THÈSE
(en cotutelle)

en vue de l’obtention du titre de docteur en
INFORMATIQUE DE GESTION
GÉNIE INFORMATIQUE ET AUTOMATIQUE

Social Network Analysis:
Link Prediction under The Belief Function Framework

par

SABRINE MALLEK

SOUTENUE LE 3 JUILLET 2018, DEVANT LE JURY COMPOSÉ DE:

Directeurs de thèse
ERIC LEFÈVRE, PROFESSEUR, UNIVERSITÉ D’ARTOIS, FRANCE
ZIED ELOUEDI, PROFESSEUR, UNIVERSITÉ DE TUNIS, TUNISIE
IMEN BOUKHRIS, MAÎTRE DE CONFÉRENCES (HDR), UNIVERSITÉ DE TUNIS, TUNISIE

Rapporteurs
DIDIER COQUIN, PROFESSEUR, UNIVERSITÉ SAVOIE-MONT-BLANC, FRANCE
MARIE-JEANNE LESOT, MAÎTRE DE CONFÉRENCES (HDR), UNIVERSITÉ PIERRE ET MARIE CURIE (PARIS 6), FRANCE

Examineurs
JEAN-LOUP GUILLAUME, PROFESSEUR, UNIVERSITÉ DE LA ROCHELLE, FRANCE
NAHLA BEN AMOR, PROFESSEUR, UNIVERSITÉ DE TUNIS, TUNISIE
DAVID MERCIER, MAÎTRE DE CONFÉRENCES (HDR), UNIVERSITÉ D’ARTOIS, FRANCE

Laboratoire de Génie Informatique et d’Automatique de l’Artois
Laboratoire de Recherche Opérationnelle, de Décision et de Contrôle des processus
Abstract

Social networks are large structures that depict social linkage between millions of actors. Social network analysis came out as a tool to study and monitor the patterning of such structures. One of the most important challenges in social network analysis is the link prediction problem. Link prediction investigates the potential existence of new associations among unlinked social entities. Most link prediction approaches focus on a single source of information, i.e. network topology (e.g. node neighborhood) assuming social data to be fully trustworthy. Yet, such data are usually noisy, missing and prone to observation errors causing distortions and likely inaccurate results. Thus, this thesis proposes to handle the link prediction problem under uncertainty. First, two new graph-based models for uniplex and multiplex social networks are introduced to address uncertainty in social data. The handled uncertainty appears at the links level and is represented and managed through the belief function theory framework. Next, we present eight link prediction methods using belief functions based on different sources of information in uniplex and multiplex social networks. Our proposals build upon the available information in data about the social network. We combine structural information to social circles information and node attributes along with supervised learning to predict new links. Tests are performed to validate the feasibility and the interest of our link prediction approaches compared to the ones from literature. Obtained results on social data from real-world demonstrate that our proposals are relevant and valid in the link prediction context.

Keywords: Social network analysis, evidential link prediction, belief function theory, structural information, social information, information fusion.

Résumé

Les réseaux sociaux sont de très grands systèmes permettant de représenter les interactions sociales entre les individus. L’analyse des réseaux sociaux est une collection de méthodes spécialement conçues pour examiner les aspects relationnels des structures sociales. L’un des défis les plus importants dans l’analyse de réseaux sociaux est le problème de prédiction de liens. La prédiction de liens étudie l’existence potentielle de nouvelles associations parmi des entités sociales non connectées. La plupart des approches de prédiction de liens se concentrent sur une seule source d’information, c’est-à-dire sur les aspects topologiques du réseau (par exemple le voisinage des nœuds) en supposant que les données sociales sont entièrement fiables. Pourtant, ces données sont généralement bruitées, manquantes et sujettes à des erreurs d’observation causant des distorsions et des résultats probablement erronés. Ainsi, cette thèse propose de gérer le problème de prédiction de liens sous incertitude. D’abord, deux nouveaux modèles de graphes de réseaux sociaux uniplexes et multiplexes sont introduits pour traiter l’incertitude dans les données sociales. L’incertitude traitée apparaît au niveau des liens et est représentée et gérée à travers le cadre de la théorie des fonctions de croyance. Ensuite, nous présentons huit méthodes de prédiction de liens utilisant les fonctions de croyance fondées sur différentes sources d’information dans les réseaux sociaux uniplexes et multiplexes. Nos contributions s’appuient sur les informations disponibles sur le réseau social. Nous combinons des informations structurelles aux informations des cercles sociaux et aux attributs des nœuds, ainsi que l’apprentissage supervisé pour prédire les nouveaux liens. Des tests sont effectués pour valider la faisabilité et l’intérêt de nos approches à celles de la littérature. Les résultats obtenus sur les données du monde réel démontrent que nos propositions sont pertinentes et valables dans le contexte de prédiction de liens.

Mots clés: Analyse des réseaux sociaux, prédiction des liens évidentielle, théorie des fonctions de croyance, informations structurelles, informations socio-spatiales, fusion d’information.
Acknowledgments

First and foremost, I would like to express my extreme gratitude to my supervisors and my sources of knowledge, Pr. Zied Elouedi, Pr. Eric Lefèvre and Dr. Imen Boukhris, for their continuous support and motivation. I gratefully acknowledge their assistance and patience throughout my dissertation. Their guidance and constant feedback made this PhD achievable.

I extend my thanks to all members of my thesis committee for agreeing to read the manuscript and to participate in the defense of this thesis: Pr. Jean-Loup Guillaume, Pr. Didier Coquin, Dr. Marie-Jeanne Lesot, Pr. Nahla Ben Amor and Dr. David Mercier.

I thank all the members of my laboratories LGI2A and LARODEC.

Finally, I owe thanks to my mother who not only supported me by all means but also encouraged me to chase my dreams. My sincere gratitude goes to my beloved husband who has always been there by my side and encouraged me to fight towards my goals.

I would like to thank Houda for her being the true friend of all good and bad times. Special thanks are given to my dear Saker who has witnessed the whole process of my dissertation writing, he has been incredibly supportive and patient in listening to my struggles and has always encouraged me to complete this work. It is an honor to have all of you in my life.

My appreciation goes out to all the persons who provided me support and encouragement during all the stages of this dissertation.
Contents

Introduction 1

I Theoretical aspects 5

1 Social Network Analysis 7

1.1 Introduction ................................................. 7
1.2 Social networks ............................................. 7
  1.2.1 Social networks representation ....................... 8
  1.2.2 Social networks graph structures .................... 10
  1.2.3 Social networks data .................................. 12
1.3 Social network analysis ................................. 12
  1.3.1 SNA measures .......................................... 13
  1.3.2 SNA visualization ...................................... 14
  1.3.3 SNA applications ..................................... 15
1.4 The link prediction problem .............................. 15
  1.4.1 Problem formulation ................................ 16
  1.4.2 Literature Review of Link Prediction Methods .... 17
## CONTENTS

1.5 Conclusion .......................................................... 24

2 Uncertainty in Social Networks .................................. 25
  2.1 Introduction ....................................................... 25
  2.2 Why uncertainty in social networks? ....................... 25
  2.3 SNA and the theories of reasoning under uncertainty ..... 27
  2.4 Belief function theory .......................................... 29
    2.4.1 Frame of discernment .................................... 30
    2.4.2 Basic belief assignment .................................. 30
    2.4.3 Combination rules ........................................ 32
    2.4.4 Vacuous extension ....................................... 34
    2.4.5 Multi-valued mapping .................................... 34
    2.4.6 Discounting ............................................... 35
    2.4.7 Reinforcement ............................................ 35
    2.4.8 Decision process ......................................... 36
  2.5 Belief function theory, social network analysis and link prediction ........................................ 37
  2.6 Conclusion ....................................................... 38

II Contributions ......................................................... 39

3 Link prediction in uniplex social networks ..................... 41
  3.1 Introduction .................................................... 41
  3.2 The model: Evidential Uniplex Social Network (EUSN) .................................................... 42
  3.3 Link Prediction based on Local Information (uLPLI) ................................................ 44
    3.3.1 The method: Belief Link Prediction .................... 45
    3.3.2 Illustrative example ...................................... 48
3.4 Link Prediction based on Group Information (uLPGI) ........................................... 52
3.4.1 The method: Evidential Link Prediction based on Group Information 52
3.4.2 Illustrative example ......................................................... 55

3.5 Link Prediction based on Node Attributes (uLPNA) ......................... 57
3.5.1 The method: Evidential Link Prediction Using Node Attributes .... 58
3.5.2 Illustrative example ......................................................... 61

3.6 Link Prediction based on Supervised Learning (uLPSL) ................. 63
3.6.1 The method: Evidential k-nearest neighbors Link Prediction .... 63
3.6.2 Illustrative example ......................................................... 66

3.7 Overview and theoretical comparison ........................................ 68

3.8 Conclusion ............................................................................ 69

4 Link Prediction in Uniplex Social Networks: Experimental Results 71
4.1 Introduction .......................................................... 71
4.2 Evidential social network creation ........................................... 72
4.2.1 Graphs generation ......................................................... 72
4.2.2 Mass functions simulation ............................................. 72
4.3 Evaluation criteria .......................................................... 73
4.4 Link Prediction based on Local Information .............................. 73
4.4.1 Network pre-processing .................................................. 74
4.4.2 Link prediction task ....................................................... 74
4.5 Link Prediction based on Group Information .............................. 75
4.5.1 Network pre-processing .................................................. 75
4.5.2 Link prediction process .................................................. 76
4.6 Link Prediction based on Node Attributes ............................... 77
## CONTENTS

4.6.1 Network pre-processing ........................................ 77  
4.6.2 Link prediction process .................................... 77  
4.7 Link Prediction based on Supervised Learning .............. 78  
4.7.1 Pre-processing ............................................. 78  
4.7.2 Link prediction ............................................ 79  
4.8 Conclusion ..................................................... 81  

5 Link Prediction in Multiplex Social Networks ................. 83  
5.1 Introduction ................................................ 83  
5.2 The model: Evidential Multiplex Social Network (EMSN) .......... 84  
5.3 Multiplex Link Prediction based on Local Information (mLPLI) ........ 86  
5.3.1 The method: Multi-relational Belief Link Prediction .......... 86  
5.3.2 Illustrative example .................................... 89  
5.4 Multiplex Link Prediction based on Group Information (mLPGI) ........ 91  
5.4.1 The method: Evidential multi-relational Link Prediction Using Group Information ........ 91  
5.4.2 Illustrative example .................................... 94  
5.5 Multiplex Link Prediction based on Node Attributes (mLPNA) ........ 98  
5.5.1 The method: Evidential Multi-relational Link Prediction Using Node Attributes ........ 98  
5.5.2 Illustrative example .................................... 102  
5.6 Multiplex Link Prediction based on Supervised Learning (mLPSL) ........ 103  
5.6.1 The method: Evidential k-nearest neighbors for multi-relational Link Prediction ........ 104  
5.6.2 Illustrative example .................................... 107  
5.7 Overview and theoretical comparison .......................... 109  
5.8 Conclusion ..................................................... 110
6 Link Prediction in Multiplex Social Networks: Experimental Results 111

6.1 Introduction ................................................. 111

6.2 Multiplex Link Prediction based on Local Information ................. 112
   6.2.1 Network pre-processing .................................... 112
   6.2.2 Experimental study ........................................ 112

6.3 Multiplex Link Prediction based on Group Information ................. 114
   6.3.1 Network pre-processing .................................... 114
   6.3.2 Experimental study ........................................ 114

6.4 Multiplex Link Prediction based on Node Attributes .................... 115
   6.4.1 Network pre-processing .................................... 116
   6.4.2 Experimental study ........................................ 116

6.5 Multiplex Link Prediction based on Supervised Learning ............... 117
   6.5.1 Data description .......................................... 117
   6.5.2 Experimental evaluation .................................... 117

6.6 Conclusion .................................................... 121

Conclusion .............................. 122

References .................................................... 127
List of Figures

1.1 (a) Social network represented as a graph (nodes are labeled by letters) (b) Adjacency list (c) Adjacency matrix ................................................................. 9
1.2 Uniplex social network of persons connected via one type relationship ........ 10
1.3 Multiplex social network of persons connected via three possible relationships 11
1.4 Heterogeneous social network of authors and papers connected via co-authorship and citation relations ................................................................. 11
3.1 An evidential uniplex social network graph with bba’s weighted edges ........ 44
3.2 Snapshots of an evidential uniplex social network in given state (a) $G(V, E)$ and in a different state (b) $G(V, E')$ ................................................................. 49
3.3 A uniplex evidential social network with overlapping social circles $C_1$, $C_2$ and $C_3$ ........................................................................................................ 56
3.4 A uniplex evidential social network, nodes attributes values are given in data. 61
3.5 A uniplex social network graph ................................................................. 67
4.1 Precision and recall of uLPLI and $CN$ on the temporal Facebook friendships network ................................................................. 75
4.2 Precision and recall results of uLPGI and $CNG$ on Facebook friendships . 76
4.3 Precision and recall of uLPNA, uLPLI and $CN$ on the Facebook network . 78
LIST OF FIGURES

4.4 Accuracy results of uLPSL and KNN according to the number of $k$ nearest neighbors .............................................................. 79
4.5 Precision results of uLPSL according to the number of negative links . . . 80
5.1 A multiplex social network graph with bba’s weighted edges .......... 85
5.2 Common neighbors’ subgraphs in each layer .............................. 89
5.3 Evidential multiplex social network graph with social circles .......... 95
5.4 Evidential multiplex social network graph .............................. 102
5.5 A multiplex social network graph ........................................ 108
6.1 Precision and recall results of mLPLI and uLPLI on students cooperation and relationships networks ..................................................... 113
6.2 Precision and recall results of mLPGI and uLPGI ...................... 115
6.3 Precision and recall results of mLPNA and uLPNA .................... 116
6.4 Results in terms of recall of mLPSL according to the number of $k$-nearest neighbors on SC and AUCS ........................................ 118
6.5 Recall results of mLPSL and baseline uniplex and multiplex similarity scores on SC and AUCS networks ........................................... 119
6.6 Class imbalance test of mLPSL and baseline methods on the SC network in terms of precision ........................................ 120
6.7 Class imbalance test of mLPSL and baseline methods on the AUCS network in terms of precision ........................................ 120
## List of Tables

1.1 Local similarity extended to multiplex networks between a pair of nodes \((u, v)\), where \(\tau(u)\) and \(\tau(v)\) are respectively their sets of neighbors across all network layers ................................................................. 19

3.1 Mass distributions of the links of the EUSN \(G(V, E)\) of Figure 3.2(a) .... 49
3.2 Vacuous extension ................................................................. 50
3.3 Conjunctive combination \(m^{N_c}\) .............................................. 50
3.4 Vacuous extension ................................................................. 50
3.5 Conjunctive combination \(m^{N_d}\) .............................................. 50
3.6 Vacuous extension ................................................................. 51
3.7 Conjunctive combination \(m^{N_e}\) .............................................. 51
3.8 Conjunctive combination \(m^{ab}\) .............................................. 52
3.9 Euclidean distance between \(ab\) and the links in the common circles illustrated in Figure 3.3 ................................................................. 56
3.10 Attributes of the nodes presented in Figure 3.4 ...................... 61
3.11 Vacuous extension ................................................................. 62
3.12 Conjunctive combination ...................................................... 62
3.13 Similarity measures for node pairs in the set \(L\) ......................... 67
3.14 Distance computation ............................................................. 67
LIST OF TABLES

3.15 Uniplex methods overview .................................................. 69

4.1 Graphs’ description ................................................................. 74

5.1 Similarity measures for node pairs in common circles sharing links of type 1 95
5.2 Distance computation ............................................................. 96
5.3 Multiplex methods overview ................................................... 109
## List of Algorithms

3.1 Link prediction in evidential uniplex social networks . . . . . . . . . . . . . 48  
3.2 Link prediction in evidential uniplex social networks using social circles . . 55  
3.3 Link prediction in uniplex social networks using node attributes . . . . . . 60  
3.4 Evidential $k$-NN link prediction in uniplex social networks . . . . . . . . . 66  

5.1 Link prediction in evidential multiplex social networks . . . . . . . . . . . . 88  
5.2 Link prediction in evidential multiplex social networks with social circles . . 94  
5.3 Link prediction in multiplex social networks using node attributes . . . . . . 101  
5.4 Evidential $k$-NN link prediction in multiplex social network . . . . . . . . . 107
Introduction

Various biological, social, and technological real-world systems can be depicted by networks. Such networks are composed of two main substructures: nodes and links. Namely, nodes describe domain-dependent entities and links relations or interactions between entities. An example of social interactions can be expressed by nodes representing persons and links depicting social relationships such as friendships.

Social network analysis is a collection of techniques specifically developed to examine the relational aspects of social linkage (Scott, 1991). Link prediction is one of the key problems investigated in social network analysis that aims to analyze the patterning of links in social networks. Simply put, link prediction is the problem of evaluating the likelihood of existence of a link between unconnected node pairs based on an observed state of the network. Link prediction is a field of promising potential deserving research and application.

Basically, the link prediction task is applied to predict future links given a current state of the network. Typical applications include the suggestion of new friends in online social networks such as Facebook\(^1\) or Google+\(^2\) (Parimi & Caragea, 2011). What is more, link prediction can be used to predict missing links in social networks. An application example would be the prediction of unobserved links between criminals in terrorist networks (Anil et al., 2015).

Social networks provide a large amount of data giving an appealing opportunity for researchers and analysts to formulate and test hypotheses. The capability to deal with social data holding any uncertainty form is definitely substantial as such data is very common in real world data mining applications. Yet, this problem makes most classical link prediction approaches incapable to achieve good performances in such context.

\(^1\)http://www.facebook.com
\(^2\)http://plus.google.com
Actually, social networks are characterized by shifting degrees of uncertainty, especially the large scale ones (Adar & Ré, 2007). Most studies assume that social networks graphs are already cleaned completely. Yet, in real world data, both nodes and links could be noisy, missing and ambiguous. Even worse, the extracted information from real world data may be unreliable which affect seriously the analysis results (Svenson, 2008). In order to overcome these drawbacks, the idea is to combine link prediction with theories managing uncertainty in order to build a trustworthy analysis and enhance the quality of the prediction task. Such theories include, among others, probability theory, fuzzy set theory, possibility theory and the belief function theory.

This thesis focuses on investigating the link prediction under a new perspective where there is still a large gap waiting to be filled. The major objective is to investigate how social data imperfection can be cast and handled in link prediction and develop models and methods that solve uncertainty issues and enhance predictability. Uncertainty in link prediction has not received much attention so far, here, it appears at the links level. To address such uncertainty form, we embrace the belief function theory (Dempster, 1967; Shafer, 1976). The choice of this non standard theory is motivated by its capability to quantify and manage uncertain knowledge numerically, partial as well as total ignorance. On top of that, the belief function theory enables fusion of information induced from different sources of information.

The aim is to handle uncertainty in social network analysis using the belief function framework. Firstly, this thesis proposes an innovated way to represent and encapsulate uncertainty found in data into social networks graph structures. Most common social network graph structures consist of nodes connected via only one possible type of relations. Such structures are called uniplex graphs. To address uncertainty, we design an evidential uniplex graph-based model for social networks based on the belief function theory. Furthermore, in many real-world systems, node pairs may be connected simultaneously via different relationships. Such systems are modeled naturally as multiplex networks which enable simultaneous heterogeneous interactions. We develop, in this context, a new multiplex graph-based model with integrated uncertainty.

