Modelling and Analysis of Probabilistic Networks

AMIN KAVEH
Abstract


As empirical data collection and inference is often an imperfect process, and many systems can be represented as networks, it is important to develop modelling and analysis methods for imperfect network data. The main focus of this dissertation is the probabilistic network model $G = (V, E, p)$ in which each edge is associated with an independent existence probability. This model can be used to represent both collected data and our understanding about it in many applications such as biological and social network analysis. A probabilistic network with $m$ probabilistic edges corresponds to $2^m$ deterministic instances, known as possible worlds, and most of the existing network analysis measures can be represented as probability distributions. This introduces three challenges. The first challenge is to find methods to calculate or estimate the required measures with non-exponential computational time complexity. The second challenge arises due to the fact that many network analysis algorithms are designed to use single number measures such as degree and cannot deal with measures that are represented as probability distributions. Therefore, the second challenge in this field is to adapt network science algorithms such that they can utilise the information represented in measures’ probability distributions. The third challenge is the aggregation of information yielded from the analysis of possible worlds. This thesis has considered these three challenges. In particular, this thesis first scrutinises fundamental local measures in probabilistic networks. It proposes measures to compare nodes’ degree distributions and it introduces a method to estimate these measures. Moreover, it extends the concepts of ego network and ego betweenness to probabilistic networks, and proposes a method to estimate ego betweenness in this context. Second, this thesis focuses on the problem of probabilistic network sparsification which is a method to generate an alternative probabilistic network whose analysis is simpler than the original one. Third, this thesis studies for the first time the problem of overlapping community detection in probabilistic networks and proposes and compares different extensions of the clique percolation method for such networks.

Keywords: Probabilistic Networks, Local Measures, Modelling, Clustering

Amin Kaveh, Department of Information Technology, Computing Science, Box 337, Uppsala University, SE-75105 Uppsala, Sweden.

© Amin Kaveh 2022

ISSN 1651-6214
URN urn:nbn:se:uu:diva-467326 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-467326)
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  
Oxford University Press.

II  **Defining and Measuring Probabilistic Ego Networks**  
Amin Kaveh, Matteo Magnani, Christian Rohner  
Social Network Analysis and Mining  
Vol. 11, No. 1, 2021.  
Springer.

III  **Probabilistic network sparsification with ego betweenness**  
Amin Kaveh, Matteo Magnani, Christian Rohner  
Applied Network Science  
Vol. 6, No. 1, 2021.  
Springer

IV  **Overlapping Clustering of Probabilistic Networks**  
Amin Kaveh, Matteo Magnani, Christian Rohner  
Under submission

Reprints were made with permission from the publishers.
List of works not included

In addition to the papers included in the dissertation, I have authored the following:

1. **Degree in Probabilistic Networks: Revisited**  
   Amin Kaveh, Matteo Magnani, Christian Rohner  
   The 6th International Conference on Complex Networks and their Applications  
   2017.

2. **Uncertainty in fMRI Functional Networks of Autism Brain Imaging Data**  
   Amin Kaveh, Matteo Magnani, Christian Rohner  
   2022.
Contents

1 Introduction ........................................................................................................... 11
   1.1 Possible worlds semantics ........................................................................... 11
       1.1.1 Challenge 1 - Computational Complexity ...................................... 13
       1.1.2 Challenge 2 - Measures as Distributions .............................. 15
       1.1.3 Challenge 3 - Aggregation ...................................................... 15
   1.2 Research Questions of the Thesis ............................................................... 16
   1.3 Structure of the Thesis ............................................................................... 17

2 Terminology and Theories ................................................................................... 19
   2.1 Terminology ................................................................................................ 19
       2.1.1 What is Information? ..................................................................... 19
       2.1.2 Sources of Imperfect Information ........................................... 21
       2.1.3 Types of Imperfect Information .............................................. 21
   2.2 Imperfection in Networks ............................................................................ 24
       2.2.1 Existence of Nodes and Edges .................................................. 25
       2.2.2 Weights/Labels of Nodes and Edges ...................................... 25
   2.3 Theories ....................................................................................................... 26
       2.3.1 Probability Theory ...................................................................... 26
       2.3.2 Dempster-Shafer Theory of Evidence .................................... 28
       2.3.3 Fuzzy Theory ............................................................................ 30
       2.3.4 Uncertainty Theory .................................................................... 32
       2.3.5 Comparing Theories ................................................................... 33

3 From Systems to Network Models ...................................................................... 43
   3.1 Uncertainty in fMRI Functional Networks .............................................. 43
   3.2 Uncertainty in Social Networks ................................................................ 45
       3.2.1 Coauthorship networks ................................................................. 45
       3.2.2 Digital communication networks ........................................... 46

4 From Network Models to Insights: Algorithms ............................................... 47
   4.1 Path Based Queries .................................................................................... 47
       4.1.1 Shortest Path Length Distribution ........................................ 47
       4.1.2 Reliability ................................................................................... 48
   4.2 Pattern Extraction Queries ....................................................................... 49
       4.2.1 Core Decomposition ................................................................ 49
       4.2.2 Clique Mining .......................................................................... 49
       4.2.3 Threshold-Based Pattern Matching ..................................... 50
1. Introduction

Scientists in many disciplines want to understand systems and their behaviour. To do so, many systems can be modelled as networks, that is, a collection of nodes that are connected via edges. Network science is an academic field about modelling a system as a network and analysing that network to understand structure and function of the system and the extent to which its structure affects its function [28, 30, 119] (Figure 1.1).

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 [157], 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, 12]. As another example in social networks, interactions are subject to some misunderstandings, due to the forgetfulness of informants [80, 81].

The most practised approach in network science consists in removing all kinds of uncertainty from data and consequently in modelling systems as deterministic networks. All developed algorithms and methods are appropriate for deterministic networks. However, the assumption of removing uncertainty before modelling systems has been challenged in recent years and thus models that can represent uncertainty have been proposed.

The most applied network model to represent imperfection of our understanding about the collected data 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 an existence probability to each edge, where the existence of an edge is independent of the others. This is called a probabilistic network model and has been used to represent imperfect network data in sensor networks [52, 82, 132], opportunistic networks [102, 103], protein-protein interaction networks [26, 136, 148, 150], social influence networks [14, 22, 129] and road networks [50]. This requires new processing and analysis methods (Figure 1.2).

1.1 Possible worlds semantics

When an edge in a probabilistic network is associated with an existence probability $p$, this means that that edge corresponds to two states, existing with
**Figure 1.1.** Many systems can be modelled with networks and measuring and analysing network models can reveal beneficial insights about the modelled systems.

**Figure 1.2.** As the modelling of a system changes, then processing and analysing methods also have to be adapted to that model.
probability $p$ and not existing with probability $1 - p$. Therefore, a probabilistic network with $m$ probabilistic edges have $2^m$ deterministic instances, so called possible worlds.

In possible world semantics, a measure in probabilistic network $\mathcal{G} = (V, E, p)$ is represented as a probability distribution over all possible worlds. Figure 1.3 (left) shows a probabilistic network $\mathcal{G} = (V, E, p)$ with 4 nodes and 4 probabilistic edges. Figure 1.4 illustrates the $2^4$ corresponding possible worlds of $\mathcal{G}$. Figure 1.3 (right) indicates the degree distribution of node $v_3$ calculated over all possible worlds.

Employing possible worlds semantics in network science results in three intrinsic challenges.

1.1.1 Challenge 1 - Computational Complexity

As the number of possible worlds is exponential in the number of edges, exact calculation of probability distributions of measures on this model is computationally prohibitive even for moderate size networks. To address this challenge, one solution is to approximate the probability distribution of measures of interest by sampling a subset of possible worlds. Although this approach takes less computational time, it still incurs high computational cost because of two reasons. First, because the number of required samples are considerable especially for networks with high entropy\(^1\) [121]. Second, many network measures are expensive to calculate even in a single possible world e.g., closeness with $\mathcal{O}(nm)$ where $n$ and $m$ are the number of nodes and edges respectively. Therefore, having $N$ samples from a probabilistic graph means that we have to calculate each measure like closeness $N$ times.

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

---

Figure 1.3. (Left) a probabilistic graph, (Right) node’s degree distribution for node $v_3$. 

\[\begin{array}{c}
0.7 \\
0.5 \\
0.3 \\
0.9 \\
\end{array}\]
Figure 1.4. (a)-(p) All possible worlds of the graph presented in Figure 1.3.
Sparsification is an approach to reduce the complexity of a network by reducing the number of edges while still maintaining a target measure. The reduction of the number of edges leads to less possible worlds, and thereby is an approach to handle the first challenge. Given a probabilistic graph \( G = (V, E, p) \), this approach generates an alternative network, \( G' = (V, E', p') \), with a different set of edges and probability assignment and with less entropy. More precisely, the algorithms proposed in this approach generate a probabilistic graph \( G' \) in which (a) \( |E'| = \alpha |E| \) where \( 0 < \alpha < 1 \), (b) the expected degree of each node in \( G' \) is 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 \( G' \) is as close as possible to the sum of expected degree of all nodes in \( G \) [121]. Instead of sampling \( G \) and calculating measures over those samples, we can sample a smaller number of possible worlds from the sparsified graph \( G' \) and calculate the measures over them. The latter is an estimation of the first with low errors.

1.1.2 Challenge 2 - Measures as Distributions

Most network science algorithms cannot deal with measures that are represented as probability distributions. One way of addressing this challenge is using summary information of probability distributions such as the expected value. For example, instead of using the degree distribution of a node we can use the expected degree of this distribution as a single number. However, using this method may lead to information loss in our analysis.

