Social Network Analysis and Mining (2021) 11:2
https://doi.org/10.1007/s13278-020-00708-w

ORIGINAL ARTICLE

Defining and measuring probabilistic ego networks

Amin Kaveh[1] · **Matteo Magnani**[1] · **Christian Rohner**[1]

Received: 31 March 2020 / Revised: 27 October 2020 / Accepted: 6 November 2020
© The Author(s) 2020

Abstract
Analyzing ego networks to investigate local properties and behaviors of individuals is a fundamental task in social network research. In this paper we show that there is not a unique way of defining ego networks when the existence of edges is uncertain, since there are two different ways of defining the neighborhood of a node in such network models. Therefore, we introduce two definitions of probabilistic ego networks, called V-Alters-Ego and F-Alters-Ego, both rooted in the literature. Following that, we investigate three fundamental measures (degree, betweenness and closeness) for each definition. We also propose a method to approximate betweenness of an ego node among the neighbors which are connected via shortest paths with length 2. We show that this approximation method is faster to compute and it has high correlation with ego betweenness under the V-Alters-Ego definition in many datasets. Therefore, it can be a reasonable alternative to represent the extent to which a node plays the role of an intermediate node among its neighbors.

Keywords Probabilistic networks · Ego networks · Local properties · Betweenness · Closeness

1 Introduction

Empirical social network data collection is often an imperfect process affected by some degree of uncertainty. Uncertainty can come from different sources. For example because of missing information and indirect measurements, as in the case when we infer social ties or influence relationships between individuals based on their interactions (Aggarwal and Wang 2010; Bernard et al. 1982). Uncertainty can be available even when we are asking about the immediate connections of an individual in social networks for example due to forgetfulness of informants (Bernard et al. 1979; Killworth and Bernard 1979). To model uncertain information in networks, probabilistic models in which each edge is associated with an independent probability are the typical choice in the literature (Asthana et al. 2004; Poisot et al. 2016; Rhodes et al. 2005). Despite the fact that uncertainty affects several types of data collection processes, the majority of works on social networks ignore it. More precisely, in data collection a thresholding approach is typically used, in which if the degree of confidence about the existence of an edge is higher than a specific value, then we draw and edge between those nodes. However, the selection of a threshold value is a subjective task. As an example, De Choudhury et al. (2010) have studied two email exchange datasets (a university email dataset and the Enron email dataset) to infer unobserved social ties using the number of exchanged emails between pairs of individuals. They have inferred the existence of a social tie between each pair of individuals if the average number of exchanged emails in a specific period of time is higher than a specific number, i.e., a threshold. As a result, they have demonstrated that different choices of the threshold lead to completely different network structures. Brugere et al. (2018) have introduced a wide variety of areas such as computational biology, neuroscience, ecology and social science in which edges between entities have been inferred using another type of interactions and the thresholding approach has been used to construct the final networks. In our opinion, the main reason why uncertainty is rarely considered in social network analysis is the lack of appropriate methods to handle it. In this paper, we thus focus on the methods to analyze probabilistic networks.

Mining and analysis of probabilistic social networks have gained a great attention during the last years and have led to formulating many problems in such networks. A large number of analytic approaches and algorithms to solve these problems are based on local properties of nodes, such as

✉ Amin Kaveh
amin.kaveh@it.uu.se

[1] InfoLab, Department of Information Technology, Uppsala University, 75105 Uppsala, Sweden

Published online: 25 November 2020

their connectivity with their one-hop (immediate) neighbors (Bonchi et al. 2014; Mukherjee et al. 2017; Parchas et al. 2015, 2018).

One of the main approaches to study the local properties of a node is to examine its ego network. In deterministic social networks, in which the existence of edges is certain, an ego network is a network consisting of a node called *ego*, its neighbors called *alters* and the edges between the alters and the ego and between the alters. Deterministic ego networks have been studied extensively following different lines of research. One direction of research is focused on studying the structural properties of ego networks to identify and predict some human behaviors in online social networks (Arnaboldi et al. 2014, 2016a, b; Roberts and Dunbar 2011). Another branch of study tries to estimate the global properties of nodes based on their corresponding properties in their ego-networks (Everett and Borgatti 2005; Marsden 2002; Pantazopoulos et al. 2013). The third branch of works attempt to focus on the differences between the egocentric properties of nodes in online and offline social networks (Arnaboldi et al. 2017; Socievole and Marano 2012).

Despite the existence of several studies on deterministic ego networks, ego measures have not been studied for probabilistic networks so far. In fact, no definition of probabilistic ego network has been proposed and evaluated yet. Considering the importance of deterministic ego networks in the field of social network analysis, the absence of a probabilistic counterpart of this theory constitutes a strong limitation. As mentioned before, in the current literature on probabilistic network analysis many methods are based on the local properties of nodes, which highlights the significance of having a clear definition of probabilistic ego network and associated measures.

Unlike the uniqueness of the definition of deterministic ego network, probabilistic ego networks can be defined in two ways. In the first one, first all possible worlds[1] are generated and then the neighborhood of an ego node in each possible world is defined independently. In the second approach, first the neighborhood of an ego node is defined in the probabilistic network and then all possible worlds corresponding to that neighborhood are generated.

1.1 Contributions and outline

In this paper, we provide three main contributions:

- As the first contribution we introduce two definitions of probabilistic ego networks, called V-Alters-Ego and F-Alters-Ego. These definitions are based on the two definitions of a node's neighborhood in probabilistic networks.
- As the second contribution, we examine degree, betweenness and closeness for both definitions of probabilistic ego networks, to see to what extent the two definitions of probabilistic ego networks lead to different sets of top-ranked nodes and to what extent they are correlated. We show that while closeness is always 1 for all nodes under V-Alters-Ego, it is represented as a probability distribution under F-Alters-Ego.
- As the third contribution, we propose a method to approximate probabilistic ego betweenness and show that this method is an acceptable alternative for the betweenness under V-Alters-Ego definition.

Section 2 presents an introduction to three concepts that are foundations of our research: probabilistic networks, ego networks and nodes' neighborhood. Section 3 describes two definitions of probabilistic ego networks based on two definitions of nodes' neighborhood in probabilistic networks. Moreover, we show how degree, betweenness and closeness apply under each of these two definitions. In Sect. 4 we propose an approximation method to estimate the extent to which an ego node plays the role of an intermediate node among its neighbors. In Sect. 5, we evaluate the extent to which different definitions of probabilistic ego networks result into different lists of most influential nodes in probabilistic networks and to what extent they are correlated. We conclude and present some opportunities for further research in Sect. 6.