Then, the thesis further puts forward new link prediction approaches that operate on uncertain social networks. Our work is motivated by the need to update link prediction solution in such uncertain context. Traditional methods compute similarity scores between node pairs based on topological patterns of the networks and predict new links between most similar unconnected nodes (Kossinets & Watts, 2006; Newman, 2001). We introduce novel approaches inspired from literature that operate merely using the belief function theory tools to address link prediction problem in social networks.

Different sources of information of the social network are considered. As such, according
to the available information recorded during data collection, we can apply a particular approach. The first framework uses structural topology based on node neighborhood. It can be applied when only basic information regarding the network is available i.e., nodes and edges. The second framework is based on information of social groups. The third approach uses nodes attribute as well as graph topology. All methods handle uncertainty in both graph structure and link prediction task. On the other hand, we propose a novel method for link prediction based on supervised learning. It handles structural network information and uncertainty at the same time. All approaches are evaluated on real social network data and compared to existing methods.

To demonstrate the capability and performance of our link prediction approaches, both for uniplex and multiplex social networks, we carry some tests on real world social data which are pre-processed to encapsulate uncertainty via belief function theory concepts. Evaluations are done based on two popular evaluation criteria which are precision and recall. Evaluations are compared with existing link prediction methods that do not handle uncertainty. Results show that handling uncertainty considerably improves the accuracy of link prediction.

To be specific, this thesis includes six chapters structured into two major parts. They are connected in some ways to propose new solutions for the problems brought out during research processes. The first part, Theoretical Aspects, is composed of two main Chapters. Chapter 1 entitled “Social Network Analysis”, gives literature survey on social network analysis and link prediction. Classical link prediction approaches are presented and reviewed. In this thesis, we tackle link prediction under uncertainty. Chapter 2: “Uncertainty in Social Networks”, correlates uncertainty to social networks. More precisely, we focus on the uncertain framework of the belief function theory. We give key concepts and important mathematical definitions that are necessary for the understanding of our proposals.

The second part of this Thesis details our contributions. It can be perceived as two main blocks. The first block includes two Chapters presenting new developed link prediction methods under the uncertain framework of the belief function theory for uniplex social networks. Accordingly, we present our proposals in Chapter 3: “Link prediction in uniplex social networks”, and we give the experimental evaluation on real-world social networks in Chapter 4: “Link prediction in uniplex social networks: experimental results”.

The second block is composed of two Chapters and gives new evidential link prediction methods for multiplex social networks. We extend our evidential link prediction methods for uniplex social network to handle multiplex networks in Chapter 5: “Link prediction in multiplex social networks”. The experimental study is presented in Chapter 6: “Link prediction in multiplex social networks: experimental results”.

Introduction
Finally, our dissertation ends with a conclusion which summarizes all the work presented in this thesis and suggests some directions for further works to be done.
Part I: Theoretical aspects

Part I gives the theoretical aspects of this thesis. It provides a literature review outlining social network analysis and the theories of reasoning under uncertainty. More precisely, existing link prediction approaches are studied to shed light on gaps in previous research. Furthermore, as the belief function theory is the adopted framework in this dissertation, an overview of key concepts and definitions essential for the understanding of our proposals are presented.
Chapter 1

Social Network Analysis

1.1 Introduction

Social Network Analysis (SNA) is a collection of specifically designed methods oriented towards an investigation of the relational aspects of social structures (Scott, 1991). It was developed, about half a century ago (Prell, 2011), mainly by sociometric analysts who altered graph theory techniques and anthropologists who benefited from these advances to investigate the structure of communities and relations.

The main objective is to determine the conditions under which the patterning of social ties arise and uncover their consequences. Therefore, the main focus of SNA is to discover the implicit relationships in a network rather than the influence and attributes of the social entities.

In this Chapter, first, in Section 1.2, we define social networks. Then, in Section 1.3, we present social network analysis, its main techniques and applications. Finally in Section 1.4, we introduce a key research area in SNA which is the Link Prediction problem (LP).

1.2 Social networks

The idea of social networks was introduced in 1954 by the sociologist J.A. Barnes (1954). Ever since, new forms of social networks have emerged especially by the development of computing technology namely Online Social Networks (OSNs) such as Facebook, LinkedIn, Youtube and Google+. They provide online platforms that enable electronic communica-
tion between the actors. Millions of users are gathered within these networks forming communities related to different relationships. They can be accessed through computers, mobile devices, tablets, etc, and provide various services such as profile management, chat, content share, comment, forums and instant messaging. The boom of the internet in the last decade has boosted the development of OSNs making them probably the central actor of the Web 2.0.

Two essential concepts are required to build a social network. First, one needs to determine the social entities of the network. These latter may be of various types such as users, organizations, firms, substrates, authors or terrorists. They can be named and have several properties joined to them. Next, one has to define the social ties according to some kinds of interdependency such as friendship, financial exchange, physical proximity, knowledge, relationships of beliefs, chemical interaction or co-authorship.

Various social networks exist (Girvan & Newman, 2002; Newman, 2001; Traud, Kelsic, Mucha, & Porte, 2011). Most popular instances would be social networks of friendships between persons (Newman, 2010), terrorist and criminal networks (Anil et al., 2015; Geffre, Deckro, & Knighton, 2009), as well as collaboration networks of co-authorships (Newman, 2001). On the other hand, there is a great deal of information that can be extracted from online social networks that enable the construction of networks such as friendships (i.e., from Facebook) and collaboration networks of co-authorships (i.e., from ArXiv and Dblp) (Newman, 2001).

Social networks are characterized by their high complexity and large sizes. Their analysis has a major implication in a variety of studies. For instance, it permits to identify influential actors, describe the mechanism of spread of information, discover the source of any given piece of information, uncover the hidden connections, predict potential new associations, detect social circles or match similar sub-graphs. These areas of research have a great potential for researchers and analysts from various disciplines. A large number of techniques, algorithms and applications have arisen to study and extract valuable information from social networks.

1.2.1 Social networks representation

From a historical standpoint, the first representation of social networks was the sociograms developed by Jacob Moreno (1933) to learn interpersonal relationships. It was drawn by a set of points representing persons and lines encoding their relationships. This representation was mathematically formalized during the 1950s and became the basic depiction for modern social and behavioral sciences, as well as social network analysis (Wasserman &
Faust, 1994). It is based on a classical graph theory concept which is the graph model representation.

From this point of view, a network of \( n \) actors and \( m \) connections is mathematically denoted by \( G(V, E) \), where \( V \) is the set of nodes \( \{v_1, v_2, v_3, \ldots, v_n\} \), and \( E \) is the set of links \( \{e_1, e_2, e_3, \ldots, e_m\} \). Nodes describe social entities and links depict the relational ties. Schematically, a node (i.e., actor, agent, object) may be represented using different colors, shades, symbols, and sizes to specify different properties or types. It can be labeled to express its name, type, number, reference, etc. To indicate different kinds of relations, the actors are linked by: dotted lines, bold lines, multi-lines, different colors, undirected lines (i.e., collaboration networks, friendship networks), directed lines (i.e., communication network, election voting), reciprocal directed lines (i.e., two way relation, friendship or marital relationship) (De Sagar & Dehuri, 2014). They can be weighted to encode their strength, probability, frequency, etc.

Graph data can be stored either using matrices or lists. The matrix representation is more convenient for representing dense networks. Conversely, data of social network are sparse especially large scale networks, thus list representation is more favored. In Figure 1.1 (a) a social network is represented using a graph with six nodes linked by eight edges. The adjacency list (Figure 1.1 (b)) of the former gives a simple list of nodes and the ones linked/adjacent to them called “neighbors”. In contrast, the adjacency matrix (Figure 1.1 (c)) is of size \( 6 \times 6 \) and contains binary values indicating the presence or absence of edges between the nodes. In the case of directed graphs, we get incidence matrices and incidence lists which are analogous to the former except that the stored information specifies nodes and edges incidence.

![Network Diagram](attachment:image.png)

(a) Social network represented as a graph (nodes are labeled by letters) (b) Adjacency list (c) Adjacency matrix

Figure 1.1: (a) Social network represented as a graph (nodes are labeled by letters) (b) Adjacency list (c) Adjacency matrix
1.2.2 Social networks graph structures

Social networks can represent various entities and relations. Their graph representation must be suitable to their shape. For example, we can have one/several type(s) of actors and one/several type(s) of relations. We discuss in the following social networks graph structures that handle social networks properties.

**Uniplex structure:** Relationships between the actors can be uniplex where only one type of connection exists. In other words, the actors can have a single kind of social tie. One such example is a network of people connected by friendship relationships or a network of employers linked by co-work relations. Figure 1.2 illustrates a uniplex social network connecting friends. Formally, the uniplex structure is represented by a simple graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is the set of actors connected via homogenous single relations in $\mathcal{E}$.

![Figure 1.2: Uniplex social network of persons connected via one type relationship](image)

**Multiplex structure:** Links between fixed set of actors are maintained by several relationship kinds where simultaneous connections are allowed. As such, multiple links are schematized in the graph structure Bródka and Kazienko (2014). For example, the social network presented in Figure 1.3 connects persons according to three kinds of relationships: friendship, colleague relation and co-authorship. In formal terms, the multiplex structure is represented by a multi-relational graph $\mathcal{G}(\mathcal{V}, \mathcal{E}_1, \ldots, \mathcal{E}_l)$ where $\mathcal{V}$ is the set of actors connected via $l$ heterogeneous simultaneous relations in $\{\mathcal{E}_1, \ldots, \mathcal{E}_l\}$.

**Homogenous structure:** Here, we focus on the type of social entities and/or links. A homogenous network handles only a single type of actors and a single type of links. For instance, the nodes are persons linked by friendship relations such as the social network presented in Figure 1.2. It is represented by a graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is the set nodes of persons and $\mathcal{E}$ is the set of their connecting friendship relations.
Section 1.2 – Social networks

Figure 1.3: Multiplex social network of persons connected via three possible relationships

**Heterogeneous structure:** A heterogeneous network takes into account multiple types of social entities and relationships (nodes and edges) each playing a distinct role. In other words, we can have various social entities and/or different social ties. For example, a network of authors and papers as nodes connected via co-authorship and citation links as presented in Figure 1.4. It can be formally defined by a graph $G(\mathcal{V}, \mathcal{E}, \mathcal{T}_\mathcal{V}, \mathcal{T}_\mathcal{E})$ where $\mathcal{V}$ is the set of $n$ couples $<v, t_v>$ where $v \in v_1, \ldots, v_n$ and $t_v \in \mathcal{T}_\mathcal{V}$, $\mathcal{E}$ is the set of $m$ tuples $<e, t_e>$ where $e \in \mathcal{E}$ and $t_e \in \mathcal{T}_\mathcal{E}$, $\mathcal{T}_\mathcal{V}$ is the set of $k$ node types $\mathcal{T}_\mathcal{V} = \{t_{v1}, \ldots, t_{vk}\}$ and $\mathcal{T}_\mathcal{E}$ is the set of $l$ edge types $\mathcal{T}_\mathcal{E} = \{t_{e1}, \ldots, t_{el}\}$.

Figure 1.4: Heterogeneous social network of authors and papers connected via co-authorship and citation relations
1.2.3 Social networks data

In traditional social science methods, social networks were collected manually via questionnaires or surveys by thoughtful observation of a group of people. Small-scale social networks were generated from such questionnaire-based approaches. More recently, automated methods of data collection have allowed to get massive social network data allowing a higher level of scales of analysis (Alhajj & Rokne, 2014).

Earlier, data were crawled manually, artificially or automatically from online social network services for analysis and scientific research purposes. In the era of Internet, an increasingly amount of valuable information becomes available thanks to cloud applications, platform services and scalable databases. Social networks dataset collections are available online. Some public social data collection sites include: Network Repository\(^1\), Snap stanford\(^2\), Github\(^3\), Networking Group UCIrvine\(^4\), BGU Social Networks Security Research Group\(^5\), Network data repository\(^6\), Minas gjoka datasets\(^7\), Dango\(^8\), Konekt\(^9\), Socialcomputing.asu\(^10\), Tore Opsahl datasets\(^11\) and CFinder\(^12\).

Research in the social network field intends to predict relationships structure among entities, as well as its effect on other social phenomena. Such research builds upon a number of measures, representations and techniques that are jointly referred to as social network analysis. We present in the following major concepts of social network analysis.

1.3 Social network analysis

Social Network Analysis is a collection of specifically designed methods oriented towards an investigation of the relational aspects of social structures (Scott, 1991). It is an interdisciplinary domain including fields such as social and behavioral sciences (i.e., sociology and social psychology), statistics, and graph theory (Butts, 2008). It has attracted consid-

---

\(^1\)http://networkrepository.com/
\(^2\)http://snap.stanford.edu/data/index.html
\(^3\)https://github.com/gephi/gephi/wiki/Datasets
\(^4\)http://odysseas.calit2.uci.edu/doku.php/public:osn_datasets
\(^5\)http://proj.ise.bgu.ac.il/sns/datasets.html
\(^6\)https://networkdata.ics.uci.edu/index.php
\(^7\)http://www.minasgjoka.com/datasets.html
\(^8\)https://dango.rocks/datasets/
\(^9\)http://konekt.uni-koblenz.de/
\(^10\)http://socialcomputing.asu.edu/pages/datasets
\(^11\)https://toreopsahl.com/datasets/#norwegianbod
\(^12\)http://www.cfinder.org/
erable interest as well in marketing (Webster & Morrison, 2004), finance (Hu, Zhao, Hua, & Wong, 2012), politics (Makazhanov & Rafiei, 2013; Ratkiewicz et al., 2011) and biology (Girvan & Newman, 2002).

SNA focuses on the structure of the social ties among the actors and on the investigation of their patterns and implications. It adds a valuable contribution to classical social and behavioral research which is mainly concerned with social units attributes (Wasserman & Faust, 1994).

Since its appearance, social network analysis has had great applicability in various fields such as political studies (Makazhanov & Rafiei, 2013), organizational studies (Stuart, 1998) or the Web (Chau & Xu, 2007). For the most part, it has been used for describing social structure, for example in studies by Leavitt (1951), Tichy, Tushman, and Fombrun (1979), Burst (1980) and Freeman (1989).

Although SNA appeared several decades ago, graphical and visual methods for SNA are still developing at an impressive speed to explore and represent network knowledge. SNA applies measures emanated from graph theory to analyze characteristics of social networks. Layout techniques are implemented to visualize social networks and facilitate analysis such as finding different communities. Popular measures are presented in the following.

1.3.1 SNA measures

In graph terminology, networks can be characterized by statistically computed measurements such as node average degree, shortest path length between a pair of nodes, network density, diameter, number of triangles, clustering coefficient, number of isomorphisms, among others. Most popular social network analysis measures include degree centrality, closeness centrality and betweenness centrality (Scott, 1991).

Centrality measures evaluate the impact of a node (i.e., an individual actor) in the network, they index the representativeness and proximity of a node towards all the others. One of the popular centrality measures is the degree centrality. It identifies the structural centrality of an actor. It represents the number of links that a node is part of. Degree centrality is easy to compute since only the local structure around the node is considered. Though simple, it is frequently effective when evaluating the importance of a node in the network. For example, in several social networks, people with many connections are likely to be more visible and important.

Another popular centrality measures is the closeness centrality. By contrast to degree centrality, closeness centrality considers the global structure of the network. It measures
how “close” a node is to all the others. It computes the average shortest between two actors in the network. For instance, the more a node is central, the lower its total distance to other nodes is.

Finally, the betweenness centrality measures the proportion of geodesics, i.e., the shortest paths between two nodes, passing through a particular node in the network. The more shortest paths are found, the more central the node is to the network. Nodes with high betweenness are often called gatekeepers information or brokers.

Another important measure is centralization. It allows to analyze the properties of the network as a whole. Centralization measures the relative difference between the number of links connected to each node divided by the sum of the maximum possible differences over all nodes in the graph. It gives an idea about the variability of centrality in the network.

1.3.2 SNA visualization

Social networks visualization is a set of mathematical and algorithmic techniques that enable graphical representation of network data to facilitate their understanding, examination, and exploration (Alhajj & Rokne, 2014). In fact, visualization permits to disclose the structure of social networks, highlight connectivity and expose the clusters and strengths of the relationships between the actors. The animated outputs reveal the evolution and dynamics of the social network over time. Visualization focuses on the spatial arrangement of actors and their connections. A well-arranged spatial layout brings closer the actors that are closely related and holds off not related actors. Such notion of proximity permits to identify communities at a brief look. In addition, visualization should be readable. Actually, the graph is hard to read when many labels and graphical elements overlap.

Various layout techniques have been implemented for visualization of networks (Crawford, Walshaw, & Soper, 2012; McGuffin, 2012), such as spectral layout, force-directed layout, layered graph drawing, tree layout, arc diagrams and circular layout. In order to specify different actors properties and types, different shades, colors and symbols are used. Similarly, different kinds of links are employed to differentiate between actors’ relations such as dotted lines, multi-lines, bold lines, or directed lines. Several visualization tools such as Gephi\(^\text{13}\), NetMiner\(^\text{14}\), SocNetV\(^\text{15}\) and Sentinel Visualizer\(^\text{16}\) are available to visualize and analyze social network data.

\(^{13}\)https://gephi.org/
\(^{14}\)http://www.netminer.com/
\(^{15}\)http://socnetv.org/
\(^{16}\)http://www.fmsasg.com/
1.3.3 SNA applications

Social networks analysis is an analytic approach with its own theoretical statements and methods that are applied in different fields. That is, depending on the objective of the analysis, one may either study the whole network (i.e., its structure) or focus on a part of it (i.e., the ties). Practical applications include information diffusion (Mannila & Terzi, 2009; Myers et al., 2012; Wei et al., 2010), behavior analysis (Bermingham et al., 2009; Nguyen et al., 2012), community detection (Onnela et al., 2011), social sharing and filtering (Ben Ismail & Bchir, 2015), recommender systems development (Chenhao et al., 2011), data aggregation and mining (J. Zhang et al., 2013). Analysis of social networks can provide perception about information flow, influencers, groups partition, actors habits, disease spread, hidden members, terrorists, voting patterns, and so on.

It is critical to learn how a network evolves; how new associations are formed over time. This is known as the “link prediction problem” (Liben-Nowell & Kleinberg, 2003), a wide research field that aims to study social network evolving and understand how the social ties are formed. Link prediction is one of the key problems treated in social network analysis. It has many applications in several domains such as politics, bioinformatics, e-commerce, security, and so on. In this thesis, we deal with the link prediction problem. We propose new frameworks to handle link prediction by considering different network information. Key concepts and a survey of link prediction approaches are given in the next section.

1.4 The link prediction problem

Link Prediction (LP) deals with predicting new links that are likely to happen among the nodes given a particular topology of the network. Link prediction has too many relevant applications. It is applied to anticipate future behavior or to identify probable relationships that are difficult or expensive to observe directly. In social networks, link prediction can be used to predict relationships that will form, uncover relationships that probably exist but have not been observed, or even to assist individuals in forming new connections (Davis et al., 2011).