One solution is first to extract information from probability distribution of measures, and second to use this information in the main task. One of the best examples of this approach is the work of Bonchi et al. [15]. The main task in this work is to find cores of probabilistic networks. The concept of core is tightly connected with nodes’ degree. As a node’s degree in probabilistic networks is represented as a probability distribution, Bonchi et al. first extract the maximum degree such that the probability of having that degree is higher than a given threshold \( 0 \leq \eta \leq 1 \), and then they use this information to extract the core of that probabilistic network.

1.1.3 Challenge 3 - Aggregation

Aggregating information yielding from the analysis of possible worlds is not straightforward if it is not expressed as a single number. One example is clustering. Figures 1.5b and 1.5c show the clustering of two possible worlds for the probabilistic graph in Figure 1.5a, where it is not obvious how these should be aggregated.

This can be addressed by narrowing down the queries to be asked from probabilistic networks. This can be performed by looking for events whose probabilities are higher than a given threshold. In other words, this approach
only considers those possible worlds in which an event is valid and the sum of their probabilities is higher than a specific threshold. Therefore, instead of looking for different events with any possible probability we are looking for specific events whose probability is higher than a specific threshold. One of the best examples of such an approach is the work of the authors of [15] in which they are looking for cores whose probability is higher than a given threshold.

1.2 Research Questions of the Thesis
In a broad perspective we followed three research questions in this thesis:

**Research Question 1.** How to extend local measures and structures to probabilistic networks?

In Paper I, we have studied the centrality of a node using the concept of degree in probabilistic networks [73]. Specifically, we proposed some measures to compare degrees of two nodes while their degree is represented as a degree distribution.

In Paper II, we have focused on the concept of ego networks in probabilistic networks [74]. Specifically, first we have explored the definition of nodes’ ego networks in probabilistic networks and second we have studied three fundamental measures, degree, betweenness and closeness in probabilistic ego networks.

Finally, in Paper IV (see Section 5.4) we have considered the concept of $k$-cliques in probabilistic networks.

**Research Question 2.** How to efficiently compute these local measures?

In paper I [73] we have proposed a method to estimate the defined comparing measures efficiently. The estimation is based on summary information of degree distributions such as expected value, variance and skewness. As the degree distribution in probabilistic networks follows Poisson-binomial distri-
bution, the aforementioned summary information can be calculated without any need to the distribution itself.

In Paper II [74] we have shown that calculating nodes’ betweenness even in their probabilistic ego network is an expensive process. Therefore, we proposed a method to estimate nodes’ ego betweenness in probabilistic ego networks.

In Paper III [75], we have proposed to sparsify probabilistic networks by preserving nodes’ ego betweenness (introduced and estimated in our second contribution [75]) in the process of sparsification. As mentioned in Section 1.1.1, instead of sampling an original probabilistic graph and calculating measures over those samples, we can sample a smaller number of instances from the sparsified graph and calculate the measures over them.

**Research Question 3.** How to discover overlapping clusters using local structures in probabilistic networks?

In Paper IV we have focused on the problem of detecting overlapping and partial community structures in probabilistic networks using the clique percolation approach. Particularly, we explored three ways of applying this approach. The first one is based on the traditional way of analysis of probabilistic networks, thresholding, that removes all edges below an input threshold and then look at the probabilistic network as a one. The second one is based on another traditional way of analysis of probabilistic networks in which edges’ probabilities are considered as edges’ weights. Our proposed method, is an extension of the first two ways.

### 1.3 Structure of the Thesis

Before addressing the aforementioned research questions, we have reviewed the terminology, the imperfect network models and the analysis theories that have been used in this field. Chapter 2 provides a terminology of imperfection. Then it briefly introduces the theories that have been used in the analysis of imperfect network data. In Chapter 3, we review a number of systems that can be properly modelled with probabilistic network models. Afterwards, we review our exploratory study in [76] that exclusively studies uncertainty in functional correlation between different regions of the brain in the autism brain imaging dataset. As the applied theory to handle imperfection in network data in this thesis is probability theory, we have reviewed the state of the art on probabilistic networks in Chapter 4. Chapter 5 reviews the motivations, contributions and findings of our articles. Chapter 6 includes discussions and potential future works.

---

2Chapters 2 and 4 are adapted and extended from the licentiate thesis published in 2019 [72].
2. Terminology and Theories

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

In the first section of this chapter (Section 2.1) our goal is to consolidate the meaning of the terms used in the literature. To do so, we focus on the definition and on the primary sources of information. Then, we present the definition of each type of information imperfection with regard to its source.

In the second section of the chapter (Section 2.2) we briefly review 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 are affected by imperfect information are generally called imperfect networks.

The third section (Section 2.3) reviews different theories that have been used to capture properties of imperfect networks within feasible and formalised mathematical frameworks. It commences with the most frequently used 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.

2.1 Terminology

2.1.1 What is Information?

This section surveys different definitions of information. The term information is often mixed with other related terms such as knowledge, fact and data. The goals are first to find out whether there is a unique definition of information in dictionaries and second, to know whether the provided definitions clearly distinguishes it from other similar terms. Having a universal and unique definition of information is the basis for a 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 fact. Specifically, Cambridge dictionary defines information as “facts about a situation, person, event, etc” and Longman dictionary defines it as “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 knowledge and intelligence, i.e. “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. “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 and what type of 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 [45, 133], 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 [11]. 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” [138]. 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: the sources from which we get information. We acquire information via communication\(^1\) and observation. Zorkovcy and Heap [172] highlight the capability of information to be communicated and argue that, as common in different definitions of information, it needs to be

\(^1\)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 and sign language.
conveyed. Moreover, Rowley [133] points out that information can be derived from observations.

2.1.2 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 categorisation of terms rely on the definitions given in Stanford Encyclopedia of Philosophy and The Dictionary of Linguistics and Phonetics. The resulting terminology is compliant with the one proposed by Smets in the database literature [146].

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 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.

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 neighbours of a node in a network, our observation is not accurate and our information is not perfect.

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

Communication
The primary meaningful linguistic object is the statement which is a sentence that conveys information [32]. A statement can include imperfection because of linguistic complexity. The two main types of linguistic imperfection are ambiguity and vagueness:
Figure 2.1. Two primary sources of imperfection

**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) [32]. 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 [110].

**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 [32]. 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.

**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 [32]. 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” [109]. Therefore, propositions always have two parts: a) unambiguous and unvague meaning of statements and b) modal properties. For instance, in the proposition corresponding to statement S = “I’m 60% sure that 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 this definition, 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 possible worlds where in the first world John is 180 cm and in the second world John is 181 cm.

*Figure 2.2. A proposition is an unvague and unambiguous statement.*
Figure 2.3. Imperfection can be in the modal property or/and in the meaning of a proposition.

**Observational Imperfection – Uncertain information:** Information about one or multiple properties is uncertain if it is stated by partial belief, i.e. \(0 < \text{belief} < 1\). 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.

2.2 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. We use the terminology that is introduced in the previous section to refer to the type of imperfect network.

2.2.1 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 [140] (see Section 2.1.3).

Mering et al. in [157] 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 [88] 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 [63]. Data delivery success between two mobile nodes in opportunistic networks can also be uncertain because we are not always assured of it [102, 103]. 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 [154, 165, 175, 177].

2.2.2 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, and we can 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 [13]. Another example is association of such labels as {professional,
friendship, family} to edges, to show different types of relationships between nodes. While weights in the first group are ratio\(^2\) attributes, labels are not necessarily ratio data and can be nominal\(^3\) 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 2.1.3). 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 in [68, 145, 158]. 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].

2.3 Theories

This section reviews four theories that have been used to model and analyse imperfect network data: probability theory, Dempster-Shafer theory of evidence, fuzzy theory and uncertainty theory. We also review the extent to which each theory has been used in networks.

2.3.1 Probability Theory

The most used and consolidated theory to handle imperfect network data is probability theory. All the possible values of the variable of interest are 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 \(0 \leq p \leq 1\). Probability theory is rigorous because it is expressed through Kolmogorov axioms:

1. The probability of an event is a non-negative real number:

\[
\forall e \subset \Omega, \quad p(e) \geq 0
\]

2. The probability of event \(\Omega\) is 1,

\[
p(\Omega) = 1
\]

\(^2\)Magnitude of assigned values to ratio attributes are meaningful and all basic mathematical operations (+, −, ×, ÷) can be performed on them.

\(^3\)Values of nominal attributes are just different names and mathematical operations are not practical on them.
3. Probability of any sequence of mutually exclusive events is the sum of the probabilities of those events,

\[ p\left( \bigcup_{i=1}^{\infty} e_i \right) = \sum_{i=1}^{\infty} p(e_i) \]

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

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, 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.

One of the most common and at the same simple models 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 which assigns an independent probability to each edge \( e \in E \). In this model, the existence of nodes is deterministic and the existence of edges is probabilistic [126, 128].

Each probabilistic network \( G \) with \( m \) edges corresponds to \( 2^m \) deterministic networks which are called **possible worlds**. The probability of each possible world is:

\[
Pr(G_i) = \prod_{e \in E_i} pr(e) \prod_{e \in E \setminus E_i} (1 - pr(e)) \quad (2.1)
\]

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

\[
\mathbb{E}(M) = \sum_{G_i \in \mathcal{G}} M_i \times Pr(G_i) \quad (2.2)
\]

where, \( G_i \) is \( i \)th possible instance of \( \mathcal{G} \) and \( M_i \) is the value of measure \( M \) in possible world \( G_i \) [16, 17, 130, 170].

