Local Measures for Probabilistic Networks

Amin Kaveh
Local Measures for Probabilistic Networks

Amin Kaveh
amin.kaveh@it.uu.se

September 2019

Division of Computing Science
Department of Information Technology
Uppsala University
Box 337
SE-751 05 Uppsala
Sweden

http://www.it.uu.se/

Dissertation for the degree of Licentiate of Philosophy in Computer Science

© Amin Kaveh 2019
ISSN 1404-5117
Printed by the Department of Information Technology, Uppsala University, Sweden
Abstract
Modeling and analysis of imperfection in network data is essential in many applications such as protein-protein interaction networks, ad-hoc networks and social influence networks. In the study of imperfect network data, three issues have to be considered: first the type of imperfection, second the aspects of networks such as existence of nodes/edges or attributes of nodes/edges in which imperfection occurs and third the theory that has been used to represent imperfection. This thesis, first, reviews the different types of imperfection and consolidates the meaning of the terms used in literature. Second, it discusses network aspects and theories through which imperfect network data is represented and analyzed. Amongst all, the most applied model is uncertainty about existence of edges which is represented using probability theory, called probabilistic networks. Third, this thesis surveys queries and algorithms which have been applied over probabilistic networks.

Forth and the main focus of this dissertation is to look deeply at nodes’ local properties in probabilistic networks. In our first contribution we have shown that two nodes with the same expected degree can have different properties. In this work we have highlighted the role of other summary information of degree distribution such as variance and skewness in addition to the expected value. In our second contribution, we have introduced two possible definitions of probabilistic ego networks and we have studied the concepts of degree, ego betweenness and ego closeness.

One of the main applications of the proposed local properties could be in the sparsification process, in which a network’s edges and the probability of the edges are altered, but nodes’ local properties are preserved.
To Saeed’s parents,
who have given new meaning to patience
List of papers

This thesis is based on the following papers:

I  **Comparing Node Degrees in Probabilistic Networks**  
Amin Kaveh, Matteo Magnani, Christian Rohner  
Journal of Complex Networks  

II  **Defining and Measuring Probabilistic Ego Networks**  
Amin Kaveh, Matteo Magnani, Christian Rohner  
Applied Network Science Journal  
SpringerOpen, Revised Version is submitted on June 16, 2019.

Reprints were made with permission from the publishers.
## Contents

1 Introduction .............................................................................................................. 9

2 Imperfect Information: Terminology ................................................................. 13
   2.1 What is Information? ..................................................................................... 13
   2.2 Sources of Imperfect Information ............................................................... 14
       2.2.1 Communication .................................................................................. 15
       2.2.2 Observation ....................................................................................... 15
   2.3 Types of Imperfect Information ................................................................... 16
       2.3.1 Communication .................................................................................. 16
       2.3.2 Observation ....................................................................................... 16

3 Imperfect Network Analysis: Theories ............................................................. 19
   3.1 Imperfection In Networks ............................................................................ 19
       3.1.1 Existence of Nodes and Edges ............................................................ 19
       3.1.2 Weights/Labels of Nodes and Edges ............................................... 20
   3.2 Theories ......................................................................................................... 21
       3.2.1 Probability Theory ........................................................................... 21
       3.2.2 Dempster-Shafer Theory of Evidence .......................................... 22
       3.2.3 Fuzzy Theory .................................................................................... 24
       3.2.4 Uncertainty Theory .......................................................................... 25
       3.2.5 Comparing Theories ........................................................................ 27

4 Probabilistic Networks: Models and Algorithms ........................................... 37
   4.1 Probabilistic Models ..................................................................................... 37
       4.1.1 Imperfect Information About Existence of Edges ....................... 37
       4.1.2 Other models ..................................................................................... 38
   4.2 Probabilistic Networks Analysis ................................................................ 39
       4.2.1 Path Based Queries ......................................................................... 39
       4.2.2 Pattern Extraction Queries ............................................................ 41
       4.2.3 Clustering ......................................................................................... 43

5 Summary of Papers ............................................................................................ 47
   5.1 Paper I ......................................................................................................... 47
   5.2 Paper II ....................................................................................................... 48

6 Discussion and Future Work .............................................................................. 49

7 Acknowledgements .............................................................................................. 51

References ............................................................................................................... 53
1. Introduction

A network or a graph is a collection of nodes which are connected via edges. Many systems in which entities interact together in some way can be modeled as a network [86]. For example, in a social system, nodes represent individuals and edges represent any kind of relationship between them.

Modeling a system as a network helps us to understand the structure of that system and to quantify how the structure of the system affects its dynamics and functions [20, 21, 87]. The study of networks has led us to a wide variety of topological measures and concepts e.g. ranking, clustering [32] and diffusion [71] with which many functions and behaviors of networks can be explained [4].

The implicit underlying assumption in a number of models is that the collected data from systems is detailed enough to support the analysis being performed. However, empirical data collection is often an imperfect process affected by different causes. For example because of noisy measurements, e.g., in biological experiments [117], or 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 [3, 7]. As another example in social networks, interactions are subject to some misunderstandings, due to the forgetfulness of informants [54, 55].

There are two main approaches to deal with the collected information and construct a network: the first is neglecting imperfection in information and building a network with the assumption that the constructed network is sufficiently detailed to represent the system [59]. The latter is to construct a network representing imperfection in information and then developing and deploying algorithms and methods to analyze them. The first approach leads to models which are simpler to analyze; however in some cases they oversimplify the modeled system. On the contrary, the second approach seems more informative because it reflects all the imperfection in information. For example, demand on a road (which can be the amount of traffic that has to traverse from one node to another one, or travel between two nodes in a limited time) can not be presented as a piece of perfect information [48, 109, 118]. As another example, we can not represent the strength of future interactions between individuals in a social network [23, 46, 108]. However these models are more complicated to analyze. Figure 1.1 shows the two approaches of representing information as network models.

The most applied imperfect network model is $G(V, E, P)$ where $V$ is the set of nodes, $E$ is the set of edges and $P : E \rightarrow (0, 1]$ is a function that assigns
existence probability to each edge where existence of each pair of edges is independent of each other. This model is called a probabilistic network model and has been used widely to represent imperfect network data in sensor networks [35, 56, 98], opportunistic networks [72, 73], protein-protein interaction networks [18, 100, 112, 113], social influence networks [9, 14, 95] and road networks [34]. In this model, a network with $m$ edges corresponds to $2^m$ deterministic networks which are called possible worlds. In possible world semantics, the expected value of any measure in $G(V,E,P)$ equals the average of that measure over all possible worlds. As the number of possible worlds is exponential in the number of edges, exact processing of measures on this model is computationally prohibitive even for moderately sized networks. Therefore, to approximate the value of desired measures on such networks, we can sample a subset of possible worlds and calculate the expected value of that measure on all samples. Although this approach takes less computational time, it still incurs high computational cost especially for networks with high entropy\(^1\) [88].

To tackle this problem two approaches have been proposed: the first approach proposes to extract a representative deterministic graph, $G'(V,E')$, in which (a) the expected degree of each node in $G$ is as close as possible to its degree in $G'$ and (b) the sum of the expected degree of all nodes in $G$ is as close as possible to the sum of the degree of nodes in $G'$ [89,90]. The second approach, called sparsification, proposes methods to generate an alternative

\(^1\)The entropy of a $G$ is defined as: $H(G) = \sum_{e \in E} (-p_e \log p_e) + (-q_e \log q_e)$, where $q_e = 1 - p_e$. 

\[\]
network to $G$ as $G''(V, E'', P'')$ which is still a probabilistic network, but with a different set of edges and probability assignment and with less entropy. The algorithms proposed in this approach seek a $G''$ in which (a) $|E''| = \alpha |E|$ where $0 < \alpha < 1$, (b) the expected degree of each node in $G$ to be as close as possible to the expected degree of that node in $G''$, and (c) the sum of the expected degree of all nodes in $G$ to be as close as possible to the sum of expected degree of all nodes in $G''$ [88].

Although these two approaches facilitate the approximation of measures in probabilistic networks, they have considered only degree (and expected degree) as the nodes’ local properties to preserve in their analysis. As the main direction of this dissertation, we aim to look deeply at other local properties that, once preserved in the process of sparsification, would help us to approximate the desired measures with more precision.

To do so, first we start to understand the local properties of nodes in probabilistic networks. In our first contribution (see Section 5.1), we have shown that two nodes with the same expected degree can have different properties. In this work we have highlighted the role of other summary information of degree distribution such as variance and skewness in addition to the expected value. More precisely, in this paper we have shown that in some cases expected degree does not allow us to evaluate the probability of two nodes having the same degree or one node having higher degree than another. Therefore, for each node we define a reference node with the same expected degree but the least possible variance, corresponding to the least uncertain degree distribution. Then, we have shown how the probability of a node’s degree being higher or equal to the degree of its reference node can be approximated by using variance and skewness of the degree distribution in addition to expected degree. In our second contribution (see Section 5.2) we have defined the concept of ego networks in probabilistic networks. Under two possible definitions of probabilistic ego networks, V-Alters-Ego and F-Alters-Ego, we have studied the concept of degree, ego betweenness and ego closeness. We have also evaluated the difference of these three measures under the selected definitions.

As future works, we pose the following questions:

- “whether the comparing measures proposed in the first paper can help us in sparsification process”.
- “whether preserving the proposed ego measures in the process of sparsification leads to more accurate sparsified networks”.

**Structure of the thesis.** To strengthen our knowledge about imperfect network models and analysis, we started our line of research with the following questions. Figure 1.2 illustrates the structure of Chapters 2-4.

1. “what is the meaning of imperfect information?”, “what are sources of imperfection?” and “which terminology can we use when we are talking about imperfection in information?”. In Chapter 2 we answer the first group of questions and hence we provide a terminology of imperfection.
This terminology is presented in such a way to consider imperfection following the sources of information.

2. “which type of imperfection does occur on which aspect of networks?”. More precisely, sometimes information about the existence of nodes and edges is imperfect and sometimes information about the attributes of nodes and edges is imperfect. Moreover, we ask “which theories have been employed to model and analyze imperfect network data?”. In Chapter 3, first we shortly review the literature that have considered imperfection on various aspects of networks. Then we briefly introduce the theories that have been used in the analysis of imperfect network data.

3. As the most common theory to handle imperfection in network data is probability theory we aim to answer these questions: “what are the main network models under probability theory?” and “what are the state of the art studies in this area?”. In Chapter 4, we review probabilistic network models, analysis methods and algorithms.

As we show in Chapter 4, most of the studies in probabilistic networks are focused on the global properties of networks such as clustering, community detection, core decomposition and so on. However, local properties of nodes are not comprehensively studied. In Chapter 5 we briefly review the motivations, contributions and findings of our articles which are focused on local structural properties of nodes in probabilistic networks. The last chapter includes discussion and future works.
2. Imperfect Information: Terminology

Imperfect information is, in literature, often referred to with terms such as \textit{random, uncertain, imprecise, vague} and \textit{ambiguous}. These terms not only indicate the type of imperfection but also sometimes reveal the theory through which imperfect information is modeled and analyzed. The challenge is that there is no consensus on the use of the aforementioned terms in scientific articles. For example, the term \textit{uncertain} in a large group of scientific articles refers to imperfect information on the existence of nodes/edges modeled by probability theory. However, in another group of articles, the term \textit{random} is used for the same purpose whilst the term \textit{uncertain} is used to specify that \textit{uncertain theory} has been used to model imperfect information.

The goal of this chapter is to consolidate the meaning of the terms used in the literature. To do so, we focus on the definition of information and one of the most important aspects of information: the primary sources of information. Then, we present the definition of each type of imperfection with regard to the source of information. We cover the theories through which imperfect information is modeled and analyzed in the next chapter.

2.1 What is Information?

This section surveys the definition of information. The term information is mixed with other related terms such as \textit{knowledge, fact} and \textit{data}. The goals are first to find out whether there is a unique definition of information in dictionary and second, to know whether the provided definitions are independent from other similar terms. Having a universal and unique definition of information leads to better understanding of the concept of imperfection in information.

First, we check generic dictionaries. Cambridge and Longman dictionaries both define information in relation to the term \textit{fact}. Specifically, Cambridge dictionary defines information as \textit{“facts about a situation, person, event, etc”} and Longman dictionary defines it as \textit{“facts or details that tell you something about a situation, person, event etc”}. Merriam-Webster dictionary presents a definition which is more intertwined with other similar terms such as \textit{knowledge} and \textit{intelligence}, i.e. \textit{“the communication or reception of knowledge or intelligence”}. Oxford dictionary has a definition independent from other similar terms, but it is still not entirely comprehensive, i.e. \textit{“What is conveyed or represented by a particular arrangement or sequence of things.”}
More technical dictionaries suggest definitions with more detail for the term information. A Dictionary of Computer Science refers to information as “whatever is capable of causing a human mind to change its opinion about the current state of the real world”. Free on-line dictionary of computing (FOLDOC) defines information as “The result of applying data processing to data, giving it context and meaning. Information can then be further processed to yield knowledge”. Although this definition highlights the difference between data, information and knowledge (i.e. information is processed data and unprocessed knowledge) it is not explicit how much processing is required to transform data to information or how much processing converts it into knowledge.

In addition to generic and technical dictionaries, several scientific works have explored the definition of information in different disciplines and in relation to other relative words such as data and knowledge [31, 99], however, none of them suggest a universal and unique definition.

Shannon’s information theory is not applicable in this case since it describes a model to quantify the amount of information in a message regardless of its semantics and of imperfection in information [6]. Shannon clearly states in his paper that the semantics of information is not addressed by his theory: “These semantics aspects of communication are irrelevant to the engineering problem” [102]. On the contrary, here we aim at considering imperfection in information with regard to its semantics.

Finding a universal and a unique definition of information is still an open issue in scientific communities but, by focusing on the aforementioned definitions, we can find two primary aspects of information that help us to present a terminology about imperfection in information. These two aspects can be seen as the sources from which we get information: we acquire information via communication and observation. Zorkovcy and Heap [129] highlight the capability of information to be communicated and argue that apart from different definitions of information, it needs to be conveyed. Moreover, Rowley [99] points out that information can be derived from observations.

2.2 Sources of Imperfect Information

In this section, we study different types of imperfection with regard to the primary sources of information: communication and observation. More precisely, we present a terminology that shows which types of deficiency in the semantics of information can occur if the source of information is communication or observation. Figure 2.1 shows two primary sources of imperfection. Our discussions about the definition and categorization of terms rely on the

\[1\text{The term communication, here is used in a general meaning and includes any way of conveying information from narrator’s mind to audience mind, such as linguistic conversation, sign language and....}\]
2.2.1 Communication
Acquired information is imperfect if it is the result of inaccuracy in communication. The most common way of conveying information is language. A piece of information is imperfect if the way that it is presented by a language includes any linguistic complexity. This means that a piece of information in the narrator’s mind can be different from what is perceived from what is written or spoken. For example, statements like “John is tall”, “Jim is the connector between new and experienced members” and “Mike and Rob have an intimate relationship” are different pieces of information with linguistic complexity, though they can be clear for the person who provides the information.

2.2.2 Observation
Information can be imperfect if there is some sort of inaccuracy in observations. For example, if there is no confidence in the numerical height of someone or the number of the neighbors of a node in a network, our observation is not accurate and our information is not perfect.
2.3 Types of Imperfect Information

In this subsection, we introduce different types of imperfection regarding our two aforementioned sources of imperfection.

2.3.1 Communication

The primary meaningful linguistic object is the *statement* which is a sentence that conveys information [22]. A statement can include imperfection because of linguistic complexity. The two main types of linguistic imperfection are *ambiguity* and *vagueness*:

**Linguistic Imperfection – Ambiguity:** Ambiguous statement refers to the statement which implies more than one meaning. To clarify, we can interpret “new bikes and cars” either as “(new bikes) and cars” (i.e. just bikes are new) or “new (bikes and cars)” (i.e. both bikes and cars are new) [22]. As another example, “I saw him with a field glass” could be understood either as “I saw him through a field glass” or “I saw him and he had a field glass”. If there is not enough information to eliminate ambiguity, multiple meanings of a statement lead to inferring multiple possible worlds for the reader/listener, though there is a unique and particular meaning for the writer/speaker [79].

**Linguistic Imperfection – Vagueness:** Vague statement refers to the statement that contains vague concepts. A concept is vague if we have doubt about which objects belong to it [22]. For instance, "fat" is a vague concept because a man who is 80 Kg is neither clearly fat nor clearly thin. Almost all linguistic adjectives such as "fat", "tall", "smart", "short" and "crazy" are vague concepts because there are different degrees of possession of these kinds of qualities between different entities. All textual and spoken information can be subject to vagueness and ambiguity.

2.3.2 Observation

Observational imperfection occurs because of deficiency in measurement. In other words, imperfection occurs in objects of thought rather than objects of language. For example, we are not sure that one’s age is 20 or 21. In order to define different kinds of observational imperfection we should, first, define *proposition*.

**Proposition:** proposition refers to the meaning of a statement where all ambiguity and vagueness is removed [22]. Figure 2.2 shows the relationship between statements and propositions. The Stanford Encyclopedia of Philosophy defines proposition as “the semantic content” of a statement. Moreover, propositions convey modal properties which are “necessity, possibility and impossibility” [78]. Therefore, propositions always have two parts: a) unambiguous and un-vague meaning of statements and b) modal properties. For instance, in the proposition corresponding to statement $S = \text{‘I’m 60% sure that}$
Figure 2.2. Proposition is an un-vague and unambiguous statement.

John is between 180 and 190 cm", the degree of confidence 60% relates to the modal property and the margin between 180 and 190 relates to the meaning of the proposition.

According to the definition of proposition, imperfection can happen in both parts of propositions: the meaning and the modal property. If imperfection happens in the meaning of a proposition, it relates to the precision of the proposition and if imperfection occurs in the modal property of a proposition it relates to its certainty:

**Observational Imperfection – Imprecise information**: Information about one or multiple properties is imprecise if it provides more than one possible values for that or those properties i.e. it corresponds to multiple possible worlds. For example, “John is 180 or 181 cm tall” is imprecise information and corresponds to two worlds where in the first world John is 180 cm and in the second world John is 181 cm.

**Observational Imperfection – Uncertain information**: Information about one or multiple properties is uncertain if it is stated by partial belief, i.e. (0 < belief < 1). For example, “I’m sure that John is 181 or 182 cm tall” is a certain proposition while “I’m 90% sure that John is 180 cm tall” is an uncertain one.

Figure 2.3 shows different types of imperfection regarding the meaning and the modal property of propositions. Meaning and modal property can be precise/imprecise and certain/uncertain respectively. On one hand, if the meaning of a proposition is precise and its modal property is certain, then it is called perfect proposition or perfect information. On the other hand, if the meaning of a proposition is precise with uncertain modal property or if it is imprecise regardless of its modal property, it is called observational imperfect information. Furthermore, there are two specific cases of observational imperfection:
Observational Imperfection – Missing information: The extreme kind of imprecise information is missing information in which the set of possible values is identical to the domain of the property. In this case, the degree of belief to the elements are equal. For example, if we do not have any information about the marital status of a person, our information will be \{married, unmarried\} with equal degree of confidence.

Observational Imperfection – Inconsistent information: A special kind of observational imperfection, related to the combination of propositions, is inconsistent information. In other words, information about one or multiple properties is inconsistent if there is no possible world compatible with given information. For example, two pieces of information “John is 190 cm tall” and “John is 175 cm tall” are inconsistent.
3. Imperfect Network Analysis: Theories

This chapter includes two main sections. The first section briefly reviews the literature in which one or multiple aspects of networks have been considered imperfect. By “aspects of networks”, we mean existence of nodes, existence of edges, attribute values of nodes and attribute values of edges. Networks in which one or multiple aspects have imperfect information are generally called imperfect networks. We use the terminology that is introduced in Chapter 2 to refer to the type of imperfect network. The main goal of the first section is to show that the subject of having imperfect information in network data is common.

The second section reviews different theories that have been used to capture properties of imperfect networks within feasible and formalized mathematical frameworks. It commences with the most practiced theory, probability theory, and continues with a short review on other applied theories such as Dempster-Shafer theory of evidence, fuzzy theory and uncertainty theory.

3.1 Imperfection In Networks

In this section, we show that imperfection in network data is common. We briefly review the literature in which existence of nodes and edges are uncertain as well as the scientific contributions in which attribute values of nodes and edges are imprecise.

3.1.1 Existence of Nodes and Edges

The simplest network model is represented as $G = (V,E)$ where $V$ is the set of nodes and $E$ is the set of edges between nodes. This model has been used extensively in various fields to show the binary relationships (existence of a specific type of connection) between nodes. However, information about the existence of interaction between nodes and also about the existence of nodes can be imperfect. When we are just capable of assigning complete confidence on the existence of nodes or edges between nodes, our information is uncertain [104] (see Section 2.3).