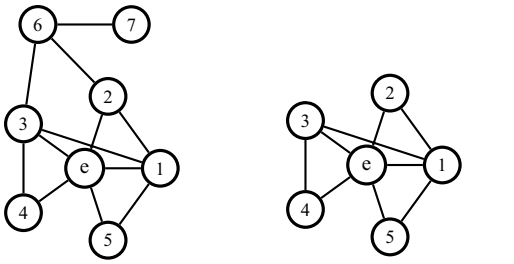
2 Preliminaries

2.1 Probabilistic networks

The most common model to represent uncertainty in networks is $\mathcal{G} = (V, E, p)$ where V and E are, respectively, sets of nodes and edges and $p : E \rightarrow (0, 1]$ is a function assigning a probability to each edge. Edge probabilities are mutually independent. This model is called a probabilistic network model and has been used widely to represent imperfect network data not only in social influence networks (Potamias et al. 2010), but also in sensor networks (Gao et al. 2017), opportunistic networks (Lu et al. 2016), **protein-protein interaction networks (Srihari and Leong 2013) and road networks (Fushimi 2018).

As each edge has two possible states (existing/non-existing) with probability p and $1 - p$, each probabilistic graph corresponds to $2^{|E|}$ deterministic graphs which are called possible worlds (or instances), where each instance G_i has an associated probability $Pr(G_i)$. Under this definition, each measure in

[1] The concept of possible worlds is explained in Sect. 3.



(a) a deterministic network **(b)** *e*'s deterministic ego network

Fig. 1 Ego network in deterministic networks

probabilistic graphs equals the expected value of that measure over all possible worlds:

$$\mathbb{E}(M) = \sum_{G_i \in G} M_i \times Pr(G_i) \tag{1}$$

where G is the set of all possible instances of $\mathcal{G}$ and M_i is the value of measure M in possible world G_i.

2.2 Deterministic ego network

In order to define probabilistic ego networks, we first look at the definition of ego network in deterministic networks. A deterministic ego network of an arbitrary node e is a network consisting of node e, called *ego*, its neighbors called *alters*, the edges between the alters and the ego and the edges between the alters (Everett and Borgatti 2005; Marsden 2002). With this network, local structural properties of nodes can be extracted (Fig. 1). The most common measures in ego networks are given below.

2.2.1 Degree

Node degree is a fundamental measure in networks. The degree of a node in its ego network is the same as the degree of that node in the whole network.

2.2.2 Ego betweenness

Ego betweenness was introduced in Marsden (2002). Following that, authors in Everett and Borgatti (2005) proposed an efficient and simple method to calculate ego betweenness based on the adjacency matrix of an ego network: the ego betweenness of node e is the sum of the reciprocal of all elements in the upper triangle of matrix $\mathbf{A}^2[\mathbf{1} - \mathbf{A}]$ without considering the diagonal elements:

$$Bw_{ego}(e) = \sum_{i>j} \frac{1}{\mathbf{A}^2[\mathbf{1} - \mathbf{A}]_{ij}} \tag{2}$$

where $\mathbf{A}$ is the ego network's adjacency matrix.

2.2.3 Ego closeness

Closeness of a node is based on the length of the shortest paths between that node and all other nodes in the network. By definition, the shortest path distance between an ego node and its alters is 1. So, the closeness of an ego node in its ego network is not meaningful.

2.3 Nodes' neighbors

The definition of ego network in deterministic networks is based on the definition of neighborhood, which is the set of nodes that are adjacent to the ego node. However, since the analysis of probabilistic networks is based on possible worlds semantics, there are two different ways of defining neighborhood in probabilistic networks: before generating possible worlds or after generating possible worlds.

Definition 1 (*after generating possible worlds*) Given a probabilistic graph $\mathcal{G} = (V, E, p)$ and an arbitrary node u, the neighbors of node u in possible world w are defined as the set of nodes adjacent to u in that possible world:

$$\forall w \in \mathbb{W}, N_w(u) = \{v \mid (u, v) \in \mathbb{E}_w\} \tag{3}$$

where $\mathbb{W}$ is the set of all possible worlds of $\mathcal{G}$ and $N_w(u)$ is the set of u's neighbors in possible world w and $\mathbb{E}_w$ is the set of edges in possible world w.

Definition 2 (*before generating possible worlds*) Given a probabilistic graph $\mathcal{G} = (V, E, p)$ and an arbitrary node u, the neighbors of node u in all possible worlds are defined as the set of nodes having a positive probability of being a neighbor of that node:

$$\forall w \in \mathbb{W}, N_w(u) = \{v \mid (u, v) \in E, p_{uv} > 0\} = N(u) \tag{4}$$

According to this definition, the set of neighbors of node u is fixed, $N(u)$, regardless of in which possible worlds they are connected and in which possible worlds they are not connected.

The two described definitions above are rooted in the probabilistic network literature. Some works implicitly use Definition 1 (Bonchi et al. 2014), while others use Definition 2 (Mukherjee et al. 2017).

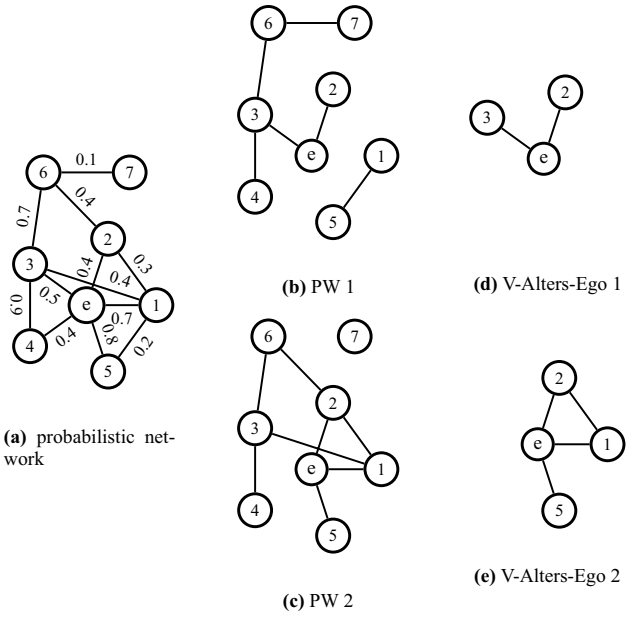


Fig. 2 **a** a probabilistic network, **b**, **c** two possible worlds of the probabilistic network in **a**, **d** and **e** two V-Alters-Ego corresponding to possible worlds PW1 and PW2 in **b**, **c**

3 Probabilistic ego networks: definitions and measures

3.1 Probabilistic ego networks with varying sets of alters

Our first definition of probabilistic ego network is based on Definition 1. Therefore, for arbitrary node e, e's ego network in a specific possible world is the network consisting of node e, its neighbors in that possible world, the edges between the neighbors and e and the edges between the neighbors. Figure 2a shows a probabilistic network and Figure 2b, c illustrate two possible worlds of it. Figure 2d, e shows two ego networks of node e, extracted from possible worlds in Fig. 2b, c respectively. Hereafter, we notate e's alters in each instance as $A_v(e)$, where subscript v denotes the variation of the set of alters in each possible world. We also use the abbreviation V-Alters-Ego to refer to this definition of probabilistic ego networks.