Many real world scenarios can be represented with this model. The authors of [148] have reviewed the research articles that have used this model to represent noisy experimental results of protein-protein interaction networks. The authors of [52, 82, 132] have modelled 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 [126, 129]. Moreover, the authors in [102, 103] have modelled opportunistic networks in which edge probabilities denote the data-delivery success rates between two mobile nodes.

While a large majority of research articles have used the aforementioned model to represent network data, some other works have deployed other variations of probabilistic network models:

- Uncertainty About The Existence of Weighted Edges:
An extension of \( \mathcal{G}(V,E,p) \) is \( \mathcal{G}(V,E,p,w) \), where \( V \) is the set of nodes, \( E \) is the set of edges, \( p : E \to [0,1] \) is a function which assigns a probability to each edge \( e \in E \) and \( W : E \to (0,\infty) \) is a function which assigns a weight to each edge [71, 129, 161, 173].

- Uncertainty about the existence of nodes and edges:
  The authors in [154, 165] 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 \( p_V : V \to (0,1] \) and \( p_E : E \to (0,1] \) assign existence probabilities to nodes and edges respectively and the existence of any pairs of nodes and any pairs of edges are independent. Moreover, Zou et al. [177] have investigated the problem of maximal clique finding in such probabilistic network models.

- Dependent edge existence:
  The authors in [24, 25, 163, 164] 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 incident to the same node are correlated.

- Uncertainty about the weight of edges:
  The authors in [68, 158] use the network model \( \mathcal{G}(V,E,W,P) \) in which the existence of nodes and edges is certain however weight on each edge follows a specific probability distribution. \( P \) 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 [68, 158] have studied different path queries on such networks.

2.3.2 Dempster-Shafer Theory of Evidence

Dempster-Shafer theory (DST) is interpreted as a generalisation 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 \( A \), which is the main difference between DST 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. \( \text{Bel}(\emptyset) = 0 \)
2. \( \text{Bel}(\Omega) = 1 \)
3. \( \text{Bel}(A_1 \cup A_2 \cup ... \cup A_n) \geq \sum_i \text{Bel}(A_i) - \sum_{i<k} \text{Bel}(A_i \cap A_k) + ... + (-1)^{n+1} \text{Bel}(A_1 \cap A_2 \cap ... \cap A_n) \)

And the plausibility function has the following properties:

1. \( \text{Pl}(\emptyset) = 0 \)
2. \( \text{Pl}(\Omega) = 1 \)
3. \( \text{Bel}(A_1 \cap A_2 \cap ... \cap A_n) \leq \sum_i \text{Pl}(A_i) - \sum_{i<k} \text{Pl}(A_i \cup A_k) + ... + (-1)^{n+1} \text{Bel}(A_1 \cup A_2 \cup ... \cup A_n) \)

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, Zadeh in [167] defined the following relation between \( \text{bel} \) and \( \text{Pl} \) functions:

\[
\text{Pl}(A) = 1 - \text{Bel}(\overline{A}) \quad (2.3)
\]

and,

\[
\text{Bel}(A) = 1 - \text{Pl}(\overline{A}) \quad (2.4)
\]

**Why DST?**

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

Firstly, when there is ignorance and we draw conclusions about subsets. In probability theory, when no information is available about the elements in a set (known as ignorance) a probability still has to be assigned, and the same value is normally assigned to all elements [59]. For example if we have statement \( s = \) “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 DST we can just set \( m(n=\{2,3\}) = 1 \) and we should not assign any value to any subset of \( \{2,3\} \).

Secondly, when having a single number probability is problematic. Smets in [147] 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, \( Bel(A) \leq m(A) \leq Pl(A) \), with well-defined rigorous mathematical rules, this theoretical framework has been used to model and to analyse imperfection in networks.

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

DST has been used to model imperfection in attributed graphs in several works [36–39, 106–108, 127]. In all these articles nodes’ attributes indicate communities of which nodes are members. The attributes on edges indicate the types of connections \{friendly, professional, family, ...\} and the attributes of the messages (passing through edges) show the categories or the contexts of messages, e.g. \{commercial, personal, ...\}. 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 supports those values. However these research articles show how DST can be used to predict spammed links [36] and to remove noise from data [38].

In summary, the works on DST on networks have focused on the elementary steps in both modelling and analysis. In the aforementioned references two main questions are not clearly answered: first, how DST improves our ability to model and to analyse imperfection in networks compared to probability theory, and second, how mass function values should be assigned to subsets of the set of discernment \( \Omega \) based on the available information.

### 2.3.3 Fuzzy Theory

Fuzzy set theory [166] is useful when we have partial information about the membership of an object to a set. It is a generalisation 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. 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 called membership function and assigns a degree of membership to all elements in \( U \) with respect to a fuzzy set \( F \):

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

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

\[
F = \{ \mu_F(u_1)/u_1, \mu_F(u_2)/u_2, ..., \mu_F(u_n)/u_n \}
\]
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 following fuzzy set operations are defined [104]:

1. Union: the union of fuzzy sets $A$ and $B$ is a new fuzzy set with membership function $\mu_{A \cup B} : U \rightarrow [0, 1]$, 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 $\mu_{A \cap B} : U \rightarrow [0, 1]$, 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 $\mu_{\overline{A}} : U \rightarrow [0, 1]$, where
   \[
   \forall u \in U : \mu_{\overline{A}} = 1 - \mu_A(u)
   \]

Why fuzzy set 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 network properties that are related to 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 associated to that set is linguistically vague (see Section 2.1), 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 [27, 44, 66, 67, 91, 111] 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 [112, 116, 144]
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.

2.3.4 Uncertainty Theory

Uncertainty theory is devised by Liu in 2007 in [93]. In [98] the author 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. if we want to combine two belief functions about tomorrow’s weather and quality of roads, Liu proposed Product Axiom in [95]:

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:

$$\mathcal{M}\left(\prod_{i=1}^{n} A_i\right) \leq \min_{1 \leq k \leq n} \mathcal{M}_k(A_i)$$

Why uncertainty theory?

Liu argues that probability theory is applicable when we have enough samples to draw probabilities. However, when there are not enough samples to accurately 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 [96] discusses why belief degree is different from subjective probability and it is inappropriate to use probability theory to analyse belief degree.

The main difference between probability theory and uncertainty theory is in the case when we want to find the probability (uncertainty measure) of the product of measures. In probability theory if two events are independent then the probability of their product 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) [94, 98].
How is uncertainty theory used to model imperfection in networks?

Gao in [55] argues that sometimes there is insufficient historical data to statistically estimate nodes’ demands in the single facility location problem. Therefore the author used uncertain variables to model the imperfection of nodes’ demands. In this model it is assumed that information on edges is perfect. The authors in [57] 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 an uncertain measure. The authors in [54,56,58,124] have used a simple network model (without any attribute on nodes and edges) in which only existence of edges is modelled with uncertain measures. 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 or unavailability of historical data, Liu in [97] 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 associated with a probability and some edges are associated with an uncertainty measure. Sheng and Mei have studied shortest path problem in uncertain random model [141]. Shi et al. in [143] studied the maximum flow problem in uncertain random networks. This model has been used to study various fundamental problems in networks such as the minimum spanning tree problem [53, 142].

2.3.5 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. DST

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 2.4 shows a network with 4 nodes and 4 edges. The length of each edge can be 1 or 2. In Figure 2.4a imperfect information about lengths of edges is modelled 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 2.4b lengths of edges are modelled 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 2.4c represents imperfect information about the length of edges with DST.
Probability theory (Figure 2.4a): Computation of the shortest paths in a network whose edges’ lengths are modelled with probability theory is straightforward. By using possible worlds semantics, the probabilistic graph in Figure 2.4a corresponds to 16 deterministic graphs demonstrated in Figure 2.5 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 $l$ is equal to the sum of the probability of all possible worlds in which the shortest path distance between $s$ and $t$ is $l$. Figure 2.5q shows the shortest path length distribution between $s$ and $t$.

Fuzzy theory (Figure 2.4b): The shortest path problem in fuzzy networks can be addressed in various ways [83]. Here we restrict ourselves to a simple\(^4\) and convenient case.

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

---

\(^4\)Solving the shortest path problem in some fuzzy network models requires complex and extended mathematical operations, e.g. see extended sum and extended min/max in Dubois and Prade model [42].
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 2.4b).

Klein in [83] 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 only mention the principles:

$$f(terminal) = (1, 1, ..., 1)$$

$$f(i) = dom_{(i,j) \in E} (e_{ij} \bigoplus f(j))$$

(2.5)

where, $e_{ij}$ are the 2 membership tuples assigned to each edge $(i, j)$ or the path from $i$ to $j$: $e_{ij} = \{ \mu_1(i, j), \mu_2(i, j), ..., \mu_R(i, j) \}$, $\bigoplus$ is the combinatorial sum and $dom$ is the domination operator.

In the fuzzy graph Figure in 2.4b 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 2.5, we obtain:

$$f(t) = (1, 1, 1, 1)$$
$$f(v_1) = (0.8, 0.8, 0, 0)$$
$$f(v_2) = (0.5, 0.6, 0, 0)$$

$$f(s) = dom \{ e_{s,v_1} \bigoplus f(v_1), e_{s,v_2} \bigoplus f(v_2) \}$$
$$= dom \{ (0, 0.6, 0.8, 0.8), (0, 0.4, 0.5, 0.6) \}$$

$$= (0, 0.6, 0.8, 0.8)$$

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.