Natural examples of link prediction include predicting co-authorships of scientific papers in a specific discipline, co-working of employees on projects in a large company, as well as monitoring activities of terrorists to deduce their involvement in the future. In other domains such as bioinformatics, methods of link prediction can be applied to infer protein interactions (Lei & Ruan, 2013). In e-commerce, it aids in constructing recommender systems to propose suggestions and provide personalized services (Li & Chen, 2013).
First works on link prediction (Sarukkai, 2000; Zhu, 2001; Zhu et al., 2002) focused on web networks for efficient web page navigation. Markov models and network structure were used for user navigation assistance from any given web page. Although link prediction was studied earlier, it was surveyed formally for the first time by Liben-Nowell and Kleinberg (2003). Since then, several surveys have been made (Getoor & Diehl, 2005; Hasan & Zaki, 2011; Lu & Zhou, 2011). In the following, we present the link prediction problem and we recall state-of-the-art link prediction methods.

1.4.1 Problem formulation

Link prediction applies two tasks: (i) predicting missing links and (ii) predicting future links. Both tasks are concerned with the prediction of potential new links to be added to the network given an observed state. However, according to the objective of the analysis and to the available data, link prediction can be formulated in two ways.

Formally, given a network $G_t(\mathcal{V}, \mathcal{E}_t)$, where $\mathcal{V}$ is the set of nodes and $\mathcal{E}_t$ is the set of edges, at a given time $t$, the aim is to predict the new set of links that includes the edges that are likely to exist. It is important to note that the set of nodes $\mathcal{V}$ does not change. When predicting missing links, the task is to predict the links which have existed at time $t - 1$ given the observed structure of the network at time $t$. Links may be missing due to missing or unobservable information from data, hidden information from the network or because of privacy settings. An application example would be to predict a friendship relation, for example in Facebook, between two unconnected friends. To predict future links, one should consider the network at time $t$ and predict the links that are likely to emerge among the unlinked nodes of the network at time $t + 1$. For example, predicting a potential friendship that may occur in the future between two unlinked users in Facebook.

Yet, future link prediction inquires social data with time information where a stream of states of the network over time is collected. Tylenda et al. (2009) predicted new and recurrent links by considering history network information derived from various snapshots of the network over time. Link prediction performance is improved when handling time-stamps of past interactions according to the reported results. Rümmele et al. (2015) solved future link prediction by considering networks over several time periods by counting 3-node graphlets. Prediction of future links is also called temporal or dynamic link prediction (Srinivas & Mitra, 2016). Yet, temporal social data are not always available. Therefore, most link prediction algorithms are tested on static data derived from the observed structure, then new links that are not visible but that are likely to exist are predicted (Taskar et al., 2003).
1.4.2 Literature Review of Link Prediction Methods

Prior art in link prediction literature can be classified into three major families: namely (i) similarity-based methods, (ii) machine-learning methods, and (iii) latent-models based methods. We present each family of methods in the following.

1.4.2.1 Similarity-based methods

Similarity-based methods are very popular in link prediction literature. Most LP algorithms compute similarity of unlinked node pairs, similarity scores are ranked and the links with highest scores are likely to exist. Similarity scores are derived from network topology. They are generic and generally simple to compute. Generally, in most state-of-the-art research, similarity-based methods for link prediction are classified into two groups (Lu & Zhou, 2011; Srinivas & Mitra, 2016; T. Zhou et al., 2009): (i) local and (ii) global information. Local information methods compute scores according to node neighborhood cohesiveness. The intuition is derived from the small-world phenomenon (Kleinberg, 2000) which conjectures that every pair of nodes is distant by a small number of nodes. In other terms, unlinked nodes are probably connected via their neighbors (Goldberg & Roth, 2003). Global methods compute metrics of network proximity based on paths. Yet, there are other similarity scores, for example group information based methods, that use more than local topological information and do not require global measures. We present in the following some popular scores based on local, global and group information of the network.

Interestingly, similarity measures are easy to compute since they are based on graph topology. In other words, they use structural information which is available and not hard to get. Furthermore, similarity score are very simple, especially local information-based ones, taking only few parameters into account. Apart from this, these methods do not depend on the domain of the social network. Structural scores can be computed to all network types, they proved their efficiency in many real-world networks (Kossinets & Watts, 2006; Newman, 2001).

Local information measures Local measures compute similarities according to the characteristics of the neighboring nodes in the network.

(1) **Common Neighbors (CN)** (Newman, 2001). One of the most popular measures is “Common Neighbors”, denoted by $CN(u, v)$. It computes the number of common neighbors shared between a pair of nodes $(u, v)$ in the network. Let $\tau(u)$ denote the
set of neighbors of a node $u$. The Common Neighbors measure is defined as:

$$CN(u, v) = |\tau(u) \cap \tau(v)|$$ (1.1)

Newman (2001) applied this measure in collaboration networks assuming that there is a correlation between the number of common co-authors of two scientists and the likelihood of a future collaboration. A part from that, the empirical analysis of an evolving social network of students and staff at a large university made by Kossinets and Watts (2006), showed that students who share many mutual friends are likely to be friends in the future.

(2) **Adamic/Adar (AA)** (Adamic & Adar, 2003). This score revises the simple counting of common neighbors by allocating more weights to less connected neighbors. It is defined as:

$$AA(u, v) = \sum_{z \in (\tau(u) \cap \tau(v))} \frac{1}{\log|\tau(z)|}$$ (1.2)

(3) **Jaccard Coefficient (JC)** (Jaccard, 1901). Another popular measure proposed by Jaccard more than a hundred years ago, it takes all the neighbors of the pair $(u, v)$. It is computed as follows:

$$JC(u, v) = \frac{|\tau(u) \cap \tau(v)|}{|\tau(u) \cup \tau(v)|}$$ (1.3)

(4) **Resource Allocation (RA)** (T. Zhou et al., 2009). This index is motivated by the resource allocation process of networks. Common neighbors of two unlinked nodes $u$ and $v$ are considered as transmitters of resources where each one supplies a single unit so that $u$ sends some resource to $v$. Thus, similarity of $(u, v)$ is the amount of resource collected from between $u$ and $v$ computed as follows:

$$RA(u, v) = \sum_{z \in (\tau(u) \cap \tau(v))} \frac{1}{|\tau(z)|}$$ (1.4)

(5) **Preferential Attachment (PA)** (Newman, 2001). The mechanism of preferential attachment speculates that a new edge has probability proportional to $|\tau(u)|$ to connect to $u$. Hence, the preferential attachment of $(u, v)$ is proportional to the number of neighbors of $u$ and $v$. It is computed as follows:

$$PA(u, v) = |\tau(u)| \cdot |\tau(v)|$$ (1.5)
(6) **Leicht-Holme-Newman (LHN)** (Leicht, Holme, & Newman, 2006). This index attaches high similarity to node pairs with many common neighbors compared to the expected number of common neighbors. It is defined as:

\[
LHN(u, v) = \frac{\left| \tau(u) \cap \tau(v) \right|}{\left| \tau(u) \right| \cdot \left| \tau(v) \right|}
\]  

Liben-Nowell and Kleinberg (2003) and T. Zhou et al. (2009) compared local similarity measures on social networks from real world. According to the reported experimental results, the RA, AA and CN scores give the best link prediction performance among compared to the presented local scores mentioned above.

Recently, some approaches have been proposed to handle link prediction in multiplex networks (Battiston et al., 2014; Davis et al., 2011; Kanawati, 2015; Pujari & Kanawati, 2013). Network topology is considered differently from one work to another. For example, node neighborhood in the network is considered in different manners: the union of all connected neighbors among all layers or more restrictively, neighbors intersection among all layers (Kanawati, 2015). From this point of view, one can extend structural local measures of network topology to multi-relational networks by taking into account the first option. Hence, we get the following measures given in Table 1.1:

<table>
<thead>
<tr>
<th>Method</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$CN_{multiplex}$</td>
<td>$\tau(u)^G \cap \tau(v)^G$</td>
</tr>
<tr>
<td>$AA_{multiplex}$</td>
<td>$\sum_{n \in (\tau(u)^G \cap \tau(v)^G)} \frac{1}{\left</td>
</tr>
<tr>
<td>$JC_{multiplex}$</td>
<td>$\frac{\tau(u)^G \cap \tau(v)^G}{\tau(u)^G \cup \tau(v)^G}$</td>
</tr>
<tr>
<td>$RA_{multiplex}$</td>
<td>$\frac{1}{\left</td>
</tr>
<tr>
<td>$PA_{multiplex}$</td>
<td>$\frac{1}{\left</td>
</tr>
<tr>
<td>$LHN_{multiplex}$</td>
<td>$\frac{\left</td>
</tr>
</tbody>
</table>

**Global information measures** Global information based methods computed scores based on network proximity. The more two unlinked nodes are close in terms of number of distant paths, the more likely they will connect.

(7) **Shortest path.** The shortest path is simply the distance in terms of number of paths to reach node $v$ from $u$. The shorter the distance, the higher $u$ and $v$ to connect.
(8) **Katz Index** (Katz, 1953). It is derived from the shortest path measure, but generally gives better results in link prediction. It sums the paths between \( u \) and \( v \) and penalizes longer paths by a factor \( \beta^l \), where \( l \) is the path length. The Katz index is computed as below:

\[
Katz(u, v) = \sum_{l=1}^{\infty} \beta^l \cdot |\text{paths}_u^v|^l
\]

(1.7)

where \( \text{paths}_u^v \) is the set of all paths of length \( l \) from \( u \) to \( v \).

(9) **SimRank** (Jeh & Widom, 2002). The intuition of the SimRank measure is very simple. Two unlinked nodes are more likely to connect if they are linked to similar nodes. It is computed using the following recursive equation:

\[
\text{SimRank}(u, v) = C \cdot \frac{\sum_{z \in \tau(u)} \sum_{z' \in \tau(v)} \text{SimRank}(z, z')}{|\tau(u)| \cdot |\tau(v)|}
\]

(1.8)

where \( \text{SimRank}(u, u) = 1 \) and \( C \in [0, 1] \) is a decay factor.

(10) **Average Commute Time (ACT)** (Lovász, 1996). Let \( N_{\text{steps}}(u, v) \) be the average number of steps necessary for a random walker to reach node \( v \) starting from \( u \). Shorter ACT means that the nodes are similar, thus they have higher possibility of connecting in the future. The average commute time between \((u, v)\) is defined by:

\[
\text{ACT}(u, v) = N_{\text{steps}}(u, v) + N_{\text{steps}}(u, v)
\]

(1.9)

Fouss et al. (2012) applied the Matrix Forest Index to evaluate similarity between nodes under a collaborative recommendation task. Reported results have shown that a simple nearest-neighbors rule based the MFI performs best.

Bródka and Kazienko (2014) investigated properties of multi-layered social networks. For instance, a multi-layered path is represented by a list of nodes connected via multi-layered edge where the total length is the sum of distances of all these edges. Hence, the shortest multiplex path between a pair of nodes is the one with the smallest length in the social network.

Global based approaches generally perform better than local information based methods. However, they are computationally costly because they require information of the
whole network. Besides, global information is not usually available (Lu & Zhou, 2011). For these reasons, most existing works on link prediction use local metrics. Meanwhile, other works extended similarity scores that do not use local information solely in order to enhance performance results such as the groups affiliation information.

**Group information measures** Group (cluster, community or circles) information based measures use the participation of the actors in social groups. In several real-world social networks, social structures may belong to several different types of groups at the same time. In fact, the idea of social circles dates back to at least 1902 (Simmel, 1902). The term was subsequently employed as a metaphor. Social circles (groups, communities, clusters) reunite social entities with common features, interests, hobbies, shared activities or beliefs. In real world social networks, actors with similar interests and experiences tend to connect unlike those that do not have common characteristics. In many online social networks, users organize accounts using social groups. For example, circles in Google+ or groups and lists in Facebook and Twitter.

Existing methods for link prediction employed information given by social circles. Some works even proposed similarity measures based on group information which have improved the link prediction task (Valverde-Rebaza & Lopes, 2012, 2014).


Let $\Lambda^G_{uv}$ denote the set of common neighbors of the pair $(u, v)$ that belong to the group $G$. The CNG describes the size of the set of common neighbors of $(u, v)$ that belong to at least one group $G$ to which $u$ or $v$ is part of. It is defined as:

$$S_{uv}^{CNG} = |\Lambda^G_{uv}|$$

(13) **Common Neighbors Within and Outside of Common Groups (WOCG)** (Valverde-Rebaza & Lopes, 2012, 2014). Let $\Lambda_{uv} = \Lambda^W_{uv} \cap \Lambda^O_{uv}$ be the set of common neighbors of $(u, v)$ such that $\Lambda^W_{uv}$ is the set of common neighbors within common groups (WCG) and $\Lambda^O_{uv}$ is the set of common neighbors outside the common groups (OCG). The WOCG measure is defined as:

$$S_{uv}^{WOCG} = \frac{|\Lambda^W_{uv}|}{|\Lambda^O_{uv}|}$$

Valverde-Rebaza and Lopes (2012) compared group information measures presented above with ten basic local similarity for link prediction. The results indicate that cluster
information improves link prediction accuracy. They did not extend these measures to multiplex networks.

Social groups are most of the time structured naturally in online social networks. Yet, when such information is not directly available, one can apply community detection algorithms to partition the network into structured clusters. A community includes many edges joining its nodes with comparatively few edges connecting nodes of different communities. Such groups include similar nodes with similar features or roles. It is possible to apply algorithms for community detection with low computation cost (Fortunato, 2010).

### 1.4.2.2 Machine-learning methods

The learning scheme for link prediction typically evaluate similarity between the query nodes using similarity scores as features to predict new links existence. The link prediction problem is usually shifted into an instance of a binary classification problem. Class labels are assigned to the links. Given the observed structure of the social network $G(V, E)$, the training set catches the observed state $s$ and the test set incorporates links of another state $s'$. Essentially, the objective is to predict the classes of the test links $e$ given the observed training set of links:

$$
\text{class}(e|G) = \begin{cases} 
1 & \text{if } e \in E \\
0 & \text{if } e \notin E
\end{cases}
$$

In this context, any supervised learning method, such as $k$-nearest neighbors (Dasarathy, 1991), naive Bayes (Russell & Norvig, 2002) or support vector machines (Cortes & Vapnik, 1995), can be used. However, the main challenge is the choice of the set of features for classification. Hasan et al. (2006) combined structural metric and external information outside the scope of topological properties for collaboration networks. Content features related to the network domain such as keywords count and sum of papers were used. The results have shown that topological similarity metrics and content information can significantly improve prediction results. Yet, content information is not usually available due to privacy and anonymization procedures. Besides, the selection of the attributes to be considered in the prediction task is crucial as it depends to the social network domain. For instance, the considered information about the researchers in a collaboration network would be different from the information about suspects in a criminal network.

Many methods have been proposed later (Bilgic et al., 2007; Doppa et al., 2009; Wang et al., 2007) for link prediction under supervised learning. Kashima and Abe (2006) pro-
posed an incremental classification method based on a probabilistic model. It computes topological properties of the network structure to predict links. Sarkar et al. (2011) provide a theoretic justification of the success of a number of popular similarity scores using familiar class of graph generation models where nodes are attached with locations in a latent metric space and links connect closer nodes. Results have shown bounds related to the usefulness of a node degree in link prediction, the relative importance of short paths compared to long paths, and the effects of increasing non-determinism in the generation of links process and the link prediction quality. Miller et al. (2009) propose a Bayesian based method for link prediction in relational data. Their framework gives the number of features and learns the feature of each entity.

1.4.2.3 Latent-models based methods

Latent-models based methods use probabilities to assess the likelihood of links existence (Kashima & Abe, 2006; Wang et al., 2007). These algorithms assume a particular organizing principles of the network structure, with the detailed rules and specific parameters retrieved by maximizing the likelihood of the observed graph structure. Then, the likelihood of any query link is computed according to those rules and parameters. Latent-models based approaches are usually applied to extract latent features and learn the classes of the entities. When the latent feature vectors are determined, latent feature matrix is computed based on a matrix factorization model (Menon & Elkan, 2011; Miller et al., 2009).

Wang et al. (2007) presented a link prediction framework based on a local probabilistic model that uses Markov Random Field (Kinderman & Snell, 1980), an undirected graphical model. To predict a new link between two unlinked nodes \( u \) and \( v \), a central neighborhood set which consists of other nodes that are present in the local neighborhood of \( u \) or \( v \) is introduced. A well constructed algorithm can handle a network with a few thousand nodes in a acceptable time, however it will definitely not scale to huge online networks with millions of nodes. Menon and Elkan (2011) presented an approach using matrix factorization combined with latent features learning for link prediction.

Yet, matrix factorization based methods are computationally costly because the networks become larger. Yu and Chu (2008) proposed stochastic relational models basically using Gaussian process models for link prediction.

One of the drawbacks of this family of methods is that they suffer from high complexity costs. Matrix decomposition, factor matrices building and latent features learning are time consuming making such algorithms intractable for large scale networks.
1.5 Conclusion

In this Chapter, we have presented social networks, the basic concepts of social network analysis and basically the link prediction problem. We have, also, discussed the major techniques for LP and their pros and cons. In the next Chapter, we will discuss the relation between social network analysis, more specifically link prediction, and the theories of reasoning under uncertainty. A special focus will be dedicated to the belief function theory as an uncertain framework to be embraced in our proposals.
Chapter 2

Uncertainty in Social Networks

2.1 Introduction

Data are usually characterized with shifted degrees of uncertainty, especially social network data. In fact, one of the inherent properties of several social network from real life is that they are characterized by different degrees of imperfection, particularly the large-scale ones (Adar & Ré, 2007). In fact, handling data imperfection when treating social networks is very crucial. Some researchers pointed out the importance of handling uncertainty in social network analysis. However, a restricted number of approaches has been proposed to tackle uncertainty problems. More precisely, most link prediction methods are restricted to treating social network under a certain framework.

Accordingly, this chapter highlights the importance of taking uncertainty into account in social network analysis methods. In Section 2.2, we explain principle causes of handling uncertainty in social networks. In Section 2.3, we present existing works that draw attention to handling uncertainty via theories of reasoning under uncertainty in social network analysis, especially the link prediction problem. Finally, in Section 2.4, we focus on the belief function theory as it is the uncertain framework embraced in this thesis as well as, existing approaches for SNA and link prediction proposed under this context in Section 2.5.

2.2 Why uncertainty in social networks?

Uncertainty can arise in many different ways. As a matter of fact, data within the intelligence domain are often unreliable, information about individuals and connections may
be imprecise and noisy. Furthermore, the objective of the analysis itself may be pervaded with uncertainty. For example, sometimes the information request may be ambiguous and unclear. Lack of information is also a source of uncertainty because data are often incomplete. As it is common in real world social data, only a partial structure of the network is given; usually, not all the links are known. On top of that, as there are various sources of social network data, there may be conflicts and inconsistency. For example, when collecting social data from different online social such as Facebook and Linkedin, we can have inconsistent information. For example, a particular user can put different birthday dates on both networks, or different hometowns.

Svenson (2008) pointed out the problem of representing uncertainty in social network analysis and gave some possible solutions to handling it. According to Svenson, uncertainty can be of two types in social networks: (i) We can be uncertain whether particular node is distinct or not, i.e., whether two social entities are the same or not, (ii) we can be uncertain about the existence of a link between a pair of nodes. For instance, if person $X$ studied in the same college as person $Y$, should we assume that they are linked?