In the following, we discuss the calculation of the most fundamental and common measures including degree, ego betweenness and ego closeness according to this definition:

3.1.1 Degree

In a probabilistic network we may not know the degree of a node with certainty; instead, we can compute degree probability distributions, where for each node a probability is associated to one or more possible values for the degree. Since the calculation and analysis of degree distributions in large networks are challenging, summary measures of degree distributions have been used as a corresponding measure for degree (Bonchi et al. 2014; Parchas et al. 2015, 2018). The most commonly used summary measure is expected degree. To calculate the expected degree of an ego node in probabilistic networks, we have to use Eq. 1 by replacing M_i with $D_i(e)$ which is the degree of node e in possible world G_i. Since a node's degree distribution in probabilistic networks is a Poisson binomial distribution (Kaveh et al. 2019) the expected degree is calculated easily by aggregating the probability of all the edges incident to e.

$$\mathbb{E}(D_e) = \sum_{u \in A_v(e)} p_{eu} \tag{5}$$

where $A_v(e)$ is the set of alters of ego e and p_{eu} is the probability of the edge between e and u. For example the expected degree of node e in Fig. 2a is 2.8.

3.1.2 Betweenness

Betweenness of an ego node in a probabilistic network equals the expected value of ego betweenness in all deterministic possible worlds. As discussed in Everett and Borgatti (2005), the shortest path length between two alters in deterministic ego networks is 1 if they are adjacent (nodes *1* and *3* in Fig. 1b) or is 2 if they are not adjacent (nodes *1* and *4* in Fig. 1b). For non-adjacent alters there is always a path with length 2 that passes through the ego node, although in addition to it, it is possible to have other paths with length 2 (e.g., two geodesic paths between alters *1* and *4* pass, respectively, through ego node e and alter node *3*). In the algorithm proposed in Everett and Borgatti (2005), if $\mathbf{A}$ is the adjacency matrix of an ego network, $\mathbf{A}^2[\mathbf{1} - \mathbf{A}]_{ij}$ is 0 if nodes i and j are adjacent, is 1 if they are not adjacent and the shortest path between them only passes through the ego node e, and is $1 + d$ if there are d paths of length 2 passing through nodes other than e.

The following matrix shows the result of $\mathbf{A}^2[\mathbf{1} - \mathbf{A}]$ for the deterministic graph presented in Fig. 1b. The matrix shows that there are 2 shortest paths with length 2 between nodes *2* and *5*. Since e is adjacent to both of them, so one of these paths is definitely passing through this node and the other path is passing through another alter (in this case node *1*).

As the ego node is represented in the first column/row of the matrix, the number of shortest paths between nodes 2 and 5 corresponds to the 3rd row and 6th column of the resulting matrix.

$$\mathbf{A}^2[\mathbf{1}-\mathbf{A}] = \begin{bmatrix} . & 0 & 0 & 0 & 0 & 0 \\ . & . & 0 & 0 & 2 & 0 \\ . & . & . & 2 & 1 & 2 \\ . & . & . & . & 0 & 2 \\ . & . & . & . & . & 1 \\ . & . & . & . & . & . \end{bmatrix}$$

If $\mathbf{A}$ is the adjacency matrix of a probabilistic network in which $\mathbf{A}_{ij}$ represents the probability of the edge between nodes i and j, $\mathbf{A}^2[\mathbf{1}-\mathbf{A}]$ does not give the same information. The following matrix presents the result of $\mathbf{A}^2[\mathbf{1}-\mathbf{A}]$ for the probabilistic graph in Fig. 2a (by removing nodes *6* and *7*). Each element of the matrix shows the expected number of paths of length 2 between the corresponding nodes. However, this value does not reflect the contribution of the ego node as the intermediate node between A and B between the considered nodes. For example, the expected number of paths of length 2 between nodes *2* and *5* in Fig. 2a is 0.38. However, the contribution of the ego node *e* as the intermediate node between nodes *2* and *5* is either 1 with probability 0.3008 (the path via node *1* does not exist) or 0.5 with probability 0.0192 (both paths exist). Then, the betweenness of the ego node is 0.3104 and it cannot be extracted from the following matrix.

$$\mathbf{A}^2[\mathbf{1}-\mathbf{A}] = \begin{bmatrix} . & 0.144 & 0.126 & 0.32 & 0.27 & 0.028 \\ . & . & 0.196 & 0.21 & 0.64 & 0.448 \\ . & . & . & 0.32 & 0.16 & 0.38 \\ . & . & . & . & 0.02 & 0.48 \\ . & . & . & . & . & 0.32 \\ . & . & . & . & . & . \end{bmatrix}$$

As a result, to obtain the probabilistic ego betweenness we can not replace the adjacency matrix of the probabilistic ego network in Eq. 2. Hence, in V-Alters-Ego, the ego betweenness has to be calculated in each possible world and the probabilistic ego betweenness is the result of Eq. 1 in which M_i is replaced by Eq. 2.

3.1.3 Closeness

Closeness in deterministic ego networks can only take the value 1, by definition, and it is thus not a meaningful measure. In V-Alters-Ego in which each measure is the mean value of that measure in all possible worlds, ego closeness is also 1.

3.2 Probabilistic ego networks with a fixed set of alters

Our second definition of probabilistic ego network is based on Definition 2. In this approach the set of neighbors of a node is fixed for all possible worlds. Therefore, for arbitrary

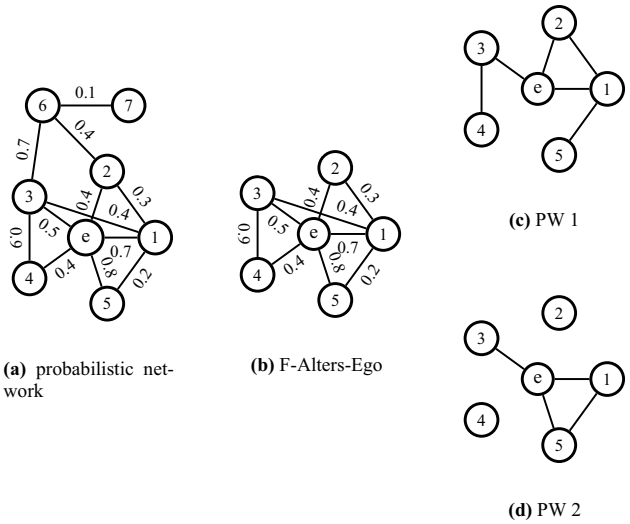


Fig. 3 **a** a probabilistic network, **b** F-Alters-Ego, **c**, **d** two possible worlds of F-Alters-Ego in **b**

node e, e's ego network in an specific possible world is the network consisting of node e and the fixed set of its neighbors, and all the edges available between the neighbors and e and the edges between the neighbors in that possible world. We notate the set of all nodes that are connected via uncertain edges to the ego node e (alters of e) as $A_f(e)$. For the sake of brevity, we use the abbreviation F-Alters-Ego to refer to the definition of probabilistic ego network based on a fixed set of alters.