Mering et al. in [117] have shown that due to noisy experiments, we can not be assured about the existence of interaction between proteins. Models that represent uncertainty about the existence of edges is also used in road
networks, telecommunication networks, opportunistic networks and wireless sensor networks. As we can not always be assured that a road is open, as traffic conditions change, our information about opening/blockage of a road between two cities is uncertain. Moreover, Lee et al. in [60] show that the connection between routers can be teared down due to failures in physical equipment. Existence of wireless communication between sensors can also be uncertain because of possibility of failures of connections [43]. Data delivery success between two mobile nodes in opportunistic networks can also be uncertain because we are not always assured of it [72, 73]. All these works elaborated this issue that the existence of interaction between nodes is not guaranteed and our information about the existence of interactions is not always certain.

In addition to uncertainty about the existence of an edge, existence of nodes in a group can be uncertain [115, 123, 132, 134].

3.1.2 Weights/Labels of Nodes and Edges

As the simplest network model – that just considers binary relationships between entities – is not immensely informative, other models have been proposed during the last two decades which we divide them into two groups. The first group of models are the models where each edge is associated with a positive number, called weight. These models are notated as $G(V, E, W)$ where $W$ represents strength or quality of interactions between nodes. The second group includes the works in which a label or a set of labels are associated with nodes and edges, $G(V, E, L_V, L_E)$, where $L_V$ and $L_E$ represent labels of nodes and edges respectively. For example, nodes’ labels can represent political tendency of users in social networks or types of blogs in web networks [8]. Another example is association of such labels as \{professional, friendship, family\} to show different types of relationships between nodes. While weights in the first group are ratio\(^1\) attributes, labels are not necessarily ratio data and can be nominal\(^2\) attributes.

In a network, we might have certain information about the existence of nodes and edges, along with imperfect information about values of edges’ weights/labels. If the number of possible values that an edge’s attribute can take (as a weight or a label) is higher than one, then our information about the weight/label of that edge is imprecise. Therefore, each possible value can be associated with a degree of confidence (see Section 2.3). Examples for applications of such models are reliability of transportation, when the existence of junctions/cities (nodes) and streets/roads (edges) is certain but the traffic conditions (weights of edges) are imprecise, has been investigated

---

\(^1\) Magnitude of assigned values to ratio attributes are meaningful and all basic mathematical operations (+, −, ×, ÷) can be performed on them.

\(^2\) Values of nominal attributes are just different names and mathematical operations are not practical on them.
in [48, 109, 118]. Moreover, Adar and Ré have introduced a diffusion model in which nodes’ labels represent users’ music preferences in a social network. Each user can have multiple preferences, however with different degree of confidence which refers to imprecise information [1].

### 3.2 Theories

This section reviews four theories that have been used to model and analyze imperfect network data: probability theory, Dempster-Shafer theory of evidence, fuzzy theory and uncertainty theory. We also review the extent to which each theory have been used in networks, except probability theory which we survey it thoroughly in Chapter 4.

#### 3.2.1 Probability Theory

The most practiced and consolidated theory to handle imperfect network data is probability theory. All the possible values of the variable of interest is called *sample space*, \( \Omega \), and any subset of the possible values is called an *event*, \( e \). The degree of belief about occurrence of each event is represented with a probability \( 0 \leq p \leq 1 \). The probability theory is rigorous because it is expressed through Kolmogorov axioms:

1. The probability of an event is a non-negative real number:
   \[
   \forall e \subseteq \Omega, \quad p(e) \geq 0
   \]
2. The probability of event \( \Omega \) is 1,
   \[
   p(\Omega) = 1
   \]
3. Probability of any sequence of mutually exclusive events is the sum of the probabilities of those events,
   \[
   p\left(\bigcup_{i=1}^{\infty} e_i\right) = \sum_{i=1}^{\infty} p(e_i)
   \]

In the case of the network data, the most common example is imperfect data about the existence of an edge between each pair of nodes. In this case, each edge is associated with a probability and the sample space corresponding to each edge includes two mutually exclusive events: \{exist, not exist\}. The degree of belief (probability) to each event either is the result of a set of frequent experiments as in the case when we infer the existence of interaction between proteins in PPI networks [117] or is the result of indirect measurement as in the case when we infer social ties or influence relationships between individuals based on their interactions [3, 7]. In chapter 4 we will discuss different models and various analysis algorithms under probability theory.
3.2.2 Dempster-Shafer Theory of Evidence

Dempster-Shafer theory (DST) is interpreted as a generalization of probability theory to argue about belief and plausibility of events. Let $\Omega$ be a finite set whose elements are mutually exclusive. Set $\Omega$ is called the frame of discernment or domain of reference and $\rho(\Omega)$ is the power set of $\Omega$. There are three important functions: (1) a mass function (or degree of belief) that shows the amount of belief that is allocated to any subset of $\Omega$, $m: \rho(\Omega) \to [0, 1]$ such that $\sum_{A \in \rho(\Omega)} m(A) = 1$ and $m(\emptyset) = 0$, (2) the Belief function and (3) the Plausibility function. It is worth to mention that the value of $m(A)$ applies only to the set $A$ and does not give any information about any subset of $A$ and it is the main difference between DS theory and probability theory. The amount of belief that is allocated to $\Omega$, $m(\Omega)$, is called the uncertainty of evidence. The belief function, notated as $Bel(A)$, is the minimal degree of belief to subset $A$ supported by available information and the plausibility function, $Pl$, is the maximum support that could be given to a subset $A$.

The belief function has the following properties:
1. $Bel(\emptyset) = 0$
2. $Bel(\Omega) = 1$
3. $Bel(A_1 \cup A_2 \cup \ldots \cup A_n) 
\geq \sum_i Bel(A_i) 
- \sum_{i<k} Bel(A_i \cap A_k) 
+ \ldots 
+ (-1)^{n+1} Bel(A_1 \cap A_2 \cap \ldots \cap A_n)$

And the plausibility function has the following properties:
1. $Pl(\emptyset) = 0$
2. $Pl(\Omega) = 1$
3. $Bel(A_1 \cap A_2 \cap \ldots \cap A_n) 
\leq \sum_i Pl(A_i) 
- \sum_{i<k} Pl(A_i \cup A_k) 
+ \ldots 
+ (-1)^{n+1} Bel(A_1 \cup A_2 \cup \ldots \cup A_n)$

The third property of Belief function is called superadditive and the third property of Plausibility function is called subadditive properties. According to the superadditive and subadditive properties that Zadeh in [125] has defined the following relation between $bel$ and $Pl$ functions:

$$Pl(A) = 1 - Bel(\bar{A}) \quad (3.1)$$

and,

$$Bel(A) = 1 - Pl(\bar{A}) \quad (3.2)$$
**Why DST?**

It is important to know what has motivated researchers to use DS theory and which shortages in probability theory has led them to apply it. We can trace back the reasons in the shortages of the probability theory in two aspects.

Firstly, when there is ignorance and we drive conclusions for subsets. In detail, when no information is available (known as ignorance) the probability has to be assigned and the probability is equal for all such cases [42]. For example if we have statement \( s = \text{“The number of members in the lab (n) is 2 or 3”} \), in probability theory we assign equal probability values to each of them: \( p(n=2) = 0.5 \) and \( p(n=3) = 0.5 \). However, in Dempster-Shafer theory we can just set \( m(n=\{2,3\})=1 \) and we can not assign any value to any subset of \( \{2,3\} \).

Secondly, when having a single number probability is problematic. Smets in [111] explains that some authors claim objective probability is not exactly known by observers and all the observer knows is that the probability is between some boundaries. This indicates that the objective probability of the event under consideration is imprecise even if it is the result of successive practical experiments and measurements because the future of the experiments is not always similar to the past [2]. As DST is a special form of upper and lower probability, \( \text{Bel}(A) \leq m(A) \leq \text{Pl}(A) \), with a well-defined rigorous mathematical rules, this theoretical framework has been used to model and to analyze imperfection in networks.

**How is DST used to model imperfection in networks?**

Evidential theory has been used to model imperfection in attributed graphs [24–27, 75–77, 93]. In all articles nodes’ attributes are communities in which nodes are members. The attributes on edges indicate the types of connections \{friendly, professional, family, \ldots\} and the attributes of the messages (passing through edges) show the categories or the contexts of messages, e.g. \{commercial, personal, \ldots\}. The set of communities, the set of types of connections and the set of the contexts of messages are called the frame of discernment of nodes, the frame of discernment of edges and the frame of discernment of messages, respectively.

All the mass function values assigned to the subsets of attributes for nodes, edges and messages are assigned randomly without considering any information that support those values. However these research articles show how evidential theory (DST) can be used to predict spammed links [24] and to remove noise from data [26].

In summary, the works on evidential theory on networks are in the elementary steps in both modeling and analysis. In the aforementioned references two main questions are not clearly answered: first, how DST empowers us to model and to analyze imperfection in networks and second, how mass function value has to be assigned to subsets of the set of discernment \( \Omega \) based on the available information.
3.2.3 Fuzzy Theory

Fuzzy set theory [124] is useful when we have partial information about the membership of an object to set. It is generalization of the traditional dual logic that an element can either belong to a set or not. For example, to define the set of height values (in cm) of tall people we are not sure if 180 cm is the height of a tall person or not. However, in fuzzy set theory we can assign a degree of truth to this height and say that the degree of membership of 180 in the fuzzy set of height values of tall people is 0.7.

Let \( U \) be a universe of discourse. The function \( \mu_F \) is the membership function that assigns a degree of membership of any element in \( U \) in the fuzzy set \( F \):

\[
\mu_F : U \to [0, 1]
\]

Then, we can represent fuzzy set \( F \) as:

\[
F = \{ \mu_F(u_1)/u_1, \mu_F(u_2)/u_2, \ldots, \mu_F(u_n)/u_n \}
\]

To extend the aforementioned example, we can define \( U = \{1, 2, 3, \ldots, 200\} \) and \( F = \{\ldots, 0/150, \ldots, 0.7/180, \ldots, 1/199, 1/200\} \) which shows the degree of membership of people with height 150cm to the fuzzy set of tall people is 0, while the degree of membership of 180cm and 200cm tall people are 0.7 and 1, respectively.

For two fuzzy sets \( A \) and \( B \) on the same universe of discourse \( U \) with the membership functions \( \mu_A \) and \( \mu_B \), the fuzzy set operations are defined [74]:

1. Union: the union of fuzzy sets \( A \) and \( B \) is a new fuzzy set with membership function \( \mu_{A\cup B} : U \to [0, 1] \), where

   \[
   \forall u \in U : \mu_{A\cup B} = \max(\mu_A(u), \mu_B(u))
   \]

2. Intersection: the intersection of fuzzy sets \( A \) and \( B \) is a new fuzzy set with membership function \( \mu_{A\cap B} : U \to [0, 1] \), where

   \[
   \forall u \in U : \mu_{A\cap B} = \min(\mu_A(u), \mu_B(u))
   \]

3. Complementation: the complementation of fuzzy set \( A \), \( \overline{A} \) is a fuzzy set with membership function \( \mu_{\overline{A}} : U \to [0, 1] \), where

   \[
   \forall u \in U : \mu_{\overline{A}} = 1 - \mu_A(u)
   \]

Why fuzzy theory?

As fuzzy set theory provides a tool to represent the degree of membership of objects in a set, it becomes a convenient way to model the properties in the network that are intertwined with linguistic adjectives. For example, good is a vague adjective. To represent the extent to which the relation between two individuals is good, fuzzy set theory can provide a quantitative measure
by proposing a membership function based on the frequency of the messages passing between individuals.

In general, fuzzy set theory is applicable in network science when we want to model the membership of a node/edge or a node’s/edge’s attribute in a set where the concept of that set is linguistically vague (see Chapter 2), but there are some membership functions and some quantitative values that can be used to find the degree of membership of that measure to that adjective. For example, the number of messages traversing between nodes can be used as a measure to denote the degree to which an edge with \( m \) traversed messages can be a member of friendship fuzzy set.

**How is fuzzy theory used to model imperfection in networks?**

The authors in [19, 30, 46, 47, 62, 80] have defined multiple fuzzy sets called *strong relation*, *weak relation* and so on. Then they have assigned to each edge the degree of membership in the fuzzy sets. The degree of membership is assigned based on predefined functions. Moreover, the authors in [81, 84, 108] have used a model in which not only edges are fuzzy, but also the association of nodes in the network under consideration is based on a membership function.

### 3.2.4 Uncertainty Theory

Uncertainty theory is devised by Liu in 2007 in [64]. In [69] he claims that uncertainty theory is the only way to model human belief mathematically. Let \( \Omega \) be a nonempty set which is called universal set and \( \mathcal{L} \) be a set of subsets of \( \Omega \) (not necessarily all subsets) that holds the three conditions: (a) \( \Omega \in \mathcal{L} \), (b) if \( A \in \mathcal{L} \) then \( \overline{A} \in \mathcal{L} \), and (c) if \( A_1, A_2, \ldots \in \mathcal{L} \) then \( \bigcup_{i=1}^{\infty} A_i \in \mathcal{L} \). Each element of \( \mathcal{L} \) is called an event.

An *uncertain measure* is a function \( \mathcal{M} : \mathcal{L} \rightarrow [0,1] \) that satisfies the following axioms:

1. (Normality Axiom): for the universal set \( \Omega \), \( \mathcal{M}(\Omega) = 1 \),
2. (Duality Axiom): for any \( A \in \mathcal{L} \), \( \mathcal{M}(A) + \mathcal{M}(\overline{A}) = 1 \)
3. (Subadditivity Axiom): for any countable sequence of events \( A_1, A_2, \ldots \),

\[
\mathcal{M}\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mathcal{M}(A_i)
\]

In order to make a compound uncertain measure, e.g. we have two belief functions about tomorrow’s weather and quality of roads, Liu has proposed Product Axiom in [66]:

4 (product Axiom): let \( \mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_n \) be sets of subsets on universal sets \( \Omega_1, \Omega_2, \ldots, \Omega_n \). Then the compound uncertain measure on the product \( \mathcal{L}_1 \times \mathcal{L}_2 \times \ldots \times \mathcal{L}_n \) holds:

\[
\mathcal{M}\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mathcal{M}(A_i)
\]
Why uncertainty theory?
Liu argues that probability theory is applicable when we have enough samples to draw probabilities. However, when there is no samples to extract the probability distribution, the only solution is to use belief which depends massively on our personal knowledge. Personal knowledge can vary in time and then our belief in different events can change. Liu in [67] discusses why belief degree is different from subjective probability and it is inappropriate to use probability theory to analyze belief degree.

The main difference between probability theory and uncertainty theory is in the case that we want to find the probability (uncertainty measure) of the product of measures. In the probability theory if two events are independent then the probability of product of them is the product of the probability of those events while in uncertainty theory, it is the minimum of the uncertain measure of them (Axiom 4) [65, 69].

How is uncertainty theory used to model imperfection in networks?
Gao in [38] argues that sometimes there is insufficient historical data to estimate statistically nodes’ demands in single facility location problem. Therefore he has used uncertain variable to model the imperfection of nodes’ demands. In this model it is assumed that information on edges is perfect. The authors in [40] have used uncertainty theory to model the variable transportation costs on railways. In their proposed model existence of nodes and edges is certain while the transportation cost (edge attribute) is not deterministic and is represented with a uncertain measure. The authors in [37, 39, 41, 91] have used a simple network model (without any attribute on nodes and edges) in which only existence of edges is modeled with uncertain measure. Similarly, they argue that in many cases having a historical record to infer probability of existence of edges is impossible. As the main reason to choose probability or uncertainty theory is the availability and unavailability of historical data, Liu in [68] considers a situation in which historical data is available just for some part of the network. Therefore, he proposes a model, called uncertain random network, in which existence of some edges are represented with a probability number and some edges are associated with an uncertainty measure. Sheng and Mei have studied shortest path problem in uncertain random model [105]. Shi et al. in [107] have studied the maximum flow problem in uncertain random networks. This model has been used to study various fundamental problems in networks such as minimum spanning tree problem [36, 106].
3.2.5 Comparing Theories

In this section we compare the aforementioned theories using two examples. In the first example (Section 3.2.5.1), we compare probability, Dempster-Shafer and fuzzy theories. In the second example (Section 3.2.5.2), probability and uncertainty theories are compared. All the applied mathematical relations in both examples are based on existing articles in literature.

3.2.5.1 Shortest path length in probability vs. fuzzy vs. DS theories

In this section we study "the shortest path" problem between two nodes in a small network, in which the existence of edges is deterministic. However, lengths of edges are not deterministic. Specifically, Figure 3.1 shows a network with 4 nodes and 4 edges. The length of each edge can be 1 or 2. In Figure 3.1a imperfect information about lengths of edges is modeled using probability theory. For example, the length of edge \((s,v_1)\) is 1 with probability 0.2 and is 2 with probability 0.8. In Figure 3.1b lengths of edges are modeled using fuzzy theory. To illustrate, edge \((s,v_1)\) would belong to the set of edges with length 1 with degree 0.6 and to the set of edges with length 2 with degree 0.8. Figure 3.1c represents imperfect information about the length of edges with DST.

\[
\begin{align*}
\text{(a) Probabilistic} & & \text{(b) Fuzzy} \\
& & \\
\text{(c) DST (Evidential)} \\
\end{align*}
\]

*Figure 3.1. (a) \((p_1, p_2)\) on an edge denote the probability that the edge’s length is 1 \((p_1)\) or is 2 \((p_2)\). (b) \((\mu_1, \mu_2)\) on an edge denote the degree of membership of that edge in the fuzzy sets (length 1) and (length 2), respectively. (c) \((m_1, m_2, m_3)\) on an edge show basic belief functions that the edge’s length is \{1\}, \{2\} or \{1,2\}.**
Probability theory (Figure 3.1a): Computation of the shortest paths in a network whose edges’ lengths are modeled with probability theory is straightforward. By using possible worlds semantics, the probabilistic graph in Figure 3.1a corresponds to 16 deterministic graphs demonstrated in Figure 3.2 whose probabilities are represented in the captions. The shortest path length between \( s \) and \( t \) is retrieved in each possible world which varies between 2 and 4. The probability that the shortest path length between \( s \) and \( t \) is \( l \) equals to the sum of the probability of all possible worlds in which the shortest path distance between \( s \) and \( t \) is \( l \). Figure 3.2q shows the shortest path length distribution between \( s \) and \( t \).

Fuzzy theory (Figure 3.1b): The shortest path problem in fuzzy networks is introduced in various ways [57]. However, we restrict ourselves to the case which is first, simple\(^3\) and second, it is convenient for our example.

Fuzziness is introduced into networks through edge lengths or edge capacities, considering each value of length/capacity as a fuzzy set. We recall that a fuzzy set is a mathematical model notated as a pair \((U, m)\) where \(U\) is a set of objects and \(m: U \rightarrow [0, 1]\) is a function that shows the degree of membership of each object to set \(U\). To clarify, \(U\) can be assumed as the set of all edges in a network with length \(l\). In this set, membership function \(m\) shows the degree of membership of edge \(e\) in the set of all edges with length \(l\). We would like to make it clear that the degree of membership of each edge to \(U\) is obtained from vague information.

In this example, the degree of membership of edge \((s, v_1)\) to fuzzy set \(edges with length 1\) is 0.6 and the degree of membership of that edge to fuzzy set \(edges with length 2\) is 0.8 (see Figure 3.1b).

Generally, in the graph presented in Figure 3.1b each edge can take an integer (1 and 2) as length. We notate the degree of membership of edge between nodes \(i\) and \(j\) to fuzzy set \(I\) as \(\mu_I(i, j)\). Therefore, each edge has 2 membership tuples, e.g. \(M(s, v_1) = \{1/\mu_1(s, v_1), 2/\mu_2(s, v_1)\} = \{1/0.6, 2/0.8\}\).

Klein in [57] has used a method called hybrid multi-criteria dynamic programming recursion to find the shortest fuzzy path length between two nodes. Here, we do not intend to explain the method and suffice to mention the principles:

\[
\begin{align*}
  f(\text{terminal}) &= (1, 1, \ldots, 1) \\
  f(i) &= \text{dom}_{(i,j) \in E}(e_{ij} \tilde{+} f(j)) \quad (3.3)
\end{align*}
\]

where, \(e_{ij}\) is a 2 membership tuple assigned to each edge \((i, j)\) or the path from \(i\) to \(j\): \(e_{ij} = \{\mu_1(i, j), \mu_2(i, j), \ldots, \mu_R(i, j)\}\), \(\tilde{+}\) is the combinatorial sum and \(\text{dom}\) is domination operator: more is better.