What is more is that intelligence information provided from signals intelligence is affected with natural uncertainty. For example, information given by a camera about a particular person seen talking to a suspicious person cannot confirm the real identity of the persons and may not confirm with certainty that they really talked to each other. Svenson proposed to combine Monte Carlo simulation, modeling via random set, and Bayesian analysis to deal with uncertainty. Bearing this in mind, researchers must be able to manipulate uncertain knowledge and formulate and test hypotheses about the available data. Analysis frameworks must have the ability to represent and reason with uncertain information.

Adar and Ré (2007) argued that collecting social network structure, and the shift in scale comes with higher degree of imprecision that must be taken into account to apply social network analysis techniques. The authors proposed to manage and manipulate uncertain social data using probabilistic databases. For the most part, uncertainty is correlated to the analyzed data. Data were collected manually through direct observations and questionnaires. Thus, data sets were generally small and social entities were mostly known. However, in contrast to old techniques, present data are very huge due to the emergence of online networking applications and the development of scalable databases. This encouraged researched from different fields to analyze the properties of large scale social data. Yet, little interest has been devoted to examination uncertainty in social networks.

Consequently, it is fundamental to take uncertainty into account when dealing with social data. In this respect, we present in the following section existing work on social network analysis and link prediction under theories of reasoning under uncertainty.
2.3 SNA and the theories of reasoning under uncertainty

In the last years, some works have been devoted to handle uncertainty in social network analysis by embracing uncertainty theories such as probability theory (Kolmogorov, 1933), fuzzy set theory (Zadeh, 1965) and possibility theory (Zadeh, 1978).

Probability theory is generally applied to represent two types of phenomena: (i) randomness by quantifying variability across repeated observations and (ii) partial knowledge.

Fuzzy set theory, introduced by Zadeh (1965), generalizes the classical notion of a set and a hypothesis to integrate fuzziness according to its sense in human language. The membership of elements in a set can be gradually assessed using a membership function valued in the interval $[0, 1]$.

Possibility theory is an extension of fuzzy sets and fuzzy logic incepted by Zadeh (1978) to handle incomplete information. It quantifies uncertain knowledge via ordinal or numerical values. It tells plausible states from less plausible ones.

We recall in the following existing work in social network analysis under uncertainty theories. As we are concerned with the link prediction problem, some works on link prediction that handled uncertainty are also reported.

**Probability theory** Various works adopted theories of reasoning under uncertainty to many social network analysis applications. Using probability theory, Foulds et al. (2011) introduced a non parametric Bayesian model for longitudinal social network data that models actors with latent features whose memberships change over time. Heaukulani and Ghahramani (2013) introduced a probabilistic model for capturing the behavior of actors in social networks over time. Akbar and Arroyo (2008) investigated and analyzed a social network using the Bayes probability theory model to depict the important actors in it. This is accomplished by the computation of the entropy of each node present in the network and the amount of resulting variation obtained due to its removal.

Other works handled link prediction under the probability theory. For instance, Tang et al. (2013) proposed a prototype for predicting the existence of links among nodes in large-scale social networks based on the Markov Logic Networks. Wang et al. (2007) presented a local probabilistic graphical method that evaluates the joint co-occurrence probability of two nodes. Tests were made by considering the co-occurrence probability feature solely and with other state topological and semantic features on real world collaboration networks.
The obtained results show performance improvement.

**Fuzzy set theory** Other works embraced fuzzy sets. Trinidad and Yager (2014) combined techniques of graph theory and fuzzy sets to characterize the Small World phenomenon features and the existence of the figure of leader in social networks. Tanwistha et al. (2011) proposed a fuzzy clustering-based approach for community detection in networks. The framework is validated on real social and biological networks. S. Zhang et al. (2007) introduced an algorithm for identifying overlapping communities in networks based on fuzzy c-means clustering. Fan et al. (2007) generalized the notion of regular equivalences to fuzzy social networks. Brunelli and Fedrizzi (2009) extended the relations in a social network to an m-dimensional matrix. The authors introduced a fuzzy m-adjacency matrix of relations and they formalized their link with fuzzy binary relations.

In link prediction, Bastani et al. (2013) proposed a new fuzzy link prediction model thanks to two new similarity indices based on granular computing method and fuzzy logic. The model was tested on collaboration networks. Results show that fuzzy analysis gives accurate predictions through better expression of the characteristics of the network compared to the crisp approach.

**Possibility theory** Other authors used possibility theory to social network analysis techniques. That is, Ben Ismail and Behir (2015) proposed an approach to automatically detect verbal offense in social network comments. It relies on a local approach that combines different regions of the feature space in order to classify comments from social networks as insult or not. Nin and Torra (2010) presented a reputation model that uses weighted minimum and maximum functions to aggregate the different trust values obtained from a trust net for access control enforcement in social networks. Wassell et al. (2011) used possibilistic analysis to fuse linguistic information and provided an approach that fuses contexts consisting of symbolic information for the prediction of appropriate actions in social networks. To the best of our knowledge, no work handled the link prediction problem under the uncertain framework of possibility theory.

All the presented theories are appealing frameworks to handle uncertainty. Yet, these theories do not handle all data imperfections. For instance, appropriate tools to represent states of ignorance are not provided. In social networks, one may have only partial information or even no information about the existence of link between two actors. Further, there are frequently distinct sources of information regarding the existence of a link between a pair nodes such as node neighborhood and social circles information. Such situations can be perfectly handled using the belief function theory (Dempster, 1967; Shafer, 1976).
Actually, the belief function theory permits to represent and manage complete or partial knowledge in a flexible manner. It is considered as the generalization of probability theory. Unlike probabilities, belief functions enable us to quantify partial evidence even when there is a state of ignorance. It also more general than possibility theory in its quantitative aspect. On the other hand, the belief function theory provides tools to combine information induced from different sources of evidence. As in many cases, we can have many sources of information that cannot be useful apart, thus we fuse the overall evidence to get a more valuable one.

We expect that the BFT provides a great potential to enhance link prediction performance. We present in the following basics of the belief function theory to facilitate the understanding of our proposals in the newt chapters.

### 2.4 Belief function theory

The belief function theory (BFT), also called the Dempster-Shafer theory of evidence, was initially introduced by Dempster (1967) then generalized by Shafer (1976). The BFT is a convenient theory for representing and managing uncertain knowledge. It permits to handle uncertainty and imprecision in data and manage it in a flexible way. For the most part, the BFT provides tools to efficiently represent imprecise and uncertain pieces of evidence.

On top of that, the BFT has well-defined mathematical rules to combine evidence collected from distinct sources of information. Such appealing properties attracted researchers to embrace the belief function theory in different research areas from computer science (Dubois et al., 2014; Elouedi et al., 2001; Fiche et al., 2012) to economics and finance (Autchariyapanitkul et al., 2014; Srivastava & Shafer, 1992) and many other domains.

Dempster-Shafer theory has many interpretations, one of which is the Transferable Belief Model (TBM) proposed by Smets and Kennes (1994). The TBM is a non probabilistic interpretation of the Dempster-Shafer theory with its own perceptions to the conditions of the later. In Smets’ TBM, degrees of uncertainty are modeled using a belief function (Shafer, 1976) and they are represented using an associated mass. It provides two levels for reasoning under uncertainty: the credal level where the information is managed and represented and the pignistic level where decisions are made.

Accordingly, in this section, we provide an overview of some concepts of the belief function theory under the Transferable Belief Model.
2.4.1 Frame of discernment

Let $\Omega$ be the frame of discernment, a set of mutually exclusive events associated to a given problem. The power set of $\Omega$ denoted by $2^\Omega$ is defined as follows:

$$2^\Omega = \{ A : A \subseteq \Omega \}$$

(2.1)

The power set $2^\Omega$ includes the empty set $\emptyset$ which matches the impossible proposition or the conflict.

Example 2.1. Let us treat a simple example, presented by Smets, where a murder case must be resolved. A witness was present at the murder scene and three persons were suspected. Thus, the frame of discernment related to this problem is defined as follows:

$\Omega = \{ Alice, Bob, Celine \}$

The powerset of $\Omega$ is:

$$2^\Omega = \{ \emptyset, \{ Alice \}, \{ Bob \}, \{ Celine \}, \{ Alice, Bob \}, \{ Alice, Celine \}, \{ Bob, Celine \}, \{ Alice, Bob, Celine \} \}$$

2.4.2 Basic belief assignment

A basic belief assignment (bba), denoted by $m$, represents the influence of a piece of evidence on subsets of the frame of discernment $\Omega$ and is defined as follows:

$$m : 2^\Omega \rightarrow [0, 1]$$

$$\sum_{A \subseteq \Omega} m(A) = 1$$

(2.2)

$m$ is the mass assigned to an event given a piece of evidence. $m(A)$, named a basic belief mass (bbm), describes the amount of belief assigned to the subset $A$. When $m(\emptyset) = 0$, then the bba is called a normalized basic belief assignment. When $m(A) > 0$, then $A$ is called a focal element.

Example 2.2. Assume $\Omega = \{ Alice, Bob, Celine \}$.

The bba related to a piece of evidence concerning the murder case is defined as follows:

$$m(\{ Alice \}) = 0.1;$$
\[ m(\{Alice, Bob\}) = 0.7; \]
\[ m(\{Celine\}) = 0.2. \]

For example, 0.1 represents the part of belief which exactly supports that the murderer is Alice.

The belief function theory depicts many types of imperfection. Accordingly, special bbas were defined. We have the following types:

- Certain bba: we have a certain bba when there is exactly one focal element that is a singleton:

\[ m(A) = 1 \text{ for one } A \in \Omega \]

- Bayesian bba: in this case, all the focal elements are singletons, which is the special case of probabilities:

\[ \text{if } m(A) > 0 \text{ then } |A| = 1; \text{ where } |A| \text{ stands for the cardinal of } A \]

- Consonant bba: a bba is consonant if all its focal elements are nested \( A_1 \subseteq A_2 \subseteq \cdots \subseteq A_n \).

- Categorical bba: it is a normalized bba having a unique focal element \( A \) different from \( \Omega \):

\[ m(A) = 1 \text{ for } A \subset \Omega \]

- Simple support function bba: the case where only \( \Omega \) and a subset \( A \subset \Omega \) are focal elements.

- Non-dogmatic bba: in this case, the frame of discernment (\( \Omega \)) is a focal element.

\[ m(\Omega) > 0 \]

- Vacuous bba: It is a normalized bba: \( m(\Omega) = 1 \) and \( m(A) = 0, \forall A \neq \Omega \). It represents a state of total ignorance.
2.4.3 Combination rules

Several combination rules have been proposed for the aggregation of the basic belief assignments (bba’s) that are provided from distinct sources of information provided by pieces of evidence from different experts.

**Combination of two sources of information**  Let $m_1$ and $m_2$ be two bba’s, defined on the same frame of discernment $\Omega$. Suppose that these two bba’s are collected by two distinct pieces of evidence and induced from two experts (information sources). The combination can be either conjunctive or disjunctive.

**Conjunctive rule of combination:** If the two sources of information are fully reliable, the resulting bba is computed using the conjunctive rule of combination. The result of combination quantifies the conjunction of the two pieces of evidence induced from the two sources. It is defined as follows (Smets, 1998):

$$ m_1 \cap \Delta m_2 (A) = \sum_{B,C \subseteq \Omega: B \cap C = A} m_1(B)m_2(C) \quad (2.3) $$

The conjunctive rule of combination is characterized by:

- The commutativity: $m_1 \cap \Delta m_2 = m_2 \cap \Delta m_1$
- The associativity: $(m_1 \cap \Delta m_2) \cap m_3 = m_1 \cap \Delta (m_2 \cap \Delta m_3)$
- The compositionality: $(m_1 \cap \Delta m_2)(A)$ is function of $A$, $m_1$ and $m_2$.
- The non-idempotency: $m \cap \Delta m \neq m$
- Neutral element: the neutral element is the vacuous basic belief assignment representing the total ignorance, $\forall m: m_0 \cap \Delta m = m$, with $m_0$ a vacuous bba.

**Example 2.3.** Assume the frame of discernment $\Omega = \{Alice, Bob, Celine\}$ of Example 2.1. Let two bba’s $m_1$ and $m_2$ relative to two pieces of evidence.

$$ m_1(\{Alice\}) = 0.6; $$

$$ m_1(\{Alice, Bob\}) = 0.2; $$

$$ m_1(\Omega) = 0.2; $$
Section 2.4 – Belief function theory

\[ m_2(\{Bob\}) = 0.4; \]
\[ m_2(\{Alice, Bob\}) = 0.3; \]
\[ m_2(\Omega) = 0.3. \]

When we apply the conjunctive rule of combination, we get:
\[ (m_1 \circ m_2)(\emptyset) = 0.24; \]
\[ (m_1 \circ m_2)(\{Alice\}) = 0.36; \]
\[ (m_1 \circ m_2)(\{Bob\}) = 0.16; \]
\[ (m_1 \circ m_2)(\{Alice, Bob\}) = 0.18; \]
\[ (m_1 \circ m_2)(\Omega) = 0.06. \]

**Dempster’s rule of combination:** The Dempster rule of combination (Dempster, 1967) is a normalized version of the conjunctive rule (i.e., \( m(\emptyset) = 0 \)), denoted by \( \oplus \) and defined as:

\[
m_1 \oplus m_2(A) = \begin{cases} 
\frac{m_1 \circ m_2(A)}{1 - m_1 \circ m_2(\emptyset)} & \text{if } A \neq \emptyset, \forall A \subseteq \Theta \\
0 & \text{otherwise} 
\end{cases}
\]  
\[ (2.4) \]

**Example 2.4.** Applying Dempster’s rule of combination to \( m_1 \) and \( m_2 \) (Example 2.3) on \( \Omega \) we get:
\[ (m_1 \oplus m_2)(\{Alice\}) = 0.47; \]
\[ (m_1 \oplus m_2)(\{Bob\}) = 0.21; \]
\[ (m_1 \oplus m_2)(\{Alice, Bob\}) = 0.24; \]
\[ (m_1 \oplus m_2)(\Omega) = 0.08. \]

**Disjunctive rule of combination:** The disjunctive rule of combination is used when we only know that at least one of the two pieces of evidence holds, but we do not know which one. It builds the bba representing the impact of the two pieces of evidence according to the following formula (Smets, 1998):

\[
m_1 \cup m_2(A) = \sum_{B,C \subseteq \Omega: B \cup C = A} m_1(B)m_2(C) \]  
\[ (2.5) \]
The disjunctive rule of combination is commutative $m_1 \circ m_2 = m_2 \circ m_1$ and associative $(m_1 \circ m_2) \circ m_3 = m_1 \circ (m_2 \circ m_3)$.

**Example 2.5.** When we combine $m_1$ and $m_2$ (Example 2.3) using the disjunctive rule of combination, we get:

$$(m_1 \circ m_2)(\{Alice, Bob\}) = 0.56;$$

$$(m_1 \circ m_2)(\Omega) = 0.44.$$  

### 2.4.4 Vacuous extension

Frequently, we get to fuse two bba’s $m_1$ and $m_2$ defined on two disjoint frames of discernment $\Theta$ and $\Omega$. To do this, we apply the vacuous extension mechanism. We proceed by extending the mass functions to the product space $\Theta \times \Omega = \{(\theta_i, \omega_k), \forall i \in \{1, \ldots, |\Theta|\}, \forall \omega \in \{1, \ldots, |\Omega|\}\}$. The vacuous extension operation, denoted by $\uparrow$, is defined as follows:

$$m^{\Theta \uparrow \Theta \times \Omega}(C) = \begin{cases} m^{\Theta}(A) & \text{if } C = A \times \Omega, \\ A \subseteq \Theta, C \subseteq \Theta \times \Omega \\ 0, & \text{otherwise.} \end{cases} \quad (2.6)$$

**Example 2.6.** Consider Example 2.2, after more investigations, a second witness is found. Two new suspects are considered, either Daniel or Eli. Let $\Theta = \{Daniel, Eli\}$ be the frame of the new suspects, and let $m'$ be the bba relative to the new pieces of evidence assigned as follows:

$m'(\{Daniel\}) = 0.6$

$m'(\Theta) = 0.4$.

Investigators suspect that the crime was committed by two accomplices. To fuse $m$ and $m'$, we apply a vacuous extension operation. For example, extending $\Theta$ to $\Omega$ we get:

$$m^{\Theta \uparrow \Theta \times \Omega}(\{Daniel\} \times \Omega) = 0.6;$$

$$m^{\Theta \uparrow \Theta \times \Omega}(\Theta \times \Omega) = 0.4.$$  

### 2.4.5 Multi-valued mapping

Let $\Theta$ and $\Omega$ be two disjoint frames of discernment. In order to establish the relation between them, one may use a multi-valued mapping (Dempster, 1967). In fact, a multi-valued mapping function denoted by $\tau$, permits to bring together to different frames of
Section 2.4 – Belief function theory

discernment the subsets \( B \subseteq \Omega \) that can possibly match under \( \tau \) to a subset \( A \subseteq \Theta \):

\[
m_{\tau}(A) = \sum_{\tau(B)=A} m(B)
\]  

(2.7)

Example 2.7. Consider Example 2.2 and the bba \( m' \) presented in Example 2.6. Let \( \tau \) be a multi-valued operation that rounds up events such that the mass of the event Daniel is transferred to the event Celine \( \subseteq \Omega \), the rest of evidence is transferred to \( \Omega \). As such we get the following bba:

\[
m_{\tau}(\{\text{Celine}\}) = 0.6
\]

\[
m_{\tau}(\Omega) = 0.4.
\]

2.4.6 Discounting

In practice, sources of evidence may not be absolutely reliable. A discounting method is applied to update the beliefs in order to reflect the knowledge of the sources. Consequently, a weight \( \alpha \in [0, 1] \), called a discounting factor, is assigned to the bba’s as follows:

\[
\begin{align*}
\alpha m(A) &= (1 - \alpha) \cdot m(A) \forall A \subset \Omega \\
\alpha m(\Omega) &= \alpha + (1 - \alpha) \cdot m(\Omega).
\end{align*}
\]  

(2.8)

- If the source is totally reliable then \( \alpha = 0 \).
- In the case of totally unreliable source of information \( \alpha = 1 \). This is the case of total ignorance, where all the mass is transferred to \( \Omega \) to obtain a vacuous function.