Figure 3b shows an arbitrarily chosen node e and all the nodes that are considered as e's neighbors in all possible worlds. Figure 3c, d demonstrates two possible worlds of it. In both possible worlds nodes $A_f(e) = \{1, 2, 3, 4, 5\}$ are treated as e's neighbors.

By defining probabilistic ego networks in V-Alters-Ego, first the distance between an ego node and its alters is always 1 and second, the distance between two alters is either 1 or 2. On the other hand, by defining probabilistic ego networks in V-Alters-Ego, first, the distance between an ego node and its alters can be longer than 1 and second, the distance between alters can be longer than 2 in some possible worlds. These two differences motivate us to re-study betweenness Sect. 3.2.1 and closeness Sect. 3.2.2 in F-Alters-Ego accordingly. However, the definition of degree is the same in the V-Alters-Ego and F-Alters-Ego cases.

3.2.1 Betweenness

Ego betweenness in deterministic networks is calculated by counting the number of the shortest paths with length 2 that traverse the ego (Everett and Borgatti 2005). Under the V-Alters-Ego definition, the ego betweenness is the expected value of deterministic ego betweenness in all

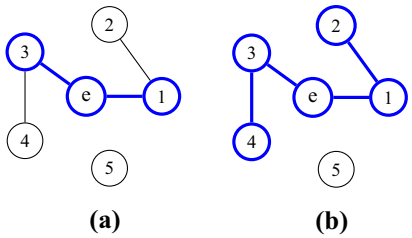


Fig. 4 **a**, **e** is an intermediate node between *1* and *3*, **b** *e* is an intermediate node between nodes *1*, *2* and *3*, *4*

possible worlds. Figure 4a shows a possible world of the probabilistic network in Fig. 3a. In this possible world in Fig. 4b, *e*'s ego betweenness under V-Alters-Ego definition is 1, however, under F-Alters-Ego, not only *e* is an intermediate node in a path with length 2 between nodes *1* and *3*, but also it is an intermediate node in the paths with length 3 (between nodes *2* and *3* as well as *1* and *4*) and length 4 (between nodes *2* and *4*).

This shows that under the F-Alters-Ego definition, the shortest paths with length higher than 2 which pass through an ego node have a contribution in the value of ego betweenness of that ego node. Therefore, the ego betweenness under F-Alters-Ego is the expected value of ego betweenness in all possible worlds in which shortest paths between alters with length higher than 2 which pass through the ego node are also counted.

3.2.2 Closeness

By defining probabilistic ego networks as in Sect. 3.1, the distance between an ego node and its alters is always 1. On the other hand, by defining it based on a fixed set of alters the distance between the ego node and each alter is represented as a shortest path distance distribution. More precisely, in some instances the distance between the ego node and an alter is higher than 1.

Having the shortest path length distribution between ego node and its alters, motivates us to study the concept of distance between an ego node and their alters to propose a new version of closeness.

3.2.3 Shortest path length distribution

The shortest path lengths between any pairs of nodes in probabilistic networks are expressed as shortest path length distributions (Potamias et al. 2010). In F-Alters-Ego, the smallest shortest path length between an ego node and its alters is 1 with the probability of the incident edge between them. The longest shortest path is in the case that there is a path between the ego and its alter by traversing all other alters. In this case, the longest shortest path length has the length equal to the number of alters. We notate the shortest path length distribution between two nodes u and v as $sp_{u,v}$ and define $sp_{u,v}(l)$ to be the probability that the shortest path length between nodes u and v is l:

$$\boldsymbol{sp}_{u,v}(l) = \sum_{G|D(u,v)=l} Pr(G) \tag{6}$$

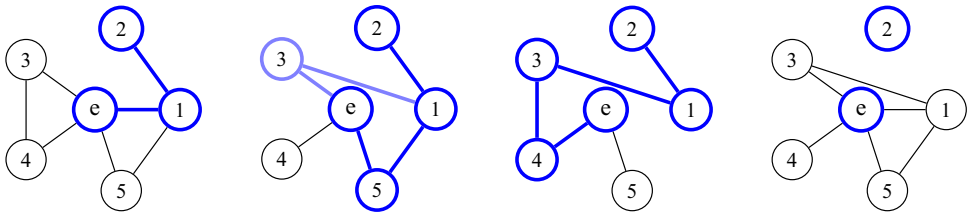
where G is the set of all possible worlds of probabilistic graph $\mathcal{G}$. To put it in another way, the probability that the shortest path length between nodes u and v is l equals the sum of the probabilities of all possible worlds in which shortest path length between these two nodes is l. For example, the shortest path length between ego e and alter *2* is 1 with probability 0.4. Moreover, alter *2* is accessible with shortest path length 2 with probability 0.126 via node *1* (Fig. 5a) and with shortest path length 3 with probability 0.018 by passing nodes *{5, 1}* or *{3, 1}* (see Fig. 5b). The highest shortest path length between e and *2* is obtained in the instance in Fig. 5c. Furthermore, node *2* is disconnected from e with probability 0.453 (Fig. 5d). Figure 5e shows the shortest path length distribution in which the event of disconnection between e and *2* is notated as ∞.

One of the most common summarizing measures of probability distributions is the expected value. As the shortest path length is presented as a probability distribution, the expected length of the shortest paths is the most desirable measure, however its calculation is problematic. The reason is that in probabilistic networks there is a probability of disconnection between each pair of nodes. Since in network science the distance between two disconnected nodes is typically assumed infinite, calculation of the expected value of the shortest path length between them is impossible. For example the expected value of the shortest paths between nodes e and *2* is: $1 \times \boldsymbol{sp}_{e,2}(1) + 2 \times \boldsymbol{sp}_{e,2}(2) + 3 \times \boldsymbol{sp}_{e,2}(3) + 4 \times \boldsymbol{sp}_{e,2}(4) + \infty \times \boldsymbol{sp}_{e,2}(\infty) = \infty$.

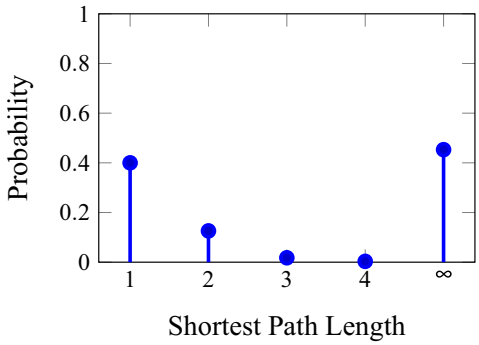
Although extracting the average distance between an ego node and an alter is implausible, still it is possible to extract useful information from the shortest path length distribution. For example, the shortest path length distribution in Fig. 5e reveals that in 52.6% of the possible worlds of the network in Fig. 3b the distance between nodes e and *2* is at most 2. Based on this intuition we define *α-distance* between two nodes in probabilistic networks.

