Nazemian A (2015) CS-ComDet: a compressive sensing approach for inter-community detection in social networks. In: IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), Paris, France, pp 89–96. doi:10.1145/2808797.2808856
Mahyar H, Rabiee HR, Movaghar A, Hasheminezhad R, Ghalebi K, E, Nazemian A (2015) A Low-cost sparse recovery framework for weighted networks under compressive sensing. In: IEEE international conference on social computing and networking (SocialCom), Chengdu, China, pp 183–190
Maia M, Almeida J, Almeida V (2008) Identifying user behavior in online social networks. In: Proceedings of the 1st workshop on Social network systems, pp 1–6. ACM
Maiya AS, Berger-Wolf TY (2011) Benefits of bias: towards better characterization of network sampling. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 105–113. ACM
Muruganantham A, Road RG, Nadu T, Road RG, Nadu T (2015) Ranking the influence users in a social networking site using an improved topsis. J Theor Appl Inf Technol 73(1):1–11
Newman ME (2001) The structure of scientific collaboration networks. Proc Natl Acad Sci USA 98(2):404–409. doi:10.1073/pnas.021544898021544898
Nikulin M (2001) Hellinger distance. Hazewinkel, Michiel, Encyclopedia of mathematics. Springer, Berlin. doi:10.1145/1361684.1361686
Okamoto K, Chen W, Li XY (2008) Ranking of closeness centrality for large-scale social networks. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 5059 LNCS, 186–195

Opsahl T (2013) Triadic closure in two-mode networks: redefining the global and local clustering coefficients. Soc Netw 35(2):159–167. doi:10.1016/j.socnet.2011.07.001

Opsahl T, Panzarasa P (2009) Clustering in weighted networks. Soc Netw 31(2):155–163. doi:10.1016/j.socnet.2009.02.002

Papagelis M, Das G, Koudas N (2013) Sampling online social networks. IEEE Trans Knowl data Eng 25(3):662–676

Robins GL, Alexander M (2004) Small worlds among interlocking directors: network structure and distance in bipartite graphs. Comput Math Organ Theory 10(1):69–94. doi:10.1023/B:CMOT.0000032580.12184.c0. http://www.springerlink.com/content/q567773rj070xvj2/

Sabidussi G (1966) The centrality index of a graph. Psychometrika 31(4):581–603

Sawant S, Pai G (2013) Yelp food recommendation system. stanford machine learning project (CS229) (2013). http://cs229.stanford.edu/proj2013/SawantPai-YelpFoodRecommendationSystem.pdf

Silva A, Guimarães S, Meira W, Zaki M (2013) ProfileRank. Proceedings of the 7th workshop on social network mining and analysis—SNAKDD '13 pp 1–9. doi:10.1145/2501025.2501033. http://dl.acm.org/citation.cfm?doid=2501025.2501033

Singla P, Richardson M (2008) Yes, there is a correlation:-from social networks to personal behavior on the web. In: Proceedings of the 17th international conference on World Wide Web, pp 655–664. ACM

Stephenson K, Zelen M (1989) Rethinking centrality: methods and examples. Soc Netw 11(1):1–37. doi:10.1016/0378-8733(89)90016-6

Sun K, Mohamed H, Marchand-Maillet S (2015) Sparsity on statistical simplexes and diversity in social ranking. In: Asian conference on machine learning

Sun K, Morrison D, Bruno E, Marchand-Maillet S (2013) Learning representative nodes in social networks. In: Pacific-Asia conference on knowledge discovery and data mining, pp 25–36. Springer

Taheri SM, Mahyar H, Firouzi M, Ghalebi K E, Grosu R, Movaghar A (2017) Extracting implicit social relation for social recommendation techniques in user rating prediction. In: Proceedings of the 26th international conference on world wide web companion (WWW), Perth, Australia, pp 1187–1196 (2017). doi:10.1145/3041021.3055363

Tang J, Sun J, Wang C, Yang Z (2009) Social influence analysis in large-scale networks. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 807–816. ACM

Tang J, Sun J, Wang C, Yang Z (2009) Social influence analysis in large-scale networks. Proceedings of SIGKDD'09 pp 807–816 (2009). doi:10.1145/1557019.1557108

Tang J, Zhang C, Cai K, Zhang L, Su Z (2015) Sampling representative users from large social networks. In: AAAI, pp 304–310. Citeseer

Tang X, Yang CC (2012) Ranking user influence in healthcare social media (). ACM Trans Intell Syst Technol 3(4):1–21. doi:10.1145/2337542.2337558

Torgersen E (1991) Comparison of statistical experiments 36. Cambridge University Press, Cambridge

Ugander J, Karrer B, Backstrom L, Kleinberg J (2013) Graph cluster randomization: network exposure to multiple universes. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 329–337. ACM

Van der Vaart AW (2000) Asymptotic statistics, vol 3. Cambridge university press

Wehmuth K, Ziviani A (2013) DACCER: distributed assessment of the closeness centrality ranking in complex networks. Comput Netw 57(13):2536–2548. doi:10.1016/j.comnet.2013.05.001

Wei D, Deng X, Zhang X, Deng Y, Mahadevan S (2013) Identifying influential nodes in weighted networks based on evidence theory. Phys A Stat Mech Appl 392(10):2564–2575. doi:10.1016/j.physa.2013.01.054

Weng J, Lim Ep, Jiang J (2010) TwitterRank : finding topic-sensitive influential twitterers. New York **Paper 504**, 261–270. doi:10.1145/1718487.1718520. http://portal.acm.org/citation.cfm?id=1718520

Xu K, Guo X, Li J, Lau RYK, Liao SSY (2012) Discovering target groups in social networking sites: an effective method for maximizing joint influential power. Electr Commer Res Appl 11(4):318–334. doi:10.1016/j.elerap.2012.01.002

You K, Tempo R, Qiu L (2015) Distributed algorithms for computation of centrality measures in complex networks pp 1–14. http://arxiv.org/abs/1507.01694

Yustiawan Y, Maharani W, Gozali AA (2015) Degree centrality for social network with opsahl method. Proc Comput Sci 59:419–426. doi:10.1016/j.procs.2015.07.559

Zhang P, Wang J, Li X, Li M, Di Z, Fan Y (2008) Clustering coefficient and community structure of bipartite networks. Phys A Stat Mech Appl 387(27):6869–6875. doi:10.1016/j.physa.2008.09.006

Zhao K, Wang X, Yu M, Gao B (2014) User recommendations in reciprocal and bipartite social networks—an online dating case study. IEEE Intell Syst 29(2):27–35. doi:10.1109/MIS.2013.104

Zhou T, Ren J, Medo M, Zhang YC (2007) Bipartite network projection and personal recommendation. Phys Rev E Stat Nonlinear Soft Matter Phys 76(4):046,115. doi:10.1103/PhysRevE.76.046115

Zhu X, Goldberg AB, Van Gael J, Andrzejewski D (2007) Improving diversity in ranking using absorbing random walks. In: HLT-NAACL, pp 97–104

Zhu Z (2013) Discovering the influential users oriented to viral marketing based on online social networks. Phys A Stat Mech Appl 392(16):3459–3469. doi:10.1016/j.physa.2013.03.035