Example 2.8. Let us discount the bba \( m \) (given in Example 2.2) with a degree of reliability \( (1 - \alpha) = 0.9 \). Thus, we get:

\[
\begin{align*}
\alpha m(\{\text{Alice}\}) &= 0.9 \times 0.1 = 0.09; \\
\alpha m(\{\text{Alice}, \text{Bob}\}) &= 0.9 \times 0.7 = 0.63; \\
\alpha m(\{\text{Celine}\}) &= 0.9 \times 0.2 = 0.18; \\
\alpha m(\Omega) &= 0.1;
\end{align*}
\]

2.4.7 Reinforcement

In some cases, one may want to update a bba \( m \) by reinforcing evidence committed to a hypothesis \( A \) of the frame. Thus, a reinforcement mechanism (Mercier, Denœux, & Masson,
Reinforcement is different from discounting as evidence concerning focal elements is recollected and redistributed to the element $A$ rather than $\Omega$. Therefore, reinforcement process towards an element $A$ using to a reinforcement coefficient $\beta \in [0,1]$ is defined by:

$$
\begin{align*}
\beta m(A) &= (1-\beta) \cdot m(A) + \beta \\
\beta m(B) &= (1-\beta) \cdot m(B), \forall B \subseteq \Omega \text{ and } B \neq A.
\end{align*}
$$

Example 2.9. Consider new evidence which states that the murderer is Celine according to a rate of confidence 0.7. A reinforcement mechanism is thereby applied to the bba $m$ presented in Example 2.2. The mass $m(\{\text{Celine}\})$ is reinforced according to the reinforcement rate $\beta = 0.7$. Thus, the updated bba is obtained as follows:

$$
\begin{align*}
0.7m(\{\text{Alice}\}) &= 0.03; \\
0.7m(\{\text{Alice, Bob}\}) &= 0.21; \\
0.7m(\{\text{Celine}\}) &= 0.76.
\end{align*}
$$

2.4.8 Decision process

Decision making within the belief function is ensured by different solutions. Within the TBM (Smets, 1989; Smets & Kennes, 1994), the decision process is performed at the pig­nistic level where beliefs are represented by pignistic measures denoted by $BetP$. Beliefs, defined at the credal level, are transformed into probabilities using the pignistic transformation. Decision is made by selecting the hypothesis with the greatest pignistic probability. It is defined as follows:

$$
BetP(A) = \sum_{B \subseteq \Theta} \frac{|A \cap B|}{|B|} \frac{m(B)}{(1-m(\emptyset))}, \text{ for all } A \in \Theta
$$

Example 2.10. In order to make a decision, one can compute the pignistic probability $BetP$ corresponding to the bba $m$ (Example 2.2). Thus, we get:

$$
\begin{align*}
BetP(\{\text{Alice}\}) &= 0.45; \\
BetP(\{\text{Bob}\}) &= 0.35; \\
BetP(\{\text{Celine}\}) &= 0.2.
\end{align*}
$$

$\Rightarrow$ It is more probable that the murderer is Alice.
2.5 Belief function theory, social network analysis and link prediction

Recently, some works have been concentrated towards the belief function theory in social network analysis. For instance, for information diffusion, Ben Dhaou et al. (2014) analyzed the nature of the transmitted messages in social networks via the BFT tools. Jendoubi et al. (2014) characterized the content of social messages using an evidential algorithm to classify the propagation of information through heterogeneous social networks. C. Gao et al. (2013) identified influential nodes in weighted networks using an evidential semi-local centrality measure based on the Dempster-Shafer theory.

For sentiment analysis, Basiri et al. (2014) detected the polarity of reviews via scores aggregation to predict overall review ratings. Nishith et al. (2006) used Dempster-Shafer theory to model individuals perceptions about knowledge in social networks in order to understand knowledge in an organizations where knowledge is represented as a collection of statements that can be true or false.

In the community detection field, K. Zhou et al. (2014) identified overlapping communities in networks by combining an evidential modularity function with a spectral mapping method and the evidential c-means clustering. Further, in data aggregation, Johansson and Svenson (2014) provided a method for the automatic construction of uncertain social networks from unstructured textual data.

Attention has been drawn only recently to handle uncertainty in social network analysis. Unfortunately, not much work has been made to deal with the link prediction problem even though the link prediction task is outlined with uncertainty. First, predicting the future is already a difficult task. Second, as explained above, social data are affected with uncertainty. Thus, the LP task itself has to accommodate such imperfection.

Recently, Yin et al. (2017) proposed an evidential approach for missing link prediction based on structural similarity. The authors proposed a new measure called evidential measure based on the Dempster-Shafer theory. Links weights are used to denote the levels of trust of neighbors. Basic belief assignments are built using links’ weights and nodes’ degrees to evaluate the levels of trust. Then, nodes are represented using the latter mass functions to evaluate similarities and predict missing links. Yet, imperfection and uncertainty found in data are not treated within the network structure.

To the best of our knowledge, there are no other works that treated the link prediction problem under the belief function theory framework. In this thesis, we propose to embrace the BFT to tackle the link prediction problem as this theory provides promising tools to
handle uncertainty.

2.6 Conclusion

In this Chapter, we discussed the importance of handling uncertainty when dealing with social data. Existing works that paid attention to data imperfection in social networks were presented with a special stick out to link prediction methods. Accurate application of uncertainty theory to the link prediction problem represents the major contribution of this thesis. New graph-based models of social networks will be presented in the second part of this dissertation where uncertainty is encapsulated into the structure thanks to the belief function theory. Furthermore, novel frameworks for link prediction will be introduced that handle imperfect social data.
Part II: Contributions

Part II details our proposals. This part presents our developed frameworks to handle the link prediction problem under uncertainty. The first two Chapters describe our approaches for link prediction in uniplex social networks and their corresponding experimental evaluation. Secondly, the next two Chapters present our solutions to tackle link prediction in multiplex social networks along with an experimental analysis. All of our proposals are based on the framework of the belief function theory.
Chapter 3

Link prediction in uniplex social networks

3.1 Introduction

In the previous chapter, we have discussed the uncertainty problem in social networks. Major issues include noisy and missing social data. Actually, real-world social data contain non-relevant information added either purposely or on account of errors. Such imperfection can be handled using mathematical formalisms of uncertainty theories. With this in mind, we introduce new approaches for link prediction as possible solutions to handle social data uncertainty. Our methods are based on state-of-the-art approaches for link prediction presented in the first Chapter.

To cope with the drawbacks of social data imperfection, we propose to embrace the belief function theory as a framework for reasoning under uncertainty. As mentioned in Chapter 2, the belief function theory offers flexibility to model uncertainty and presents rich tools to quantify and manage different kinds of imperfection. Firstly, in this chapter, we present a novel graph-based model for uniplex social networks that encapsulates uncertainty within its structure. Integrating uncertainty in the social network allows us to better represent uncertain data. Belief function theory tools are used to represent uncertain information regarding the existence of links in the network. Secondly, we develop new approaches to predict new links under uncertainty. Our proposals are based on the belief function theory concepts.

The rest of the chapter details our proposals for link prediction under the uncertain framework of the belief function theory. Four new approaches that operate merely using
the BFT tools are proposed. The first method predicts links in social networks taking structural local information into account while handling uncertainty within both the graph structure and in the prediction task. The second method upgrades the first approach by taking into account social circles information for link prediction. On the other hand, the third method adds semantic information to link prediction by considering the attributes of the nodes in the prediction task. As such, both network topology and valuable information regarding the nodes are considered. As for the last method, it takes into account the assets of supervised learning to enhance link prediction under uncertainty. All methods are inspired from existing methods described in Chapter 1. We illustrate our proposals via detailed examples.

In Section 3.2, we introduce the new graphical model for social networks with uniplex structure that incorporates uncertainty at the links level. Sections 3.3, 3.4, 3.5 and 3.6 present our proposals for link prediction under the belief function framework. Each section describes a novel framework based on a different source of information for link prediction. Actually, each method is applied when particular information regarding the network is available.

## 3.2 The model: Evidential Uniplex Social Network (EUSN)

Networks are usually modeled via a classical graph representation. The formal graph representation of a social network is given by $G = (\mathcal{V}, \mathcal{E})$. The set $\mathcal{V}$ includes all the nodes depicting the social entities and $\mathcal{E}$ encompasses their connecting links that represent social relationships. Such representation assumes that nodes and edges certainly exist in the network.

Yet, as observed by Frantz et al. (2009), social data are often missing and prone to observation errors, e.g., due to inherent ambiguity of human informants, bias and dodged responses when collecting social data manually via interviews, or due to unreliability of data collecting tools and anonymization for privacy preservation concerns when the data are collected automatically. Adar and Ré (2007) asserted that one of the inherent properties of several social networks from real world is the shifted degrees of uncertainty characterizing their structures, especially the large-scale ones. Thus, the network units cannot exist with certainty in social networks given that these latter are generally very large. As a result, incorporating uncertainty into social network structures can be argued to be more substantial.
Social network analysis focuses on the inter-relatedness of social units. It is explicitly interested to the patterning of the social ties, how links are formed and what are their consequences. Most existing works assume links with binary relation values, e.g., either the value 1 to say that the link exists in the network or 0 to describe its absence. On the other hand, social network structures critically rely on the precise quality of data. Yet, such depiction does not picture the reality of social data which are usually affected with noise.

Johansson and Svenson (2014) suggested to weight links with values in $[0, 1]$ to express uncertainty degrees regarding links existence. The authors asserted that handling uncertainty in social networks is of high importance. They suggested a method that incorporates uncertainty into social networks that are automatically constructed from textual data using the belief function theory. Each link is weighted by a strength value given by probability mass function generated from a so called “ramp shaped membership function” that uses the number of words found in text between a pair of named entities. The belief function theory is principally used because it offers ways to quantify and represent uncertainty and define the degree of ignorance in the problem creating social networks automatically from textual data. On account of the fact that techniques for named entity recognition and extraction of relations are generally imperfect. Added to that, several types of relations are not explicit in unstructured text. However, the approach is only suitable to the special context of the framework. On the other hand, Ben Dhaou et al. (2014) presented what they called a “belief social network” where nodes, edges and messages are given uncertainty degrees using basic belief assignments to depict their types. The objective is to detect the nature of the exchanged messages across the network. Yet, what we are mainly concerned here is handling uncertainty concerning links existence.

From this perspective, we introduce our new model for uniplex social networks (networks with one type of relationship and only one possible link between a particular pair of nodes). Actually, since we treat the link prediction problem, it is meaningful to focus on the links structure, their existence in the network and the probability of their potential existence in a different state of the network. That is, we encode uncertainty regarding links existence in the network using mass distributions from the belief function theory.

To this end, we formally define our new graph model that we call “evidential uniplex social network” (EUSN) (Mallek et al., 2015b) where each link $uv$ has assigned a bba $m_{uv}$ defined on the frame of discernment $\Theta_{uv} = \{E_{uv}, \neg E_{uv}\}$. The hypothesis $E_{uv}$ speculates that a link exists between $u$ and $v$ and $\neg E_{uv}$ encodes its absence. The mass on $\Theta_{uv}$ depicts the degree of ignorance concerning $uv$ existence in the network. Hence, the bba $m_{uv}$ quantifies the degree of uncertainty about whether or not $uv$ exits. Our graph-based model is defined as $G(\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ and $\mathcal{E}$ are respectively the sets of nodes and edges.
Chapter 3: Link prediction in uniplex social networks

A couple \((uv, m_{uv})\) is attached to each link \(uv \in \mathcal{E}\), \(u, v \in V\), \(u \neq v\), and \(m_{uv}\) is its corresponding bba.

Figure 3.1 illustrates an example of such a graphical model. Assume that the graph describes a social network of friendships. Each node represents a person and links connect two friends. As represented in Figure 3.1, each link is weighted by a mass distribution. For example, the link between Eli and Cai has assigned a bba \(m_{EC}\) (for simplicity, we use the first letter of the nodes labels). That is, instead of coding links with binary weight values (either 1 (exist) or 0 (absent)), we build mass functions, with values in \([0, 1]\), that measure the degree of uncertainty about links existence. For clarity, a link \(uv\) is schematized in the network if its corresponding pignistic probability on the hypothesis “exists” is higher than 50%. In other terms, when \(BetP_{uv}(E_{uv}) > 0.5\), it means that the probability that the link exists is greater than 50%.

Thanks to our new graph-based model, we are able to better model uncertain social data, and encapsulate uncertainty within the network structure. In the rest of this Chapter, we present our proposals for link prediction under the uncertain framework of the belief function theory that operate using our novel model.

3.3 Link Prediction based on Local Information (uLPLI)

Motivated by state-of-the-art link prediction approaches based on network topology. We propose a new approach for link prediction which takes uncertainty into account during all the procedural steps and uses local structural information (Mallek et al., 2015b).

Our method is inspired from local information similarity-based methods as they present numerous advantages. For instance, they are simple to compute and they are not compu-
tionally costly. Yet, they do not take uncertainty into account. Accordingly, we propose in this section a novel approach for link prediction based on local information methods in uniplex social networks. In order to handle imperfect knowledge about the links in social networks, we use the belief function theory assets to build an overall mechanism that fulfills the task of predicting potential links.

3.3.1 The method: Belief Link Prediction

Our new method for link prediction, draws on methods based on local information, more specifically, the common neighbors method (Newman, 2001). The intuition of common neighbors is very interesting as it pictures a real world matter. In fact, it states that the more a pair of nodes shares common neighbors, the more its likelihood to connect. This is a real life phenomenon. For example, two persons having many common friends are very likely to become friends.

To this end, we solve the link prediction problem by using the common neighbors intuition and taking uncertainty into account. To do this, we consider the graph-based model presented in the previous Section. To predict a new link between an unlinked pair of nodes, we consider their common neighbors as independent sources of information. We build a mechanism to get neighborhood information and fuse it to get an overview about the existence of a new link. Given an observed state of the evidential uniplex social network $G(V,E)$, the task is to predict the potential existence of a link $uv$ in the new set of edges $E'$ between a pair of nodes $(u,v)$. The shared links between $u$ and $v$ and their common neighbors are considered. We detail our method in the following four steps.

3.3.1.1 Step 1: Information acquiring

Let $CN_{uw}$ be the set of common neighbors of $u$ and $v$. For each common neighbor $n_k \in CN_{uw}$, we consider the shared links between $u$ and $v$ and $n_k$: $un_k$ and $vn_k$. To consolidate the information given by the common neighbor, we apply a vacuous extension operation on the joint frame denoted by $\Theta^{N_k}$ where $\Theta^{N_k} = \Theta^{un_k} \times \Theta^{vn_k}$ using Equation 2.6.

During this first step, we bring together information given by each common neighbor across its two shared links. The vacuous extension allows us to work on a unified referential. It important to notice that extension to the joint frame is a conservative process of reallocation of evidence. That is, the degree of belief committed to $A \subseteq \Theta^{un_k}$ is reallocated to $A \times \Theta^{vn_k}$ after the application of the vacuous extension. Therefore, it minimizes any a priori on $\Theta^{un_k}$ and does not promote any hypothesis $B \subseteq \Theta^{un_k}$. Next, we combine the
Chapter 3: Link prediction in uniplex social networks

induced bba’s by applying the conjunctive rule of combination (Equation 2.3). Hence, we get the bba’s of the possible event pairs included in the product space \( \Theta^{N_k} \).

### 3.3.1.2 Step 2: Information transfer

The next step is to transfer the collected information from the neighboring nodes to the frame of discernment \( \Theta^{uv} \) of the query link \( uv \). For that, a multi-valued operation (Equation 2.7), denoted by \( \tau \), is applied. The function \( \tau: \Theta^{N_k} \rightarrow 2^{\Theta^{uv}} \) rounds up event pairs as follows:

- **Masses of the event couples with at least an element in \( \{E_{un_k}, E_{vn_k}\} \) and not in \( \{-E_{vn_k}, -E_{un_k}\} \)** are transferred to \( E_{uv} \subseteq \Theta^{uv} \) such that:
  \[
  m_\tau(\{E_{uv}\}) = \sum_{\tau(S_i) = E_{uv}} m(S_i), S_i \subseteq \Theta^{N_k} \tag{3.1}
  \]

- **Masses of the event couples with at least an element in \( \{-E_{vn_k}, -E_{un_k}\} \) and no element in \( \{E_{vn_k}, E_{un_k}\} \)** are transferred to \( -E_{uv} \subseteq \Theta^{uv} \) as:
  \[
  m_\tau(\{-E_{uv}\}) = \sum_{\tau(S_i) = -E_{uv}} m(S_i), S_i \subseteq \Theta^{N_k} \tag{3.2}
  \]

- **Masses of the event couples with at least an element in \( \{E_{vn_k}, E_{un_k}\} \) and an element in \( \{-E_{vn_k}, -E_{un_k}\} \)** are transferred to \( \Theta^{uv} \) such that:
  \[
  m_\tau(\Theta^{uv}) = \sum_{\tau(S_i) = \Theta^{uv}} m(S_i), S_i \subseteq \Theta^{N_k} \tag{3.3}
  \]

### 3.3.1.3 Step 3: Pieces of evidence fusion

Now that we transferred all information collected from the common neighbors to the frame of discernment of \( uv \). We fuse the overall evidence in order to quantify the bba \( m^{uv} \). The mass functions \( m^{uv}_{N_k} \), considering all the common neighbors, are fuse by the conjunctive rule of combination as follows:

\[
 m^{uv} = m^{uv}_{n_1} \odot m^{uv}_{n_2} \odot \ldots \odot m^{uv}_{[|CN_{uv}|]} \tag{3.4}
\]

This step is important, as it allows us to fuse evidence given by the neighbors and treat the links shared with them as independent sources of information.
3.3.1.4 Step 4: Decision making

In the final step, we evaluate our evidence about the query link $uv$. The belief function theory provides tools to make decisions according to the available evidence. For that, we compute the pignistic probability $BetP_{uv}$ using Equation 2.10. The value of the pignistic probability on the hypothesis “$uv$ exists in the network” $BetP_{uv}(E_{uv})$ allows us to decide whether $uv$ is a potential existing link. Similarity-based methods from literature rank the top query links according to similarity scores values. In the same manner, we rank pignistic probabilities on the hypothesis “exists” in decreasing order and cast the top $k$ predicted links.

Note that the presented approach is incremental. When new links are added to the network, the new evidence can be easily combined and integrated into the updated state after transferring information using the $\tau$ function. Actually, when we got new common neighbors, we go through the first and second step, then the information is fused with the existing evidence which is already processed considering the former set of common neighbors. Thus, steps 1 and 2 are not applied again for these latter.

What is more is that our belief link prediction method is generic. It operates merely using the belief function theory tools and takes structural information of the network into account. There is no reliance on the type of the social network. Hence, it is applicable to several social networks e.g., criminal networks, collaboration networks, professional networks, citation networks, etc.

Algorithm 3.1 gives the detailed algorithm of our approach. It consists of an iterative loop on the query links. Information is collected from neighbors at a first stage. Then, it is pooled to get decision regarding potential existence.
Chapter 3: Link prediction in uniplex social networks

Algorithm 3.1: Link prediction in evidential uniplex social networks

Input: Evidential uniplex social network graph $G(V,E)$, Query links $Q, k$

Output: List of predicted links

foreach $uv \in Q$ do
  get $CN_{uv}$ in $G$;
  foreach $n_k \in CN_{uv}$ do
    /* step 1 */
    apply vacuous extension to $m^u_{nk}$ and $m^v_{nk}$;
    combine bba’s conjunctively;
    /* step 2 */
    transfer bba’s using $\tau: \Theta^k \rightarrow 2^{\Theta^{uv}}$;
    /* step 3 */
    pool evidence using conjunctive rule to get $m^{uv}$;
    /* step 4 */
    compute pignistic probability $BetP^{uv}$;
  end
  rank all $BetP(E)$;
end return top $k$ links with highest $BetP(E)$;

3.3.2 Illustrative example

We explain the belief link prediction method using an illustrative example. Let the graph $G(V,E)$ presented in Figure 3.2 be a friendship network connecting persons. Assume that the network has been constructed from data gathered manually via interviews. Uncertainty may come from implicit vagueness, unreliability and bias of the questioned persons. For example, a person may say “I saw person A walking with person B”. This statement does not mean that persons A and B are linked. In addition to that, uncertainty may occur because of missing information, i.e., the questioned person is not sure whether it was A or C walking with B.