Definition 3 *α-distance* is the minimum shortest path length where the probability of having this length or less is higher than α:

$$d_\alpha(v,u) = \arg\min_k \left\{ \sum_{l=1}^{k} sp_{v,u}(l) \geq \alpha \right\} \tag{7}$$



(a) $l = 2$ **(b)** $l = 3$ **(c)** $l = 4$ **(d)** $l = \infty$



(e) shortest path length distribution between and *2*

Fig. 5 **a–d** An example of possible worlds where the shortest path length between nodes *e* and *2* has different values in F-Alters-Ego definition, **e** shortest path length distribution between nodes *e* and *2*

where $0 < \alpha \leq 1$. In other words, α-distance between two nodes is k if at least in $\alpha \times |PW(\mathcal{G})|$ of possible worlds, the shortest path length between them is at most k. By replacing α with $\frac{1}{2}$, we will have the median distance which is similar to the definition of median distance introduced in Potamias et al. (2010). As an example, $d_{0.5}(e, 2) = 2$ and shows that at least in 50% of possible worlds, nodes e and 2 are connected with paths with length at most 2 (see Fig. 5e).

As discussed before, the concept of closeness is meaningless in deterministic ego networks and in V-Alters-Ego, because the shortest path length between an ego node and all its alters is 1. However by defining the probabilistic ego networks based on a fixed set of alters and having the shortest path length distribution between each alter and ego node, the notion of closeness becomes relevant. We define *α-closeness* of an ego node to be the sum of the reciprocal of the *α-distance* between the ego node and each of its alters:

$$C_\alpha(e) = \sum_{v \in A_f(e)} \frac{1}{d_\alpha(e, v)} \tag{8}$$

where $A_f(e)$ is the set of alters of the ego node e.

4 Approximating ego betweenness in V/F-Alters-Ego

Here, we outline a method to calculate the contribution of an ego node as the intermediate node in paths with length 2. To this end, for each pair of alters u and v, we consider the probability of existence of edges (e, u) and (e, v) and at the same time the probability of nonexistence of an edge between u and v, i.e., the probability of an open triplet made by (e, u) and (e, v). Hence, we define $b_e(u, v)$ to be the probability that ego node e is the intermediate node in shortest paths with length 2 between its alters u and v:

$$b_e(u, v) = p_{eu}\, p_{ev}\, (1 - p_{uv}) \tag{9}$$

where p_{eu} is the probability of the edge between nodes e and u and so on. As a result, we define betweenness of an ego node e to be the sum of the shortest paths with length 2 crossing e (the sum of the probability of all open triplets, centered on e):

$$B_e = \sum_{u,v \in A_f(e)} b_e(u, v) \tag{10}$$

where $A_f(e)$ is the set of alters of ego node e. As an example in Fig. 3b, $b_e(1, 2) = 0.7 \times 0.4 \times (1 - 0.3)$, $b_e(1, 3) = 0.7 \times 0.5 \times (1 - 0.4)$ and $b_e(1, 4) = 0.7 \times 0.4$ and so on so forth, and then $B_e = 2.554$.

We aim to call attention to three points: first, Eq. 10 aggregates the probability of all shortest paths of length 2 that cross node e, between all pairs of its possible alters, regardless of whether there are other geodesic paths of length 2 in the ego network between u and v or not. Second, Eq. 10 takes into accounts all shortest paths of length 2 between alters, however, there could be paths of length higher than 2 between alters that cross through the ego

Table 1 Characteristics of datasets, $|\mathbf{V}|$ is the number of nodes, $|\mathbf{E}|$ is the number of edges, $\overline{\mathbf{p}}$ is the mean of the edge probabilities and $\overline{\mathbf{D}}$ is the mean of nodes' degree

Dataset	$\|\mathbf{V}\|$	$\|\mathbf{E}\|$	$\overline{\mathbf{p}}$	$\overline{\mathbf{D}}$
Enron	805	3956	0.173	9.83
Facebook	1976	1809	0.179	1.83
FriendFeed	150	619	0.547	8.25
DBLP ($\mu = 0.05$)	2763	3268	0.074	2.37
DBLP ($\mu = 0.5$)			0.486	

node. Third, Eq. 10 violates possible worlds semantics and includes intersections among open triplet.

Regarding the last point, it is worth to mention that our approximation approach is consistent with the method proposed in Pfeiffer and Neville (2011) to approximate clustering coefficient of nodes in probabilistic networks. The author of this paper outlined that their approximation method is based on the first-order Taylor expansion, though they did not provide any mathematical proof. Therefore, we use experimental/empirical analysis approach to see whether the proposed approximation method is an appropriate method to estimate ego betweenness in either V-Alters-Ego approach or F-Alters-Ego approach or not. If the number of incident edges to the ego node is D_e then the computational complexity is $O(\binom{D_e}{2}) = O(D_e^2)$ which is tractable even for nodes with large D_e.

5 Evaluation

In this section we want to investigate whether the two ways of defining probabilistic ego networks (V-Alters-Ego and F-Alters-Ego) lead to different local properties of the nodes. Answering this question is important because, as mentioned before, the result of many algorithms and analytical approaches in the analysis of probabilistic networks depend on the nodes' local properties.

As a method of evaluation, we study the association among the aforementioned measures by first calculating the Pearson, Spearman and Kendall correlation coefficients and then calculating the proportion of common top-k nodes obtained by using the centrality measures.

5.1 Datasets

For the evaluation, we use four probabilistic social networks from the literature. Table 1 summarizes the characteristics of these datasets.

5.1.1 Enron

The first dataset is a snowball sample of the Enron email network which consists of emails sent between employees of Enron between 1999 and 2001. Nodes represent employees and there is an edge between two nodes if at least one email has been exchanged between them. The probabilities of the edges are set using equation $p_{i,j} = 1 - \prod_k (1 - \exp(-\mu(t_{now} - t_k)))$ quantifying the probability that a new email will be exchanged between a pair of nodes at time t_{now}. μ is the scaling parameter, and t_k is the time when message k has been exchanged between nodes i and j (Pfeiffer and Neville 2011). The Enron dataset is denser than the others.

5.1.2 Facebook

The second dataset contains two years of wall-to-wall postings between a snowball sample of users in Facebook. There is an edge between two nodes if at least one of them has posted at least one message on another person's wall. The probabilities on the edges come from the same equation in the Enron dataset and represent the likelihood of having an active relationship at time t_{now} (Pfeiffer and Neville 2011).

5.1.3 FriendFeed

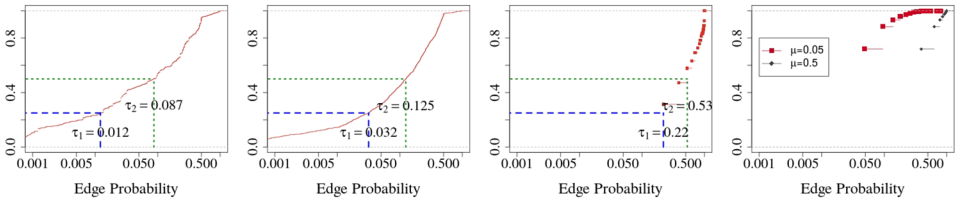
The third dataset is a snowball sample of the FriendFeed network (Magnani et al. 2010) with 150 nodes and 619 edges. We draw an edge between two nodes if they mutually follow each other. The probabilities of an edge is the likelihood that two nodes will exchange a message in the future. This probability is quantified by the exponential function $p_{ij} = 1 - \exp(-\mu n)$, where n is the number of messages exchanges between them in any direction and μ is the scaling parameter with the value of 0.25.