---

\(^3\)Shortest path problem in some fuzzy network models require complex and extended mathematical operations, e.g. see extended sum and extended min/max in Dubois and Prade model [29].
(a) $p = 0.008$

(b) $p = 0.072$

(c) $p = 0.032$

(d) $p = 0.008$

(e) $p = 0.002$

(f) $p = 0.288$

(g) $p = 0.072$

(h) $p = 0.018$

(i) $p = 0.032$

(j) $p = 0.008$

(k) $p = 0.002$

(l) $p = 0.288$

(m) $p = 0.072$

(n) $p = 0.018$

(o) $p = 0.008$

(p) $p = 0.072$

(q) Shortest path length distribution

Figure 3.2. (a)-(p) All possible worlds of the graph presented in Figure 3.1a, (q) Shortest path length distribution between nodes $s$ and $t$. 
In the fuzzy graph Figure in 3.1b each edge can take a length of 1 or 2. Then the shortest path length would be a minimum of 2 and a maximum of 4. **Calculation.** By applying principles in Equation 3.3, we obtain:

\[
\begin{align*}
f(t) &= (1,1,1,1) \\
f(v_1) &= (0.8,0,8,0,0) \\
f(v_2) &= (0.5,0.6,0,0) \\
f(s) &= dom\{ e_{s,v_1} \cdot f(v_1), e_{s,v_2} \cdot f(v_2) \} \\
&= dom\{ (0,0.6,0.8), (0,0.4,0.5,0.6) \} \\
&= (0,0.6,0.8,0.8)
\end{align*}
\]

Therefore, the fuzzy shortest path length between \(s\) and \(t\) is:

\[
\tilde{P}_{s,t} = (1/0,2/0.6,3/0.8,4/0.8).
\]

The shortest possible path with length 2 corresponds to the path \(s-v_1-t\) and with length 4, \(s-v_1-t\). The shortest possible path with length 3 corresponds to the path \(s-v_1-t\) in which the length of \((s,v_1)\) is 2 and the length of \((v_1,t)\) is 1.

**DST theory (Figure 3.1c):** Figure 3.1c shows a network in which the imperfect information about the length of edges is represented with basic belief functions, \(m\), defined in Dempster-Shafer theory. Similar to the two previous examples, existence of all edges is certain, however the length of each edge can take values 1 or 2. In other words, the degree of belief that edge \((s,v_1)\) has length 1 is \(m(\{1\}) = 0.7\) and length 2 is \(m(\{2\}) = 0.1\). Additionally, the degree of belief to both values is \(m(\Omega) = 0.2\) which is called uncertainty of evidence.

**Calculating edges lengths.** The following calculations show the flow of retrieving the length of each edge and path according to the canonical procedures described in the Dempster-Shafer theory.

**Edge:** \((s,v_1)\):

\[
\begin{align*}
\Omega &= \{1,2\}, \quad P(\Omega) = \emptyset, \{1\}, \{2\}, \Omega \\
m(\{1\}) &= 0.7, \quad m(\{2\}) = 0.1, \quad m(\Omega) = 0.2 \\
Bel(\{1\}) &= 0.7, \\
Bel(\{2\}) &= 0.1, \\
Pl(\{1\}) &= 1 - Bel(\{1\}) = 1 - Bel(\{2\}) = 1 - 0.1 = 0.9, \\
Pl(\{2\}) &= 1 - Bel(\{2\}) = 1 - Bel(\{1\}) = 1 - 0.7 = 0.3, \\
length(s,v_1) &= [1 \times 0.7, 1 \times 0.9] + [2 \times 0.1, 2 \times 0.3] = [0.9,1.5]
\end{align*}
\]
Edge \((s, v_2)\):

\[
\Omega = \{1, 2\}, \quad P(\Omega) = \{\emptyset, \{1\}, \{2\}, \Omega\}
\]

\(m(\{1\}) = 0.15, \quad m(\{2\}) = 0.8, \quad m(\{\Omega\}) = 0.05\)

\(Bel(\{1\}) = 0.15, \quad Bel(\{2\}) = 0.8, \quad Bel(\{\Omega\}) = 0.2\)

\(Pl(\{1\}) = 1 - Bel(\{1\}) = 1 - Bel(\{2\}) = 1 - 0.8 = 0.2, \quad Pl(\{2\}) = 1 - Bel(\{2\}) = 1 - Bel(\{1\}) = 1 - 0.15 = 0.85\)

\(length(s, v_2) = [1 \times 0.15, 1 \times 0.2] + [2 \times 0.8, 2 \times 0.85] = [1.75, 1.9]\)

Edge \((v_1, t)\):

\[
\Omega = \{1, 2\}, \quad P(\Omega) = \{\emptyset, \{1\}, \{2\}, \Omega\}
\]

\(m(\{1\}) = 0.25, \quad m(\{2\}) = 0.65, \quad m(\{\Omega\}) = 0.1\)

\(Bel(\{1\}) = 0.25, \quad Bel(\{2\}) = 0.65, \quad Bel(\{\Omega\}) = 0.1\)

\(Pl(\{1\}) = 1 - Bel(\{1\}) = 1 - Bel(\{2\}) = 1 - 0.65 = 0.35, \quad Pl(\{2\}) = 1 - Bel(\{2\}) = 1 - Bel(\{1\}) = 1 - 0.25 = 0.75, \quad length(v_1, t) = [1 \times 0.25, 1 \times 0.35] + [2 \times 0.65, 2 \times 0.75] = [1.55, 1.85]\)

Edge \((v_2, t)\):

\[
\Omega = \{1, 2\}, \quad P(\Omega) = \{\emptyset, \{1\}, \{2\}, \Omega\}
\]

\(m(\{1\}) = 0.5, \quad m(\{2\}) = 0.3, \quad m(\{\Omega\}) = 0.2\)

\(Bel(\{1\}) = 0.5, \quad Bel(\{2\}) = 0.3, \quad Bel(\{\Omega\}) = 0.2\)

\(Pl(\{1\}) = 1 - Bel(\{1\}) = 1 - Bel(\{2\}) = 1 - 0.3 = 0.7, \quad Pl(\{2\}) = 1 - Bel(\{2\}) = 1 - Bel(\{1\}) = 1 - 0.5 = 0.5, \quad length(v_2, t) = [1 \times 0.5, 1 \times 0.7] + [2 \times 0.3, 2 \times 0.5] = [1.1, 1.7]\)

It is worth mentioning the direct relation between the uncertainty of evidence \((m(\{\Omega\}))\) and width of the obtained intervals: for edge \((s, v_2)\) whose \((m(\{\Omega\}))\) is 0.05, the width of length interval is \(1.9 - 1.75 = 0.15\). While for edge \((v_2, t)\) whose \((m(\{\Omega\}))\) is larger (0.2) that width is also larger, \(1.7 - 1.1 = 0.6\).

Calculating paths lengths. We can compute the length interval of a path based on the obtained length intervals of its constituent edges: \(path\_length(s, v_1, t) = [0.9, 1.5] + [1.55, 1.85] = [2.45, 3.35]\) and \(path\_length(s, v_2, t) = [1.75, 1.9] + [1.1, 1.7] = [2.85, 3.6]\). To find the shortest path we have to decide based on the length intervals of obtained paths. To compare two paths \(P_1\) and \(P_2\), if
the lowest value of $P_1$’s interval is higher than the largest value of $P_2$’s interval, then $P_2$ is always preferred. However, if intervals overlap then making a decision is critical. In this calculation, we use the methods applied in [85, 114] which compare the central points of two intervals. The centers of $\text{path}_\text{length}(s, v_1, t)$ and $\text{path}_\text{length}(s, v_1, t)$ are 2.9 and 3.225 respectively and path $s, v_1, t$ is chosen as the shortest path between $s$ and $t$.

3.2.5.2 Shortest path length in probability vs. uncertainty theories

In this section we compare two theories to handle imperfect information in networks: probability and uncertainty theories. Similar to the previous example, we consider the shortest path problem in networks where the existence of edges is uncertain. Figure 3.3 shows a similar graph to that which we utilized in Section 3.2.5. In Figure 3.3a, imperfect information about the existence of edges (i.e., uncertainty) is modeled with probability values and, similarly in Figure 3.3b, uncertainty is represented with uncertainty theory. In both models it is assumed that the existence of any pairs of edges is independent of each other. To clarify the similarities and differences between these two theories, we have used the equal values on corresponding edges. We note that according to Liu [69], probability values are retrieved from the results of sequential experiments and "uncertain measure" is just "occurrence possibility" of an event obtained by the judgment of experts.

Figure 3.3. Uncertainty about the existence of edges are modeled with (a) probability values, (b) uncertain measures in uncertainty theory.

Similar to possible worlds semantics, an uncertain graph (existence of edges are modeled with uncertainty measure) corresponds to $2^m$ deterministic networks where, $m$ is the number of edges. In probability theory each deterministic network is called a possible world while in uncertainty theory it is called realization.

Probability theory (Figure 3.3a): All possible worlds corresponding to the probabilistic network in Figure 3.3a are depicted in Figure 3.4. The probability that only path $s - v_1 - t$ is the shortest path between $s$ and $t$ is 0.448, the probability that only path $s - v_2 - t$ is the shortest path is 0.088 and the probability that both paths are the shortest paths between $s$ and $t$ is 0.112.
Figure 3.4. (a)-(p) All possible worlds of the graph presented in Figure 3.3a.

Uncertainty theory (Figure 3.3b): To retrieve the uncertain measure $\mathcal{M}$ of each path under uncertainty theory, we use the theorem that has been described in [37, 39]:

**Theorem 1** Suppose that $\xi_1, \xi_2, \ldots, \xi_n$ are independent Boolean uncertain variables (only receive 0 and 1 values). If $f$ is an increasing Boolean function, then $\xi = f(\xi_1, \xi_2, \ldots, \xi_n)$ is Boolean uncertain variable such that

$$\mathcal{M}\{\xi = 1\} = \sup_{f(B_1, B_2, \ldots, B_n)} \min_{1 \leq i \leq n} \mathcal{M}\{\xi_i \in B_i\},$$

where $B_i$ is a subset of $\{0, 1\}, i = 1, 2, \ldots, n.$
Assume that $P_1$ is a function that denotes only path $s - v_1 - t$ is the shortest path between $s$ and $t$: 

$$P_1 = \begin{cases} 
1 & \text{if } (s - v_1 - t) \text{ is the only shortest path between } s \text{ and } t \\
0 & \text{otherwise}
\end{cases}$$

Similarly we define $P_2$ as the function that denotes only path $s - v_2 - t$ is the shortest path between $s$ and $t$ and $P_{1,2}$ stands as the function that both paths are available.

In the network in Figure 3.3b, each edge is associated with a Boolean uncertain variable which are independent from each other. The values on edges in Figure 3.3b indicate $\mathcal{M} \{ \xi_i = 1 \}$. Following Duality Axiom, $\mathcal{M} \{ \xi_i = 0 \} =$
Based on Theorem 1, uncertain measure of $P_1$, $P_2$ and $P_{1, 2}$ are as follows:

$$
\mathcal{M}\{P_1 = 1\} = \sup_{P_1 = 1} \min_{1 \leq i \leq 4} \mathcal{M}\{\xi_i \in B_i\},
$$

$$
\mathcal{M}\{P_2 = 1\} = \sup_{P_2 = 1} \min_{1 \leq i \leq 4} \mathcal{M}\{\xi_i \in B_i\},
$$

$$
\mathcal{M}\{P_{1, 2} = 1\} = \sup_{P_{1, 2} = 1} \min_{1 \leq i \leq 4} \mathcal{M}\{\xi_i \in B_i\},
$$

where $B_i$ is a subset of $\{0, 1\}, i = 1, 2, 3, 4$. R7, R12 and R14 corresponding to Figures 3.5g, 3.5i and 3.5m are the realizations that holds $P_1$. In the following we calculate $\mathcal{M}\{P_1 = 1\}$:

$$
\mathcal{M}\{P_1 = 1\} = \sup_{P_1 = 1} \min_{1 \leq i \leq 4} \mathcal{M}\{\xi_i \in B_i\}
= \sup_{R7, R12, R14} \min_{1 \leq i \leq 4} \mathcal{M}\{\xi_i \in B_i\}
= \sup \left( \min\{0.8, 0.7, 0.6, 0.5\}, \min\{0.8, 0.7, 0.4, 0.5\}, \min\{0.8, 0.7, 0.6, 0.5\} \right)
= \max \{0.5, 0.4, 0.5\}
= 0.5
$$

Following the similar procedure, $\mathcal{M}\{P_2 = 1\} = 0.3$ and $\mathcal{M}\{P_{1, 2} = 1\} = 0.4$. Although probability and uncertain theories give different probability and uncertain measures values for events $P_1$, $P_2$ and $P_{1, 2}$, they have similar ordering, i.e. $pr(P_1) > pr(P_{1, 2}) > pr(P_2)$ and $\mathcal{M}\{P_1 = 1\} > \mathcal{M}\{P_{1, 2} = 1\} > \mathcal{M}\{P_2 = 1\}$.

To the best of our knowledge, there is no extensive work to thoroughly compare probability and uncertainty theories on fundamental problems in networks and it is, potentially, an interesting issue to be investigated.
4. Probabilistic Networks: Models and Algorithms

In this chapter first we review the models that have been used to represent imperfect network data under probability theory. Then, we review the literature that have devised methods and algorithms to analyze probabilistic networks.

4.1 Probabilistic Models

Deterministic networks model different aspects of data, from only existence of interactions between nodes to including labels to nodes and edges such as time, type, preference and etc. Similarly, imperfect data can be represented by different models which encompass imperfect information on existence of nodes and edges or imperfect information on labels of nodes and edges. Here, we review the main models under probability theory.

4.1.1 Imperfect Information About Existence of Edges

One of the most common and at the same time simple representation of probabilistic networks is $\mathcal{G}(V, E, P)$ where $V$ is the set of nodes, $E$ is the set of edges and $P : E \rightarrow (0, 1]$ is a function which assigns a probability to each edge $e \in E$. In this model, the existence of nodes is deterministic and the existence of edges is probabilistic [92, 94].

Another model which is widely used by researchers is $\mathcal{G}(V, E, P, W)$ where $V$ is the set of nodes, $E$ is the set of edges, $P : E \rightarrow (0, 1]$ is a function which assigns a probability to each edge $e \in E$ and $W : E \rightarrow (0, \infty)$ is a function which assigns a weight to each edge [50, 95, 119, 130].

In these two models the existence of any pairs of edges is independent of each other:

$$\forall e, l \in E, \ pr(e|l) = pr(e)$$

Each probabilistic network $\mathcal{G}$ with $m$ edges corresponds to $2^m$ deterministic networks which are called possible worlds. The probability of each possible world is:

$$Pr(G_i) = \prod_{e \in E_{G_i}} pr(e) \prod_{e \in E \setminus E_{G_i}} (1 - pr(e))$$

(4.1)
In possible world semantics, the expected value of each measure in a probabilistic network equals the average of that measure over all possible worlds:

\[
E(M) = \sum_{G_i \in \mathcal{G}} M_i \times Pr(G_i)
\]  

(4.2)

where, \(G_i\) is \(i\)th possible instance of \(\mathcal{G}\) and \(M_i\) is the value of measure \(M\) in possible world \(G_i\) \([11, 12, 96, 127]\).

Many real world scenarios can be represented with these two models. The authors of [112] have reviewed the research articles that have used the first model to represent noisy experimental results of protein-protein interaction networks. The authors of [35,56,98] have modeled reliability of links between sensors in wireless sensor networks with a probability number and have used this model to find the most reliable sinks of information in such networks. These models have been applied to represent the probability of interaction between individuals \([92, 95]\). Moreover, the authors in \([72, 73]\) have modeled opportunistic networks in which edge probabilities denote the data-delivery success rates between two mobile nodes.

4.1.2 Other models

While majority of research articles have used the aforementioned models to represent network data, some other works have deployed other variations of probabilistic network models, although they are in minority:
• Imperfect information about existence of nodes and edges:
The authors in [115, 123] have studied the subgraph search problem in networks in which the existence of both nodes and edges is uncertain, $\mathcal{G}(V, E, P_V, P_E)$ where $P_V : V \rightarrow (0, 1]$ and $P_E : E \rightarrow (0, 1]$ assign existence probability to nodes and edges respectively and the existence of any pairs of nodes and any pairs of edges are independent. Moreover, Zou et al. [134] have investigated the problem of maximal clique finding in such probabilistic network models.

• Dependent edge existence:
The authors in [16,17,121,122] argue that the assumption of independent edge existence probability is invalid in some scenarios such as communication and road networks. For example highly loaded roads or communication links often block traffics in nearby roads and links. Therefore, they have applied a model in which the existence probability of edges connecting to the same node are correlated.

• Imperfect information about the weight of edges:
The authors in [48, 118] use the network model $\mathcal{G}(V, E, W, \mathcal{P})$ in which the existence of nodes and edges is certain however weight on each edge follows a specific probability distribution. $\mathcal{P}$ is a set of probability distributions. The weight in these models can represent load of traffic, traverse time or the demand of each edge. The authors in [48, 118] have studied different path queries on such networks.

4.2 Probabilistic Networks Analysis
This section reviews the literature which have proposed methods to analyze probabilistic networks.

4.2.1 Path Based Queries
A path is a sequence of distinct edges that join two nodes in a network. Path is a fundamental concept in graph theory and many other measures and notions are defined based on it [33, 116, 126]. The two most essential path based queries in probabilistic networks are shortest path and reliability. This section reviews the research articles that have considered different variations of these two problems.

Shortest Path Length Distribution
Shortest path length between any pairs of nodes in probabilistic networks are expressed as a shortest path length distributions [95]. The shortest path length distribution between two nodes $s$ and $d$ notated as $sp_{s,d}(l)$ is the sum of the probability of all possible worlds in which the shortest path length between nodes $s$ and $d$ is $l$:
where $G$ is the set of all possible worlds of probabilistic graph $\mathcal{G}$. Equation 4.3 gives the exact probability of shortest paths with different length, however it is computationally prohibitive for large probabilistic networks. Therefore, approximating shortest path length distribution using sampling is a way to reduce the computation. Potamias et al. have defined three new distance functions derived from standard statistical properties of shortest path length distribution: median distance, majority distance and expected reliable distance [95]. They have used these three distance functions to explore the $K$-nearest neighbors of nodes in a probabilistic network. Zou et al. have studied top-K possible shortest paths problem which is a combination of two independent phases: first they find top-k candidate shortest paths using Yen’s algorithm and then they have approximated the probability of each candidate path [130]. Yuan et al. propose a method that returns the shortest paths between two nodes whose probability is higher than a specific threshold [119]. The authors in [50] have proposed two sampling methods to efficiently calculate the probability that the distance between two nodes is less than or equal to a given threshold $d$.

Reliability

The second fundamental path based problem in probabilistic networks is reliability problem which measures the probability that two given nodes are reachable. This problem has been extended to more general problems such as the probability that all pairs of nodes in a network are connected [103], the probability that all pairs of nodes in a subset of nodes are connected [45] and the probability that there is at least one path between a node and a set of nodes [52].

Zhu et al. in [128] have studied the problem of identifying the top-k target nodes that have the highest reliability from a given source node. Their method combines two techniques: a) offline sampling technique which samples required number of possible worlds at the same time and stores them in a bit-vector and b) BFS sharing technique that performs a single BFS on all sampled possible worlds simultaneously. Khan et al. in [52] have investigated the problem of finding all nodes in a probabilistic graph that are reachable from a given source node $s$ with probability higher than or equal to a given threshold $\eta$. They proposed a method to cluster hierarchically nodes, called $RQ$-tree, and online sampling method which means that an edge is sampled in the current possible world if it is requested. The authors in [53] have considered a specific probabilistic network model in which edge probabilities depend on some external conditions. Then they have studied the problem of finding top-k conditions that maximize the reliability between two given nodes. Lin et al. in [63] have studied the problem of data cleaning for reliability queries.
under limited budget. In their problem they remove uncertainty about the existence of $k$ edges using crowd-sourcing. $k$ edges are selected such that cleaning them maximizes the precision of approximation of reliability in probabilistic networks. The authors in [51] have examined approximation reliability algorithms that are based on various sampling methods. They have evaluated the performance of the algorithms by comparing their accuracy, running time and memory consumption.

A specific reliability query is the problem of finding the most reliable path that has been studied in [101]. This path can be calculated easily by transforming a probabilistic graph into a weighted graph by replacing each edge probability $p_e$ with $w_e = -\log(p_e)$ and running Dijkstra shortest path algorithm on the transformed network.

4.2.2 Pattern Extraction Queries

Core Decomposition

The $k$-core of a network is the maximal subgraph in which each node is connected to at least $k$ nodes all in the same subgraph. The operation of specifying cores with different $k$s is called $k$-core decomposition. Batagelj and Zaveršnik proposed a core decomposition method for deterministic graphs based on the degree of nodes [5]. In this method, in each iteration one node with the smallest degree is removed and the number of the core that the node belongs to will be set.

Bonchi et al. extended this algorithm to be used in probabilistic networks by introducing notions $\eta$-degree and $(k, \eta)$-core [10]. $\eta$-degree is the maximum degree such that the probability to have that degree is more than or equal to $\eta$. $(k, \eta)$-core is the set of nodes in which each node has at least the degree of $k$ with the probability equal or greater than $\eta$. Since computation of $\eta$-degree in each iteration is expensive, Bonchi et al. have utilized a sampling method to approximate it. The utilized sampling method is called sample finite population with unequal probabilities which has been introduced in [15].