Let $ab$ be the query link which we aim to predict existence in the new state of the graph. To apply the belief link prediction approach, we first proceed by considering the common neighbors $CN_{ab}$ of $a$ and $b$ in the network. Hence, we extract the neighbors sharing links with $a$ and $b$ which mass functions on the event “not exist” is not equal one. In other words, we must not be certain that the link is absent. Assume that the bba’s are assigned as given in Table 3.1, where $\Theta^{link} = \{E^{link}, \neg E^{link}\}$. Therefore, the set of common neighbors of $(a,b)$ is $CN_{ab} = \{c, d, e\}$. The shared links with each common neighbor are: $ac, bc, ad, bd, ae$ and $be$. 

Section 3.3 – Link Prediction based on Local Information (uLPLI)

49

Figure 3.2: Snapshots of an evidential uniplex social network in given state (a) $\mathcal{G}(V, \mathcal{E})$ and in a different state (b) $\mathcal{G}(V, \mathcal{E}')$.

Table 3.1: Mass distributions of the links of the EUSN $\mathcal{G}(V, \mathcal{E})$ of Figure 3.2(a)

<table>
<thead>
<tr>
<th>Link</th>
<th>$m^{\text{link}}{(E_{\text{link}})}$</th>
<th>$m^{\text{link}}{\neg E_{\text{link}}}$</th>
<th>$m^{\text{link}}(\Theta_{\text{link}})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ac</td>
<td>0.6</td>
<td>0.1</td>
<td>0.3</td>
</tr>
<tr>
<td>bc</td>
<td>0.3</td>
<td>0.4</td>
<td>0.3</td>
</tr>
<tr>
<td>ad</td>
<td>0.42</td>
<td>0</td>
<td>0.58</td>
</tr>
<tr>
<td>bd</td>
<td>0.4</td>
<td>0.2</td>
<td>0.4</td>
</tr>
<tr>
<td>ae</td>
<td>0.2</td>
<td>0.6</td>
<td>0.2</td>
</tr>
<tr>
<td>be</td>
<td>0.4</td>
<td>0.2</td>
<td>0.4</td>
</tr>
<tr>
<td>ce</td>
<td>0.65</td>
<td>0.2</td>
<td>0.15</td>
</tr>
<tr>
<td>ey</td>
<td>0.5</td>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>yb</td>
<td>0.45</td>
<td>0.25</td>
<td>0.3</td>
</tr>
</tbody>
</table>

The first two steps presented in the previous Subsection are applied to each common neighbor. We start by the node c. The product space $\Theta^N_c = \Theta^{ac} \times \Theta^{bc}$ includes the couples \{(E_{ac}, E_{bc}), (E_{ac}, \neg E_{bc}), (\neg E_{ac}, E_{bc}), (\neg E_{ac}, \neg E_{bc})\}.

**Step 1 (node c)** We first extend the frame of discernment of the shared links $\Theta^{ac}$ and $\Theta^{bc}$ to the product space $\Theta^{N_c}$ as shown in Table 3.2. Then, the bba’s on $\Theta^{N_c}$ are fused using the conjunctive rule of combination (Equation 2.3). We get $m^{\cap}_{N_c} = m^{ac\uparrow N_c} \Theta \Theta^{bc\uparrow N_c}$ as illustrated in Table 3.3.

**Step 2 (node c)** Subsequently, we transfer the obtained bba’s considering the node c by applying a multi-valued mapping operation using the function $\tau$ (Equation 3.1, Equa-
Chapter 3 : Link prediction in uniplex social networks

Table 3.2: Vacuous extension

\[
\begin{array}{|c|c|c|}
\hline
m^{ac\uparrow N_c} & \{E_{ac}\} \times \Theta^{bc} & 0.6 \\
& \{-E_{ac}\} \times \Theta^{bc} & 0.1 \\
& \Theta^{ac} \times \Theta^{bc} & 0.3 \\
\hline
m^{bc\uparrow N_c} & \Theta^{ac} \times \{E_{bc}\} & 0.3 \\
& \Theta^{ac} \times \{-E_{bc}\} & 0.4 \\
& \Theta^{ac} \times \Theta^{bc} & 0.3 \\
\hline
\end{array}
\]

Table 3.3: Conjunctive combination \(m^{N_c}\)

\[
\begin{array}{|c|c|c|c|}
\hline
m^{N_c} & \{E_{ac}\} & \{-E_{ac}\} & \Theta^{ac} \\
\hline
\{E_{bc}\} & 0.18 & 0.03 & 0.09 \\
\{-E_{bc}\} & 0.24 & 0.04 & 0.12 \\
\Theta^{bc} & 0.18 & 0.03 & 0.09 \\
\hline
\end{array}
\]

\[\text{tion 3.2, and Equation 3.3). Thus, we get the following masses:}\]

- \(m^{ab}_c(\{E_{ab}\}) = m^{N_c}(E_{ac}, E_{bc}) + m^{N_c}(E_{ac}, \Theta^{bc}) + m^{N_c}(\Theta^{ac}, E_{bc}) = 0.45\)
- \(m^{ab}_c(\{-E_{ab}\}) = m^{N_c}(\neg E_{ac}, \neg E_{bc}) + m^{N_c}(\Theta^{ac}, \neg E_{bc}) + m^{N_c}(\neg E_{ac}, \Theta^{bc}) = 0.19\)
- \(m^{ab}_c(\Theta^{ab}) = m^{N_c}(E_{ac}, \neg E_{bc}) + m^{N_c}(\neg E_{ac}, E_{bc}) + m^{N_c}(\Theta^{ac}, \Theta^{bc}) = 0.36\)

\[\text{Similarly, we follow the same process to the common neighbor } D. \text{ The product space: } \Theta^{N_d} = \Theta^{ad} \times \Theta^{bd} = \{(E_{ad}, E_{bd}), (E_{ad}, \neg E_{bd}), (\neg E_{ad}, E_{bd}), (\neg E_{ad}, \neg E_{bd})\}.\]

**Step 1** (node d) The vacuous extension of \(\Theta^{ad}\) and \(\Theta^{bd}\) gives \(\Theta^{N_d}\) as reported in Table 3.4. Application of the conjunctive rule (Equation 2.3) provides the bba’s defined on \(\Theta^{N_d}\) shown in Table 3.5.

Table 3.4: Vacuous extension

\[
\begin{array}{|c|c|}
\hline
m^{ad\uparrow N_d} & \{E_{ad}\} \times \Theta^{bd} & 0.42 \\
& \Theta^{ad} \times \Theta^{bd} & 0.58 \\
\hline
m^{bd\uparrow N_d} & \Theta^{ad} \times \{E_{bd}\} & 0.4 \\
& \Theta^{ad} \times \{-E_{bd}\} & 0.2 \\
& \Theta^{ad} \times \Theta^{bd} & 0.4 \\
\hline
\end{array}
\]

Table 3.5: Conjunctive combination \(m^{N_d}\)

\[
\begin{array}{|c|c|c|}
\hline
m^{N_d} & \{E_{ad}\} & \Theta^{ad} \\
\hline
\{E_{bd}\} & 0.168 & 0.232 \\
\{-E_{bd}\} & 0.084 & 0.116 \\
\Theta^{bd} & 0.168 & 0.232 \\
\hline
\end{array}
\]

**Step 2** (node d) Application of the multi-valued mapping (Equation 3.1, Equation 3.2 and Equation 3.3) gives the following results:

- \(m^{ab}_d(\{E_{ab}\}) = m^{N_d}(E_{ad}, E_{bd}) + m^{N_d}(E_{ad}, \Theta^{bd}) + m^{N_d}(\Theta^{ad}, E_{bd}) = 0.568\)
Section 3.3 – Link Prediction based on Local Information (uLPLI)

- \( m_{ad}^{ab}(\{\neg E_{ab}\}) = m^{N_e}(\Theta^{ad}, \neg E_{bd}) = 0.116 \)
- \( m_{ad}^{ab}(\Theta^{ab}) = m^{N_e}(E_{ad}, \neg E_{bd}) + m^{N_e}(\neg E_{ad}, E_{bd}) + m^{N_e}(\Theta^{ad}, \Theta^{bd}) = 0.316 \)

Finally, we consider the third common neighbor \( E \). The product space: \( \Theta^N_e = \Theta^{ae} \times \Theta^{be} = \{(E_{ae}, E_{be}), (E_{ae}, \neg E_{be}), (\neg E_{ae}, E_{be}), (\neg E_{ae}, \neg E_{be})\} \).

**Step 1** (node e) The vacuous extension is reported in Table 3.6. Conjunctive combination (Equation 2.3) of the obtained bba’s is given in Table 3.7.

<table>
<thead>
<tr>
<th>( m^{ae\uparrow N_e} )</th>
<th>( {E_{ae}} \times \Theta^{be} )</th>
<th>0.2</th>
<th>( {\neg E_{ae}} \times \Theta^{be} )</th>
<th>0.6</th>
<th>( \Theta^{ae} \times \Theta^{be} )</th>
<th>0.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>( m^{be\uparrow N_e} )</td>
<td>( \Theta^{ae} \times {E_{be}} )</td>
<td>0.4</td>
<td>( \Theta^{ae} \times {\neg E_{be}} )</td>
<td>0.2</td>
<td>( \Theta^{ae} \times \Theta^{be} )</td>
<td>0.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( m^{N_e} )</th>
<th>( {E_{ae}} \times \Theta^{ae} )</th>
<th>( {\neg E_{ae}} \times \Theta^{ae} )</th>
<th>( \Theta^{ae} \times \Theta^{be} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( {E_{be}} )</td>
<td>0.08</td>
<td>0.24</td>
<td>0.08</td>
</tr>
<tr>
<td>( {\neg E_{be}} )</td>
<td>0.04</td>
<td>0.12</td>
<td>0.04</td>
</tr>
<tr>
<td>( \Theta^{be} )</td>
<td>0.08</td>
<td>0.24</td>
<td>0.08</td>
</tr>
</tbody>
</table>

**Step 2** (node e) Application of the muti-valued mapping \( \tau \) (Equation 3.1, Equation 3.2 and Equation 3.3) gives:

- \( m_e^{ab}(\{E_{ab}\}) = m^{N_e}(E_{ae}, E_{be}) + m^{N_e}(E_{ae}, \Theta^{be}) + m^{N_e}(\Theta^{ae}, E_{be}) = 0.24 \)
- \( m_e^{ab}(\{\neg E_{ab}\}) = m^{N_e}(\neg E_{ae}, \neg E_{be}) + m^{N_e}(\Theta^{ae}, \neg E_{be}) + m^{N_e}(\neg E_{ae}, \Theta^{be}) = 0.4 \)
- \( m_e^{ab}(\Theta^{ab}) = m^{N_e}(E_{ae}, \neg E_{be}) + m^{N_e}(\neg E_{ae}, E_{be}) + m^{N_e}(\Theta^{ae}, \Theta^{be}) = 0.36 \)

**Step 3** Once the first and second steps are applied to all common neighbors, information given by all common neighbors \( m_e^{ab}, m_d^{ab} \) and \( m_e^{ab} \) is fused using the conjunctive rule. The resulting evidence is given in Table 3.8.

**Step 4** At last, we evaluate the pignistic probability (Equation 2.10) of the two hypotheseses \( \{E_{ab}\} \) and \( \{\neg E_{ab}\} \):

\[
\text{Bet}P^{ab}(E_{ab}) = 0.625 \quad \text{and} \quad \text{Bet}P^{ab}(\neg E_{ab}) = 0.375
\]

Thus, there is 62% chance that the node pair \((a, b)\) connects.
### Table 3.8: Conjunctive combination $m_{ab}$

<table>
<thead>
<tr>
<th>$m_{ab}$</th>
<th>Mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>${E_{ab}}$</td>
<td>0.39</td>
</tr>
<tr>
<td>${\neg E_{ab}}$</td>
<td>0.14</td>
</tr>
<tr>
<td>$\Theta_{ab}$</td>
<td>0.04</td>
</tr>
<tr>
<td>$\emptyset$</td>
<td>0.43</td>
</tr>
</tbody>
</table>

#### 3.4 Link Prediction based on Group Information (uLPGI)

In the previous Section, we presented what motivated us to adopt similarity-based approaches to handle the link prediction problem. We introduced a new framework that draws on the common neighbors method. Obviously, the latter method has its assets, but another source of information worthy of consideration can also be used which is the participation of social entities in social groups.

In this section, we propose a framework for link prediction that takes information given by social circles into account while handling uncertainty at the same time thanks to the belief function theory tools (Mallek et al., 2015a). The main intuition is that in several real social networks, social entities with analogous experiences, interests and characteristics tend to connect more than those that do not have common properties.

#### 3.4.1 The method: Evidential Link Prediction based on Group Information

We introduce a novel framework for link prediction based on group information. Popular structural measures based on both local and group information, presented in the first Chapter, are used to predict new links. Note that this framework assumes the query links to have existing bba’s. In other terms, the goal is to update the mass distributions with pignistic probabilities on the event “exists” are lower than 50%.

A feature vector is attached to each link $uv$, it consists of local and global structural similarity measures. The feature vector is used to evaluate the correlation between the query link and its neighboring links of common circles. We extract the most similar link to use as a source of information concerning the existence of the query link. To this end, our framework for link prediction based on group information is detailed on the following
3.4 - Link Prediction based on Group Information

3.4.1 Step 1: Distance computation

The first step is to compute the Euclidean distance $D(\text{uv}, \text{u'v'})$ between the link $\text{uv}$ and each link $\text{u'v'}$ enclosed in the mutual circles. As features, we use structural similarities based on local and group information. Therefore, similarity between $\text{uv}$ and the adjacent links can be evaluated based on $D(\text{uv}, \text{u'v'})$. In the prediction task, we consider the most similar link to $\text{uv}$ which has the shortest distance. We need to divide the distance metric by its maximum value to obtain values in $[0, 1]$. It is computed as follows:

$$D(\text{uv}, \text{u'v'}) = \frac{\sqrt{\sum_{s=1}^{n} (x_{s}^{uv} - y_{s}^{u'v'})^2}}{D_{\text{max}}}$$

Where $s$ is the index of a similarity measure, $x_{uv}$ and $y_{u'v'}$ are respectively its values for $\text{uv}$ and $\text{u'v'}$. $D_{\text{max}}$ is the maximum value of the Euclidean distance used to get values in $[0, 1]$.

3.4.1.2 Step 2: Reliability computation

A discounting mechanism (Equation 2.8) allows us to measure up the reliability of the most akin link, where $\alpha = D(\text{uv}, \text{u'v'})$ is the discount coefficient given by the distance measure. The accuracy of the most similar link depends on the degree of likeliness between the two links. Hence, if the two links are similar, i.e., $D(\text{uv}, \text{u'v'})$ is equal to 0, then $\text{u'v'}$ is perceived as an entirely reliable link ($\alpha = 0$). That is, $m_{u'v'}$ is discounted as follows:

$$\begin{align*}
\alpha m_{u'v'}(\{E_{u'v'}\}) &= (1 - \alpha) \cdot m_{u'v'}(\{E_{u'v'}\}) \\
\alpha m_{u'v'}(\{-E_{u'v'}\}) &= (1 - \alpha) \cdot m_{u'v'}(\{-E_{u'v'}\}) \\
\alpha m_{u'v'}(\Theta_{u'v'}) &= \alpha + (1 - \alpha) \cdot m_{u'v'}(\Theta_{u'v'})
\end{align*}$$

$$\tau: \Theta_{u'v'} \rightarrow 2^{\Theta_{uv}}$$

It is important to note that when there are two or more similar links, i.e., comparable smallest distances, we choose the link which has the highest evidence on the hypothesis “exists” for the simple reason that the degree of certainty concerning its existence is higher.

3.4.1.3 Step 3: Information transfer and fusion

To transfer the discounted mass of the most similar link $u'v'$ to the frame $\Theta_{uv}$, we perform a multi-valued operation function (Equation 2.7) $\tau: \Theta_{u'v'} \rightarrow 2^{\Theta_{uv}}$ to draw out items such as:
• The discounted mass $\alpha_{m^{u'u'}}(\{E_{u'v'}\})$ is transferred to $m_{u'u'}^{uv}(\{E_{uv}\})$;

• The discounted mass $\alpha_{m^{u'u'}}(\{\neg E_{u'v'}\})$ is transferred to $m_{u'u'}^{uv}(\{\neg E_{uv}\})$;

• The discounted mass $\alpha_{m^{u'u'}}(\Theta_{u'v'})$ is transferred to $m_{u'u'}^{uv}(\Theta_{uv})$.

Where $\alpha = D(uv, u'v')$ and $m_{u'u'}^{uv}$ denotes the bba of $uv$ on $\Theta_{uv}$ given the most similar link, here $u'v'$. Upon transferring $\alpha_{m^{u'u'}}$ to $\Theta_{uv}^{uv}$, the bba of $uv$ is revised given the new evidence obtained from the most similar link. To do this, the original bba $m^{uv}$ and $m_{u'u'}^{uv}$ are combined using the conjunctive rule of combination (Equation 2.3). This step is fundamental, because it allows us to update evidence concerning the query link. This evidence is valuated thanks to the most similar link considered as an independent source of information.

3.4.1.4 Step 4: Decision making

Finally, the pignistic probability $BetP^{uv}(E_{uv})$ is computed (Equation 2.10) to determine the potential existence of the link $uv$ in the new state of the network. The top $k$ links with highest pignistic probabilities $BetP^{uv}(E_{uv})$ are considered.

In the following, we present the global algorithm in Algorithm 3.2 which details our approach.
Algorithm 3.2: Link prediction in evidential uniplex social networks using social circles

**Input:** Evidential uniplex social network graph with social circles $G(V, E, C)$, query links $Q$.

**Output:** List of predicted links.

```plaintext
foreach $uv \in Q$ do
    get common circles $C_{uv}$ in $G$;
    compute features of $uv$;
    foreach $u'v' \in C_{uv}$ do
        compute features of $u'v'$;
        /* step 1 */
        compute $D(uv, u'v')$;
        get most similar link that minimizes $D$;
        /* step 2 */
        revise $m^{u'v'}$ using $\alpha = D(uv, u'v') / D_{max}$;
        /* step 3 */
        transfer bba’s using $\tau$: $\Theta^{u'v'} \rightarrow 2^{\Theta^{uv}}$;
        combine $m^{uv}$ and $m^{u'v'}$ conjunctively;
        /* step 4 */
        compute pignistic probability $BetP^{uv}$;
    rank all $BetP(E)$;
return top $k$ links with highest $BetP(E)$;
```

### 3.4.2 Illustrative example

To illustrate our proposals, we consider the evidential uniplex social network presented in Figure 3.3. Nodes are grouped into overlapping communities $C_1$, $C_2$ and $C_3$. The objective is to predict the existence of a link between the unlinked pair of nodes $(a, b)$. For that, structural attributes of $ab$ are computed in order to compare them to their adjacent links in the shared circles $C_2$ and $C_3$. For instance, we use the structural measures: CN (Equation 1.1), JC (Equation 1.3), AA (Equation 1.2), CNG (Equation 1.10) and WOCG (Equation 1.11). As shown, $a_i, a_c, a_e, b_g, b_h, h_g, b_e, b_c, c_e$ and $g_i$ are the adjacent links to nodes $a$ and $b$ in the common communities. The steps of our method are applied as follows.