35
(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 2.5. (a)-(p) All possible worlds of the graph presented in Figure 2.4a, (q) Shortest path length distribution between nodes $s$ and $t$. 
**DST (Figure 2.4c):** Figure 2.4c shows a network in which the imperfect information about the length of edges is represented with basic belief functions, \( m \), defined in DST. 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 DST.

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

\[
\Omega = \{1, 2\}, \quad P(\Omega) = \{\emptyset, \{1\}, \{2\}, \Omega\} \\
m(\{1\}) = 0.7, \quad m(\{2\}) = 0.1, \quad m(\{\Omega\}) = 0.2 \\
Bel(\{1\}) = 0.7, \\
Bel(\{2\}) = 0.1, \\
Pl(\{1\}) = 1 - Bel(\{T\}) = 1 - Bel(\{2\}) = 1 - 0.1 = 0.9, \\
Pl(\{2\}) = 1 - Bel(\{T\}) = 1 - Bel(\{1\}) = 1 - 0.7 = 0.3, \\
length(s, v_1) = [1 \times 0.7, 1 \times 0.9] + [2 \times 0.1, 2 \times 0.3] = [0.9, 1.5]
\]

**Edge \((s, v_2)\):**

\[
\Omega = \{1, 2\}, \quad P(\Omega) = \{\emptyset, \{1\}, \{2\}, \Omega\} \\
m(\{1\}) = 0.15, \quad m(\{2\}) = 0.8, \quad m(\{\Omega\}) = 0.05 \\
Bel(\{1\}) = 0.15, \\
Bel(\{2\}) = 0.8, \\
Pl(\{1\}) = 1 - Bel(\{T\}) = 1 - Bel(\{2\}) = 1 - 0.8 = 0.2, \\
Pl(\{2\}) = 1 - Bel(\{T\}) = 1 - Bel(\{1\}) = 1 - 0.15 = 0.85, \\
length(s, v_2) = [1 \times 0.15, 1 \times 0.2] + [2 \times 0.8, 2 \times 0.85] = [1.75, 1.9]
\]

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

\[
\Omega = \{1, 2\}, \quad P(\Omega) = \{\emptyset, \{1\}, \{2\}, \Omega\} \\
m(\{1\}) = 0.25, \quad m(\{2\}) = 0.65, \quad m(\{\Omega\}) = 0.1 \\
Bel(\{1\}) = 0.25, \\
Bel(\{2\}) = 0.65, \\
Pl(\{1\}) = 1 - Bel(\{T\}) = 1 - Bel(\{2\}) = 1 - 0.65 = 0.35, \\
Pl(\{2\}) = 1 - Bel(\{T\}) = 1 - Bel(\{1\}) = 1 - 0.25 = 0.75, \\
length(v_1, t) = [1 \times 0.25, 1 \times 0.35] + [2 \times 0.65, 2 \times 0.75] = [1.55, 1.85]
\]
obtained paths. To compare two paths we have to decide based on the length intervals of its constituent edges:

Calculating paths lengths. We can compute the length interval of a path based on the obtained length intervals of its constituent edges:

\[\text{path}_1 = \{0.9, 1.5\} + \{1.55, 1.85\} = \{2.45, 3.35\}\]

\[\text{path}_2 = \{1.75, 1.9\} + \{1.1, 1.7\} = \{2.85, 3.6\}\]

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 [117, 151] which compare the central points of two intervals. The centres of \(\text{path}_1\) and \(\text{path}_2\) 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. We now consider the shortest path problem in networks where the existence of edges is uncertain. Figure 2.6 shows a similar graph to that which we utilised in Section 2.3.5. In Figure 2.6a, imperfect information about the existence of edges (i.e., uncertainty) is modelled with probability values and in Figure 2.6b 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 equal values on corresponding edges. We note that according to Liu [98], probability values are retrieved from the results of sequential experiments and "uncertain measure" is just "occurrence possibility" of an event obtained by the judgement of experts.

Similar to possible worlds semantics, an uncertain graph (where existence of edges is modelled with uncertainty measures) corresponds to \(2^m\) determin-
Figure 2.6. Uncertainty about the existence of edges is modelled with (a) probability values, (b) uncertain measures in uncertainty theory.


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 realisation.

**Probability theory (Figure 2.6a):** All possible worlds corresponding to the probabilistic network in Figure 2.6a are depicted in Figure 2.7. 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.

**Uncertainty theory (Figure 2.6b):** To retrieve the uncertain measure \( M \) of each path under uncertainty theory, we use the theorem that has been described in [54, 56]:

**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 a Boolean uncertain variable such that

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

where \( B_i \) is a subset of \( \{0, 1\}, i = 1, 2, \ldots, n \).

Assume that \( P_1 \) is a function denoting that 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 \( s - v_2 - t \) as the shortest path between \( s \) and \( t \) and \( P_{1,2} \) indicates that both paths are available.

In the network in Figure 2.6b, each edge is associated with a Boolean uncertain variable, and these variables are independent from each other. The values on edges in Figure 2.6b indicate \( M\{\xi_i = 1\} \). Following Duality Axiom, \( M\{\xi_i = 0\} = 1 - M\{\xi_i = 1\} \). Based on Theorem 1, uncertain measure
(a) $p = 0.018$   (b) $p = 0.072$   (c) $p = 0.012$   (d) $p = 0.042$

(e) $p = 0.018$   (f) $p = 0.048$   (g) $p = 0.168$   (h) $p = 0.072$

(i) $p = 0.028$   (j) $p = 0.012$   (k) $p = 0.042$   (l) $p = 0.112$

(m) $p = 0.048$   (n) $p = 0.168$   (o) $p = 0.028$   (p) $p = 0.112$

Figure 2.7. (a)-(p) All possible worlds of the graph presented in Figure 2.6a.
(a) $\mathcal{M}\{R_1\} = 0.2$  
(b) $\mathcal{M}\{R_2\} = 0.3$  
(c) $\mathcal{M}\{R_3\} = 0.2$  
(d) $\mathcal{M}\{R_4\} = 0.2$

Figure 2.8. (a)-(p) All realisations of the graph presented in Figure 2.6b.
of $P_1$, $P_2$ and $P_{1,2}$ can be computed 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 2.8g, 2.8l and 2.8m are the realisations for which $P_1$ holds. In the following we calculate $\mathcal{M}\{P_1 = 1\}$:

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

$$= \sup_{R7, R12, R14} \min_{1 \leq i \leq 4} \mathcal{M}\{\xi_i \in B_i\}$$

$$= \sup (\min\{0.8, 0.7, 0.6, 0.5\}, \min\{0.8, 0.7, 0.4, 0.5\}, \min\{0.8, 0.7, 0.6, 0.5\})$$

$$= \max \{0.5, 0.4, 0.5\}$$

$$= 0.5$$

Following a 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 produce similar orderings, 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.
3. From Systems to Network Models

In this chapter we look at the process to model a system with probabilistic networks, by focusing on selected datasets. While all the other datasets reviewed in this chapter have been introduced in published scientific articles, the brain fMRI imaging dataset originally studied for the first time in this thesis as a system that can be modelled appropriately with the probabilistic network model, $G = (V, E, p)$.

3.1 Uncertainty in fMRI Functional Networks

Analysis of functional neuroimaging data has attained extensive attention during the last decade. Among all functional imaging methods, resting state functional Magnetic Resonance Imaging (fMRI) has had a prominent role not only in identifying the intrinsic organisation of the brain, but also in detecting the changes caused by psychiatric disorders. One way of analysing fMRI data is to transform it into a brain network [85].

In the current approaches of modelling and analysis of resting state functional brain networks, similarity between different regions of the brain based on their activation during the scanning period is measured. If the similarity value between two regions is larger than or equal to a given threshold, then those two regions are assumed to be connected in the resulting functional brain network. Otherwise, those two regions are not connected.

While it may sound intuitive to represent brain activity using a network, it is important to remember that these networks are the result of a complex data transformation process. Therefore, there is a risk that what we observe in a network is partly an artefact of the design choices behind the generation process, and thus not an accurate representation of the underlying system under study.

In the technical report in [76], we experimentally have quantified the impact of some design choices on various features of the resulting networks. Figure 3.1 shows the preprocessing pipeline through which fMRI data is transformed into a network. This pipeline includes five stages that we concisely describe in the following.

Modelling Pipeline

1. Imaging: fMRI is an imaging method that measures the changes in blood flow in the brain. The blood flow in the brain creates some signals which are referred to as Blood-Oxygen-Level-Dependent (BOLD)
signals. Functional imaging can be performed in two manners. The first is to stimulate the brain regional interactions by giving some task to subjects to be performed during the scanning [8]. These are called task-based fMRI. The second way of performing fMRI is called resting state fMRI in which the subjects’ brains are in a task-free state meaning that no specific task is assigned to them during the brain scanning.

2. **Preprocessing:** fMRI signals inherently suffer from several noise sources, such as head movement, cardiac/respiratory pulsation or scanner-induced artefacts [7]. Therefore, several preprocessing stages are required to alleviate the effects of noise in fMRI signals. Preprocessing steps can be divided into two groups. First, the group of standard preprocessing steps on which there is a consensus among scholars such as slice timing correction, head motion correction, (magnetic field) distortion correction [51, 134], registration, spatial normalisation and spatial smoothing [61, 160]. The second group includes preprocessing steps which are controversial and there is no agreement on applying or not applying them. This group includes filtering and global signal regression (GSR).