Clique Mining

A clique in a network is a set of nodes where each pair of them are adjacent. A maximal clique is a clique that is not a member of any other cliques. Finding densely connected nodes in a network is an important graph mining task and researchers look at cliques as cores of densely connected components in networks. Therefore, finding cliques and maximal cliques in probabilistic networks is also an essential problem that has attracted many researchers. Zou et al. [134] have investigated the problem of finding top-k maximal cliques in probabilistic networks. They have defined the problem as follows: Given a probabilistic graph $G$ and two positive integers $k$ and $s$, top-k maximal cliques are a collection of $\mathcal{F}$ of $k$ node sets such that:
1. the size of each node set in \( F \) is at least \( s \) (i.e. \( \forall C \in F, |C| \geq s \)), and
2. there is no node sets out of \( F \) whose maximal clique probability is higher than node sets in \( F \) (i.e. \( \forall C \in F, C' \notin F : P_{\text{max-clq}}(C) \geq P_{\text{max-clq}}(C') \)).

In their approach first they find maximal cliques in the underlying deterministic graph of probabilistic graph \( G \) (by transforming all probabilities to 1) and then compute the probability of each clique and include top-k maximal cliques in \( F \) that have the aforementioned properties.

Mukherjee et al. have studied another variation of maximal clique problem in probabilistic graphs, called maximal clique enumeration [82, 83]. Given a probabilistic graph \( G = (V, E, P) \) and a probability threshold \( 0 < \alpha \leq 1 \), a set of nodes \( C \subseteq V \) is an \( \alpha \)-maximal clique if its probability is higher than or equal to \( \alpha \) and there is no node \( v \in (V \setminus C) \) such that \( C \cup \{v\} \) is a clique whose probability higher than or equal to \( \alpha \) (\( \alpha \)-clique).

While [134] has focused on finding top-k maximal cliques whose size is higher than a threshold and [82, 83] have investigated the problem of finding all maximal cliques whose probability is higher than a threshold, Li et al. [61] have studied the problem of finding all maximal cliques whose size is larger than \( k \) and their probability is higher than or equal to \( \tau \). To reduce computational complexity in their proposed solution they have introduced two core-based pruning algorithms to reduce the size of probabilistic graphs by removing nodes that are not included in any clique.

**Threshold-Based Pattern Matching**

This subsection revises the pattern mining queries that do not have any specific definition of pattern as an input (as we had for core and clique decomposition), but they extract the patterns whose probability of existence is higher than a specific given threshold.

According to our knowledge, Zou et al. [133] studied the problem of frequent subgraph matching in probabilistic networks for the first time. Given a probability threshold \( s \), they have defined the output of the frequent subgraph matching problem as a set of all subgraphs whose probability of occurrence is higher than or equal to \( s \). The same authors have extended their work to the probabilistic networks where not only their edges are associated with a probability but also the existence of nodes is uncertain [132]. As reliability is a fundamental issue in probabilistic graphs, Jin et al. [49] recalled the concept of subgraph reliability (\( R \)) which is the probability that a probabilistic subgraph is a connected component. Based on subgraph reliability they have investigated the problem of finding all subgraphs whose reliability is higher than a given threshold \( \alpha \). They have shown that for two subgraphs \( V_1 \) and \( V_2 \) in probabilistic graph \( G \) where \( V_1 \subset V_2 \subset G \), we can not claim neither \( R(V_1) < R(V_2) \) nor \( R(V_1) > R(V_2) \). Zou has investigated the problem of determining the densest subgraph that contains a given set of nodes [131]. He
has shown that this problem is solvable in polynomial time by extending Goldberg’s Algorithm [28].

**Similarity Matching**

This subsection reviews the group of articles that generally search for a specific given pattern in uncertain graphs. More technically, the algorithms proposed for similarity search receive a deterministic graph as input query, \( q \), and return back the probability of occurrence of \( q \) in uncertain graph \( G \) or returns true if \( q \)'s probability is higher than specific threshold.

Yuan et al. have studied the problem of finding a subgraph in a set of probabilistic networks notated as \( D = \{ G_1, G_2, ..., G_n \} \) [121, 123] whose both nodes and edges are associated with a probability. More specifically, given a set of probabilistic graphs \( D \), a query graph \( q \) and probability threshold \( 0 < \eta \leq 1 \), their proposed algorithm returns those uncertain graphs in \( D \) where the probability of existence of \( q \) is higher than or equal to \( \eta \). Tong et al. [115] have considered the similar problem which is applicable for cases where the size of probabilistic networks (number of nodes) is small and at the same time size of \( q \) is large. The problem is called supergraph matching. Given a set of probabilistic graphs \( D \), a query graph \( q \) and probability threshold \( 0 < \eta \leq 1 \), this problem is to find uncertain graphs in \( D \) such that the probability of occurrence of their possible worlds in \( q \) is higher than or equal to \( \eta \).

The authors in [120] generalized the definition of match in deterministic graph pattern matching. Their proposed algorithm have three input: a deterministic graph \( G \), a query graph \( q (\{v_1, v_2, ... v_n\}) \) which in both of them nodes are labeled, and the shortest path threshold \( \gamma \). Then a set of \( n \) nodes in \( G \), \( \{u_1, u_2, ... u_n\} \) are called match of \( q \) if and only if (1) node labels in both sets are equal, \( L(u_i) = L(v_i) \) for \( i \in \{1, 2, ..., n\} \) and (2) if there is an edge between nodes \( v_i \) and \( v_j \) in \( q \), the shortest path distance between \( u_i \) and \( u_j \) in \( G \) is smaller or equal to \( \gamma \). Then they define pattern matching problem in probabilistic networks: given a probabilistic graph \( \mathcal{G} \), deterministic query graph \( q \), distance threshold \( \gamma \) and probability threshold \( \eta \), this problem retrieves all node sets in \( \mathcal{G} \) whose probability to match with \( q \) (with regard to \( \gamma \)) is higher than or equal to \( \eta \).

### 4.2.3 Clustering

Network clustering is to divide network’s nodes into a number of groups, called clusters, such that the nodes in the same cluster have the most similarity together. Similarity measure is based on some node-distance measures. Clustering is a fundamental problem in network mining because of its applications in detecting protein complexes in PPI networks, sink placement in wireless sensor networks, or finding groups of people in social networks who get most influence from each other. Clusters in general can overlap, but clustering in
the literature on probabilistic networks has been used as partitioning, i.e. there is no common node between any pairs of clusters. This section reviews the articles which have investigated clustering problem in probabilistic network.

Liu et al. in [70] have defined two desired qualities that has to be kept in the process of clustering: first, for each connected component of each possible world of $\mathcal{G}$ the number of distinct clusters to which the different nodes belong should to be as small as possible. Second, the size of clusters should be balanced. To measure these two qualities the authors have defined two functions: the entropy of cluster labels notated as $\mathcal{F}_p$ and the entropy of cluster size notated as $\mathcal{F}_e$. As the objective function they minimize $\mathcal{F} = \mathcal{F}_p - \mathcal{F}_e$ and to reduce the computational time complexity they use Monte Carlo sampling method.

Kollios et al. have proposed a clustering method by extending the concept of edit distance [58]. Edit distance between two deterministic graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ is the number of edges that need to be added or deleted from $G_1$ in order to be transformed to $G_2$. Edit distance between a probabilistic graph $\mathcal{G} = (V, E, P)$ and a deterministic graph $G = (V, E)$ is defined as the expected edit distance between every possible world of $\mathcal{G}$ and $G$ which can be computed in polynomial time. The second central notion in this article is cluster graph which is a deterministic $G_c = (V, E_C)$ graph with the following characteristics:

1. $G_c$ partitions the nodes in $V$ into $k$ partitions, $V = \{V_1, V_2, ..., V_k\}$ such that $V = \bigcup_{i=1}^{k} V_k$ and $V_i \cap V_j = \emptyset$,
2. all nodes in $i$th partition are connected together in the form of a clique with size $|V_i|$ and,
3. for each pair of nodes that are in distinct partitions, there is no edge between them.

Based on expected edit distance and cluster graph, Kollios et al. have formulated the clustering problem in probabilistic graphs $\mathcal{G} = (V, E, P)$ as finding a cluster graph whose expected distance to $\mathcal{G}$ is minimized. In this problem, the number of clusters, $k$, is not given in advance and it is the output of the clustering algorithms. They have shown the similarity between this problem with other existing problems such as CorrelationClustering and have employed the proposed algorithms for them to solve their stated problem.

Ceccarello et al. in [13] have considered $k$-clustering in probabilistic networks. Given $k$, a k-clustering of a probabilistic network $\mathcal{G} = (V, E, P)$ is a partitioning of $V$ into $k$ clusters $\{C_1, C_2, ..., C_k\}$ and a set of cluster centers $\{c_1, c_2, ..., c_k\}$ where $c_i \in C_i$. Ceccarello et al. have defined two clustering objective functions as:

$$\mathcal{F}_1(C) = \min_{1 \leq i \leq k} \min_{v \in C_i} Pr(v \sim c_i) \quad (4.4)$$

$$\mathcal{F}_2(C) = \frac{1}{|V|} \sum_{1 \leq i \leq k} \sum_{v \in C_i} Pr(v \sim c_i) \quad (4.5)$$
where \( v \sim c_i \) is the event that two nodes \( v \) and \( c_i \) are connected. The main goal of k-clustering algorithms is to maximize \( \mathcal{F}_1 \) and \( \mathcal{F}_2 \) in Equations 4.4 and 4.5 respectively. The first method is called \( k\text{-}center \) and the second one is called \( k\text{-}median \) clustering. Han et al. [44] have increased computational efficiency and approximation precision of \( k\text{-}center \) and \( k\text{-}median \) clustering methods studied in [13].

Qiu et al. in [97] have studied the problem of structural clustering by defining notions structural similarity, reliable structural similarity and reliable neighborhood. The structural similarity between two nodes \( u \) and \( v \) is the Jaccard coefficient between the set of \( u \)’s neighbors and itself and the set of \( v \)’s neighbors including \( u \) itself. In probabilistic network \( G \) for a given edge \( e \), a similarity threshold \( 0 < \varepsilon \leq 1 \) and probability threshold \( 0 < \eta \leq 1 \), \( u \) is reliably structural similar to \( v \) if the probability that the structural similarity between \( u \) and \( v \) is higher than or equal to \( \varepsilon \) is higher than or equal to \( \eta \) (\( u \) and \( v \) are incident nodes to edge \( e \)). Given a similarity threshold \( 0 < \varepsilon \leq 1 \) and probability threshold \( 0 < \eta \leq 1 \), the \( (\varepsilon, \eta) \)-reliable neighborhood of node \( v \) is defined as the set including \( v \) and the subset of \( v \)’s neighbors that are reliably structural similar to \( v \). Therefore, the authors define clustering problem in probabilistic graphs as: Given parameters \( \varepsilon \) and \( \eta \), the problem of probabilistic graph clustering is to compute a set of reliable clusters \( C \) that for each \( c \in C \):

- \( |c| \geq 2 \)
- for each node \( v \) in \( c \in C \) whose reliable neighborhood size is larger than a specific given number \( \mu \), all nodes that are in its reliable neighborhood are in cluster \( c \).
- for any two nodes \( v_1, v_2 \in c \), there exists a node \( u \) such that both of them are in its reliable neighborhood set.

Therefore, the authors have adapted an existing deterministic clustering framework called PSCAN for clustering probabilistic graphs based on defined structural similarity measures.
5. Summary of Papers

5.1 Paper I

Comparing Node Degrees in Probabilistic Networks
Amin Kaveh, Matteo Magnani, Christian Rohner
Journal of Complex Networks

Summary. A node’s degree in a probabilistic network is a probability distribution rather than a single number. Computing the degree distribution of nodes in probabilistic networks is an expensive task for large networks. To overcome this difficulty, expected degree which is easy to compute is commonly utilized as the summary information of a degree distribution. However, more summary information, such as variance and skewness, have to be included to evaluate the probability of two nodes having the same degree or one having higher degree than another. In order to compare degree distribution of nodes using summary information (expected degree, variance and skewness), without computing degree distribution, we have defined the concept of reference nodes. Reference node for expected degree $E$ is a node with the same expected degree but the least possible variance. Then, we show how the probability of a node’s degree being higher or equal to the degree of its reference node can be approximated by using variance and skewness of the degree distribution in addition to expected degree.

Contribution. In summary we have three contributions in this paper:
- We have defined comparison measures to calculate the probabilities of two nodes having the same degree or one node having higher degree than another,
- We have defined reference nodes as criteria to quantify the equality and inequality of degree of nodes which have the same expected degree.
- We have proposed approximation functions for the defined comparison measures based on variance and skewness.

My Contribution. I conceived the study, performed the experiments, analyzed the data and wrote the manuscript. M.M. and C.R. analyzed the results and helped writing the paper.
5.2 Paper II

Defining and Measuring Probabilistic Ego Networks
Amin Kaveh, Matteo Magnani, Christian Rohner
Applied Network Science Journal
SpringerOpen, Revised Version is submitted on June 16, 2019.

Summary. Probabilistic ego networks have been considered. This paper shows there is no unique way of defining probabilistic ego networks. Therefore, the paper introduces two definitions of probabilistic ego networks, called V-Alters-Ego and F-Alters-Ego and investigates three fundamental measures (degree, betweenness and closeness) for each definition. Betweenness and closeness in F-Alters-Ego are defined for the first time in this paper. The paper compares the proposed definitions and measures. The findings highlight that F-Alters-Ego measures are faster to compute, but are not compatible with a possible-world semantics, showing different reasons to choose one or the other definition.

Contribution. In summary we have three main contributions in this paper:
- we have specified two definitions of probabilistic ego networks: V-Alters-Ego and F-Alters-Ego,
- We have investigated betweenness and closeness under both aforementioned definitions. Betweenness and closeness under F-Alters-Ego are defined for the first time in this paper. As a result of our examination, we show the extent to which different definitions of probabilistic ego networks lead to different sets of top-ranked nodes in networks,
- we have examined the association between the sets of top-ranked nodes obtained from the proposed probabilistic egocentric measures and those in the corresponding deterministic networks.

My Contribution. I initiated the study and implemented and performed the experiments. All authors contributed equally in conceiving and developing the study and analyzing the results.
6. Discussion and Future Work

The most commonly applied theory to model and analyze imperfect network data is probability theory. The main reason behind this broad application is its formalized mathematical axioms that have been consolidated over the last few centuries. Among all probabilistic network models, $G(V,E,P)$ is compliant with a wide variety of real world scenarios such as wireless sensor networks, protein-protein interaction networks, opportunistic network, road networks and so on. In this model, $V$ is the set of nodes, $E$ is the set of edges and $P : E \rightarrow (0,1]$ is a function that assigns existence probability to each edge and existence of each pair of edges is independent of each other.

In this model, a network with $m$ edges corresponds to $2^m$ deterministic networks. Each deterministic network is called a possible world with a probability derived from multiplication of probability of present/non-present edges at that possible world. In possible world semantics, the expected value of any measure in $G(V,E,P)$ equals the average of that measure over all possible worlds. Since the number of possible worlds is exponential in the number of edges, exact processing of measures on this model is computationally prohibitive even for moderate size networks.

Therefore, approximation via sampling is a utilized replacement for exact computation of measures in such networks. In sampling, we sample a subset of possible worlds and calculate the expected value of that measure on all samples. Although this approach takes less computational time, it still incurs high computational cost especially for networks with high entropy (see Section Chapter 1).

Decreasing entropy of a probabilistic network causes needs for the number of sampled possible worlds to be decreased. As the first effort to reduce the entropy of probabilistic networks, Pargas et al. in [88] have proposed sparsification. Specifically, an sparsified probabilistic network, $G'(V,E',P')$, is an alternative network to $G(V,E,P)$ where $|E'| = \alpha |E|$, $0 < \alpha < 1$.

$G'$ is still a probabilistic network, but with different set of edges and probability assignment and with less entropy. The algorithms proposed in this approach seek a $G'$ in which (a) the expected degree of each node in $G$ to be as close as possible to the expected degree of that node in $G'$ and (b) the sum of the expected degree of all nodes in $G$ to be as close as possible to the sum of expected degree of all nodes in $G'$.

In fact in [88] just the expected degree as a local property is preserved. As we discussed in our first paper (see Section 5.1) expected degree is not
very informative and infinite number of nodes with different degree distribution can have the same expected degree. In the first paper we proposed several comparing measures to compare a node in a probabilistic network with its corresponding reference node. One approach to extend the existing sparsification method is preserving other characteristics of degree distribution. For example, we can preserve the association/relation between a node with its corresponding reference node. Specifically, a sparsified node is a node with the same expected degree, less variance (entropy) and the same $P_> - P_<$.

However, node’s degree distribution and its summary measures such as expected degree are not the only local properties. In our second contribution (see Section 5.2) we have studied other local properties in the context of probabilistic ego networks. In this/that work, we have defined node’s ego betweenness and ego closeness which highlight the relationship of a node among its neighbors. As a result, another approach to extend the existing sparsification method is preserving other local properties of nodes examined in our work (see Section 5.2) during the process of sparsification and it might lead to better and more precise sparsified networks.

**Future work.** As the future work we will try to repeat the process of sparsification by first keeping the relation of nodes with their reference node and second by keeping the examined local measures discussed in 5.2.
I would like to express sincere gratitude to my advisor, Matteo and my co-supervisor, Christian for their patience, encouragement and continued support, which has been invaluable.

I thank my colleagues, Davide Vega D’aurelio and Khalid Mahmood for sharing their experience in research, teaching and team-working activities. Thank you for your help and friendship.

In addition, I would like to thank my parents, my uncle Hassan, and my parents in law, Nadereh and Rahbar, for their endless love and support.

Many thanks to Emilia, Daniel, Elnaz and Alaleh for their support and friendship.

Last but not the least, I genuinely appreciate my wife, Alaa, for being in my life; Your existence have brought peace to my life.
References


[14] Chen Chen, Xifeng Yan, Philip S Yu, Jiawei Han, Dong-Qing Zhang, and Xiaohui Gu. Towards graph containment search and indexing. In *Proceedings of the 33rd international conference on Very large data bases*, pages 926–937. VLDB Endowment, 2007.


[99] Jennifer Rowley. What is information? Information Services & Use,


59


Paper I
Comparing node degrees in probabilistic networks

AMIN KAVEH*, MATTEO MAGNANI AND CHRISTIAN ROHNER

InfoLab, Department of Information Technology, Uppsala University, Lägerhyddsvägen 7, House 19, SE-752 37, 75105 Uppsala, Sweden

*Corresponding author. Email: amin.kaveh@it.uu.se

[Received on 21 May 2018; editorial decision on 15 January 2019; accepted on 18 January 2019]

Degree is a fundamental property of nodes in networks. However, computing the degree distribution of
nodes in probabilistic networks is an expensive task for large networks. To overcome this difficulty, expected
degree is commonly utilized in the literature. However, in this article, we show that in some cases expected
degree does not allow us to evaluate the probability of two nodes having the same degree or one node
having higher degree than another. This suggests that expected degree in probabilistic networks does not
completely play the same role as degree in deterministic networks. For each node, we define a reference node
with the same expected degree but the least possible variance, corresponding to the least uncertain degree
distribution. Then, we show how the probability of a node’s degree being higher or equal to the degree of its
reference node can be approximated by using variance and skewness of the degree distribution in addition
to expected degree. Experimental results on a real dataset show that our approximation functions produce
accurate probability estimations in linear computational complexity, while computing exact probabilities
is polynomial with order of 3.

Keywords: probabilistic networks; degree; expected degree; edge probability; variance; skewness.

1. Introduction

Degree, which is defined as the number of incident edges to a node, is a fundamental property in networks.
Despite its simplicity in calculation, degree is widely applied to characterize both individual nodes (degree
centrality) and whole networks (degree distribution). In addition, many centrality measures are based on
degree, such as pagerank [1], leverage centrality [2] and degree-mass centrality [3] which represent a
node’s relative importance in the network based on the node’s degree and the degree of its neighbours.
Many measures used in community detection, such as modularity [4], are also defined based on degree.

Nevertheless, to compute a node degree we need to count edges, while in many cases it is difficult
to determine if an edge between two nodes exists or not. For example, Mering et al. [5] have found that
more than 50% of the edges in the yeast protein–protein interaction network [6] have been mistakenly
considered as an edge. In metabolic networks, one metabolite may be converted to another one or not
[7] and this means that edges between metabolites are not certain. In social networks, interactions are
subject to sampling noise, inference [8] and inaccuracy due to informant’s cognition about their own
behaviour [9].
As a consequence, scholars have used probabilities to represent uncertainty about edge existence and the resulting networks are known as probabilistic networks. The main difference between deterministic and probabilistic network models is that many structural properties that are expressed as a number in deterministic networks, such as the degree of a node or the length of the shortest path between two nodes, correspond to a probability distribution in probabilistic networks. For example, each node in a probabilistic network has a corresponding degree probability distribution.

In this context, expected degree, that is the sum of the probabilities on a node’s incident edges, is vastly used as a summary of the node’s degree distribution, so that a single value can be associated to the node as it happens in deterministic networks. Poisot et al. [10] have used expected values not only to calculate node degree but also to represent various structural properties of probabilistic networks. Expected degree has also been used to extract a certain representative instance from probabilistic graphs [11, 12] and to summarize probabilistic graphs [13].