5.1.4 DBLP

The fourth dataset is a snowball sample of the computer science bibliography DBLP dataset. In this network, nodes are authors of papers and two authors have an uncertain edge if they have co-authored at least one paper. The probabilities of the edges are obtained from exponential function $p_{ij} = 1 - \exp(-\mu n)$ determining the probability that two authors will co-author a paper in the future. n is the number of papers that two authors have co-authored in the past and μ is the scaling factor (Parchas et al. 2015).

Figure 6 shows the CDF[2] of edge probabilities of our datasets. The blue dashed lines show the probability

[2] Cumulative distribution function.



(a) Enron **(b)** Facebook **(c)** FriendFeed **(d)** DBLP

Fig. 6 CDF of edge probabilities

Table 2 Correlation coefficients between probabilistic ego betweenness in V-Alters-Ego, F-Alters-Ego and the approximation method

Dataset	$\rho_{V,F}$	$s_{V,F}$	$\tau_{V,F}$	$\rho_{V,APP}$	$s_{V,APP}$	$\tau_{V,APP}$	$\rho_{F,APP}$	$s_{F,APP}$	$\tau_{F,APP}$
Enron	0.98	0.96	0.91	1	0.93	0.88	0.98	0.93	0.86
Facebook	0.98	0.76	0.73	0.99	0.71	0.65	0.99	0.72	0.66
FriendFeed	1	0.99	0.95	1	1	0.97	1	0.99	0.94
DBLP (0.05)	0.91	0.8	0.76	0.99	0.75	0.68	0.91	0.76	0.68
DBLP (0.5)	0.95	0.92	0.86	0.99	0.95	0.9	0.9	0.95	0.89

ρ, s and τ are, respectively, Pearson, Spearman and Kendall correlation coefficients and subscripts V, F and APP, respectively, refer to probabilistic ego betweenness in V-Alters-Ego and F-Alters-Ego definitions and the approximation method

threshold from which 25% of the edges have lower probability (τ_1). Likewise, the green dotted lines indicate the threshold from which 50% of edges have lower probability (τ_2). The deterministic graphs for each dataset is obtained by removing all probabilities from the edges, or by removing all edges with probability lower than the threshold and then considering all the remaining edges as certain edges. For the DBLP dataset, since more than 72% of the edges have the same probability, finding a threshold to remove 25% and 50% low probability edges is impossible. So, instead of using a DBLP dataset that does not include 25% (50%) of its edges, we use two complete DBLP datasets with different scaling parameters $\mu = \{0.05, 0.5\}$.

5.2 Comparing measures in V-Alters-Ego and F-Alters-Ego

5.2.1 Degree

The notion of degree in deterministic networks is replaced by the notion of node degree distribution in probabilistic networks. However, in practice instead of computing the whole distribution its expected value is used: the expected degree, in both V-Alters-Ego and F-Alters-Ego, is the sum of the probabilities of all edges incident to the ego node. The computational complexity of these measures (degree and expected degree) is $O(|V|)$, where V is the number of nodes in network $\mathcal{G}$.

5.2.2 Betweenness

Ego betweenness in each definition of the probabilistic ego network has a different interpretation. In V-Alters-Ego, probabilistic ego betweenness is the expected value of deterministic ego betweenness in all possible worlds. The number of possible worlds increases exponentially as the number of edges in ego network increases. Hence, the calculation of ego betweenness in V-Alters-Ego for even average size ego networks is intractable. Similarly, ego betweenness in F-Alters-Ego is the expected value of deterministic betweenness of nodes in all possible worlds. In the first, just shortest paths with length 2 are counted while in the latter, shortest paths with length higher than 2 have also input in the value of betweenness.

Columns 2 to 4 in Table 2 show high correlation coefficients between probabilistic ego betweenness in V-Alters-Ego and F-Alters-Ego. However, in all datasets Pearson correlation coefficient is higher than Spearman and Kendall. High value for Pearson correlation coefficient reveals that ego betweenness increases/decreases in V-Alters-Ego when it increases/decreases in F-Alters-Ego. Spearman and Kendall correlation coefficients expose the association between two centrality measures regarding ranking, not necessarily the value of centrality measures. Therefore, the lower values of rank correlation coefficients, in comparison to Pearson, show that increase and decrease in the value of ego betweenness in both definitions are not with the same proportion/rate. This motivates us to study the proportion of common top-k ranked nodes obtained by

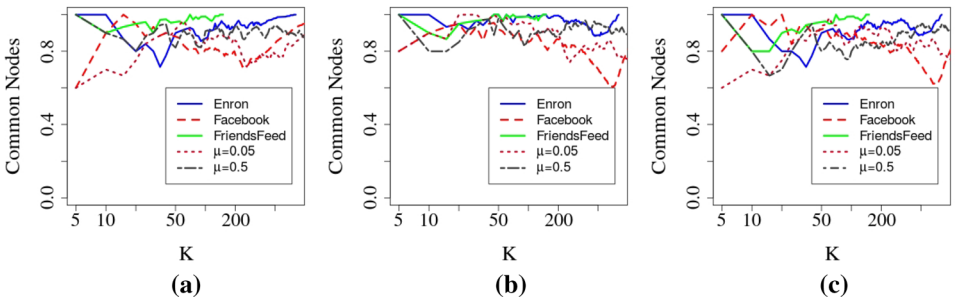


Fig. 7 Proportion of common top-k nodes obtained using probabilistic ego betweenness in **a** F-Alters-Ego and V-Alters-Ego, *b* V-Alters-Ego and the approximation method and *c* F-Alters-Ego and the approximation method

Table 3 Correlation coefficients (Pearson ρ, Spearman s and Kendall τ) between α-*closeness* and expected degree (columns 2–4), between α-*closeness* and the expected betweenness under F-Alters-Ego definition (columns 5–7), between α -*closeness* and the expected betweenness under V-Alters-Ego definition (columns 8–10), and between α-*closeness* and the value of the approximation method for betweenness (columns 11–13)