3. **Parcellating:** the resolution of neuroimages with the current available scanning machines is at least $1\text{mm} \times 1\text{mm} \times 1\text{mm}$. Thus we need a method to aggregate cubes of voxels\(^1\) based on some criterion to find out the regions that can be seen as a single unit. This processing step is called parcellating.

4. **Associating:** this is the forth step of the pipeline which evaluates the similarity between two regions’ timeseries [35]. There are multiple methods to find associations between two parcels such as Pearson correlation coefficient, mutual information [20] and phase coherence [21]. The most applied method is calculating the absolute value of Pearson correlation coefficient of two timeseries which is a value between 0 and 1. The higher correlation value between a pair of parcels indicates the stronger functional connectivity between those regions.

---

\(^1\)A 3-dimensional fMRI image is made of units called voxels. Each voxel represents a small cube of brain tissue.
5. **Thresholding:** the next step is thresholding which acts as a filtering step, filtering out the correlation values that are lower than a threshold and passing those values that are higher and represent significant functional connectivity. The survived correlations are then set to 1. There has been a long debate about selecting an appropriate threshold value in the literature [35].

The output of this modelling pipeline, is an $N \times N$ matrix\(^2\) with values 1 and 0, and defines the adjacency matrix of a brain network with $N$ nodes.

Among all steps of the modelling pipeline, the effect of thresholding on the resulting networks has been studied the most. Yet, many studies do not account for the uncertainty due to preprocessing and imaging time length. In our exploratory study in [76] we have examined these two stages that mostly affect our understanding about the existence of functional correlations between the brain regions.

In our experiments we have addressed three questions:

1. How can the imaging time length and preprocessing stages affect our understanding about the existence of functional correlation between two parcels?
2. How can the aforementioned stages affect our understanding about the existence of functional correlation in the whole brain network?
3. How do they change our understanding about the difference between the brain networks of a sick and a control case?

In summary in this exploratory analysis we showed that choosing different preprocessing parameters and imaging time lengths can lead to completely different networks modelling the brain data. As a consequence, producing a single deterministic network as the end of the pipeline only provides a partial representation of the information available in the fMRI data. A possible direction that we believe is worth exploring to address this problem is to use probabilistic network models instead of deterministic ones to represent uncertainty in functional interactions between brain regions in fMRI data.

### 3.2 Uncertainty in Social Networks

In this section, we review some social network datasets that have been modelled as probabilistic networks, showing how probabilities have been assigned.

#### 3.2.1 Coauthorship networks

Science coauthorship datasets, such as the computer science bibliography DBLP dataset, have been modelled as probabilistic networks. In this network, nodes

\(^2N\) is the number of parcels in the selected parcellating scheme.
are authors of papers and two authors have a probabilistic edge if they have co-authored at least one paper. The probabilities of the edges are obtained from exponential function $p_{ij} = 1 - \exp(-\mu n)$ determining the probability that two authors will co-author a paper in the future. $n$ is the number of papers that two authors have co-authored in the past and $\mu$ is the scaling factor [122].

3.2.2 Digital communication networks

The Enron email network is another dataset that has been modelled as a probabilistic network. Nodes in the Enron dataset represent employees and there is an edge between two nodes if at least one email has been exchanged between them. Edge probabilities quantify the likelihood that a new email will be exchanged between a pair of nodes at time $t$, $p_{i,j} = 1 - \prod_k (1 - \exp(-\mu (t - t_k)))$. $\mu$ is the scaling parameter, and $t_k$ is the time when message $k$ has been exchanged between nodes $i$ and $j$ [126].

Wall-to-wall postings between users in Facebook and other social media sites have also been modelled as probabilistic networks. There is an edge between two nodes if at least one of them has posted at least one message on another person’s wall. The probabilities on the edges come from the same equation in the Enron dataset and represent the likelihood of having an active relationship at time $t_{now}$ between those two users [105, 126].
4. From Network Models to Insights: Algorithms

In this chapter we review the literature on methods and algorithms to analyse probabilistic network models \( G = (V, E, p) \) where \( p : E \rightarrow (0, 1] \) assigns independent existence probability to each edge \( e \in E \). The term query refers to the analysis task or the research question that has been asked about a network.

4.1 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 measures and notions are defined based on it [49, 156, 168]. 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 queries.

4.1.1 Shortest Path Length Distribution

The shortest path length between any pairs of nodes in a probabilistic network is expressed as a shortest path length distribution [129]. 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 \( l \):

\[
sp_{s,d}(l) = \sum_{G|D(s,d)=l} Pr(G) \tag{4.1}
\]

where \( G \) is a possible world of probabilistic graph \( G \). Equation 4.1 gives the exact probability of shortest paths with different length, however it is computationally prohibitive to directly compute this formula for large probabilistic networks. Therefore, approximating shortest path length distributions 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 distributions: median distance, majority distance and expected reliable distance [129]. They have used these three distance functions to explore the \( K \)-nearest neighbours of nodes in a probabilistic network.

47
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 approximate the probability of each candidate path [173]. Yuan et al. propose a method that returns the shortest paths between two nodes whose probability is higher than a specific threshold [161]. The authors in [71] 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 $d$.

4.1.2 Reliability

The second fundamental path based problem in probabilistic networks is the 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 [139], the probability that all pairs of nodes in a subset of all nodes are connected [65] and the probability that there is at least one path between a node and a set of nodes [78].

Zhu et al. in [171] 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 the 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 [78] 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 hierarchically cluster nodes, called $RQ$-tree, and an online sampling method which means that an edge is sampled in the current possible world if it is requested. The authors in [79] 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 maximise the reliability between two given nodes. Lin et al. in [92] 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 maximises the precision of approximation of reliability in probabilistic networks. The authors in [77] 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 [137]. 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 Pattern Extraction Queries

4.2.1 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 [10]. 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 the notions of $\eta$-degree and $(k, \eta)$-core [15]. $\eta$-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 utilised a sampling method to approximate it. The utilised sampling method is called sample finite population with unequal probabilities which has been introduced in [23].

4.2.2 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. [177] 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 $F$ 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 set 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 [113, 114]. 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 is higher than or equal to $\alpha$ ($\alpha$-clique).

While [177] has focused on finding top-$k$ maximal cliques whose size is higher than a threshold and [113, 114] have investigated the problem of finding all maximal cliques whose probability is higher than a threshold, Li et al. [90] 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.

4.2.3 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 extract the patterns whose probability of existence is higher than a specific given threshold.

According to our knowledge, Zou et al. [176] 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 [175]. As reliability is a fundamental issue in probabilistic graphs, Jin et al. [70] 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 cannot 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 [174]. He has shown that this problem is solvable in polynomial time by extending Goldberg’s Algorithm [41].

4.2.4 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 return true if $q$’s probability is higher than a specific threshold.

Yuan et al. have studied the problem of finding a subgraph in a set of probabilistic networks notated as $D = \{G_1, G_2, \ldots, G_n\}$ [163, 165] 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. [154] 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 [162] generalised the definition of match in deterministic graph pattern matching. Their proposed algorithm takes as input: a deterministic graph $G$, a query graph $q (\{v_1, v_2, \ldots, v_n\})$, both with labelled nodes, and the shortest path threshold $\gamma$. Then a set of $n$ nodes in $G$, $\{u_1, u_2, \ldots, u_n\}$ are called match of $q$ if and only if (1) node labels in both sets are equal, $L(u_i) = L(v_i)$ for $i \in \{1, 2, \ldots, n\}$ 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 a pattern matching problem in probabilistic networks: given a probabilistic graph $G$, deterministic query graph $q$, distance threshold $\gamma$ and probability threshold $\eta$, this problem consists in retrieving all node sets in $G$ whose probability to match with $q$ (with regard to $\gamma$) is higher than or equal to $\eta$.

4.3 Clustering

The objective of network clustering is to divide a network’s nodes into a number of groups, called clusters, such that the nodes in the same cluster are more similar to each other than to nodes in other clusters. The similarity measure is often 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 the clustering problem in probabilistic networks.

Liu et al. in [99] have defined two desired qualities that have to be considered in the process of clustering: first, for each connected component of each possible world of $G$ the number of distinct clusters to which the different nodes
belong should 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 \( F_p \) and the entropy of cluster size notated as \( F_e \). As the objective function they minimise \( F = F_p - F_e \) and to reduce the computational time complexity they use Monte Carlo sampling.