However, in this article, we show that expected degree does not allow us to perform basic tasks such as evaluating the chances for two nodes to have the same degree or for one node to have a higher degree than the other. For example, nodes A and D in Fig. 1a have the same expected degree (0.9), but node A is more likely to have a higher degree than node D, that is \( P(\text{deg}_A > \text{deg}_D) > P(\text{deg}_D > \text{deg}_A) \), albeit D’s degree can be up to four times higher than A. Moreover, the actual probability that the two nodes have the same degree, \( P(\text{deg}_A = \text{deg}_D) \), is only 0.43. Hence, expected degree fails in capturing these basic properties.

The problem with expected degree is that despite it being very fast to compute, it loses information about the original degree distribution. To address this problem, we can perform the aforementioned comparison tasks if we compute the full degree distribution, but doing so is an expensive task. The time complexity to compute the degree distribution of node \( i \) is \( O(L_i^3) \) [14], where \( L_i \) is the number of edges incident to node \( i \).

In this article, we show that it is possible to estimate \( P(\text{deg}_{A} = \text{deg}_{B}) \) and \( P(\text{deg}_{A} > \text{deg}_{B})/P(\text{deg}_{A} < \text{deg}_{B}) \) for two nodes with the same expected degree without any increase in computational complexity in comparison with the calculation of expected degree. In particular, the second and the third moments of the degree distributions of the two nodes (that are, variance and skewness, noted respectively \( \sigma^2 \) and \( \gamma \) ) are predictors, respectively, for \( P(\text{deg}_A = \text{deg}_B) \) and \( P(\text{deg}_A > \text{deg}_B)/P(\text{deg}_A < \text{deg}_B) \) and they can be computed linearly in time with respect to the number of edges connected to the node, that is \( L_i \).

2. Results

In this section, we define three measures \( P_\approx(A, B) \), \( P_\geq(A, B) \) and \( P_\leq(A, B) \) to compare the degrees of two arbitrary nodes A and B. Then, for each unique expected degree \( E \), we define a reference node \( R_E \) to compare all nodes with expected degree \( E \) with it. Exact comparisons are also computed from the node degree distributions. Then, we use the second (\( \sigma^2 \)) and the third (\( \gamma \)) moments of A’s degree distribution, together with its first moment (\( E \)), to estimate \( P_\approx(A, R_E) \), \( P_\geq(A, R_E) \) and \( P_\leq(A, R_E) \). We conclude by testing our approximations on a real dataset, both with regard to accuracy and computational complexity.

In summary, we:

- Define three comparison measures to calculate the probabilities of two nodes having the same degree or one node having higher degree than another,
- Define reference nodes as criteria to quantify the equality and inequality of degree of nodes which have the same expected degree.
Find the correlation between our comparison measures and variance and skewness of node degree distributions and then propose approximation functions for our comparison measures based on variance and skewness.

Some coefficients in the proposed approximation functions depend on the minimum and the maximum values of the probability that a node and its reference node have the same degree. So, we analytically calculate these values. Then, we apply the proposed approximation functions on a real dataset to test both their efficiency and accuracy.

2.1 Preliminary 1—comparison measures

We define three measures: \( P_e(A, B) \) as the probability that nodes \( A \) and \( B \) have equal degree, \( P_>(A, B) \) as the probability that node \( A \) has higher degree than \( B \) and \( P_<(A, B) \) as the probability that \( A \) has lower degree than \( B \). In deterministic networks, in which all edges are associated with probability 1, if nodes \( A \) and \( B \) have the same degree, then \( P_e(A, B) \) is 1 and \( P_>(A, B) \) and \( P_<(A, B) \) are 0. Moreover, if they do not have the same degree, either one of \( P_>(A, B) \) and \( P_<(A, B) \) is 1 and the other two comparison measures are 0.

In probabilistic networks, each node \( i \) has a degree distribution which we notate \( \pi_i \). The degree distribution \( \pi_i \) is determined by the edge probability set which we notate as \( \mathcal{P}_i \). For example, for node 2 in Fig. 1c, \( \mathcal{P}_2 = \{1, 1, 0.5, 0.5\} \) and its degree distribution \( \pi_2 \) is plotted in Fig. 1e. Note that in probabilistic networks the comparison measures for two arbitrary nodes with the same \( \mathcal{P} \) do not necessarily equal 0 and 1. For example, by comparing two nodes \( A \) and \( B \) with \( \mathcal{P}_A = \mathcal{P}_B = \{1, 1, 0.5, 0.5\} \), \( P_e(A, B) = 0.38 \), \( P_>(A, B) = 0.31 \) and \( P_<(A, B) = 0.31 \). The small value of probability of having the same degree, that is \( P_e(A, B) = 0.38 \), for two equal nodes in probabilistic networks, where equal means that \( \mathcal{P}_A = \mathcal{P}_B \), can be perceived as counter-intuitive.
For two arbitrary nodes \( A \) and \( B \) with \( L_A \) and \( L_B \) incident edges and degree distributions \( \pi_A \) and \( \pi_B \):

\[
P_\leq (A, B) = \sum_{k=0}^{\min(L_A,L_B)} \pi_A(k)\pi_B(k) \tag{2.1}
\]

\[
P_\geq (A, B) = \sum_{k=1}^{L_A} \sum_{h=0}^{k-1} \pi_A(k)\pi_B(h) \tag{2.2}
\]

\[
P_\prec (A, B) = 1 - P_\leq (A, B) - P_\geq (A, B) \tag{2.3}
\]

### 2.2 Preliminary 2—reference nodes

As mentioned above, expected degree is the most commonly used counterpart of node degree in probabilistic networks. Let us first concentrate on the case in which expected degree takes an integer value. Then, all nodes with expected degree \( E \) are indistinguishable from a node in a deterministic network with \( E \) edges when we look at the expected degree. For example, the expected degree of nodes 2 and 3 in Fig. 1 is equal to the degree of node 1. Therefore, comparing a node in a probabilistic network with its reference node allows us to characterize the uncertainty in the node’s degree distribution.

However, reference nodes with non-integer expected degrees cannot be found in deterministic networks. Therefore, we extend the idea of finding reference nodes with no uncertainty in their degree distribution to the idea of finding hypothetical nodes with the least uncertainty in degree distribution. Since nodes with the lowest variance have the least uncertainty in their degree distributions [15, 16], we define the reference node for all nodes with expected degree \( E \) as a node with expected degree \( E \) and the lowest possible value of variance for its degree distribution and notate it as \( R_E \). For example, the reference node for all nodes with expected degree 3.4 is a node with edge probability set \( \mathcal{P} = \{1, 1, 1, 0.4\} \) (see Fig. 1f–i). The reference node is a hypothetical node and is not necessarily a member of the network.

Table 1 illustrates how different nodes with the same expected degree but different degree distributions have different values when compared with their corresponding reference node. For the sake of brevity, we use notations \( P_\leq(A) \), \( P_\geq(A) \) and \( P_\prec(A) \) instead of \( P_\leq(A, R_E) \), \( P_\geq(A, R_E) \) and \( P_\prec(A, R_E) \), respectively, where \( R_E \) is the corresponding reference node for nodes with expected degree \( E \). Take, as an example, nodes 1, 2 and 3 which have the same expected degree \( E = 3 \). Node 1’s degree is certain and the degrees of nodes 2 and 3 are presented in the form of degree distributions. The number of incident edges to nodes 2 and 3 is 4. However, their edge probability sets, \( \mathcal{P} \), are different. As the expected degree is 3, we compare all three nodes with the corresponding reference node which is a node with degree of 3 and the minimum possible value of variance, that is \( \sigma^2 = 0 \) (in this case node 1 is equal to the reference node).

The first three lines in Table 1 represent what extent the degrees of nodes 1, 2 and 3 are equal to the degree of the reference node, \( P_\leq(A) \), where \( A \) corresponds to nodes 1, 2 and 3. Node 1, as is expected, is equal to the deterministic node. However, comparison measures for nodes 2 and 3 are not the same. In more details, \( P_\leq(A) \) decreases as the variance of node degree distribution of node \( A \) increases. Moreover, \( P_\leq(A) \) is equal to \( P_\leq(A) \) when skewness is 0. As another example, take nodes 4, 5, 6 and 7 with expected degree \( E = 3.4 \). According to our definition of reference node, node 4 is equivalent to \( R_{3.4} \). The node degree distributions are plotted in Fig. 1j. We compare all nodes with the corresponding reference node. Lines 4–7 in Table 1 show the probability of degree of nodes 4, 5, 6 and 7 being equal or higher/lower than the degree of the reference node. Although node 4 is equivalent to \( R_{3.4} \), the probability of equality is not 1. However, the probability of having a higher degree is equal to the probability of having a lower degree, that is \( P_\geq(A) = P_\prec(A) \). The probability of equality declines as variance increases. Except for
node 4 that is exactly equal to the reference node, when $\gamma$ is negative, $P_>$ tends to be higher than $P_<$ (nodes 5 and 7) and when $\gamma$ is small (close to 0), $P_>$ tends to be equal to $P_<$ (node 6). These values suggest that by considering the value of $\gamma$ we can estimate (but not exactly calculate) the coefficient $P_>/P_<$. As a summary of Table 1, (a) there is a dependency between $P_=$ and $\sigma^2$ and (b) there is a dependency between $P_>/P_<$ and $\gamma$, and we intend to estimate these measures by using the second and the third moments.

### 2.3 Correlation of $P_=$ and $P_>/P_<$ with $\sigma^2$ and $\gamma$

In this section, we show how $P_=$, $P_>$ and $P_<$ change with respect to variance and skewness of node degree distributions. For various expected degrees, $E = \{0.1, 0.2, ..., 15.9, 16\}$, we have calculated $P_=(A)$, $P_>(A)$ and $P_<(A)$ by changing the values of variance and skewness. The different values of variance and skewness (for each $E$) have been obtained from changing the number of edges, $L_i$, and also changing the edge probability sets, $P$, on $L$ edges (for each value of $E$, we have generated 200 edge probability sets, so that a wide range of possible values of variance for each value of $E$ is covered). For example, the number of incident edges in node 1 in Fig. 1 is 3 while this value for nodes 2 and 3 is 4. At the same time, the edge probability sets for nodes 2 and 3 are $P_2 = \{1, 1, 0.5, 0.5\}$ and $P_3 = \{0.75, 0.75, 0.75, 0.75\}$ which leads to different values for $\sigma^2$ and $\gamma$. Black dots in Figs 2a–d and 3a–d show how $P_=(A)$ and the ratio of $P_>(A)/P_<(A)$ change when variance and skewness of $A$’s degree distribution increase, respectively.

Additionally, the maximum and the minimum values of $\sigma^2$ and also the maximum and the minimum values of $P_=$ depend on expected degree. For example, the maximum and minimum values of $\sigma^2$ for the nodes with $E = 25$ are $\sigma^2_{\min} = 0$ and $\sigma^2_{\max} = 25$, while these values for nodes with $E = 4.7$ are $\sigma^2_{\min} = 0.21$ and $\sigma^2_{\max} = 4.7$. Additionally, the highest and the lowest values for $P_=(A)$ when $E = 25$ are 1 and 0.08, while these values are 0.58 and 0.177, respectively, when $E = 4.7$ (see Fig. 2a, b). This means that $P_=$ and $P_>/P_<$ change not only with $\sigma^2$ and $\gamma$, respectively, but also with $E$.

The exact values of $P_=$, $P_>$ and $P_<$ are obtained from the node degree distributions, $\pi$. The computational complexity of calculation of $\pi$ for node $i$ with $L_i$ incident edges is $O(L_i^3)$ which is expensive for nodes with large number of incident edges. However, variance and skewness of the node degree distribution can be obtained directly from the edge probability set, $P$, with $O(L_i)$ and there is no need to compute the degree distribution if we can estimate $P_=$, $P_>$ and $P_<$ as a function of $\sigma^2$, $\gamma$ and $E$. In the next subsection, we propose two functions to approximate $P_=$ as a function of $(E, \sigma^2)$ and the ratio of $P_>/P_<$ as a function of $(E, \gamma)$. By knowing that $P_+ + P_ - P_ = 1$ and by approximating $P_=$ and the ratio $P_>/P_<$.
Fig. 2. Black dots: the exact values of $P_\omega$, that is the probability that a node with variance $\sigma^2$ and its corresponding reference have the same degree obtained from degree distributions for randomly selected values of $E$: (a) $E = 25$, (b) $E = 4.7$, (c) $E = 15.5$, (d) $E = 19.1$. Solid red lines: estimated values of $P_\omega$ based on Equation 2.4. Green dashed lines: the maximum and the minimum values of $\sigma^2$ calculated based on Equations 2.10 and 2.11. Blue dashed lines: the maximum and the minimum values of $P_\omega$ determined by Equations 2.12 and 2.13. For the case of $E = 25$, the maximum and the minimum values of $P_\omega$ is calculated using Equations 2.14 and 2.15.

Fig. 3. Black dots: exact values of $P_\geq$ for nodes with skewness $\gamma$ obtained from the nodes’ degree distributions for (a) $E = 7$, (b) $E = 10.8$, (c) $E = 15.5$ and (d) $E = 18.2$. Solid red lines: estimated values of $P_\geq$ based on Equation 2.8. Blue dashed lines pass through the point ($\gamma = 0, \frac{P_\geq}{P_\leq} = 1$) and show that the estimated values are crossing this point.

we will have the approximated values of $P_\geq$ and $P_\leq$. Some coefficients in the proposed approximation functions are based on the maximum and minimum possible values of $P_\omega$ and $\sigma^2$. In Section 2.5, we calculate them analytically as a function of $E$.

2.4 Approximation of $P_\omega$ and $P_\geq/P_\leq$

Based on the experimental results, we propose Equation 2.4 as the approximation function for $P_\omega$. Since the proposed function is an approximation, we notate it as $\tilde{P}_\omega$. Moreover, the proposed approximation function takes the variance and the expected degree of the node under study as input. So, we use the notation of $\tilde{P}_\omega(A(\sigma^2,E))$ to refer to the estimated value of $P_\omega$ for an arbitrary node $A$ when the variance of its degree distribution is $\sigma^2$ and its expected degree is $E$:

$$\tilde{P}_\omega(A(\sigma^2,E)) = \begin{cases} P_\omega^{max}(E) & \text{if } \sigma^2 < \tau_E \\ C_E \sigma^{2\tilde{\beta}_E} & \text{if } \sigma^2 \geq \tau_E \end{cases} \tag{2.4}$$

where,

$$C_E = \left( \frac{P_\omega^{min}(E)}{E^{\beta_E}} \right) \tag{2.5}$$

$$\tilde{\beta}_E = -0.42 - 0.0024 E \tag{2.6}$$
\[
\tau_E = E \left( \frac{P^\text{max}(E)}{P^\text{min}(E)} \right)^{\frac{1}{E}} 
\] (2.7)

Red lines in Fig. 2a–d show the approximation values of \(P_\gamma\) for 4 arbitrary values of \(E\). \(C_E\) and \(\tau_E\) are functions of \(P^\text{min}(E)\) and \(P^\text{max}(E)\) which both of them depend on \(E\) and are analytically calculated in the Section 2.5. The value of \(\beta_E\) is an approximated value that is discussed in detail in Appendix A.1.

We also notate the approximation of \(P_\gamma\) for an arbitrary node \(A\) having expected degree \(E\) and skewness \(\gamma\) as \(\tilde{P}_\gamma(A_{(y,E)})\). To approximate the ratio \(\frac{P_{\gamma}(A_{(y,E)})}{P_{<}(A_{(y,E)})}\), we propose Equation 2.8:

\[
\frac{\tilde{P}_{\gamma}(A_{(y,E)})}{P_{<}(A_{(y,E)})} \approx \frac{P_{\gamma}(A_{(y,E)})}{P_{<}(A_{(y,E)})} = 1 + \tilde{C}_E \gamma
\] (2.8)

where,

\[
\tilde{C}_E = \begin{cases} 
-0.202 - 0.0045 & \text{if } \lfloor 2E \rfloor \text{ is odd} \\
-0.271 - 0.0037 & \text{if } \lfloor 2E \rfloor \text{ is even}
\end{cases}
\] (2.9)

Red lines in Fig. 3a–d show the approximation values for 4 arbitrary values of \(E\). The value of \(\tilde{C}_E\) is also an approximated value explained in Appendix A.2. The estimation of the ratio \(\frac{P_{\gamma}(A_{(y,E)})}{P_{<}(A_{(y,E)})}\) is 1 if \(\gamma\) is zero.

As it is shown in Fig. A.1b with black dots, the real (exact) values of \(\frac{P_{\gamma}}{P_{<}}\) do not always cross the point \((\gamma = 0, P_{\gamma}/P_{<} = 1)\) and \(P_{\gamma}/P_{<}\) can be higher/lower than 1 for non-symmetric distributions, but the estimation line crosses that point with the minimum mean square error. The reason for the estimation line to pass through the point \((\gamma = 0, P_{\gamma}/P_{<} = 1)\) is that a node with small \(\sigma^2\) has a large value of \(\gamma\) (\(\sigma^2\) is in the denominator in the calculation of \(\gamma\)), but that node is assumed to be equal to the reference node (see Equation 2.4, \(P_\gamma\) for \(\sigma^2 < \tau_E\)) and then \(P_{\gamma}/P_{<} = 1\). Since the nodes with small variance (\(\sigma^2 < \tau_E\)) and very large skewness are assumed to be equal to their reference nodes, they do not have any impact on the estimation of \(P_{\gamma}/P_{<}\) vs. \(\gamma\) and are not plotted in Fig. 3.

### 2.5 Lower and upper bounds

The reference node for all nodes with expected degree \(E_i\) is the node with the least uncertainty in its degree distribution, which is also the node with the minimum variance \([15, 16]\). The lowest variance is achieved when the node has \(\lfloor E_i \rfloor\) edges from which \(\lfloor E_i \rfloor\) edges are certain and just one edge is uncertain with probability \(E - \lfloor E_i \rfloor\). Then, the minimum variance is \((E_i - \lfloor E_i \rfloor) (\lfloor E_i \rfloor - E_i)\).

The maximum value of variance is obtained when the probabilities of incident edges are equal and the number of edges goes to infinity. To find the maximum value of variance, we have to take two points into consideration:

- first: the highest value of variance for node \(i\) with expected degree \(E_i\) and \(L_i\) incident edges is achieved when the probability of existence of all edges are equal, that is \(\frac{E_i}{L_i}\) \([17]\). In this case, the node degree distribution is a binomial distribution and the value of variance is \(\sigma^2_{\text{max}}(E_i, L_i) = L_i \frac{E_i}{L_i} (1 - \frac{E_i}{L_i}) = E (1 - \frac{E_i}{L_i})\).
second: between two nodes with the same expected degree \( E_1 = E_2 = E \) but different number of edges \( L_1 > L_2 \), the maximum value of variance for the first node is higher than the maximum value of variance for the latter, that is \( \sigma_{\max}^2(E, L_1) > \sigma_{\max}^2(E, L_2) \).

From these two facts, we infer that the maximum value of variance is obtained when the probabilities of the edges are equal and the number of incident edges goes to infinity, that is \( \sigma_{\max}^2(E_i) = \lim_{l_i \to \infty} E_i(1 - \frac{E_i}{l_i}) = E_i \). As a result, the lower and the upper bounds of variance of node degree distributions with expected degree \( E_i \) are \( (E_i - [E_i]) \) and \( E_i \), respectively:

\[
\sigma_{\min}^2(E_i) = (E_i - [E_i]) ([E_i] - E_i) \tag{2.10}
\]

\[
\sigma_{\max}^2(E_i) = E_i \tag{2.11}
\]

Figure 2 shows that the maximum value of \( P_\max \) is obtained when the variance is the minimum and the minimum value of \( P_\min \) is achieved when the variance of a node degree distribution is maximum, that is \( E_i \). As the reference node can only have degree \( [E_i] \) with probability \( E - [E_i] \) or degree \( [E_i] \) with probability \( [E_i] - E_i \), the maximum value of \( P_\min \) is the probability that both nodes have degree \( [E_i] \) or \( [E_i] \):

\[
P_{\max}(E_i) = P(\text{deg} = [E_i] \lor \text{deg} = [E_i] | \sigma^2 = \sigma_{\min}^2(E_i)) = (E_i - [E_i])^2 + ([E_i] - E_i)^2 \tag{2.12}
\]

Moreover, to calculate the minimum value of \( P_\min \), we need to consider the node with the highest variance. As mentioned above, the maximum variance corresponds to the case when the degree distribution is binomial with a large number of edges. In this situation, a binomial distribution behaves like a Poisson distribution with \( \lambda = E_i \). Thus, the probability of equality between a node with the maximum variance and the reference node \( R_{E_i} \) is:

\[
P_{\min}(E_i) = P(\text{deg} = [E_i] \lor \text{deg} = [E_i] | \sigma^2 = \sigma_{\max}^2(E_i)) = \frac{e^{-E_i} E_i^{[E_i]} [E_i]!}{[E_i]!} \left( ([E_i] - E_i) + \frac{(E_i - [E_i]) E_i}{[E_i]} \right) \tag{2.13}
\]

\( P_{\max}(E_i) \) and \( P_{\min}(E_i) \) in Equations 2.12 and 2.13 are 0 if \( E_i \) is an integer. In this case, the reference node is the node with \( L = E_i \) certain edges and the maximum value of \( P_\max \) is 1 (when two certain nodes have been compared) and the minimum value of \( P_\min \) is achieved based on Equation 2.15:

\[
P_{\max}(\text{deg} = E_i | \sigma^2 = 0) = 1 \tag{2.14}
\]

\[
P_{\min}(E_i) = P(\text{deg} = E_i | \sigma^2 = \sigma_{\max}^2(E_i)) = \frac{e^{-E_i} E_i^{E_i}}{E_i!} \tag{2.15}
\]

Dashed lines in Fig. 2a–d show the upper and the lower bounds for variance and \( P_\max \) for 4 randomly selected \( E_i \)s.