**Step 1** To start, the Euclidean distance of $ab$ and each adjacent link included within their common circles $C_1$ and $C_2$ is computed based on Equation 3.5. Results are shown in Table 3.9.
Chapter 3 : Link prediction in uniplex social networks

Figure 3.3: A uniplex evidential social network with overlapping social circles $C_1$, $C_2$ and $C_3$

Table 3.9: Euclidean distance between $ab$ and the links in the common circles illustrated in Figure 3.3

<table>
<thead>
<tr>
<th>Distance</th>
<th>$ac$</th>
<th>$ae$</th>
<th>$bc$</th>
<th>$be$</th>
<th>$bg$</th>
<th>$bh$</th>
<th>$ce$</th>
<th>$hg$</th>
<th>$ai$</th>
<th>$gi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$ab$</td>
<td>0.282</td>
<td>0.282</td>
<td>0.283</td>
<td>0.283</td>
<td>0.551</td>
<td>0.599</td>
<td>0.357</td>
<td>0.638</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

According to Table 3.9, $ac$ and $ae$ are the most similar links to $ab$. Yet, only one of them should be considered to update $m_{ab}$. Assume we have bba’s committed as follows:

\[
\begin{align*}
&m_{ab}(\{E_{ab}\}) = 0.35 \\
&m_{ab}(\{\neg E_{ab}\}) = 0.42 \\
&m_{ab}(\Theta_{ab}) = 0.23
\end{align*} \quad \begin{align*}
&m_{ac}(\{E_{ac}\}) = 0.65 \\
&m_{ac}(\{\neg E_{ac}\}) = 0.2 \\
&m_{ac}(\Theta_{ac}) = 0.15
\end{align*} \quad \begin{align*}
&m_{ae}(\{E_{ae}\}) = 0.55 \\
&m_{ae}(\{\neg E_{ae}\}) = 0.25 \\
&m_{ae}(\Theta_{ac}) = 0.2
\end{align*}
\]

Hence, we choose the link $ac$ as a source of evidence as $m_{ac}(\{E_{ac}\})$ is higher than $m_{ae}(\{E_{ae}\})$.

**Step 2** We proceed by evaluating the degree of reliability of the bba $m_{ac}$ using $\alpha = D(ab, ac)$ as a discounting factor. Hence, by applying the discounting operation to the bba $\alpha m_{ac}$, we get:

- $\alpha m_{ac}(\{E_{ac}\}) = (1 - 0.282) \cdot 0.65 = 0.4667$;
Step 3 Once we transfer the discounted mass of the most similar link using the $\tau$ function (Equation 2.7), evidence on the link $ab$ considering $ac$ is given as follows:

- $m_{ac}(\{E_{ab}\}) = 0.4667$;
- $m_{ac}(\{-E_{ab}\}) = 0.1436$;
- $m_{ac}(\Theta_{ab}) = 0.3897$.

Next, we update the mass of the link $ab$. For that, we combine $m_{ab}$ and $m_{ac}$ using the conjunctive rule of combination (Equation 2.3). Hence, we get:

- $m_{ab} \cap m_{ac}(\{E_{ab}\}) = 0.407$;
- $m_{ab} \cap m_{ac}(\{-E_{ab}\}) = 0.257$;
- $m_{ab} \cap m_{ac}(\Theta_{ab}) = 0.09$;
- $m_{ab} \cap m_{ac}(\emptyset) = 0.246$.

Step 4 At the final step, we compute the pignistic probability $BetP_{ab}$ (Equation 2.10) to evaluate the likelihood of existence of the link $ab$. Therefore, we get: $BetP(E_{ab}) = 0.575$ and $BetP(\neg E_{ab}) = 0.425$. That is, new evidence given by the most similar link $ac$ to $ab$ shows that there is 57% chance that $ab$ might exist. Decision concerning the predicted links is made latter according to the ranking of the top $k$ links with highest pignistic probabilities on the event “exists”.

3.5 Link Prediction based on Node Attributes (uLPNA)

In this Section, we develop a new approach for link prediction that uses a new source of information. In fact, in some cases, information about the attributes of nodes is available. This information is valuable as it provides semantic concerning social linkage formation. Even though network topology provides substantial information, semantic information regarding the network units gives rich knowledge which may certainly inform the link prediction process.
Yet, such data are not usually available. Most of the time, we do not have information concerning the questioned persons in surveys and interviews. Besides, data collected from online social networks are frequently anonymized and do not contain information about the social entities to conserve privacy. Actually, such networks appoint confidentiality policies related to privacy, ethical, legal and practical issues. For instance, Facebook users can set their privacy settings to restrict access to their personnel information. On top of that, information submitted by users is not fully reliable because they can possibly set false and misleading information, or even set up fake profiles. For these reasons, we must integrate uncertainty to resolve the link prediction problem. Therefore, we adopt the belief function theory framework.

In this section, we propose a framework that combines the attributes of the nodes with network topology and handles uncertainty to enhance the link prediction task (Mallek et al., 2017b).

3.5.1 The method: Evidential Link Prediction Using Node Attributes

We continue to draw on the intuition of local methods, we propose a link prediction using node attributes. The method takes the neighborhood of nodes and evaluates their features. Let \( u \) and \( v \) be the query unlinked nodes, the attributes of each node are compared to those of the second. This is motivated by the following idea: when a social entity is very similar to another’s entity connection, then they are likely to connect. For example, if a person is very similar to another person’s friend e.g., they are from the same hometown, work in the same company and went to the same school, then they have more chances to become friends.

We consider the evidential uniplex social network from Figure 3.1. We present in the following the detailed steps of our evidential link prediction approach based node attributes.

3.5.1.1 Step 1: Similarity assessment

To begin with, the sets of neighbors \( \mathcal{N}_u \) and \( \mathcal{N}_v \) of respectively \( u \) and \( v \) are computed and the characteristic of each node and the neighbors of the second are compared such that \( u \) is correlated to the neighbors \( v_n \in \mathcal{N}_v \) of \( v \) and vice versa. The attributes of the nodes are supposed to have categorical values without missing values. The similarity between \( node_1 \)
and node\textsubscript{2} is assessed as follows:

\[
S_{\text{node}_{1},\text{node}_{2}} = \frac{\text{# matched attributes}}{\text{# total attributes}}
\]  

(3.7)

For example, if we have three attributes: company, study, hometown, and the two nodes share the same attribute values for two features then \(S_{\text{node}_{1},\text{node}_{2}} = \frac{2}{3}\).

Then, we have to consider the most similar node to \(u\) denoted by \(v_{s}\) and that to \(v\) denoted by \(u_{s}\). On the one hand, when the number of most similar nodes exceeds two, we choose the common neighbor with the highest mass on the event “exist in the network”, otherwise it is selected randomly. On the other hand, if there is no similar node to \(u\) or to \(v\), analysis is stopped at this point for the link \(uv\).

### 3.5.1.2 Step 2: Reliability evaluation

Evidence given by the most similar nodes is revised according to their reliability using a discounting operation (Equation 2.8). If the most similar node is not a common neighbor, it is not considered as a fully reliable source. We set a discount coefficient: \(\beta = 1 - S_{\text{node}_{1},\text{node}_{2}}\) based on the similarity score. Indeed, in the case where the nodes share all attributes i.e., \(S_{\text{node}_{1},\text{node}_{2}} = 1\), then the most similar node is completely reliable i.e., \(\alpha = 0\). Accordingly, we get the discounted mass \(\alpha m^{uu}_{s}\) using Equation 2.8.

\[
\begin{align*}
\alpha m^{uu}_{s}(\{E_{uu_{s}}\}) & = (1 - \alpha) \cdot m^{uu}_{s}(\{E_{uu_{s}}\}) \\
\alpha m^{uu}_{s}(\{-E_{uu_{s}}\}) & = (1 - \alpha) \cdot m^{uu}_{s}(\{-E_{uu_{s}}\}) \\
\alpha m^{uu}_{s}(\Theta^{uu_{s}}) & = \alpha + (1 - \alpha) \cdot m^{uu}_{s}(\Theta^{uu_{s}})
\end{align*}
\]  

(3.8)

### 3.5.1.3 Step 3: Fusion and prediction

In order to be able to work on the same referential, we apply a vacuous extension (Equation 2.6) on the product space \(\Theta^{uu_{s}} \times \Theta^{vv_{s}}\). The induced bba’s are combined thanks to the conjunctive rule and we compute the global mass \(m^{P\Theta}_{\bar{S}}\) as follows:

\[
m^{P\Theta}_{\bar{S}} = m^{uu_{s} \uparrow P\Theta} \ominus m^{vv_{s} \uparrow P\Theta}
\]  

(3.9)

The obtained bba’s are then transferred to the frame \(\Theta^{uv}\) using a multi-valued mapping mechanism (Equation 2.7) according to the technique presented in subsection 3.3.1 (Step 2).
At last, to select existing links, pignistic probability are computed and are ranked in a decreasing order of confidence. Links with highest pignistic probability values on the hypothesis “exist in the network” are selected.

The algorithm of our framework is described in Algorithm 3.3.

---

**Algorithm 3.3: Link prediction in uniplex social networks using node attributes**

**Input:** Uniplex social network graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, Query links $\mathcal{Q}$, Node attributes $\mathcal{A}$, $k$

**Output:** List of predicted links

```plaintext
/* step 1 */
foreach $uv \in \mathcal{Q}$ do
    get $\mathcal{N}_u$ in $\mathcal{G}$;
    get $\mathcal{N}_v$ in $\mathcal{G}$;
    foreach $x \in \mathcal{N}_u$ do
        compute $S_{v,x}$;
        get most similar node $u_s$ to $v$;
    endforeach
    foreach $x \in \mathcal{N}_v$ do
        compute $S_{u,x}$;
        get most similar node $v_s$ to $u$;
    endforeach
    /* step 2 */
    if $u_s \notin CN_{uv}$ then
        revise $m_{us}$ using $\alpha = 1 - S_{v,u_s}$;
    if $v_s \notin CN_{uv}$ then
        revise $m_{vs}$ using $\alpha = 1 - S_{u,v_s}$;
    /* step 3 */
    apply vacuous extension to $m_{us}$ and $m_{vs}$;
    pool evidence using conjunctive rule to get $m_{uv}$;
    transfer bba’s to the frame $\Theta_{uv}$;
    compute $BetP_{uv}$;
rank all $BetP(E)$;
return top $k$ links with highest $BetP(E)$;
```
3.5.2 Illustrative example

The presented approach is illustrated using a simple example. Consider the graph presented in Figure 3.4. Suppose that we aim to evaluate the bba of the query link $ab$. First, we extract the sets of their neighbors $\mathcal{N}_a = \{c, d, e, f\}$ and $\mathcal{N}_b = \{g, f\}$. The steps of the framework are performed as follows.

![A uniplex evidential social network, nodes attributes values are given in data.](image)

Table 3.10: Attributes of the nodes presented in Figure 3.4

<table>
<thead>
<tr>
<th>Node</th>
<th>Company</th>
<th>Study</th>
<th>Hometown</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>IBM</td>
<td>Informatics</td>
<td>NY</td>
</tr>
<tr>
<td>b</td>
<td>Tesla</td>
<td>Informatics</td>
<td>Cal</td>
</tr>
<tr>
<td>c</td>
<td>IBM</td>
<td>Informatics</td>
<td>NY</td>
</tr>
<tr>
<td>d</td>
<td>Cisco</td>
<td>Marketing</td>
<td>Cal</td>
</tr>
<tr>
<td>e</td>
<td>Tesla</td>
<td>Finance</td>
<td>Cal</td>
</tr>
<tr>
<td>f</td>
<td>IBM</td>
<td>Informatics</td>
<td>NY</td>
</tr>
<tr>
<td>g</td>
<td>Tesla</td>
<td>Informatics</td>
<td>Cal</td>
</tr>
</tbody>
</table>

**Step 1** Similarity is evaluated using simple matching of the attributes of each node and the neighbors of the second presented in Table 3.4. Thus, we get:

- $S_{a,g} = \frac{1}{5}$, $S_{a,f} = 1$;
• $S_{b,c} = \frac{1}{3}$; $S_{b,d} = \frac{1}{3}$ and $S_{b,e} = \frac{2}{3}$.

Accordingly, the most similar node to $a$ is $f$ and the most similar node to $b$ is $e$.

**Step 2** Suppose that we have bba’s $m^{ae}$ and $m^{bf}$ committed as follows:

\[
\begin{align*}
    m^{ae}(\{E_{ae}\}) &= 0.55 \\
    m^{ae}(\{\neg E_{ae}\}) &= 0.20 \\
    m^{ae}(\Theta^{ae}) &= 0.25 \\
    m^{bf}(\{E_{bf}\}) &= 0.60 \\
    m^{bf}(\{\neg E_{bf}\}) &= 0.10 \\
    m^{bf}(\Theta^{bf}) &= 0.30
\end{align*}
\]

Reliability of the most similar nodes is evaluated using a discounting operation. Actually, the node $f$ is a fully reliable node as $S_{a,f}$ is equal 1, it is also a common neighbor. Thus, the bba $m^{bf}$ is not discounted. Hence, we only discount the bba $m^{ae}$ using $\alpha = 1 - S_{b,e}$ as follows:

\[
\begin{align*}
    a^{m^{ae}}(\{E_{ae}\}) &= 0.367 \\
    a^{m^{ae}}(\{\neg E_{ae}\}) &= 0.133 \\
    a^{m^{ae}}(\Theta^{ae}) &= 0.5
\end{align*}
\] $(3.10)$

**Step 3** New evidence given by the most similar nodes is extended to the product space $\mathcal{PS} = \Theta^{ae} \times \Theta^{bf}$ using a vacuous extension operation as shown in Table 3.11. Then, the obtained bba’s are combined using the conjunctive rule of combination (Equation 2.3). We get $m^{\mathcal{PS}} = m^{ae\uparrow\mathcal{PS}} \odot m^{bf\uparrow\mathcal{PS}}$ as reported in Table 3.12.

<table>
<thead>
<tr>
<th>$m^{ae\uparrow\mathcal{PS}}$</th>
<th>${E_{ae}} \times \Theta^{bf}$</th>
<th>0.367</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>${\neg E_{ae}} \times \Theta^{bf}$</td>
<td>0.133</td>
</tr>
<tr>
<td></td>
<td>$\Theta^{ae} \times \Theta^{bf}$</td>
<td>0.5</td>
</tr>
<tr>
<td>$m^{bf\uparrow\mathcal{PS}}$</td>
<td>$\Theta^{ae} \times {E_{bf}}$</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>$\Theta^{ae} \times {\neg E_{bf}}$</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>$\Theta^{ae} \times \Theta^{bf}$</td>
<td>0.30</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$m^{\mathcal{NC}}$</th>
<th>${E_{ae}}$</th>
<th>${\neg E_{ae}}$</th>
<th>$\Theta^{ae}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.22</td>
<td>0.08</td>
<td>0.3</td>
</tr>
<tr>
<td>${E_{bf}}$</td>
<td>0.037</td>
<td>0.013</td>
<td>0.05</td>
</tr>
<tr>
<td>$\Theta^{bf}$</td>
<td>0.110</td>
<td>0.04</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Application of the multi-valued mapping operation gives the following results:
3.6 Link Prediction based on Supervised Learning (uLPSL)

In this Section, we tackle the link prediction problem from another perspective using a supervised learning paradigm. As presented in the first Chapter, several methods tackled the link prediction problem using classification techniques.

We extend the $k$-nearest neighbor ($k$-NN) classification method to link prediction. The intuition is that non-observed links similar to links that already exist in the network are likely to exist. The challenge is how to check out the degree of support concerning the class memberships of the query links. Essentially, one has to pay attention to the reliability of the nearest neighbors as they cannot be considered fully trustworthy due to data imperfection. Therefore, we use the evidential $k$-nearest neighbor as a classification method that operates under the uncertain framework of the belief function theory.

We propose a framework that uses local structural information to extend the evidential $k$ nearest neighbor (EKNN) classifier (Denoeux, 1995) to handle link prediction (Mallek et al., 2017a).

3.6.1 The method: Evidential $k$-nearest neighbors Link Prediction

We aim to classify new links that do not exist in a graph network $G(V, E)$ in a given state. Let $\Theta = \{E, \neg E\}$ be the set of classes, where $E$ expresses the existence of a link in $G$ and $\neg E$ depicts its absence. The class of each edge in $E$ is supposed to be known with certainty. Let $\mathcal{T} = \{(e^i, \theta^i), \ldots, (e^{\mid E\mid}, \theta^{E\mid})\}$ be the set of single labeled links, where $e^i \in E$, $i \in \{0, \ldots, \mid E\mid\}$ and $\theta^i \in \Theta$. $\mathcal{T}$ is its corresponding class label. $\mathcal{T}$ includes couples of links and their classes.
Suppose that we aim to classify a link $e$ that connects $u$ and $v$. Let $\mathcal{L}^1$ be the set of links connecting $u$ and $v$ with their direct ($1^{st}$ level) neighbors in $\mathcal{G}$ (e.g., the links of $u$ and its neighbors in $\tau(u)$), where $|\mathcal{L}^1| = |\tau(u)| + |\tau(v)|$. In addition, let $\mathcal{L}^2$ be the set of links that are not shared with $2^{nd}$ level neighbors of the pair $(u, v)$ in $\mathcal{G}$, where $|\mathcal{L}^2| = |\tau(u)^2| + |\tau(v)^2|$. The set $\mathcal{L}$ includes $\mathcal{L}^1$ and $\mathcal{L}^2$.

Obviously, links included in the set $\mathcal{L}^1$ have classes $E$ ("exists") and those in $\mathcal{L}^2$ have classes $\neg E$ as this latter set incorporates links that are not shared with $2^{nd}$ level neighbors. Actually, for each query link, we take the neighborhood of its adjacent nodes to extract the nearest neighbors. Rather than analyzing each unobserved link with all the possible instances of the network, which is very costly, or even not possible because there are $\frac{|V| \times (|V| - 1)}{2}$ imaginable links, it is compared to only the links in the neighborhood. Accordingly, we roughly reduce the search space. On top of that, the inherent intuition of local information-based methods is naturally employed since we use neighborhood information and extract similar nodes.

To put all in a nutshell, our proposed framework uses the evidential $k$-NN classifier (Denoeux, 1995) by operating in two stages. Firstly, we evaluate similarity between the query link $e$ and its neighborhood in $\mathcal{L}$ using a distance measure. Then, we preserve the smallest $k$ distances. In the second stage, we fuse knowledge given by the $k$-nearest neighbors to get the global evidence concerning the existence of $e$. The steps are presented in the following.