Kollios et al. have proposed a clustering method by extending the concept of edit distance [84]. Edit distance between two deterministic graphs \( G_1 = (V, E_1) \) and \( G_2 = (V, E_2) \) 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 \( G = (V, E, p) \) and a deterministic graph \( G = (V, E) \) is defined as the expected edit distance between every possible world of \( G \) and \( G \). This 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 the \( i \)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 \( G = (V, E, p) \) as finding a cluster graph whose expected distance to \( G \) is minimised. 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 [19] have considered \( k \)-clustering in probabilistic networks. Given \( k \), a \( k \)-clustering of a probabilistic network \( G = (V, E, p) \) is a partitioning of \( V \) into \( k \) clusters \( \{C_1, C_2, ..., C_k\} \) and a set of cluster centres \( \{c_1, c_2, ..., c_k\} \) where \( c_i \in C_i \). Ceccarello et al. have defined two clustering objective functions as:

\[
F_1(C) = \min_{1 \leq i \leq k} \min_{v \in C_i} Pr(v \sim c_i) \\
F_2(C) = \frac{1}{|V|} \sum_{1 \leq i \leq k} \sum_{v \in C_i} Pr(v \sim c_i)
\]

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 \( F_1 \) and \( F_2 \) in Equations 4.2 and 4.3 respectively. The first method is called \( k \)-center and the second one is called \( k \)-median clustering. Han et al. [64] have increased computational efficiency and approximation precision of the \( k \)-center and \( k \)-median clustering methods studied in [19].
Qiu et al. in [131] have studied the problem of structural clustering by defining the notions of structural similarity, reliable structural similarity and reliable neighbourhood. The structural similarity between two nodes $u$ and $v$ is the Jaccard coefficient between the set of $u$’s neighbours and itself and the set of $v$’s neighbours 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 structurally 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 neighbourhood of node $v$ is defined as the set including $v$ and the subset of $v$’s neighbours that are reliably structurally similar to $v$. Therefore, the authors define a 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 $C$ such that for each $c \in C$:

• $|c| \geq 2$

• for each node $v$ in $c \in C$ whose reliable neighbourhood size is larger than a specific given number $\mu$, all nodes that are in its reliable neighbourhood 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 neighbourhood 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 - Contributions

5.1 Paper I

Comparing Node Degrees in Probabilistic Networks

Amin Kaveh, Matteo Magnani, Christian Rohner

Journal of Complex Networks

Oxford University Press.

Summary. One of the most fundamental micro/local properties of nodes is degree. Degree of a node in a probabilistic network is represented as a probability distribution, called node’s degree distribution. In deterministic networks two nodes with the same degree will have the same rank, if ranking is based on degree centrality. However, comparing centrality of a node using the concept of degree is not straightforward in probabilistic networks. To compare degrees of two nodes $u$ and $v$ in a probabilistic network, the three events $\text{deg}(u) = \text{deg}(v)$, $\text{deg}(u) > \text{deg}(v)$ and $\text{deg}(u) < \text{deg}(v)$ are valid and their probabilities have real values such that $p[\text{deg}(u) = \text{deg}(v)] + p[\text{deg}(u) > \text{deg}(v)] + p[\text{deg}(u) < \text{deg}(v)] = 1$. To calculate these values we need to consider all possible outcomes in all possible worlds. In other words, we need to calculate nodes’ degree distributions using the possible worlds semantics. This approach makes the analysis of probabilistic networks computationally expensive even for moderate size networks. In this case, we have two solutions: either to use only summary information such as expected degree of nodes (which can be computed without calculating degree distributions), or to estimate $p[\text{deg}(u) = \text{deg}(v)]$, $p[\text{deg}(u) > \text{deg}(v)]$, and $p[\text{deg}(u) < \text{deg}(v)]$ by using other summary information (that does not require degree distributions to be computed).

In the first paper [73], we have highlighted the role of other summary information, i.e., variance and skewness of nodes degree distributions, to estimate the aforementioned probabilities.

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.

This paper addressed the first and second challenges explained in Sections 1.1.1 and 1.1.2. The second challenge has been addressed as we have proposed a method to compare the degree distribution of nodes in probabilistic networks. The first challenge has been addressed as our approximations do not require to calculate nodes’ degree distributions while would incur a high computational cost.

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

5.2 Paper II
Defining and Measuring Probabilistic Ego Networks
Amin Kaveh, Matteo Magnani, Christian Rohner
Social Network Analysis and Mining
Vol. 11, No. 1, 2021.
Springer.

Summary. Our second contribution is focused on micro level analysis. An ego network and its related measures such as ego betweenness represent the local properties of a node among its immediate neighbours. Ego network of a node is a unique and consolidated concept in deterministic networks, which is defined as a network containing the ego node, its immediate neighbours, the edges between the ego node and its neighbours and the edges between ego’s immediate neighbours. Some measures such as a node’s degree are the same in its ego network and in the whole network. Some measures such as closeness are meaningless in ego networks as a node is always directly connected to its neighbours. Some measures represent a different type of importance of a node in a network. For example, betweenness represents the extent to which a node plays an intermediate role among all pairs of nodes in a network, while ego betweenness measures the intermediation of a node among its neighbours. These two measures can be contradictory for some nodes such that while a node has a low ego betweenness, it can have a high betweenness and vice versa.

In our second paper [74] we have shown that contrary to deterministic networks in which the concept of ego network is unique, there is not only one definition for ego networks in probabilistic networks. Therefore, we have defined two possible definitions of ego network in probabilistic networks, V-Alters-Ego and F-Alters-Ego. Moreover, under these two definitions, we have studied the concept of degree, ego betweenness and ego closeness. The two definitions of ego networks and their corresponding measures are based on
the possible worlds semantics. Henceforth, calculating the probability distributions of these measures is computationally expensive. Therefore, as the second contribution of the paper, we have proposed a method to estimate ego betweenness under the V-Alters-Ego definition. This estimation directly addresses the first challenge discussed in Section 1.1.1.

**Contribution.** In summary we have two main contributions in this paper:
- We have specified two definitions of probabilistic ego networks: V-Alters-Ego and F-Alters-Ego, and investigated degree, betweenness and closeness under both definitions.
- We have proposed a method to estimate ego betweenness under the V-Alters-Ego definition.

**My Contribution.** I initiated the study and implemented and performed the experiments. All authors contributed equally in conceiving and developing the study and analysing the results.

### 5.3 Paper III

**Probabilistic network sparsification with ego betweenness**

Amin Kaveh, Matteo Magnani, Christian Rohner

Applied Network Science

Vol. 6, No. 1, 2021.

Springer

**Summary.** In our third contribution, we have revisited the problem of sparsification in probabilistic networks.

Sparsification is an approach to reduce the complexity of a network by reducing the number of edges while still maintaining the original graph’s properties. The reduction of the number of edges in probabilistic networks leads to less possible worlds, and thereby it is an approach that can reduce the computational complexity of measures and algorithms in probabilistic networks. Given a probabilistic graph $G = (V, E, p)$, a sparsification method generates an alternative probabilistic network, $G' = (V, E', p')$, with a different set of edges and probability assignment.

In the initial definition of probabilistic network sparsification, the expected degree of each node is the property that has to be preserved in the process of sparsification [121].

**Contribution.** In summary we have two main contributions in this paper:
- Sparsification is defined by Parchas et al. as the problem of preserving expected degree [121]. We generalized that definition to the problem of preserving a generic function.
• As a specific case of generic function we use expected ego betweenness that we have considered in Paper II [74], and we formulated how the change of an edge probability alters the ego betweenness of other nodes. This paper directly addresses the first challenge discussed in Section 1.1.1.

**My Contribution.** I conceived the study, performed the experiments, analysed the data, and wrote the manuscript. M.M. and C.R. analysed the results and wrote the manuscript. All authors approved the final version of the manuscript.

### 5.4 Paper IV

**Overlapping Clustering of Probabilistic Networks**

Amin Kaveh, Matteo Magnani, Christian Rohner
Under Submission

**Summary.** In our forth contribution, for the first time we have considered the problem of identifying overlapping and partial clusters in probabilistic networks with focus on the clique percolation approach. We explored three ways of applying this approach that we call CPM$_t$, CPM$_w$ and CPM$_p$. CPM$_t$ is based on the most traditional way of analysis of probabilistic networks, thresholding, that removes all edges below an input threshold and then looks at the probabilistic network as a deterministic one. CPM$_w$ is based on another traditional way of analysis of probabilistic networks, and considers edges’ existence probabilities as edges’ weights. Our proposed method, CPM$_p$, is an extension of both CPM$_t$ and CPM$_w$.

**Contribution.** In summary we have three main contributions in this paper:

- We introduced an extension of the clique percolation method to identify overlapping clusters in probabilistic networks.
- We showed how this method can identify ground-truth clusters that would be lost without considering the uncertainty on the existence of the edges.
- We experimentally showed that CPM$_w$ can be easily affected just by adding low-probability false positive edges, and CPM$_t$ suffers from the presence of both false positive and false negative edges, while CPM$_p$ would be able to perform well using the right thresholds.

This paper has addressed the third challenge discussed in Section 1.1.3.

**My Contribution.** I initiated the study and implemented the code. All authors contributed equally in conceiving and developing the study and analysing the results.
6. Summary and Future Work

The idea of including uncertainty about systems into models and analysing those models has gained attention in recent years. The most commonly applied theory to model and analyse imperfect network data has been probability theory. The main reason behind this broad application is its formalised mathematical axioms that have been consolidated over the last few centuries. Among all models, the probabilistic network model $G(V, E, p)$ in which each edge is associated with an independent existence probability is compliant with a wide variety of real world scenarios such as wireless sensor networks, protein-protein interaction networks, opportunistic networks, road networks and so on.

In this model, a network with $m$ probabilistic 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, each measure is represented as a probability distribution over all possible worlds that leads to three challenges:

1. As the number of possible worlds is exponential in the number of edges, exact calculation of probability distribution of measures on this model is computationally expensive.
2. The most of current algorithms and methods in network science cannot deal with measures that are represented as probability distributions.
3. Aggregating information yielding from the analysis of possible worlds is not straightforward.

This thesis and the included papers, first have highlighted these challenges in this research field and second have addressed them in specific scenarios and settings.