2.6 Experimental results

We use the computer science bibliography DBLP dataset [18] that has been widely used in the literature on probabilistic networks. In this network, nodes are authors of scientific publications and two authors have an uncertain edge if they have co-authored at least one paper. The probabilities of the edges are obtained
COMPARING NODE DEGREES IN PROBABILISTIC NETWORKS

Fig. 4. (a) The distance between the estimated and the exact values is high if the ratio of variance of node degree distribution to the maximum possible value of variance ($\sigma^2/E$ see Equation 2.11) is lower than 0.2. (b) The real dataset used in our experiments the variances are higher than 0.2$E$. (c) The accuracy of the proposed approximation functions calculated for all nodes in the real dataset. The average estimation distance is higher than 0.01 for 3.9% of the nodes and higher than 0.02 for 0.8% of the nodes. (d) (Blue line) The cdf of nodes whose probability of having equal degree to their corresponding reference node is less than or equal to $P_\approx$. (red line) the cdf of nodes with more than 1 incident edge whose probability of having equal degree to their corresponding reference node is less than or equal to $P_\approx$. (green line) cdf of all pairs of nodes with the same expected degree whose probability of having equal degree is less than or equal to $P_\approx$.

From an exponential function and determine the probability that two authors will co-author a paper in the future. The dataset includes 684,911 nodes and 2,284,991 edges. Nodes have 2,299 unique $E$s between $E_{\text{min}} = 0.05$ and $E_{\text{max}} = 50.7$. However, by classifying nodes into different classes for distinct values of the triplet ($E, \sigma^2, \gamma$) they are divided into 47,560 classes. This means that there are 2,299 hypothetical reference nodes (which are not necessarily members of the network), and we have to calculate 47,560 times $P_\approx, P_\gg$ and $P_\lessgtr$ for each unique triplet ($E, \sigma^2, \gamma$).

Before calculating the comparison measures, we evaluate the estimation error of our approximation functions. We use the estimation distance $|\hat{P}_\approx - P_\approx|$ to measure the error of approximation. The experimental results show that the maximum distance between the estimated value and the real value, $|\hat{P}_\approx - P_\approx|$, occurs when variance is in the range $0 < \sigma^2 < 0.1E$ (see Fig. 4a). On the other hand, for the rest of the possible values of variance, that is $\sigma^2 \approx 0$ and $0.1E \leq \sigma^2 \leq E$, the error (the estimation distance) is low. Figure 4a illustrates the error for five randomly selected $E$s. However, our analysis of the DBLP dataset shows that the variance of node degree distributions for all $E$s is always higher than 0.2$E$ (see Fig. 4b).

To illustrate the accuracy of our approximation functions, we plot the error of our estimation functions based on the value of expected degree for nodes in the aforementioned DBLP dataset in Fig. 4c. The error rate is the average of absolute values of estimation distances. We use the average because each expected degree corresponds to multiple nodes with different ($\sigma^2, \gamma$) and each of them has a different distance between the real and the estimated values. Additionally, we use absolute values to highlight the values of estimation distances. If we calculate the average value without absolute values, there is the risk
that negative and positive estimation distances eliminate each other. Figure 4c shows that the average estimation distance is higher than 0.01 for just 3.9% of nodes and is higher than 0.02 for only 0.08% of nodes. Moreover, the exact values of $P_e$, $P_{>}$ and $P_{<}$ for 47,560 unique triplets $(E, \sigma^2, \gamma)$ are calculated in 6 h and 25 min while our approximated values, that is $\tilde{P}_e$, $\tilde{P}_{>}$ and $\tilde{P}_{<}$ for all of them are calculated in 69 s which shows the efficiency of the proposed approximation functions.

In Fig. 4c, there are some oscillating pattern in the value $|\tilde{\beta}_E - \beta_E|$. As discussed in Appendix A, the oscillation depends on the variance of the reference node.

The blue line in Fig. 4d shows the cdf of nodes whose probability of having equal degree to their corresponding reference node is less than or equal to $P_e$. These data show that the probability of equality is never higher that 0.9 and for 30% of the nodes this value is less than 0.5. Moreover, almost 17% of nodes in the DBLP dataset have one incident edge, which means that they are equal to their reference node. By removing the nodes with just one incident edge from the dataset, the maximum value of $P_e$ is 0.81 and for 37% of the nodes this value is <0.5 (red line).

The probability of having equal degree for all pairs of nodes with the same expected degree is shown in Fig. 4d (green line). This result illustrates that in more than 99% of pairs of nodes with the same $E$ the probability of having equal degree is lower than 0.5.

3. Discussion and Conclusion

The most common model of probabilistic networks is the model in which edges are associated with a probability, more specifically independent and identically distributed (i.i.d.) random variables. Existence/non-existence of each edge is analogous to the success/failure of an independent Bernoulli trial and as a result, the degree distribution of node $i$ with $L_i$ incident edges is corresponding to the repetition of $L_i$ independent Bernoulli trials which is called Poisson Binomial distribution.

The probability density function (p.d.f.) of this distribution with $n$ trials is calculated with time complexity $O(n^3)$. Therefore, calculating node degree distributions for big networks and nodes with high degrees is an expensive task. On the other hand, different moments of the p.d.f. such as expected value $(E)$, variance ($\sigma^2$) and skewness ($\gamma$) of node $i$ with $L_i$ independent Bernoulli trials which is called Poisson Binomial distribution.

The probability density function (p.d.f.) of this distribution with $n$ trials is calculated with time complexity $O(n^3)$. Therefore, calculating node degree distributions for big networks and nodes with high degrees is an expensive task. On the other hand, different moments of the p.d.f. such as expected value $(E)$, variance ($\sigma^2$) and skewness ($\gamma$) of node $i$ with $L_i$ independent Bernoulli trials which is called Poisson Binomial distribution. Therefore, calculating node degree distributions for big networks and nodes with high degrees is an expensive task. On the other hand, different moments of the p.d.f. such as expected value $(E)$, variance ($\sigma^2$) and skewness ($\gamma$) of node $i$ with $L_i$ independent Bernoulli trials which is called Poisson Binomial distribution. Therefore, calculating node degree distributions for big networks and nodes with high degrees is an expensive task. On the other hand, different moments of the p.d.f. such as expected value $(E)$, variance ($\sigma^2$) and skewness ($\gamma$) of node $i$ with $L_i$ incident edges are computed linearly with the following relations [19]:

$$E_i = \sum_{j=1}^{L_i} p_j$$

$$\sigma_i^2 = \sum_{j=1}^{L_i} (p_j)(1 - p_j)$$

$$\gamma_i = \frac{1}{\sigma_i^3} \sum_{j=1}^{L_i} (p_j)(1 - p_j)(1 - 2p_j)$$

In real datasets [18, 20–22], nodes with the same expected degree have different configurations, that is number of edges and dispersion of probabilities on edges. Hence, having nodes with the same expected degree but different degree distributions is common.

In this work, we have focused on the concept of degree in probabilistic networks. Expected degree, which is the most common way of representing the degree of nodes in probabilistic networks, loses information with respect to the degree distribution: all nodes with the same expected degree but different
values for variance and skewness are assumed to be equal in degree which is an oversimplification. To keep the simplicity in calculation and at the same time to have more information about node degrees, we have shown that we can estimate \( P = \), \( P > \) and \( P < \) in comparison with a reference node as a function of variance and skewness, respectively.

The calculation of \( P = \), \( P > \) and \( P < \) helps us to make a distinction between nodes which are considered equal according to their expected degree. These values can be useful in studying degree assortativity [23, 24], which determines to what extent nodes with similar degrees are connected to each other. By using expected degree as a way to measure the similarity of nodes, all nodes with the same expected degree are assumed identical. For example, if two nodes with completely different degree distributions but the same expected degree are connected via an uncertain edge, these two nodes positively contribute to the assortativity of the network, weighted based on the probability of the edge between them. However, by comparing them with the reference node, we can find that they give a more limited contribution on the value of degree assortativity. We believe that a definition of assortativity for probabilistic networks based on the concepts introduced in this article, and potentially based on a direct comparison between nodes (without using reference nodes) is an interesting future direction.

\( P = \), \( P > \) and \( P < \) can also improve our understanding about inter-community connectivity in probabilistic networks. For example in Fig. 5a, the contribution of \( C_2 \) and \( C_3 \) in the calculation of network modularity is equal if we use expected degree. However, according to our comparison measures, \( C_2 \) and \( C_3 \) do not have the same effect on network modularity as \( C_1 \) is more likely to be more connected to \( C_2 \) than to \( C_3 \).

Expected degree is also used to measure the connectivity of sets of nodes in uncertain graphs. Given a set of nodes \( S \subseteq V \), the corresponding cut is the set of edges for which one end is in \( S \) and the other end is not in \( S \). Hence, we can compute the expected degree associated to the cut as the sum of the probabilities of all edges in the cut. Parchas et al. [25] have used this concept to compare different cuts in uncertain graphs. Moreover, they have tried to maintain the expected degree of subsets of nodes while reducing uncertainty (variance). By rewiring the edges in the cut, the resulting expected degree can be a little different from the original one because rewiring is being done on subsets of nodes with different sizes [25].

Figure 5b shows a cut with expected degree 2.4 and Fig. 5c–e illustrates multiple sparse cuts (with reduced uncertainty while keeping the expected degree) for it. Although all sparse (rewired) cuts have the same expected degree, our proposed measures \( P > \) and \( P < \) show that the best candidate for substitution is (e), because both (b) and (e) have the same value \( P > - P < \) (see Table 2).

Another application of the proposed measures is in ranking nodes based on their degree. Expected degree can be used as a way to reflect the importance of nodes based on their degree. However, as we have
Table 2 Cut (e) in Fig. 5 has expected degree 2.4 and $P_\rightarrow - P_\leftarrow = 0.02$. Only cut (e) has the same value for both $E$ and $P_\rightarrow - P_\leftarrow$.

<table>
<thead>
<tr>
<th>Cut</th>
<th>$E$</th>
<th>$P_\rightarrow - P_\leftarrow$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(b)</td>
<td>2.4</td>
<td>0.02</td>
</tr>
<tr>
<td>(c)</td>
<td>2.4</td>
<td>0.05</td>
</tr>
<tr>
<td>(d)</td>
<td>2.4</td>
<td>0.037</td>
</tr>
<tr>
<td>(e)</td>
<td>2.4</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Table 3 All nodes have the same expected degree, but the values $P_\rightarrow - P_\leftarrow$ are different. Considering both $E$ and $P_\rightarrow - P_\leftarrow$, a possible alternative ranking is Node_1 > Node_2 = Node_4 > Node_3.

<table>
<thead>
<tr>
<th>Node</th>
<th>$E$</th>
<th>$P_\rightarrow - P_\leftarrow$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>0.014</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>-0.0156</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

shown in Fig. 1 there is a plethora of configurations with the same expected degree. Hence, in order to draw a distinction between nodes with the same expected degree, we may consider using the value $P_\rightarrow - P_\leftarrow$ in addition to $E$ and then sort nodes based on both values. By using this metric, nodes with the same expected degree but with higher/lower tendency to have higher degree ($P_\rightarrow - P_\leftarrow$) will have higher/lower ranks. For example, the rank of four nodes with edge probabilities: $\mathcal{P}_1 = \{0.8, 0.9, 0.3\}$, $\mathcal{P}_2 = \{0.5, 0.6, 0.5, 0.4\}$, $\mathcal{P}_3 = \{0.5, 0.6, 0.2, 0.4, 0.3\}$ and $\mathcal{P}_4 = \{0.5, 0.5, 0.5, 0.5\}$ is the same. Table 3 shows different values of $P_\rightarrow - P_\leftarrow$. According to these values, ranking an alternative can be Node_1 > Node_2 = Node_4 > Node_3.

References

In order to measure the equality and inequality of node degree in probabilistic networks, we employ three measures:

\[ P_{=}(A, B) = \min_{i} \sum_{i=0}^{\min(L_A, L_B)} P(\text{deg}_A = i, \text{deg}_B = i) \]

where edges are i.i.d. random variables

\[ P_{>}(A, B) = \sum_{i=0}^{\min(L_A, L_B)} \pi_A(i)\pi_B(i) \]

\[ P_{<}(A, B) = 1 - P_{=}(A, B) - P_{>}(A, B) \]

And \( P_{<}(A, B) = 1 - P_{=}(A, B) - P_{>}(A, B) \).

Reference nodes are the nodes with the same expected degree of the node under consideration and the minimum variance. For integer \( E \) the minimum value of variance is 0. Henceforth, the reference node is a node with \( E \) edges. For non-integer \( E \), the reference node is a node with \( \lfloor E \rfloor \) certain edges and 1 edge with probability \( E - \lfloor E \rfloor \).
A.1 Fitted curve for \( P_\sigma = (A_{(\sigma^2, E)}) \)

To find the fitted curve for \( P_\sigma = (A_{(\sigma^2, E)}) \), we have generated various degree distributions for each expected degree in the set \( E = \{0.1, 0.2, \ldots, 15.9, 16\} \). We have chosen this set to cover significantly different values of \( E \) (from a very small value to 16), but also to have values close to each other (with a difference of 0.1) so that we can observe small changes in \( P_\sigma \). For each unique expected degree there is an infinite number of probability sets, and we have generated 200 of them. These sets are chosen to cover a wide range of possible variance values (see Equations 2.10 and 2.11). By computing \( P_\sigma \), we found a linear correlation between \( P_\sigma \) and \( \sigma^2 \) in the log-log plot. So we used a power law function as the fitted curve that describes how \( P_\sigma \) changes with respect to the variance (see Fig. 2). To estimate the exponent from empirical data (i.e. \( \beta_E \) in Equation 2.4), we impose two constraints. The first constraint appoints to integer \( E \), and enforces the fitted curve to cross point \((\sigma^2, P_\sigma) = (0.5, 0.5)\). The second constraint is for the case of non-integer \( E \)s that is crossing through point \((\sigma^2, P_\sigma) = (pq, p^2 + q^2)\), where \( p = E - \lfloor E \rfloor \) and \( q = \lceil E \rceil - E \).

The first constraint comes from the facts that both degree distribution variance and \( P_\sigma \) for a node with \( E - 1 \) certain edges and two edges with probability 0.5 are always 0.5. The second constraint is based on the fact that the maximum \( P_\sigma \) is attained when a node has the minimum degree distribution variance.

To find the coefficient \( C_E \), we enforce passing through the point having the maximum value of variance and, as a result, the minimum value of \( P_\sigma \). As discussed before the maximum value of variance is \( E \) and the minimum value of \( P_\sigma \) is calculated in Equations 2.13 and 2.15. Hence,

\[
P_\sigma(A_{(\sigma^2, E)}) = C_E E^{\tilde{\beta}_E} = P_{\sigma}^{\min}(E) \quad \rightarrow \quad C_E = \left( \frac{P_{\sigma}^{\min}(E)}{E^{\tilde{\beta}_E}} \right).
\]  

(A.3)

Moreover, by invoking of the fact that \( P_\sigma \) has to be not greater than \( P_{\sigma}^{\max}(E) \), we can find \( \tau_E \),

\[
P_\sigma(A_{(\sigma^2, E)}) \leq P_{\sigma}^{\max}(E) \quad \rightarrow \quad C_E \sigma^{2\tilde{\beta}_E} \leq P_{\sigma}^{\max}(E) \quad \rightarrow \quad \tilde{\beta}_E \ln \sigma^2 \leq \ln \left( \frac{P_{\sigma}^{\max}(E)}{C_E} \right)\]

\[\tilde{\beta}_E < 0 \quad \rightarrow \quad \sigma^2 \geq E \left( \frac{P_{\sigma}^{\max}(E)}{P_{\sigma}^{\min}(E)} \right)^{\frac{1}{\tilde{\beta}_E}} = \tau_E
\]  

(A.4)

and for \( \sigma^2 < \tau_E \), we regard \( P_\sigma(A_{(\sigma^2, E)}) = P_{\sigma}^{\max}(E) \). Which means that we look at the nodes with small variance like the reference node.

We plotted \( \beta_E \) based on different \( E \)s in Fig. A.1a, which are obtained by performing linear regression. Then, by carrying out linear regression, the relation \( \beta_E = -0.42 - 0.0024E \) is obtained.

A.2 Fitted curve for \( \frac{P_\gamma(A_{(\gamma, E)})}{P_\sigma(A_{(\gamma, E)})} \)

To find the relation for the ratio of \( P_\gamma / P_\sigma \) based on skewness, first we remove nodes whose degree distribution variance is less than \( \tau_E \) from the 200 previously generated edge probability sets, because these nodes are deemed to be the same as the reference. Then, we perform linear regression to find \( P_\gamma / P_\sigma \) as a function of skewness (see Equation 2.8). We plotted \( C'_E \) as well as intercept of the function for different \( E \)s in Fig. A.1. The intercept is obtained as 1 (see Fig. A.1b), while \( C'_E \) is estimated based on two different fitted curves (see Fig. A.1c). The green line is the fitted curve for the value of \( C'_E \) when
COMPARING NODE DEGREES IN PROBABILISTIC NETWORKS

Fig. A.1. (a) Black line: estimated value of $\beta_E$ obtained for various $E$s. Red line: estimation of $\beta_E$ as a function of $E$. (b) Black line: estimated value of intercept of Equation 2.8 obtained for various $E$s. Red line: estimation of intercept which is 1 for all $E$s. (c) Black dots: estimated value for $C_E'$ for $E$s that $E - \lfloor E \rfloor < 0.5$. Green line: estimation of $C_E'$ as a function of $E$ if $E - \lfloor E \rfloor < 0.5$. Red dots: estimated values for $C_E'$ for $E$s that $E - \lfloor E \rfloor \geq 0.5$. Blue line: estimation of $C_E'$ as a function of $E$ for stated range. (d) Possible values of variance based on the value of the uncertain edge of the reference nodes.

$E - \lfloor E \rfloor$ is less than 0.5, that is $\lfloor 2E \rfloor$ is even (black points in Fig. A.1c) and the blue line is the estimation of $C_E'$ when $E - \lfloor E \rfloor$ is more than or equal to 0.5, that is $\lfloor 2E \rfloor$ is odd (red points in Fig. A.1c). Employing two different estimation functions instead of one decreases the error rate of estimation for $P_>/P_<$.

For various expected degrees, $E = \{0.1, 0.2, ..., 15.9, 16\}$, we have estimated the corresponding values of $\beta_E$ which are shown with black dots in Fig. A.1a. There are repeated oscillations for the value $\beta_E$ for various $E$s (see Fig. A.1a). The two green squares are $\beta_E$ for $E = 3$ and $E = 4$. In these cases, the reference nodes are nodes with 3 and 4 certain edges, respectively and the reference nodes have no uncertainty, that is the $\sigma^2 = 0$ and the minimum value of $\beta_E$ is obtained. The blue crosses, $+$, show the value of $\beta_E$ for $E = 3.1$ and $E = 3.9$. In these two cases, the reference nodes are the nodes with edge probability sets $\{1, 1, 1, 0.1\}$ and $\{1, 1, 1, 0.9\}$ with equal variance $\sigma^2 = 0.09$. The value $\beta_E$ for both of them is higher than the previous cases. Finally, the highest value $\beta_E$ is obtained when $E = 3.5$ (the red triangle) in which the edge probability set of the reference node is $\{1, 1, 1, 0.5\}$ with the highest variance $\sigma^2 = 0.25$. This pattern is repeated for all $E$s.

This oscillation depends on the variance of the reference node. Higher values of variance of the reference node lead to higher values of $\beta_E$. When the reference node has the highest uncertainty, the probability of equality $P_\equiv$ decreases fast. For example, if we consider a reference node with $E = 3.5$ and $\sigma^2 = 0.25$, for two arbitrary nodes A and B with variance $\sigma^2(A) = 0.75$ and $\sigma^2(B) = 0.87$ we have $P_\equiv(A) = 0.38$ and $P_\equiv(B) = 0.36$. On the other hand, if we consider a reference node with less uncertainty in its degree distribution then $P_\equiv$ decreases more slowly. For example, if we consider a reference node with expected degree 3.1 and $\sigma^2 = 0.09$, for two arbitrary nodes with variance $\sigma^2 = 0.75$ and $\sigma^2 = 0.87$, we have, respectively, $P_\equiv = 0.415$ and $P_\equiv = 0.404$.