Dataset	Cl_α, E			$Cl_\alpha, F-btw$			$Cl_\alpha, V-btw$			$Cl_\alpha, A-btw$		
	ρ	s	τ	ρ	s	τ	ρ	s	τ	ρ	s	τ
Enron ($\alpha = 0.03$)	.9	.95	.84	.74	.91	.82	.69	.9	.8	.69	.91	.8
Enron ($\alpha = 0.05$)	.93	.96	.87	.8	.89	.82	.76	.89	.81	.75	.91	.81
Facebook ($\alpha = 0.03$)	.82	.79	.66	.66	.66	.6	.66	.65	.59	.66	.79	.7
Facebook ($\alpha = 0.05$)	.85	.84	.7	.67	.65	.59	.67	.64	.59	.67	.75	.66
FriendFeed ($\alpha = 0.05$)	.51	.73	.66	.3	.74	.68	.35	.73	.67	.37	.72	.67
FriendFeed ($\alpha = 0.1$)	.52	.73	.66	.31	.74	.68	.36	.73	.67	.38	.72	.67
DBLP-0.5 ($\alpha = 0.1$)	.95	.92	.84	.65	.94	.88	.78	.93	.87	.8	.98	.93
DBLP-0.5 ($\alpha = 0.25$)	.95	.92	.84	.65	.94	.88	.78	.93	.87	.8	.98	.93
DBLP-0.5 ($\alpha = 0.5$)	.97	.88	.81	.59	.73	.63	.73	.73	.63	.75	.74	.64

using probabilistic ego betweenness in the two definitions to verify whether the difference in ranking occurs among the top ranked nodes or the medium/low ranked nodes. Figure 7a shows that this difference happens with higher proportion among top-k ranked nodes when k is smaller. Hence, probabilistic ego betweenness in V-Alters-Ego and F-Alters-Ego are not replaceable.

We repeated the same experiments to investigate the difference and similarity of probabilistic ego betweenness in V/F-Alters-Ego with the proposed approximation method in 4. In general, Table 2 shows that the proposed approximation method for probabilistic ego betweenness has a very high Pearson correlation with probabilistic ego betweenness for V-Alters-Ego, however, rank correlation coefficients are low. Again we examine the proportion of common top-k ranked nodes obtained by using probabilistic ego betweenness in V-Alters-Ego and the approximation method. Figure 7b indicates that the difference between the two ranking methods happens when k is a large number.

Generally, the results shown in Table 2 and Fig. 7 suggest that the approximation method for betweenness in Sect. 4 is an appropriate method to approximate probabilistic ego betweenness in V-Alters-Ego.

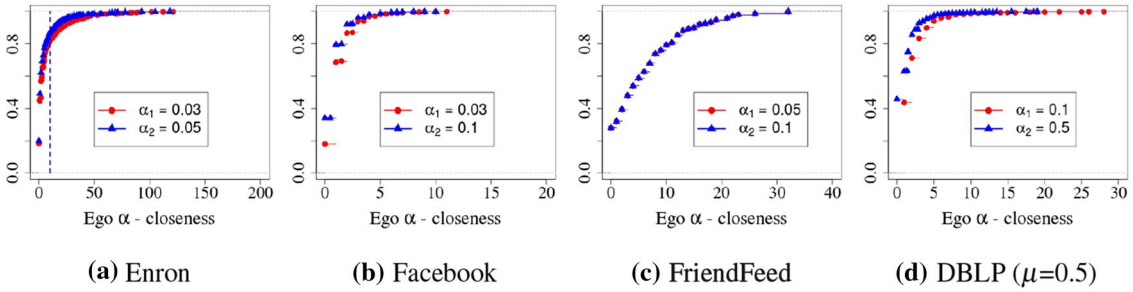
The ego betweenness in V-Alters-Ego and F-Alters-Ego has been obtained by averaging on 15,000 samples from each node's ego networks.

5.2.3 Closeness

In V-Alters-Ego, probabilistic ego closeness is 1 for all nodes by definition. However, in F-Alters-Ego α-*closeness* is capable of making a distinction among nodes in a network. The shorter the distance is between an ego and its alters in at least $\alpha|\mathcal{G}|$ of the possible worlds, the higher value of α -*closeness* this node has. The time complexity of α-*closeness* depends on the time complexity of shortest path length distribution. The calculation of the complete shortest path length distribution needs to generate all possible worlds in F-Alters-Ego. However, α prunes many possible worlds and just considers those possible worlds where the distance between ego node and its alter is as short as possible and the sum of the probability of those possible worlds is greater than or equal to α. Therefore, the smaller α is, the less possible worlds are needed to be generated.

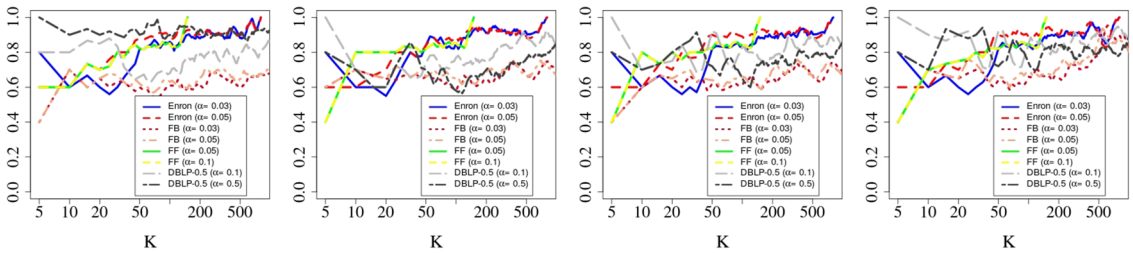
Figure 8 shows the CDF of ego α-closeness in our datasets. According to the definition of α-closeness in Eq. 8 the higher α leads to the lower α-closeness. Figure 8 confirms this property in all the datasets. For example, the dashed line in Fig. 8a demonstrates that 143 nodes have 0.03-closeness higher than 10, while just 105 nodes have 0.05-closeness higher than 10.

To evaluate the proposed ego closeness, we examine the correlation between it and the expected degree which is the same in both probabilistic ego definitions, probabilistic ego



(a) Enron (b) Facebook (c) FriendFeed (d) DBLP (μ=0.5)

Fig. 8 CDF of *α-closeness*



(a) Exp. Degree (b) F_btw (c) V_btw (d) approx_btw

Fig. 9 Proportion of common top-k nodes obtained using *α-closeness* and **a** expected degree, **b** ego betweenness for F-Alters-Ego, **c** ego betweenness for V-Alters-Ego, and ego betweenness for V-Alters-Ego, and **d** approximated value for betweenness

betweenness in V-Alters-Ego, probabilistic ego betweenness in F-Alters-Ego and probabilistic ego betweenness calculated using the approximation method. Table 3 shows high Pearson as well as Spearman and Kendall correlation coefficients between α-closeness and expected degree in all the datasets.

The results in Table 3 show that among all measures (expected degree, V-Ego betweenness, F-Ego betweenness and approximated betweenness), α-closeness has high correlation coefficients just with expected degree. However, Fig. 9a reveals that α-closeness and expected degree do not have high intersection of top-k nodes except for DBLP-0.5 with $\alpha = 0.5$.

Generally for all datasets, the intersection between sets of top-k nodes obtained using α-closeness and other four measures, for small values of k, is neither close to 1, which would have shown that those measure are good replacements for α -closeness, nor close to 0, which would have implied that α -closeness is reflecting completely different local structural properties in comparison to the other four measures (Fig. 9).