First this thesis has scrutinised fundamental local measures in probabilistic networks. We have shown that to compare two nodes in probabilistic networks we need to know their degree distributions which is computationally expensive. On the other hand, as a node’s degree distribution in a probabilistic network is a Poisson-binomial distribution, it is possible to calculate its summary information such as expected value, variance and skewness without having the distribution itself. Therefore, we proposed approximation functions to compare degree distributions based on their expected value, variance and skewness. As a result, we showed that nodes with the same expected degree can have different ranking in a probabilistic network, while it was unrevealed without our proposed measures.

In addition, this thesis has focused on the definition of ego networks in probabilistic networks. We have shown that contrary to deterministic networks
in which the concept of ego network is unique, there is not only one definition for ego networks in probabilistic networks. Therefore, we have defined two possible definitions of ego networks in probabilistic networks, V-Alters-Ego and F-Alters-Ego. Moreover, we have proposed a method to estimate ego betweenness under the V-Alters-Ego definition.

While the aforementioned estimation functions address Challenge 1, this thesis has addressed this challenge from another angle. Particularly, we have revisited the problem of sparsification in probabilistic networks. In more detail, first we have generalised the definition of probabilistic network sparsification such that any generic function can be preserved. Then, we have applied the estimated ego betweenness (introduced in this thesis) as a specific case of generic function and have formulated how the change of an edge probability alters the ego betweenness of other nodes. We have shown that instead of sampling the original probabilistic graph and calculating measures over those samples, we can sample a smaller number of possible worlds from the sparsified graph and calculate the measures over them. The latter is an estimation of the first with low errors.

Furthermore, in this thesis we have explored the problem of identifying overlapping and partial clusters in probabilistic networks with focus on the clique percolation approach. The three ways of applying this approach are called CPM$_t$, CPM$_w$ and CPM$_p$.

CPM$_t$, that is based on the most traditional way of analysis of probabilistic networks, removes all edges below an input threshold and then looks for deterministic $k$-cliques in the resulting network. A potential pitfall of this method is that this edge-based filter may not be able to discriminate between false positive edges (the edges between clusters that get some low probability but do not exist) and false negative edges (the edges in the deterministic ground-truth graph that also get a low probability).

CPM$_w$ is the weighted version of the clique percolation method and considers edges’ existence probabilities as edges’ weights and is not considering the existential semantics of the probabilities. This method only applies filters to $k$–cliques. A potential problem with this approach is that it cannot apply filters at lower (e.g., edge) level, so false positive edges across clusters risk to result in the clusters being erroneously merged, even when they have a low probability.

CPM$_p$ is an extension of CPM$_t$. CPM$_p$ performs exactly the same thresholding operation that CPM$_t$ does on edges. However, after this process CPM$_t$ considers the resulting graph as a deterministic graph and looks for deterministic $k$–cliques while CPM$_p$ has this advantage that can filter 3-cliques to $k$-cliques based on their existence probabilities. CPM$_p$ is also an extension of CPM$_w$. Given $k$, CPM$_w$ considers the intensity of $k$-cliques: if they are higher than a given threshold, they are considered acceptable, otherwise they are rejected. CPM$_p$ performs the same process when considering $k$-cliques, however at the same time considers the existence probability of constituent cliques, i.e.,
(k − 1)-clique, (k − 2)-clique, . . . , down to 2−cliques (edges). Having multiple thresholds allows CPM\textsubscript{p} to accept or reject cliques of different sizes based on different thresholds.

Our results show that on the one hand CPM\textsubscript{w} can be easily affected just by adding low-probability false positive edges, where CPM\textsubscript{r} and CPM\textsubscript{p} would be able to filter them out using the right thresholds. On the other hand, CPM\textsubscript{r} suffers from the presence of both false positive and false negative edges, confirming its inability to discriminate between them using a single threshold. CPM\textsubscript{p}, being an extension of both CPM\textsubscript{r} and CPM\textsubscript{w}, can be executed with settings that make it perform well under both scenarios.

**Future work.** This thesis can be extended in some directions. However, I only mention some of the potential extensions.

The initial probabilistic network sparsification method in [121], takes degree distribution of each node as input and generates a new degree distribution with the same (or as close as possible) mean and lower variance for that node. In our first paper, we have shown that not only the mean and variance of degree distribution is effective in ranking of nodes, but also the skewness of degree distribution is important [73]. Therefore, one potential approach to extend the sparsification method in [121] is to generate a new degree distribution with the same (or as close as possible) mean and skewness and lower variance for each given node.

The current algorithms in [75], called the gradient descent algorithm (GD) and the expectation maximisation algorithm (EM), run several iterations and in each iteration parameters (edge probabilities) are updated one by one. This approach of updating parameters has two drawbacks: (1) in each iteration all parameters become updated regardless of which one has higher priority to be updated, (2) the order of parameters to be updated in each iteration is the same. One potential extension of these algorithms to accelerate the sparsification process is to update the parameters (edge probabilities) that can affect the top-k discrepancy values. This seems logical, as each parameter does not affect the loss for all data points (discrepancies of all nodes).

The available methods to identify overlapping clusters in deterministic networks can be divided into four categories known as *clique percolation* [120], **line graphs partitioning** [125], **local expansion** [87] and **label propagation** [6,62,159]. Paper IV is the extension of the first category, the clique percolation method, for probabilistic networks. One potential extension of this thesis is to extend the other categories. In particular, in an ongoing research we are extending the line graph partitioning method for probabilistic networks.

In line graph partitioning method, first we transform an input probabilistic graph \( G \) to a line graph \( L(G) \). Then, we partition the line graph \( L(G) \) that leads to overlapping clustering in the original graph. In our ongoing research, as the first step we have provided a definition for probabilistic line graph:
**Definition 1 - probabilistic line graph.**
given a probabilistic graph $\mathcal{G} = (V, E, p)$, a probabilistic line graph $L(\mathcal{G}) = (V_L, E_L, p_L)$ is a graph whose nodes represent edges in $\mathcal{G}$ ($|V_L| = |E|$), edges’ existence probabilities in $\mathcal{G}$ are associated with nodes in $L(\mathcal{G})$ as nodes existence probabilities, and there is an edge between nodes $u$ and $v$ in $L(\mathcal{G})$ if the edges $u$ and $v$ are incident to a common node in $\mathcal{G}$.

As the second step we have defined the probabilistic line graph partitioning method as follows:

**Definition 2 - probabilistic line graph partitioning.**
Given a probabilistic line graph $L(\mathcal{G})$, find the partitioning of $L(\mathcal{G})$ such that the expected number of edges between partitions (inter-partition) are minimised and the expected number of edges inside partitions (intra-partition) are maximised.

In order to optimise the objective function described in Definition 2 we need to formulate the expected number of edges incident to each node (expected degree) in a probabilistic line graph, $L(\mathcal{G})$. Therefore, we have defined and mathematically proved that the expected degree of a node in a probabilistic line graph is as follows:

**Definition 3 - expected degree of a node in a probabilistic line graph.**
Given a probabilistic line graph $L(\mathcal{G})$, the expected degree of node $u$ is:

$$E(u) = p_u \cdot \sum_{v \in N_u} p_v \quad (6.1)$$

where $p_u$ is the existence probability of node $u$, and $N_u$ is the set neighbours of $u$ in $L(\mathcal{G})$.

Therefore, the objective function described in Definition 2 has been formulated as follows:

**Definition 4 - objective function of probabilistic line graph partitioning.**
Given a probabilistic line graph $L(\mathcal{G})$, the objective function of line graph partitioning is defined as:

$$\sum_{(u,v) \in E, \mathcal{P}(v) = \mathcal{P}(u)} (1 - p_u p_v) + \sum_{(u,v) \in E, \mathcal{P}(v) \neq \mathcal{P}(u)} p_u p_v \quad (6.2)$$

This objective function is similar to the objective function in correlation clustering defined in [60]. Therefore, correlation partitioning methods can be applied or adapted to partition probabilistic line graphs. In our current ongoing research we are working on extending such algorithms to partition probabilistic line graphs. To the best of our knowledge, the other two overlapping
clustering categories, *local expansion* and *label propagation*, have not been studied for probabilistic networks so far. One additional potential extension of this thesis is to extend these two categories.

The subject of overlapping clustering in probabilistic networks can be extended from another angle as well. Particularly, when we are not even sure about the existence of an edge between each pair of nodes (representing it as a probability), how can we conclude with 100% confidence that a node is a member of one cluster and not member of any other communities? This means that we have to go one step further from the proposed method in Paper IV in which if a node is identified as a member of one or multiple clusters, this membership is represented with 100% confidence.

As it has been mentioned before, the idea of including uncertainty about systems into models and analysing those models has gained attention recently. However, to the best of our knowledge the majority of the research has been devoted to analysis of models, with less attention on the sources and the ways of modelling data as uncertain models. This thesis in Chapter 2 has provided a terminology and briefly has introduced the theories that have been used in the analysis of imperfect network data. However, we think that there has not been a sufficient number of studies on identifying and modelling the different kinds of imperfection in real datasets. This is evidenced by the fact that there are only a few real datasets to be analysed in this research field.

Henceforth, as a case study we have studied the uncertainty in functional networks of the brain fMRI imaging data that has been summarised in Chapter 3. However, we suggest that the definition of a data preparation process and exploratory analysis that first reveals the uncertainty in data and second includes this uncertainty in models is an open challenge for many application areas.
Forskare inom många ämnen har behov av att förstå system och deras beteende. För att göra det kan många system modelleras som nätverk, det vill säga en samling noder som ansluter till varandra i kanter. Nätverksvetenskap är ett akademiskt område som handlar om att modellera system som nätverk och analysera dessa nätverk för att förstå systemets struktur och funktion och i vilken utsträckning dess struktur påverkar dess funktion.