Figure A.1d shows different possible values of variance based on the value of the uncertain edge of the reference nodes. The maximum value of variance for reference nodes is obtained when the $E - \lfloor E \rfloor = 0.5$ or the reference node has an uncertain edge with probability 0.5. The minimum value is attained when the reference node does not have an uncertain edge, or $E - \lfloor E \rfloor = 0$. This argumentation is valid for the oscillations in Fig. A.1b, c.
Defining and Measuring Probabilistic Ego Networks

Amin Kaveh, Matteo Magnani, and Christian Rohner

InfoLab, Department of Information Technology, Uppsala University, 75105 Uppsala, Sweden
amin.kaveh@it.uu.se
matteo.magnani@it.uu.se
christian.rohner@it.uu.se

Abstract. Ego networks constitute a prominent sub-field of social network analysis. Investigating empirical ego networks has resulted in an increased understanding of human behaviors. Ego networks have also been used to estimate global network properties in those cases where the collection of the whole network data is impossible or too costly. However, despite the existence of several studies on the egocentric properties of deterministic networks, this subject has been neglected as a line of research in probabilistic graphs. In the current literature on probabilistic network analysis many methods are based on the local properties of nodes. This highlights the significance of having a clear definition of probabilistic ego network and associated measures. In this paper we show that there is not a unique way of defining probabilistic ego networks, because there are different ways of defining the neighborhood of a node when edges are uncertain. Therefore, we introduce two definitions of probabilistic ego networks, called V-Alters-Ego and F-Alters-Ego. Moreover, we investigate three fundamental measures (degree, betweenness and closeness) for each definition including two new ego measures that are only meaningful for F-Alters-Ego: probabilistic ego betweenness and $\alpha$-closeness. We compare the proposed definitions and measures, and we also compare them to the corresponding measures computed on deterministic versions of the ego networks to investigate the role of uncertainty on the results of the analysis. Our findings highlight the important role played by edge probabilities in the identification of the most influential nodes, showing that discarding the uncertainty may lead to the discovery of different sets of nodes. We also show that F-Alters-Ego measures are faster to compute, but are not compatible with a possible-world semantics, highlighting different reasons to choose one or the other definition.

1 Introduction

Empirical data collection is often an imperfect process affected by some degree of uncertainty. Uncertainty can come from different sources. For example because of noisy measurements, e.g., in biological experiments [1], or 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 [2,3]. 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 [4,5,6].
In deterministic networks, 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 [7,8,9,10]. Another branch of study tries to estimate the global properties of nodes based on their corresponding properties in their ego-networks [11,12,13]. The third branch of works attempt to focus on the differences between the egocentric properties of nodes in online and offline social networks [14,15].

The main motivation to study ego networks is the ease and accuracy of data collection. Examining the immediate neighbors of selected nodes can be a very reliable way of gathering data [12]. For example we may ask specific individuals about others with whom they have interactions. However, Bernard et al. have pointed out that this data collection method is not free from uncertainty, for example due to forgetfulness of informants [3,16,17]. More in general, when social networks are collected using questionnaires it is not unusual that different individuals provide incompatible information about the presence of some social ties. If we instead consider online ego networks (and online social networks in general) we typically have to rely on the simple data provided by Application Programming Interfaces (APIs) to measure complex types of ties, such as whether two individuals share a strong social tie, or if they have recent and significant interactions. Therefore, models allowing us to represent uncertainty may provide a more faithful representation of the collected data.

However, 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 network analysis, the absence of a probabilistic counterpart of this theory constitutes a strong limitation. For example, ego networks are important when analytic approaches and algorithms are devised based on local properties of the nodes, such as their connectivity with their one-hop (immediate) neighbors. This is true for several methods designed for probabilistic networks [18,19,20,21].

To use ego networks in probabilistic network studies we must answer two important questions. How can we extend the definition of ego networks and ego network measures to the probabilistic case? Is it beneficial to use probabilistic ego networks compared to an analysis where uncertainty is not considered? The first question may sound easy, because ego networks are simple concepts in deterministic models. However, the deterministic definition of ego network is based on the definition of neighborhood and there are different ways of defining neighborhood when edges are uncertain. Therefore, in this paper we introduce two alternative definitions and we compare them, showing that one allows a faster computation of node measures but does not satisfy the so-called possible-worlds semantics [1].

The second question is important to justify the usage of probabilistic ego networks. It is not obvious that a probabilistic model can provide better results than just discarding uncertainty and using more traditional deterministic models and methods. It could well

---

1The concept of possible-worlds is explained in Section [3]
be the case that for some centrality measures the top-ranked nodes would be the same whether we discard or we keep edge probabilities. For example, top-degree nodes in real networks often have a significantly higher number of neighbors if compared with less central nodes, so one might expect that even if some edges are uncertain these nodes will still emerge as the most central ones. However, our findings highlight the important role played by edge probabilities in the identification of the most influential nodes, showing that discarding uncertainty may lead to the discovery of different sets of nodes.

1.1 Contributions and outline

In this study, we provide three main contributions:

– As the first contribution we specify two definitions of probabilistic ego networks. In the first definition, the sets of alters vary in each possible world and are defined according to the incident edges to the ego present in that possible world. In the second definition, the set of alters is fixed. More specifically, in this definition a node is an alter if it has a positive probability of being a neighbor of the ego node. The two definitions of probabilistic ego network investigated in this work are the result of the fact that there are two answers to the question “what is node v’s neighborhood in a probabilistic network?”. Some works implicitly define v’s neighbors as a different set for each possible world, more specifically those nodes having an edge incident to v in that world. This interpretation is used in our first definition of ego network and V-Alters-Ego. As an alternative answer, some works have defined v’s neighborhood as the set of nodes having a (possibly uncertain) edge incident to v. Our second definition, and the associated concept of F-Alters-Ego, is based on this interpretation [21].

– As the second contribution, we define betweenness and closeness for both types of probabilistic ego networks and we examine to what extent the two definitions of probabilistic ego networks lead to different sets of top-ranked nodes. Notice that in the case of deterministic networks, as well as for the definition based on possible worlds, the original whole-network definitions of betweenness and closeness are used, with the only difference that they are applied to the ego network instead of the whole network. We also recall the definition of expected degree, which however is the same for both types of probabilistic ego networks.

– While network data collection processes are often imperfect and the collected data would be more accurately represented using uncertain (e.g., probabilistic) models, uncertainty is typically discarded at some point during the analysis process. That is, the main approach used in practice to analyze an uncertain network is to transform it into a deterministic network, for example by considering all the uncertain edges as certain or by keeping the edges whose probabilities are higher than a specific threshold and dropping the edges with lower probabilities [4,22,23]. In all these cases we lose information, but it is unclear whether this loss has any practical consequences, e.g., to what extent it affects the ranking of the top nodes according to the main centrality measures. Therefore, to quantify how the lost information affects our understanding of the most important nodes in the network, as the third
contribution, we examine the association between the sets of top-k nodes obtained from the proposed probabilistic egocentric measures and those in the corresponding deterministic networks.

Section 2 reviews the traditional concept of ego network and some fundamental measures that can be applied to ego networks: degree, betweenness and closeness. Section 3 describes our extensions to the probabilistic case. In particular, Section 3.1 introduces our first definition of probabilistic ego network and shows how degree, betweenness and closeness apply to this definition. Section 3.2 presents our second definition of probabilistic ego network. In this section we also propose two new measures, ego-betweenness and $\alpha$-closeness, that are only meaningful for this definition. In Section 4, we evaluate the extent to which different definitions of probabilistic ego networks result into different lists of most influential nodes in probabilistic networks. Moreover, we evaluate to what extent transforming a probabilistic network to a deterministic network (by removing edge probabilities or by thresholding) affects the list of top-ranked nodes. We conclude and present some opportunities for further research in Section 4.4.

2 Preliminaries

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 [11,12]. With this network, local structural properties of nodes can be extracted (see Figure 1a). The most common measures in ego networks are:

**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.

**Ego betweenness** Ego betweenness was introduced in [11]. Following that, the authors in [12] 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 $A^2[1-A]$ without considering the diagonal elements:

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

(1)

where, $A$ is the ego network’s adjacency matrix.

**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.
3 Probabilistic Ego Networks: Definitions and Measures

Probabilistic networks are notated as $\mathcal{G} = (V, E, p)$ 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. 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 probabilistic graphs equals the expected value of that measure over all possible worlds:

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

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

To the best of our knowledge there is no definition of probabilistic ego network in the literature. In this section we provide two definitions based on the answer to the question “what is the set of a node’s neighbors in a probabilistic network?”. One possible answer is that the set of neighbors is defined independently in each possible world as a set of the nodes that have an incident edge to the node under consideration in that possible world. Another approach is to define the set of neighbors as all nodes having an incident edge to the node under consideration with probability $p > 0$.

These two alternative definitions lead to different definitions of ego-betweenness and ego-closeness in probabilistic graphs.

![Fig. 1: Ego networks in deterministic and probabilistic networks: (a) a deterministic ego network, (b) a probabilistic network, (c) a possible world for the probabilistic network in (b), where according to the definition of V-Alters-Ego network the set of alters of $e$ is $A_v(e) = \{2, 3\}$, (d) the same possible world with the alters chosen according to the F-Alters-Ego definition that is $A_f(e) = \{1, 2, 3, 4, 5\}$.

3.1 Probabilistic Ego Networks with Varying Sets of Alters
As mentioned above, there are two ways to interpret the neighborhood of a node in probabilistic networks. In the first interpretation, the neighbors of an arbitrary node $e$ are evaluated in each possible world. In our first definition of probabilistic ego network, we take this interpretation of neighbors into consideration: for a node $e$, $e$’s ego network 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 1b shows a probabilistic network and Figure 1c illustrates a possible ego network for node $e$. In this ego network $e$’s alters are $\{2,3\}$. 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:

**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 is challenging, summary measures of degree distributions have been used as a correspondent measure for degree [18,19,20]. The most commonly used summary measure is the expected degree. To calculate the expected degree of an ego node in probabilistic networks, we have to use Equation 2 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 [24] the expected degree is calculated easily by aggregating the probability of all the edges incident to $e$.

$$E(D_e) = \sum_{u \in A_v(e)} p_{eu}$$  \hspace{1cm} (3)

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 Figure 1b is 2.8.

**Betweenness:** the probabilistic betweenness of an ego node equals the expected value of ego betweenness in all deterministic possible worlds. As discussed in [12], the shortest path length between two alters in deterministic ego networks is 1 if they are adjacent (nodes 1 and 3 in Figure 1a) or is 2 if they are not adjacent (nodes 1 and 4 in Figure 1a). 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 [12], if $A$ is the adjacency matrix of an ego network, $A^2(1 - A)_{ij}$ 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 $A^2(1 - A)$ for the deterministic graph presented in Figure 1a. 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 I).

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.

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

If A is the adjacency matrix of a probabilistic network in which \(A_{ij}\) represents the probability of the edge between nodes i and j, \(A^2[1 - A]\) does not give the same information. The following matrix presents the result of \(A^2[1 - A]\) for the probabilistic graph in Figure 1b. 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 Figure 1b 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 I 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.

\[
A^2[1 - A] =
\begin{bmatrix}
.144 & 0.126 & 0.32 & 0.27 & 0.028 \\
. . & 0.196 & 0.21 & 0.64 & 0.448 \\
. . & . & 0.32 & 0.16 & 0.38 \\
. . & . & . & 0.02 & 0.48 \\
. . & . & . & . & 0.32 \\
. . & . & . & . & .
\end{bmatrix}
\]

As a result, to obtain the probabilistic ego betweenness we cannot replace the adjacency matrix of the probabilistic ego network in Equation 1. 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 Equation 2 in which \(M_i\) is replaced by Equation 1.

**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

We recall that the second possible answer to the question: “what is the set of a node’s neighbors in a probabilistic network?” is the set of nodes having an uncertain incident
edge to e. Therefore, our second definition of probabilistic ego network defines the alters as the set of nodes that may be adjacent to the ego node. In this approach, we define the ego network of a node e to be a probabilistic network consisting of node e called ego, all the nodes having uncertain edges with the ego node, called alters, and all the uncertain edges between alters and between the ego and the alters. We notate the set of all nodes that are connected via uncertain edges to the ego node e (alters of e) as $A_f(e)$. The subscript f denotes that the set of alters is fixed in all the possible worlds. As an example for the probabilistic network presented in Figure 1b, $A_f(e) = \{1, 2, 3, 4, 5\}$. 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.

The important point in this definition of probabilistic ego networks is that while in the deterministic case the distance between the ego node and its alters is always 1, this becomes a probability distribution over possible distances in F-Alters-Ego networks. More precisely, the distance between the ego node and its alters can be higher than 1 if the edge connecting the ego and the alter has a probability lower than 1.0. As a result, we need to contemplate redefining some of the fundamental measures.

**Degree:** the notion of expected degree does not change in our two definitions of probabilistic ego networks. The expected degree of ego node e is the sum of the probabilities of the edges between the ego node and its alters (see Equation 3). For example the expected degree of e in Figure 1b in both definitions of ego network is the same, that is 2.8.

**Betweenness:** we remind the reader that ego betweenness in deterministic networks is calculated by counting the number of shortest paths with length 2 that traverse the ego node in paths with length 2 we have to consider, for each pair of alters u and v, the probability of existence of an edge between the ego and alters u and v and at the same time the probability of non-existence of that edge between them.

To define the probabilistic ego betweenness in F-Alters-Ego, first 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:

$$b_e(u,v) = \sum_{u,v \in A_f(e)} p_{eu} p_{ev} (1 - p_{uv}) \quad (4)$$

where, $p_{eu}$ is the probability of the edge between nodes e and u. As a result, we define betweenness of an ego node e to be the expected number of the shortest paths with length 2 crossing e:

$$\mathbb{E}(B_e) = \sum_{u,v \in A_f(e)} b_e(u,v) \quad (5)$$

where, $A_f(e)$ is the set of alters of ego node e. As an example in Figure 1b $b_e(1,2) = 0.7 \times 0.4 \times (1 - 0.3)$, $b_e(1,3) = 0.7 \times 0.5 \times (1 - 0.4)$ 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 two points: first, Equation 5 counts the expected number of shortest paths of length 2 between nodes \( u \) and \( v \) that cross node \( e \), regardless of whether there are other geodesic paths of length 2 in the ego network between \( u \) and \( v \) or not. Second, Equation 5 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 node. Nevertheless, we use Equation 5 as probabilistic ego betweenness in F-Alters-Ego in order to first refrain from calculating the effect of the other geodesic paths with length 2 (whose time complexity is exponential), and second to be compatible with the length of the shortest paths between alters in deterministic ego networks (which is 2).

**Closeness:** by defining probabilistic ego networks as in Section 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:

**Shortest Path Length Distribution** Shortest path lengths between any pairs of nodes in probabilistic networks are expressed as shortest path length distributions \[25\]. 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 and the longest shortest path is in the case that there is a path between the ego and its alter by traversing all other alters, and in this case the shortest path length will be 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 \( l \):

\[
sp_{u,v}(l) = \sum_{G \in \mathcal{G} : D(u,v) = l} Pr(G) \tag{6}
\]

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 probability 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 (Figure 2a) and with shortest path length 3 with probability 0.018 by passing nodes \( \{5,1\} \) or \( \{3,1\} \) (see Figure 2b). The highest shortest path length between \( e \) and 2 is obtained in the instance in Figure 2c. Furthermore, node 2 is disconnected from \( e \) with probability 0.453 (Figure 2d). Figure 2e shows the shortest path length distribution in which the event of disconnection between \( e \) and 2 is notated as \( \infty \).

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 the calculation of it 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: $1 \times sp_{e,2}(1) + 2 \times sp_{e,2}(2) + 3 \times sp_{e,2}(3) + 4 \times sp_{e,2}(4) + \infty \times sp_{e,2}(\infty) = \infty$.

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 Figure 2e reveals that in 52.6% of the possible worlds of the network in Figure 1b the distance between nodes $e$ and 2 is at most 2. Based on this intuition we define $\alpha$-distance between two nodes in probabilistic networks.

Definition: we define $\alpha$-distance to be the minimum shortest path length where the probability of having this length or less is higher than $\alpha$.

![Diagram](image)

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

Fig. 2: (a)-(d): An example of possible worlds where the shortest path length between nodes $e$ and 2 has different values, (e) shortest path length distribution between nodes $e$ and 2.
\[ d_\alpha(v, u) = \arg \min_k \left\{ \sum_{l=1}^k sp_{v,u}(l) \geq \alpha \right\} \]  

(7)

where, \( 0 < \alpha \leq 1 \). In other words, \( \alpha \)-distance between two nodes is \( k \) iff at least in \( \alpha \times |\text{PW}(\mathcal{G})| \) of possible worlds, the shortest path length between them is at most \( k \). By replacing \( \alpha \) with \( \frac{1}{2} \), we will have the median distance which is similar to the definition of median distance introduced in [25]. 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 Figure 2e).

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 \( \alpha\)-closeness of an ego node to be the sum of the reciprocal of the \( \alpha \)-distance between the ego node and each of its alters:

\[ C_\alpha(e) = \sum_{v \in A_f(e)} \frac{1}{d_\alpha(e,v)} \]  

(8)

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

4 Evaluation

In this section we want to answer three questions:

- first, what are the similarities and differences between the two definitions of probabilistic ego networks (V-Alters-Ego and F-Alters-Ego)? We answer this question in Section 4.2.
- As the second question we want to know how much information is lost if we deal with probabilistic networks as deterministic ones. We ask this question, because one of the main approaches to cope with probabilistic networks consists in looking at them as deterministic [4,22,23]. We present an answer to this question in Section 4.3.
- As the third question, we ask if the proposed probabilistic ego closeness (\( \alpha \)-closeness) is related to the other two measures in F-Alters-Ego, i.e. expected degree and probabilistic ego betweenness. The results are presented in Section 4.4.

As a method to answer the three questions above, we study the association among centrality measures by first calculating the Pearson and Spearman correlation coefficients and then calculating the proportion of common top-k nodes obtained by using the centrality measures.

4.1 Datasets

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

| Dataset | $|V|$ | $|E|$ | $p$ | $D$ |
|---------|-----|-----|-----|-----|
| Enron   | 805 | 3956| 0.173| 9.83 |
| Facebook| 1976| 1809| 0.179| 1.83 |
| Fruit-Fly| 3751| 3692| 0.148| 1.97 |
| DBLP ($\mu = 0.05$) | 2763| 3268| 0.074| 2.37 |
| DBLP ($\mu = 0.1$) | 2763| 3268| 0.139| |
| DBLP ($\mu = 0.25$) | 2763| 3268| 0.298| |
| DBLP ($\mu = 0.5$) | 2763| 3268| 0.486| |

$|V|$ is the number of nodes, $|E|$ is the number of edges, $p$ is the mean of the edge probabilities and $D$ is the mean of nodes’ degree.

**Enron** The first dataset is a random 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 $p_{i,j} = 1 - \prod\left(1 - e^{\mu (t_{now} - t_k)}\right)$ quantifying the probability that a new email will be exchanged between a pair of nodes at time $t_{now}$. $\mu$ is the scaling parameter, and $t_k$ is the time when message $k$ has been exchanged between nodes $i$ and $j$ [26]. The Enron dataset is denser than the others.

**Facebook** The second dataset contains two years of wall-to-wall postings between a sample of users in Facebook. There is an edge between two nodes if the 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 $t_{now}$ [26].

**Fruit-Fly** The third dataset is a protein-protein interaction (PPI) network of a fruit-fly. The network consists of 3751 nodes and 3692 edges [21].

**DBLP** The fourth dataset is a random 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 $p_{ij} = 1 - e^{\mu n}$ 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 $\mu$ is the scaling factor [20].

Figure 3 shows the CDF of edge probabilities of our datasets. The blue dashed lines show the probability threshold from which 25% of the edges have lower probability ($\tau_1$). Likewise, the green dotted lines indicate the threshold from which 50% of edges have lower probability ($\tau_2$). 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

---

2 Cumulative Distribution Function
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 four complete DBLP datasets with different scaling parameters $\mu = \{0.05, 0.1, 0.25, 0.5\}$.