6 Conclusions and future works

In this paper, we investigated two definitions of ego networks in probabilistic graphs that we call V-Alters-Ego and F-Alters-Ego. In V-Alters-Ego, first possible worlds are generated and then in each possible world the neighbors of the ego node and the corresponding ego network are defined independently. In F-Alters-Ego, the set of neighbors of an ego node is defined in the initial step and the possible worlds are generated. We examined notions of degree, betweenness and closeness in both definitions. Both V-Alters-Ego and F-Alters-Ego are based on alternative definitions of neighborhood in the literature on probabilistic networks.

We also proposed an approximation method to calculate the extent to which an ego node plays the role of intermediate node among its neighbors in shortest paths with length 2. This approximation method, is not only very close to ego betweenness in the V-Alters-Ego definition, but also computationally simple, i.e., $O(D_e^2)$ where D_e is the number of incident edges to an arbitrary node e.

We believe that this study paves the path for studying more structural properties in probabilistic networks. More precisely, in the future we aim to investigate the approximation of global structural properties of nodes in the network by using their local properties, which is something that has already been done for deterministic ego networks but not investigated for the more general probabilistic case. Moreover, the approximation method to calculate ego betweenness in V-Alters-Ego can be used as a fast-computing local property for nodes in algorithms that aim to maintain local properties of nodes for further processing (Parchas et al. 2018).

Funding Open access funding provided by Uppsala University..

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Aggarwal CC, Wang H (2010) Graph data management and mining: a survey of algorithms and applications. In: Managing and mining graph data. Springer, pp 13–68

Arnaboldi V, La Gala M, Passarella A, Conti M (2014) The role of trusted relationships on content spread in distributed online social networks. In: European Conference on Parallel Processing. Springer, pp 287–298

Arnaboldi V, Conti M, La Gala M, Passarella A, Pezzoni F (2016a) Ego network structure in online social networks and its impact on information diffusion. Comput Commun 76:26–41

Arnaboldi V, La Gala M, Passarella A, Conti M (2016b) Information diffusion in distributed OSN: the impact of trusted relationships. Peer-to-Peer Netw Appl 9(6):1195–1208

Arnaboldi V, Conti M, Passarella A, Dunbar RI (2017) Online social networks and information diffusion: the role of ego networks. Online Soc Netw Media 1:44–55

Asthana S, King OD, Gibbons FD, Roth FP (2004) Predicting protein complex membership using probabilistic network reliability. Genome Res 14(6):1170–1175

Bernard HR, Killworth PD, Sailer L (1979) Informant accuracy in social network data IV: a comparison of clique-level structure in behavioral and cognitive network data. Soc Netw 2(3):191–218

Bernard HR, Killworth PD, Sailer L (1982) Informant accuracy in social-network data V. An experimental attempt to predict actual communication from recall data. Soc Sci Res 11(1):30–66

Bonchi F, Gullo F, Kaltenbrunner A, Volkovich Y (2014) Core decomposition of uncertain graphs. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp 1316–1325

Brugere I, Gallagher B, Berger-Wolf TY (2018) Network structure inference, a survey: motivations, methods, and applications. ACM Comput Surv (CSUR) 51(2):1–39

De Choudhury M, Mason WA, Hofman JM, Watts DJ (2010) Inferring relevant social networks from interpersonal communication. In: Proceedings of the 19th international conference on World wide web, pp 301–310

Everett M, Borgatti SP (2005) Ego network betweenness. Soc Netw 27(1):31–38

Fushimi T, Saito K, Ikeda T, Kazama, K (2018) A new group centrality measure for maximizing the connectedness of network under uncertain connectivity. In: International conference on complex networks and their applications. Springer, pp 3–14

Gao X, Chen Z, Wu F, Chen G (2017) Energy efficient algorithms for *k*-sink minimum movement target coverage problem in mobile sensor network. IEEE/ACM Trans Netw 25(6):3616–3627

Kaveh A, Magnani M, Rohner C (2019) Comparing node degrees in probabilistic networks. J Complex Netw 1:1. https://doi.org/10.1093/comnet/cnz003

Killworth PD, Bernard HR (1979) Informant accuracy in social network data III: a comparison of triadic structure in behavioral and cognitive data. Soc Netw 2(1):19–46

Lu Z, Sun X, La Porta T (2016) Cooperative data offloading in opportunistic mobile networks. In: IEEE INFOCOM 2016-The 35th annual IEEE international conference on computer communications. IEEE, pp 1–9

Magnani M, Montesi D, Rossi L (2010) Friendfeed breaking news: death of a public figure. In: 2010 IEEE second international conference on social computing. IEEE, pp 528–533

Marsden PV (2002) Egocentric and sociocentric measures of network centrality. Soc Netw 24(4):407–422

Mukherjee AP, Xu P, Tirthapura S (2017) Enumeration of maximal cliques from an uncertain graph. IEEE Trans Knowl Data Eng 29(3):543–555

Pantazopoulos P, Karaliopoulos M, Stavrakakis I (2013) On the local approximations of node centrality in internet router-level topologies. In: International workshop on self-organizing systems. Springer, pp 115–126

Parchas P, Gullo F, Papadias D, Bonchi F (2015) Uncertain graph processing through representative instances. ACM Trans Database Syst (TODS) 40(3):1–39

Parchas P, Papailiou N, Papadias D, Bonchi F (2018) Uncertain graph sparsification. IEEE Trans Knowl Data Eng 30(12):2435–2449. https://doi.org/10.1109/TKDE.2018.2819651

Pfeiffer JJ, Neville J (2011) Methods to determine node centrality and clustering in graphs with uncertain structure. In: Fifth international AAAI conference on weblogs and social media

Poisot T, Cirtwill AR, Cazelles K, Gravel D, Fortin MJ, Stouffer DB (2016) The structure of probabilistic networks. Methods Ecol Evol 7(3):303–312

Potamias M, Bonchi F, Gionis A, Kollios G (2010) K-nearest neighbors in uncertain graphs. Proc VLDB Endow 3(1–2):997–1008

Rhodes DR, Tomlins SA, Varambally S, Mahavisno V, Barrette T, Kalyana-Sundaram S, Ghosh D, Pandey A, Chinnaiyan AM (2005) Probabilistic model of the human protein–protein interaction network. Nat Biotechnol 23(8):951

Roberts SG, Dunbar RI (2011) Communication in social networks: effects of kinship, network size, and emotional closeness. Pers Relationsh 18(3):439–452

Socievole A, Marano S (2012) Exploring user sociocentric and egocentric behaviors in online and detected social networks. In: 2012 2nd Baltic Congress on Future Internet Communications (BCFIC). IEEE, pp 140–147

Srihari S, Leong HW (2013) A survey of computational methods for protein complex prediction from protein interaction networks. J Bioinform Comput Biol 11(02):1230002

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.