Det underförstådda antagandet som ligger till grund för ett antal modeller är att insamlad data från system är tillräckligt detaljerad för att stödja den analys som görs. Emellertid är empirisk datainsamling ofta en ofullkomlig process som påverkas av olika saker. Exempel kan vara bullriga mätningar, t.ex. i biologiska experiment, eller på grund av saknad information och indirekta mätningar, som i fallet när vi härleder sociala band eller påverkar relationer mellan individer baserat på deras interaktioner. Som ett annat exempel, i sociala nätverk, innefattar insamlade interaktioner vissa missförstånd på grund av informanternas glömska.

Det mest använda tillvägagångssättet inom nätverksvetenskap består i att ta bort all slags osäkerhet från data och följaktligen i att modellera system som deterministiska nätverk. Alla utvecklade algoritmer och metoder är lämpliga för deterministiska nätverk. Antagandet om att ta bort osäkerhet innan modellering har dock ifrågasatts de senaste åren och därför har modeller som kan representera osäkerhet föreslagits.


Sannolikhetsteori och semantik för möjliga världar säger oss att om ett nätverk har $m$ probabilistiska kanter måste vi analysera det $2^m$ gånger och vår resulterande analys representeras som en sannolikhetsfördelning. Detta leder till tre utmaningar:

1. Om antalet kanter av ett nätverk ($m$) är stort, är en fullständig analys av det nätverket mycket beräkningsmässigt kostsam.
2. Även om vi kan hitta ett sätt att analysera ett probabilistiskt nätverk med en acceptabel beräkningskomplexitet, representeras analysresultaten som en sannolikhetsfördelning, inte ett enda tal. De flesta av de nuvarande
algoritmerna och metoderna inom nätverksvetenskap kan endast hantera enskilda tal och kan inte hantera mått som representeras som sannolikhetsfördelningar. Det betyder att vi behöver utöka många algoritmer och metoder till att omfatta dessa nätverk.

3. Även om vi har metoder och algoritmer som först kan analysera probabilistiska nätverk på ett tidseffektivt sätt och också är kapabla att hantera sannolikhetsfördelningar, är det inte okomplicerat att aggregera den information som produceras av denna analys.

Den här avhandlingen och de inkluderade artiklarna har för det första belyst dessa utmaningar inom detta forskningsfält och för det andra har de behandlat dem i specifika scenarier och miljöer.

Först har denna avhandling granskat grundläggande lokala åtgärder i probabilistiska nätverk. Vi har visat att för att jämföra två noder i probabilistiska nätverk måste vi känna till deras gradfördelningar, vilket är beräkningsmässigt kostsamt. Å andra sidan, eftersom en nods gradfördelning i ett probabilistiskt nätverk är en Poisson-binomialfördelning, är det möjligt att beräkna dess sammanfattande information såsom förväntat värde, varians och skevhet utan att ha själva fördelningen. Därför föreslog vi approximationsfunktioner för att jämföra gradfördelningar baserat på deras förväntade värde, varians och skevhet. Som ett resultat visade vi att noder med samma förväntade grad kan ha olika rangordning i ett probabilistiskt nätverk, vilket inte syns utan våra föreslagna åtgärder.


I denna avhandling har vi också undersökt problemet med att identifiera överlappande och partiella kluster i probabilistiska nätverk med fokus på klickperkolationsmetoden. Vi har testat tre metoder för att tillämpa detta tillvägagångssätt.

Den första metoden, som är baserad på det mest traditionella sättet att analysera probabilistiska nätverk, tar bort alla kanter under en ingångströskel och letar sedan efter deterministiska klickar i det resulterande nätverket. En poten-
tiell fallgrop med denna metod är att detta kantbaserade filter kanske inte kan skilja mellan falskt positiva kanter (kanterna mellan kluster som får en viss låg sannolikhet men som inte finns) och falska negativa kanter (kanterna i den deterministiska marken- sanningsgraf som också får en låg sannolikhet).

Den andra är den viktade versionen av klickperkolationsmetoden och betraktar kanternas existenssannolikheter som kanternas vikter och tar inte hänsyn till sannolikheternas existentiella semantik. Denna metod tillämpar bara filter på de största klickarna. Ett potentiellt problem med detta tillvägagångssätt är att det inte kan tillämpa filter på lägre (t.ex. kant) nivå, så falska positiva kanter som korsar flera kluster riskerar att resultera i att klustren felaktigt slås samman, även när de har låg sannolikhet.


Våra resultat visar att vår föreslagna metod överträffar de två första metoderna för att identifiera marksanningsgemenskaper när ett nätverk har mätts i en bullrig miljö och inkluderar falskt positiva och falskt negativa kanter.
8. Acknowledgements

I would like to express my gratitude to all the people without whom I would not have been able to complete this research, and without whom I would not have made it through:

To my main supervisor Matteo Magnani, thank you for your dedicated support, guidance, patience and advice. You introduced me to the vast world of uncertain networks and helped me through many difficult periods in my research. I truly appreciate all these shared years of research, studies and collaboration.
To my co-supervisor Christian Rohner, thanks for leading me through this journey. You have taught me how to become a researcher. Thank you for your consistent words of encouragement and all the emotional support during those critical moments.
I have been extremely lucky to have supervisors like you two.
To Francesco Bonchi, thank you for welcoming me to the ISI foundation. Even though the pandemic made it impossible for us to have the expected collaboration, our engaging conversations have been a great inspiration in my research.
To Onerva Korhonen from Aalto University, thank you for all the scientific discussions and your valuable inputs.

To all my colleagues in Uppsala University. All the scientific and non-scientific discussions within these 5 years has made my time more exciting.
To Davide Vega D’Aurelio, thanks for not only being my office-mate but also a good friend. Your ability to explain every experience in a fun detailed story has made the workplace a more friendly environment and our time together will never be forgotten.
To Khalid Mahmood and Georgios Kalamatianos, my great Ph.D. fellows, thanks for the friendship and the shared memories that we had.
To Pierre Flener, thanks for your support, friendly and approachable attitude, tea-breaks and biking advice.

To many inspiring researchers and teachers that I had the pleasure to meet with during my scientific journey up to this moment. Special thanks to Mohammad Khansari from University of Tehran for giving me the opportunity to do research at CITC and ITRC prior to my Ph.D.
During these years of doing my Ph.D., Sweden has become my home with the help of many friends who cared for me and the unforgettable memories that we shared. Thanks to all of you for being there when I needed it.

To Eliza and Behrooz, with you I could experience my hometown culture in Sweden. Thank you for your hospitality and I appreciate all the kindness you have towards me.

To Ghazale and Gustav, you have been with me during all the ups and downs through these years. Your friendship has been a great comfort. I thank you for all the kindness you have been giving me and the most delicious home cuisine.

To Emilia, Daniel and little Elliot, for welcoming me in your lovely home. Thank you for sharing many Christmases with me and always giving me a wonderful time.

To Natalie and Rasam, thanks for all the good times and many fun experiences. Thank you for all the barbecues, paddle and squash games that we played together.

To Stine and my three little friends Alessandro, Anna-Sofia, and Vincent, thank you all for the game nights, dinners and birthday surprises.

To Daniel and Per, thanks for being around and offering help when I needed it. Mingling with you guys made the pandemic feel less lonely.

To Chiara, thank you for all the laughter, encouragement and bikepacking trips.

To Mojtaba and Masoud, you guys bridged the distance by staying in touch with me via phone through these years. Thank you for always reminding me of all the wonderful memories of our youth.

I have come a long way thanks to my family. In fact, all my family have been supportive and I feel blessed everyday to have them.

I wish to thank my wife Alaa, you have been my best friend and a great companion ever since we met. You have loved, encouraged, supported, entertained, and helped me get through this period in the best way possible.

I deeply thank my parents, Zari and Hossein. You offered me the best life and the best education I could possibly have. Without your support I would not be the person I am today.

My parents-in-law, Nadere and Rahbar, thank you for your unfailing emotional support and timely motivation and encouragement. Thank you for being so caring and generous towards me.

To Alaleh and Elnaz, thanks a lot for being supportive, helpful, and caring. I always appreciate how you are always there for me through the bad and good times.

My sister, Fahime and her husband Soroush, you have been generous with your love and encouragement despite the long distance between us. Thanks to my adorable niece, Nikkie who always kept me motivated with her fun songs, stories and video clips.

My Aunt Taherch and Uncle Reza, I am thankful to have you supporting me.
through the years of my studies in Mashhad. You helped me find some much-needed peace these past years and I truly appreciate it. Special thanks to my cousins, Elahe and Saman, who designed the cover of this thesis. Your talent brought my imagination on paper perfectly. Last but not least, I would like to thank my late grandparents, Mamani and Babaie, who have loved me, took care of me and gave me the best memories of my childhood. All these happy memories have kept me going through the hard times and I am forever grateful.
References


[22] 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.


[61] Matthew F Glasser, Stamatios N Sotiropoulos, J Anthony Wilson, Timothy S Coalson, Bruce Fischl, Jesper L Andersson, Junqian Xu, Saad Jbabdi, Matthew Webster, Jonathan R Polimeni, et al. The minimal preprocessing pipelines for


[175] Zhaonian Zou, Hong Gao, and Jianzhong Li. Discovering frequent subgraphs


Acta Universitatis Upsaliensis

*Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 2117

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology*. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology”.)