Fig. 3: CDF of edge probabilities

![CDF of edge probabilities](image)

<table>
<thead>
<tr>
<th>dataset</th>
<th>$\rho_{V,F}$</th>
<th>$s_{V,F}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enron</td>
<td>0.98</td>
<td>0.87</td>
</tr>
<tr>
<td>Facebook</td>
<td>0.9</td>
<td>0.44</td>
</tr>
<tr>
<td>Fruit-Fly</td>
<td>0.81</td>
<td>0.49</td>
</tr>
<tr>
<td>DBLP (0.05)</td>
<td>0.56</td>
<td>0.65</td>
</tr>
<tr>
<td>DBLP (0.1)</td>
<td>0.77</td>
<td>0.7</td>
</tr>
<tr>
<td>DBLP (0.25)</td>
<td>0.94</td>
<td>0.76</td>
</tr>
<tr>
<td>DBLP (0.5)</td>
<td>0.92</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Table 2: Correlation coefficients between probabilistic ego betweenness in V-Alters-Ego and F-Alters-Ego. $\rho$ and $s$ are respectively Pearson and Spearman correlation coefficients and subscripts $V$ and $F$ respectively refer to probabilistic ego betweenness in V-Alters-Ego and F-Alters-Ego definitions of probabilistic ego networks.
4.2 Comparing Measures in V-Alters-Ego and F-Alters-Ego

The notion of degree in deterministic networks is replaced by the notion of node degree probability distribution in probabilistic networks. However, in practice instead of computing the whole distribution its expected value is used: the expected degree 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 edges in network $G$.

Ego betweenness in each definition of the probabilistic ego network has different definitions. In V-Alters-Ego, ego betweenness is the mean of ego betweenness in each possible world. The number of possible worlds increases exponentially as the number of edges in the ego network increases. Hence, calculation of the ego betweenness in V-Alters-Ego for even average size ego networks is intractable. Ego betweenness in F-Alters-Ego is based on the calculation of the probability of open triangles in probabilistic ego networks. If the number of incident edges to the ego node is $D_e$ then the computational complexity is $O\left(\binom{D_e}{2}\right) = O(D_e^2)$ which is tractable even for nodes with large $D_e$. Table 2 shows high correlation coefficients between probabilistic ego betweenness in V-Alters-Ego and F-Alters-Ego. Moreover, Figure 4 shows the proportion of common top-k nodes obtained by using probabilistic ego betweenness in the two definitions. The ego betweenness in V-Alters-Ego has been obtained by averaging the ego betweenness of 15000 samples from each node’s ego networks.

In V-Alters-Ego, probabilistic ego closeness is 1 for all nodes by definition. However, in F-Alters-Ego $\alpha$-closeness is capable of making a distinction between the nodes in a network, based on $\alpha$-distances. 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 $\alpha$-closeness this node has. The time complexity of $\alpha$-closeness depends on the time complexity of the shortest path length distribution. The calculation of the complete shortest path length distribution needs to generate all possible worlds in F-Alters-Ego. However, $\alpha$ prunes many possible worlds and just considers the possible worlds where the distance between the

---

Fig. 4: Proportion of common top-k nodes obtained using the two definitions of probabilistic ego betweenness for V-Alters-Ego and F-Alters-Ego networks.
ego node and its alter is as short as possible and the sum of those possible worlds is greater or equal to $\alpha$. Then, the smaller $\alpha$ is, the less possible worlds are needed to be generated.

4.3 Ego Measures in Probabilistic vs. Deterministic Networks

The analysis of probabilistic networks is mainly based on the possible worlds semantics \[25,19,18\]. However, processing of all possible graphs is intractable since there are $2^{\mid E\mid}$ possible deterministic graphs. One direct solution to this challenge is the sampling approach, but still the number of sampled worlds can be considerable \[18\]. In practice, uncertainty is typically discarded at some point during the data collection or analysis process. One common approach is to discard all the uncertainty during the definition of the network. In other words, the network is dealt with as a deterministic network by considering all the uncertain edges as certain. In the second approach a probabilistic network is transformed to a deterministic network by keeping the edges whose probability is higher than a specific threshold and dropping the edges with lower probabilities. Both approaches lead to information loss about the structural properties of the network \[4,22,23\].

To evaluate the effect of the lost information on our perception about the most important nodes in networks, we study the association between the degree and the expected degree as well as deterministic ego betweenness and the two ego betweenness in V-Alters-Ego and F-Alters-Ego.

<table>
<thead>
<tr>
<th>dataset</th>
<th>$\rho_{D,E}$</th>
<th>$\rho_{D_{T1},E}$</th>
<th>$\rho_{D_{T2},E}$</th>
<th>$\sigma_{D,E}$</th>
<th>$\sigma_{D_{T1},E}$</th>
<th>$\sigma_{D_{T2},E}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enron</td>
<td>0.81</td>
<td>0.86</td>
<td>0.93</td>
<td>0.79</td>
<td>0.92</td>
<td>0.96</td>
</tr>
<tr>
<td>Facebook</td>
<td>0.71</td>
<td>0.82</td>
<td>0.9</td>
<td>0.52</td>
<td>0.8</td>
<td>0.9</td>
</tr>
<tr>
<td>Fruit-Fly</td>
<td>0.78</td>
<td>0.8</td>
<td>0.81</td>
<td>0.81</td>
<td>0.82</td>
<td>0.75</td>
</tr>
<tr>
<td>DBLP (0.05)</td>
<td>0.79</td>
<td>-</td>
<td>-</td>
<td>0.85</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DBLP (0.1)</td>
<td>0.83</td>
<td>-</td>
<td>-</td>
<td>0.85</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DBLP (0.25)</td>
<td>0.89</td>
<td>-</td>
<td>-</td>
<td>0.88</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DBLP (0.5)</td>
<td>0.95</td>
<td>-</td>
<td>-</td>
<td>0.92</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3: Correlation coefficients between expected degree in probabilistic ego networks and degree in their corresponding deterministic networks. Deterministic networks are obtained by either removing all probabilities or thresholding. $\tau_1$ is threshold 1 and $\tau_2$ is threshold 2 in Figure [5].

Table 3 shows high correlation values between degree and expected degree. The Pearson correlation coefficient between expected degree in a probabilistic graph and degree in the corresponding deterministic graph increases after thresholding. The reason is that the edges with low probability values are removed from the network and the edges with high probability values have a contribution in calculating both nodes expected degree and degree. Then the difference between degree and expected degree of nodes decreases. This pattern appears also when the Spearman correlation coefficient is used except for $\sigma_{D_{T1},E}$ in Fruit-Fly.
Table 4: Correlation coefficients between the two definitions of probabilistic ego betweenness for V-Alters-Ego (Section 3.1) and F-Alters-Ego networks (Section 3.2) in probabilistic ego networks and ego betweenness in their corresponding deterministic networks. Deterministic networks are obtained by either removing all probabilities or thresholding. $\sigma$ is deterministic ego betweenness, $B_V$ is probabilistic ego betweenness in V-Alters-Ego networks and $B_F$ is probabilistic ego betweenness in F-Alters-Ego networks.

<table>
<thead>
<tr>
<th>dataset</th>
<th>$\rho_{\sigma, B_V}$</th>
<th>$\rho_{\sigma_1, B_V}$</th>
<th>$\rho_{\sigma_2, B_V}$</th>
<th>$\rho_{\sigma, B_F}$</th>
<th>$\rho_{\sigma_1, B_F}$</th>
<th>$\rho_{\sigma_2, B_F}$</th>
<th>$\rho_{\sigma, \sigma_2, B_V}$</th>
<th>$\rho_{\sigma_1, \sigma_2, B_V}$</th>
<th>$\rho_{\sigma, \sigma_1, B_V}$</th>
<th>$\rho_{\sigma, \sigma_2, B_V}$</th>
<th>$\rho_{\sigma_1, \sigma_2, B_V}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enron</td>
<td>0.57</td>
<td>0.68</td>
<td>0.71</td>
<td>0.81</td>
<td>0.84</td>
<td>0.82</td>
<td>0.58</td>
<td>0.71</td>
<td>0.76</td>
<td>0.88</td>
<td>0.9</td>
</tr>
<tr>
<td>Facebook</td>
<td>0.56</td>
<td>0.68</td>
<td>0.74</td>
<td>0.45</td>
<td>0.49</td>
<td>0.54</td>
<td>0.71</td>
<td>0.84</td>
<td>0.9</td>
<td>0.93</td>
<td>0.87</td>
</tr>
<tr>
<td>Fruit-Fly</td>
<td>0.48</td>
<td>0.55</td>
<td>0.61</td>
<td>0.46</td>
<td>0.47</td>
<td>0.44</td>
<td>0.77</td>
<td>0.85</td>
<td>0.93</td>
<td>0.96</td>
<td>0.85</td>
</tr>
<tr>
<td>DBLP (0.05)</td>
<td>0.1</td>
<td>-</td>
<td>-</td>
<td>0.3</td>
<td>-</td>
<td>-</td>
<td>0.62</td>
<td>-</td>
<td>-</td>
<td>0.43</td>
<td>-</td>
</tr>
<tr>
<td>DBLP (0.1)</td>
<td>0.33</td>
<td>-</td>
<td>-</td>
<td>0.33</td>
<td>-</td>
<td>-</td>
<td>0.68</td>
<td>-</td>
<td>-</td>
<td>0.43</td>
<td>-</td>
</tr>
<tr>
<td>DBLP (0.25)</td>
<td>0.85</td>
<td>-</td>
<td>-</td>
<td>0.34</td>
<td>-</td>
<td>-</td>
<td>0.76</td>
<td>-</td>
<td>-</td>
<td>0.44</td>
<td>-</td>
</tr>
<tr>
<td>DBLP (0.5)</td>
<td>0.95</td>
<td>-</td>
<td>-</td>
<td>0.38</td>
<td>-</td>
<td>-</td>
<td>0.81</td>
<td>-</td>
<td>-</td>
<td>0.47</td>
<td>-</td>
</tr>
</tbody>
</table>

Fig. 5: Proportion of common top-k nodes obtained using degree and expected degree.

In Table 4, although the correlation coefficients between the two definitions of probabilistic ego betweenness and deterministic ego betweenness are considerable, they are lower than the corresponding values in comparing degree and expected degree. This implies that
more information will be lost in measuring the importance of nodes in connecting alters, by considering a probabilistic graph as a deterministic one.

The low correlation coefficient between ego betweenness measures in the DBLP dataset arises from the fact that this dataset contains many cliques of different sizes. The deterministic ego betweenness of a node in a clique is zero, because there is an edge between each pair of alters. However, the ego betweenness of nodes in the probabilistic version of DBLP is not zero and nodes in cliques with different sizes and different edge probabilities have different probabilistic ego betweenness values.

As the second method to evaluate the relationship among centrality measures in deterministic and probabilistic ego networks, we examine the top-k nodes extracted using different measures. Figure 5 shows the proportion of common top-k nodes chosen using expected degree and degree. The red solid lines represent the proportion of the common top-k nodes chosen by the expected degree in probabilistic networks and degree in the deterministic version of the same networks (considering all uncertain edges as certain). The dotted green and the dashed blue lines demonstrate the same proportion in the cases of comparing the probabilistic graphs with deterministic graphs after removing 25% and 50% of low probability edges respectively. Subfigures 5a, 5b, and 5c show that the proportion of common nodes for small k values is low and illustrate that by not looking at a probabilistic network as probabilistic, it is highly likely to select the set of top nodes differently. Figure 6a shows the expected degree distribution in the Enron dataset. One node has expected degree higher than 40, which is the top-1 node chosen using expected degree. However, this node is not the node with the largest degree. Figure 6 shows the distribution of all probabilistic ego measures in the two definitions in all the datasets. In all subfigures, the number of nodes with high values of x-axis (probabilistic ego measures) are small. Similarly our experiments show that the number of nodes with large values of counterpart measures in deterministic ego networks is also small. However, the number of common nodes between these two sets is much lower than the values that the correlation coefficients would suggest.

We have repeated the same experiment to understand whether our understanding about the ego betweenness of nodes is influenced by dealing with a probabilistic network as deterministic. Figure 7 shows the proportion of common top-k nodes chosen by ego betweenness introduced in Equation 5 and ego betweenness in the corresponding deterministic graphs. In all datasets, the proportion of common top-k nodes chosen by using deterministic and probabilistic ego betweenness is lower than the same proportion, in the case of comparing degree and expected degree.

Figures 5d and 7d show the proportion of common top-k nodes chosen using degree and ego betweenness, respectively. Since edge probabilities in the DBLP dataset with $\mu = 0.5$ are higher than other DBLP datasets, then this graph has higher value for the proportion of common top-k nodes obtained from degree and the expected degree. Moreover, Pearson and Spearman correlation coefficients for the DBLP datasets also show that as the probability of edges increases (based on the scaling parameter $\mu$), the graph is more behaving like a deterministic network in measuring node degree. The correlation coefficients between the probabilistic and deterministic ego betweenness show the higher correlation values for the datasets with higher probabilities. As mentioned before, the DBLP dataset contains a large number of cliques. In deterministic cliques
the ego betweenness is zero and in probabilistic cliques the ego betweenness is higher based on the edge probabilities. In the DBLP datasets with higher scaling parameters.
the probability of edges between alters increases and as a result the betweenness of the ego node decreases and the results of probabilistic and deterministic ego betweenness become closer.

It is worth mentioning that more than 70% of the edges in DBLP datasets have the lowest probability (see Figure 3d). So, finding a threshold that removes a small percentage of the edges with the lowest probability is impossible. For example, by choosing the lowest edge probability as the threshold more than 70% of the edges are dropped and the structure of the network completely changes. This issue confirms that in some probabilistic networks, applying the thresholding approach to transform a probabilistic dataset to a deterministic one is implausible.

4.4 Evaluating $\alpha$-closeness

Figure 8 shows the CDF of ego $\alpha$-closeness in our datasets. According to the definition of $\alpha$-closeness in Equation 8, the higher $\alpha$ leads to the lower $\alpha$-closeness. Figure 8 confirms this property in all the datasets. For example, the dashed line in Figure 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 study the correlation between it and the other two probabilistic ego measures: expected degree and probabilistic ego betweenness. Table 5 shows high Pearson correlation between $\alpha$-closeness and expected degree in all the datasets. Moreover, the correlation is higher as $\alpha$ gets higher values.

Although the correlation coefficients between probabilistic ego betweenness and $\alpha$-closeness are lower than those coefficients between expected degree and $\alpha$-closeness, still they are considerable. The results in Table 5 and Figure 9a show that as $\alpha$ increases both correlation coefficients and intersection of top-k nodes between $\alpha$-closeness and the other two probabilistic measures increase. The same pattern has been observed for the Facebook dataset, however, the intersection of top-k nodes is higher when we are comparing $\alpha$-closeness with probabilistic ego betweenness. Generally for all datasets, the intersection between sets of top-k nodes for small values of $k$ chosen using expected degree (or probabilistic ego betweenness) and $\alpha$-closeness is neither close to 1, which would have shown that expected degree or probabilistic ego betweenness are good replacements for $\alpha$-closeness, nor close to 0, which would have implied that $\alpha$-closeness is reflecting completely different local structural properties in comparison to the other two measures, i.e. expected degree and probabilistic ego betweenness.

### Conclusions and Future Works

In this paper, we propose and investigate two alternative definitions of ego networks in probabilistic graphs that we call V-Alters-Ego and F-Alters-Ego. We reviewed notions of degree, betweenness and closeness in V-Alters-Ego and proposed two new measures for F-Alters-Ego, namely ego-betweenness and $\alpha$-closeness.

V-Alters-Ego is a straightforward application of the possible worlds semantics. The F-Alters-Ego definition is appealing because not only it is rooted in the literature on probabilistic networks, but it is also computationally simpler than the approach based on possible worlds. At the same time, we also show that the saving in computational complexity may correspond to results that are not compatible with the possible worlds semantics. In particular we show that only some of the top nodes identified using this definition and some popular centrality metrics correspond to the nodes we would identify using V-Alters-Ego. In summary both definitions have pros and cons, that we have examined both in theory and experimentally.

We also looked at the correlations between rankings of nodes obtained using deterministic ego networks and our definitions of probabilistic ego networks. We have examined these correlations because one of the typical approaches to analyze probabilistic

<table>
<thead>
<tr>
<th>dataset</th>
<th>$\rho_{Cl_{\alpha}E}$</th>
<th>$\rho_{Cl_{\alpha}E}$</th>
<th>$\rho_{Cl_{\alpha}E}$</th>
<th>$\rho_{Cl_{\alpha}E}$</th>
<th>$\rho_{Cl_{\alpha}E}$</th>
<th>$\rho_{Cl_{\alpha}E}$</th>
<th>$\rho_{Cl_{\alpha}E}$</th>
<th>$\rho_{Cl_{\alpha}E}$</th>
<th>$\rho_{Cl_{\alpha}E}$</th>
<th>$\rho_{Cl_{\alpha}E}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enron</td>
<td>0.92</td>
<td>0.93</td>
<td>0.95</td>
<td>0.96</td>
<td>0.69</td>
<td>0.75</td>
<td>0.91</td>
<td>0.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Facebook</td>
<td>0.82</td>
<td>0.79</td>
<td>0.79</td>
<td>0.78</td>
<td>0.51</td>
<td>0.57</td>
<td>0.98</td>
<td>0.69</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fruit-Fly</td>
<td>0.78</td>
<td>0.79</td>
<td>0.81</td>
<td>0.78</td>
<td>0.8</td>
<td>0.75</td>
<td>0.98</td>
<td>0.74</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DBLP (0.5)</td>
<td>0.95</td>
<td>0.97</td>
<td>0.92</td>
<td>0.88</td>
<td>0.8</td>
<td>0.75</td>
<td>0.98</td>
<td>0.74</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Correlation coefficients between $\alpha$-closeness, expected degree and probabilistic ego betweenness in F-Alters-Ego
networks in the literature is to translate the probabilistic network into a deterministic one by neglecting all probability values or by thresholding. Our results show that, for some well-known data used in the literature, transforming a probabilistic network into a deterministic one does lead to a loss of information, where some of the most central nodes can no longer be identified for degree, betweenness and closeness, equally. This is both surprising and significant. Surprising because the centrality measures typically follow a heavy tail distribution with clear distinctions among the top-k nodes; significant because top-k nodes have an important role in many applications such as network immunization and viral marketing.

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.

**Declarations**

**Abbreviations.**

V-Alters-Ego: Varying set of Alters Ego network; F-Alters-Ego: Fixed set of Alters Ego network; API: Application Programming Interface; PW: Possible World; SP: Short-
Fig. 9: Proportion of common top-k nodes obtained using $\alpha$-closeness and expected degree as well as $\alpha$-closeness and probabilistic ego betweenness in F-Alters-Ego.

**Availability of data and material.**

The data used during the experimental evaluation of the work is publicly available through the references cited in the paper. DBLP dataset can be found at [20]. Fruit-Fly dataset at [21]. Facebook and Enron datasets are available in the Koblenz Network collection at [http://konect.uni-koblenz.de/networks/facebook-wosn-wall](http://konect.uni-koblenz.de/networks/facebook-wosn-wall) and [http://konect.uni-koblenz.de/networks/enron](http://konect.uni-koblenz.de/networks/enron) respectively. The analysis was performed using a code developed in R.

**Competing interests.**

The authors declare that they have no competing interests.

**Funding.**

Not applicable

**Authors’ contributions.** The first author (A.K.) initiated the study and implemented and performed the experiments. All authors contributed equally in conceiving and developing the study and analyzing the results. All authors approved the final version of the manuscript.

**Acknowledgements.** Not applicable.
References

20. Parchas, P., Gullo, F., Papadias, D., Bonchi, F.: Uncertain graph processing through repre-
21. Mukherjee, A.P., Xu, P., Tirthapura, S.: Enumeration of maximal cliques from an uncertain
graph. IEEE Transactions on Knowledge and Data Engineering 29(3) (2017) 543–555
N., Tikuisis, A.P., et al.: Global landscape of protein complexes in the yeast saccharomyces
Review E 93(1) (2016) 012306
Journal of Complex Networks (2019 (to appear))
26. Pfeiffer, J.J., Neville, J.: Methods to determine node centrality and clustering in graphs with
uncertain structure. In: Fifth International AAAI Conference on Weblogs and Social Media.
(2011)
Recent licentiate theses from the Department of Information Technology

2019-003 Viktor Bro: *Volterra Modeling of the Human Smooth Pursuit System in Health and Disease*


2019-001 Diane Golay: *An Invisible Burden: An Experience-Based Approach to Nurses’ Daily Work Life with Healthcare Information Technology*

2018-004 Charalampos Orfanidis: *Robustness in Low Power Wide Area Networks*

2018-003 Fredrik Olsson: *Modeling and Assessment of Human Balance and Movement Disorders Using Inertial Sensors*

2018-002 Tatiana Chistiakova: *Ammonium Based Aeration Control in Wastewater Treatment Plants - Modelling and Controller Design*

2018-001 Kim-Anh Tran: *Static Instruction Scheduling for High Performance on Energy-Efficient Processors*

2017-003 Oscar Samuelsson: *Fault Detection in Water Resource Recovery Facilities*

2017-002 Germán Ceballos: *Modeling the Interactions Between Tasks and the Memory System*

2017-001 Diana Yamalova: *Hybrid Observers for Systems with Intrinsic Pulse-Modulated Feedback*

2016-012 Peter Backeman: *New Techniques for Handling Quantifiers in Boolean and First-Order Logic*

2016-011 Andreas Svensson: *Learning Probabilistic Models of Dynamical Phenomena Using Particle Filters*