### 3.6.1.1 Step 1: $k$ nearest neighbors extraction

In the first step, we have to extract the $k$ nearest neighbors. To do this, it is necessary to choose a distance to figure out the similarity between the links in the test set and those in the train set. We use the Euclidean distance, denoted by $d$. The distance between the link $e$ and its neighbors $e^i \in \mathcal{L}$, $d(e, e^i)$, is computed to assess similarities as such:

$$d(e, e^i) = \sqrt{\sum_{j=1}^{n} (s_{e}^j - s_{e^i}^j)^2} \quad (3.11)$$

where $j$ is the index of a local similarity measure, e.g, $CN$, $AA$, $JC$, $RA$ and $PA$. The parameters $s_e$ and $s_{e^i}$ are respectively its values for $e$ and $e^i$, and $n$ is the total number of structural similarities.
3.6.1.2 Step 2: Evidence pooling and class prediction

Each link \( e^i \) in \( \mathcal{L} \) induces a piece of evidence that builds up our belief about \( e \) also belonging to \( \theta^i \). However, this information does not only supply certain knowledge about the class of \( e \). In the belief function theory, this case is shaped by simple support functions, where only a part of belief is committed to \( \theta^i \) and the rest is assigned to \( \Theta \). As a result, we obtain the following bba:

\[
\begin{aligned}
    m_i(\{\theta^i\}) &= \alpha \phi(d_i) \\
    m_i(\Theta) &= 1 - \alpha \phi(d_i).
\end{aligned}
\]  

(3.12)

where \( d_i = d(e, e^i) \), \( \alpha \) is a parameter such that \( 0 < \alpha < 1 \) and \( \phi \) is a decreasing function. Actually, the closer \( e \) is to \( e^i \) according to the distance \( d \), the more probable for \( e \) to have the same class as \( e^i \). Nevertheless, when \( e \) is not close (in terms of distance) to \( e^i \), then \( e^i \) would provide little knowledge about the class of \( e \). Therefore, the function \( \phi \) must verify \( \phi(0) = 1 \) and \( \lim_{d \to \infty} \phi(d) = 0 \). The following decreasing function is proposed by the authors in (Denoeux, 1995):

\[
\phi(d_i) = e^{(-\gamma d_i^\beta)}
\]

(3.13)

where \( \gamma > 0 \) and \( \beta \in \{1, 2, \ldots\} \). \( \beta \) can be arbitrarily set to a small value, e.g., 1 or 2.

We select the \( k \) nearest nearest neighbor in \( \mathcal{L} \) to get the set \( \mathcal{K}_{near} \) of nearest neighbors. Each one is considered as an independent source of evidence concerning the class of \( e \), we obtain \( k \) bba’s that can be fused using the Dempter rule of combination (Equation 2.4). Hence, a global mass function \( m \) is obtained as such:

\[
m^e = m_1 \oplus \cdots \oplus m_k
\]

(3.14)

At the end, to determine the membership of \( e \) to one of the classes in \( \Theta \), we compare \( m(\{E_e\}) \) and \( m(\{\neg E_e\}) \). If \( m(\{E_e\}) \) is higher than \( m(\{\neg E_e\}) \) then \( e \) exists, if not, it is classified not existing.

We detail the algorithm of our framework in Algorithm 3.4.
Algorithm 3.4: Evidential $k$-NN link prediction in uniplex social networks

**Input:** Uniplex social network graph $G(V, E)$, Query links $Q$, $k$, $\sigma$, $\tau$

**Output:** List of predicted links

```plaintext
foreach $e \in Q$ do
    compute structural features of $e$; /* step 1 */
    get the list of neighboring links $L$ in $G$
    foreach $e' \in L$ do
        get the structural features of $e'$;
        compute the Euclidean distance $d(e, e')$;
        select the $k$ links that minimize $d$; /* step 2 */
    foreach $e' \in K_{\text{near}}$ do
        if $e'$ has class exists then
            $m_{e'}(\{\text{exists}\}) = \sigma \phi(d(e, e'))$
            $m_{e'}(\Theta) = 1 - \sigma \phi(d(e, e'))$;
        else
            $m_{e'}(\{\neg\text{exists}\}) = \sigma \phi(d(e, e'))$
            $m_{e'}(\Theta) = 1 - \sigma \phi(d(e, e'))$;
        pool evidence using Dempster’s rule to get $m_e$;
        compare $m_e(\{\text{exists}\})$ and $m_e(\{\neg\text{exists}\})$ and get class of $e$;
    return class of $e$;
```

### 3.6.2 Illustrative example

To illustrate our approach, consider the graph given in Figure 3.5. Assume we have the link $ab$ between the nodes $a$ and $b$ in the test set. Thus, the set of neighbors of $a$ and $b$ are respectively $\tau(a) = \{c, d, e\}$ and $\tau(b) = \{f, g, e\}$. Hence, the set of links shared with 1st neighbors is $L^1 = \{ac, ad, ae, bf, bg, be\}$. The sets of 2nd level neighbors of respectively $a$ and $b$ include $\{f, b, g\}$ and $\{a, c, d\}$. Thus, $L^2 = \{af, ac\}$. Therefore, $L = \{ac, ad, ae, bf, bg, be, af, ag, bc, bd\}$.

**Step 1** The first step is to compute local similarity measures. Table 3.13 shows structural scores values.

Next, we compute distance between $ab$ and their neighboring links. Table 3.14 reports
Figure 3.5: A uniplex social network graph

Table 3.13: Similarity measures for node pairs in the set $L$

<table>
<thead>
<tr>
<th></th>
<th>$CN$</th>
<th>$AA$</th>
<th>$JC$</th>
<th>$RA$</th>
<th>$PA$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$ab$</td>
<td>1</td>
<td>2.10</td>
<td>0.2</td>
<td>0.2</td>
<td>9</td>
</tr>
<tr>
<td>$ac$</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>$ad$</td>
<td>1</td>
<td>2.10</td>
<td>0.25</td>
<td>0.25</td>
<td>6</td>
</tr>
<tr>
<td>$ae$</td>
<td>1</td>
<td>3.32</td>
<td>0.2</td>
<td>0.2</td>
<td>9</td>
</tr>
<tr>
<td>$be$</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>$bf$</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>$bg$</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>$af$</td>
<td>1</td>
<td>3.32</td>
<td>0.25</td>
<td>0.25</td>
<td>6</td>
</tr>
<tr>
<td>$bc$</td>
<td>1</td>
<td>3.32</td>
<td>0.25</td>
<td>0.25</td>
<td>6</td>
</tr>
</tbody>
</table>

Euclidean distance values.

Table 3.14: Distance computation

<table>
<thead>
<tr>
<th>Link</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>$ab - ac$</td>
<td>3.80</td>
</tr>
<tr>
<td>$ab - ad$</td>
<td>3.00</td>
</tr>
<tr>
<td>$ab - ae$</td>
<td>1.23</td>
</tr>
<tr>
<td>$ab - be$</td>
<td>2.34</td>
</tr>
<tr>
<td>$ab - bf$</td>
<td>2.34</td>
</tr>
<tr>
<td>$ab - bg$</td>
<td>6.36</td>
</tr>
<tr>
<td>$ab - af$</td>
<td>3.24</td>
</tr>
<tr>
<td>$ab - bc$</td>
<td>3.24</td>
</tr>
</tbody>
</table>
Suppose that $k$ is equal 3. Thus, we extract the 3 nearest neighbors to $ab$ that minimize the Euclidean distance. Hence, we get the links $ae$, $be$ and $bf$.

**Step 2** Next, we get evidence from the $k$ nearest neighbors. Let $\alpha$ be fixed to 0.95, $\beta$ to 1 and $\phi$ to 0.12. Therefore, we get the following bba’s:

\[
\begin{align*}
    m_{ae}^{ab}(\{E_{ab}\}) &= 0.82 \\
    m_{ae}(\Theta) &= 0.18 \\
    m_{be}^{ab}(\{E_{ab}\}) &= 0.72 \\
    m_{be}(\Theta) &= 0.28 \\
    m_{bf}^{ab}(\{E_{ab}\}) &= 0.72 \\
    m_{bf}(\Theta) &= 0.28
\end{align*}
\]

Applying Dempster’s rule of combination we get:

\[
\begin{align*}
    m_{ab}(\{E_{ab}\}) &= 0.985 \\
    m_{ab}(\Theta) &= 0.015
\end{align*}
\]

Since $m_{ab}(\{E_{ab}\}) > m_{ab}(\{\neg E_{ab}\})$, then the class of $ab$ is “exists”. Hence, it is predicted.

### 3.7 Overview and theoretical comparison

In this section, we summarize the characteristics of the different methods presented in this Chapter. Furthermore, we report estimated algorithmic complexity of all our methods in Table 3.15.

We presented, in this Chapter, four link prediction in uniplex social network that handle uncertainty. The first three methods: uLPLI, uLPGI and uLPNA belong to the family of similarity-based link prediction approaches. By contrast to state-of-the-art methods, our approaches handle uncertainty in social data. Each approach can be applied when particular information regarding the network is included in data. The first method can be applied when we only have basic information of the network: nodes and edges. The second method can be applied when we have social circles information in data. On the other hand, we can apply the third method when we have node attributes information. Most of the time, we do not have all information in data. Therefore, we can apply each method according to the available information. The fourth method, uLPSL, belongs to the family of link prediction methods based on supervised learning. We adapted the the evidential $k$ nearest neighbor classification method to handle the link prediction problem using network topology.

Actually, it is a well-known fact that link prediction methods based on local information have the lowest computational complexity (Yang et al., 2015) compared to structural global
methods. Local measures are simple and very easy to compute. Since our approaches are based on local methods, theoretical complexity would be almost the same as these methods. All methods are based on the common neighbors approach. The complexity of the latter is $O(N.k^2)$ (F. Gao et al., 2015), where $N$ is the number of nodes and $k$ denotes the average degree of nodes in the network. Computational costs due to representing and combining belief masses are also very low since we manipulate frames of discernment with just two elements.

Table 3.15: Uniplex methods overview

<table>
<thead>
<tr>
<th>Method</th>
<th>Input</th>
<th>Output</th>
<th>Source of information</th>
<th>Theoretical complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>uLPLI</td>
<td>EUSN</td>
<td>Predicted links</td>
<td>Local information of neighboring nodes</td>
<td>$O(N.k^2)$</td>
</tr>
<tr>
<td>uLPGI</td>
<td>EUSN</td>
<td>Predicted links</td>
<td>Social circles information and local measures</td>
<td>$O(N.k^2)$</td>
</tr>
<tr>
<td>uLPNA</td>
<td>EUSN</td>
<td>Predicted links</td>
<td>Node attributes and structural local information</td>
<td>$O(N.k^2)$</td>
</tr>
<tr>
<td>uLPSL</td>
<td>uniplex SN</td>
<td>Classified links</td>
<td>Structural information</td>
<td>$O(N.k^3)$</td>
</tr>
</tbody>
</table>

### 3.8 Conclusion

In this Chapter, we presented our proposals for link prediction in uni-directional social networks. First, we introduced a novel graph-based model for uniplex social networks that incorporates uncertainty within its structure. Uncertainty regarding links existence is quantified using mass distributions. Each link has assigned a mass function encoding whether it is present or absent in the network. Accordingly, data imperfection and ambiguity regarding relationships existence is quantified and integrated into the network structure.

Subsequently, we proposed four different frameworks that handle the link prediction problem under uncertainty. The first method uses local information of nodes neighborhood. Belief function theory assets for data fusion are explored to collect and fuse information regarding the existence of unobserved links. The second approach takes social circles
information into account where actors participating to common groups are more likely to share links as they have similar interests and experiences. The third framework considers actors features by matching neighboring nodes with similar characteristics. As such, we add semantic to network topology. The fourth method extends supervised learning to link prediction and combines local topology information.

In the next Chapter, we present experiments conducted to evaluate our proposals. We use real-world social network data. In addition, we provide a technique for the generation of evidential social networks.
Chapter 4

Link Prediction in Uniplex Social Networks: Experimental Results

4.1 Introduction

In this chapter, we evaluate our link prediction frameworks in uniplex social networks. To this end, we develop programs of our algorithms using object-oriented programming language Java under the integrated development environment Netbeans. Implementation of our link prediction approaches under the belief function framework seems necessary since it allows us to get an insight concerning the feasibility of our proposals and their efficiency.

Experiments are performed on real world social network data. To the best of our knowledge, there are no social data of evidential social networks where edges are weighted by basic belief assignments encoding their existence. We present a new technique for evidential social network generation from existing social data. The proposed technique is based on graph sampling and simulation approaches from literature.

We further report the results carried out from the experiments applying our link prediction approaches on evidential social networks in order to demonstrate the applicability of our frameworks using the belief function tools. To expose the effectiveness of our proposals, we compare our methods with those from state of the art using popular evaluation measures such as precision and accuracy.

This Chapter is organized as follows: in Section 4.2, we present network pre-processing technique for the creation of belief social networks. Section 4.3 provides evaluations metrics used to measure prediction quality. Sections 4.4, 4.5 and 4.6 detail experimental results of
our four frameworks for link prediction under uncertainty.

4.2 Evidential social network creation

What we are mainly concerned with in this thesis is how to solve the link prediction problem under uncertainty accurately. Actually, to the best of our knowledge, no data of evidential social networks exist. Therefore, it is fundamental to develop a technique to generate evidential social networks to be able to test our proposals. Therefore, we present a graph generation procedure inspired from traditional methods of graph sampling. At first, social data are pre-processed by sampling the social network. A simulation procedure is subsequently conducted to transform the samples into an evidential social networks. Mass functions are simulated and attached to the links to encode their existence.

As pointed out in (Svenson, 2008), sampling mechanisms and simulation techniques are promising approaches to model and evaluate social networks with uncertain data. Actually, sampling methods are frequently practiced to deal with missing or partially observed data (Clauset et al., 2008; Koskinen et al., 2013). It is often considered that treating missing data and processing sampled data are very similar as what is not sampled can be considered as not observed. Particularly, many link tracing and mining techniques are based on network sampling (Gile & Handcock, 2010; Heckathorn, 2011).

In the following, we present a novel procedure for evidential social graph generation. It performs following two steps: (1) graphs generation then (2) mass functions simulation.

4.2.1 Graphs generation

First, we generate three graph samples from data by eliminating randomly a number of links. As such, we get three graphs that we call $G_2$, $G_1$ and $G_0$. Actually, such technique is popular in link prediction literature. Various link prediction methods prune a number of links from the network to be considered lately in the prediction process (Valverde-Rebaza & Lopes, 2013; Q.-M. Zhang et al., 2013).

4.2.2 Mass functions simulation

To create the evidential version of the social network, we attach each link $uv$ by a simulated bba that encodes its existence according to the graph samples as follows:
• If \( uv \) exists in the three graph samples \( G_2, G_1 \) and \( G_0 \) then a simple support function \( m_{uv} \) is created such that \( m_{uv}(\{E_{uv}\}) \in [2/3, 1] \);

• If \( uv \) exists in \( G_2 \) and \( G_0 \) or \( G_1 \) and \( G_0 \) then a mass \( m_{uv} \) is generated such that \( m_{uv}(\{E_{uv}\}) \in [1/3, 2/3], m_{uv}(\{\neg E_{uv}\}) \in [0, 1/3] \);

• If \( uv \) exists only in \( G_0 \) then a mass function \( m_{uv} \) is assigned such that \( m_{uv}(\{E_{uv}\}) \in [0, 1/3], m_{uv}(\{\neg E_{uv}\}) \in [1/3, 2/3] \);

• If \( uv \) does not exist in \( G_0 \) and exists in \( G_2 \) and \( G_1 \) then a simple support function \( m_{uv} \) is attached such that \( m_{uv}(\{\neg E_{uv}\}) \in [1/3, 2/3] \);

• If \( uv \) exists only in \( G_2 \) or in \( G_1 \) then a simple support function \( m_{uv} \) is generated such that \( m_{uv}(\{\neg E_{uv}\}) \in [0, 1/3] \).

At the end, we get an evidential social network graph with bba’s weighted links. Our approaches for link prediction can be henceforth applied.

4.3 Evaluation criteria

In order to evaluate the performance of our proposals, we use well-known metrics such as precision and recall. The precision computes the number of relevant predicted links \( \epsilon \) according to the total number of predicted links \( \delta \). It is defined as follows:

\[
\text{Precision} = \frac{\epsilon}{\delta}. \tag{4.1}
\]

On the other hand, recall gives the number of relevant predicted links \( \epsilon \) against the number of relevant ones \( \gamma \). It is computed as follows:

\[
\text{Recall} = \frac{\epsilon}{\gamma}. \tag{4.2}
\]

4.4 Link Prediction based on Local Information

In this Section, we present the conducted experiments to evaluate our proposals described in Chapter 3, in Section 3.3.
4.4.1 Network pre-processing

At a first stage, we pre-process data of the Facebook online social network. A component of this real-world dynamic social network composed of 10K nodes and 146K edges is considered. We obtained data from the network repository (Rossi & Ahmed, 2015). Nodes represent users and links between them are friendship relations. Edges are attached with timestamps (unix timestamp). However, some timestamps of links are missing and are committed null value.

This dataset presents a good illustration of the relevance of integrating uncertainty in social network analysis since it incorporates missing information. We apply our technique for evidential social network generation presented in Section 4.2.

Accordingly, four snapshots of the network are generated from data. Yet, since we have edges timestamps in data, graph samples are not generated randomly but rather by taking into account links time formation. Table 4.1 gives a description of the four graph samples $G(t - 2)$, $G(t - 1)$, $G(t)$ and $G(t + 1)$. One should note that $G(t + 1)$ represents the whole component of the social network.

<table>
<thead>
<tr>
<th>Graph</th>
<th>#edges with timestamps</th>
<th>#edges with missing timestamps</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G(t - 2)$</td>
<td>10,250</td>
<td>26,250</td>
</tr>
<tr>
<td>$G(t - 1)$</td>
<td>20,500</td>
<td>26,250</td>
</tr>
<tr>
<td>$G(t)$</td>
<td>30,750</td>
<td>26,250</td>
</tr>
<tr>
<td>$G(t + 1)$</td>
<td>41,000</td>
<td>105,000</td>
</tr>
</tbody>
</table>

4.4.2 Link prediction task

Once the evidential social network is constructed, we can conduct the link prediction process. We apply our method for link prediction in uniplex social networks based on local information denoted by uLPLI presented in Chapter 3, Section 3.3. In order to predict new links, we use local information based on common neighbors to quantify bba’s regarding their existence. Mass functions are transformed into pignistic probabilities to make decisions about the links existence in $t + 1$. Results are compared against $G(t + 1)$. We compare our method to the Common Neighbors (CN) approach from literature. Performance of our link prediction approach is evaluated using precision and recall.

We conduct three experiments according to three different values of analyzed links:
100K, 150K and 170K. Figure 4.1 displays precision and recall variations according to the number of correctly predicted links.

Figure 4.1: Precision and recall of uLPLI and CN on the temporal Facebook friendships network

As shown in Figure 4.1, our proposed approach for link prediction gives acceptable precision performance. Prediction accuracy measured by precision hands out values higher than 50% reaching a maximum performance of 60% when \( n = 100K \). Additionally, uLPLI gives higher prediction accuracy in terms of both precision and recall than the CN method. Therefore, we can say that the validity of the proposed framework is empirically confirmed. Taking uncertainty into account in social data enhances link prediction.

### 4.5 Link Prediction based on Group Information

In this Section, we evaluate the performance of our approach for link prediction based on social circles information presented in the previous Chapter, in Section 3.4.

#### 4.5.1 Network pre-processing

We use a real-world social network of 4K nodes and 88K edges of Facebook friendships obtained from the Stanford Large Network Dataset Collection (McAuley & Leskovec, 2012).
This dataset includes social circles information. Nodes are Facebook users connected by friendship relations. The dataset contains information such as features of nodes and social circles.

In order to get the evidential version of the Facebook dataset, we pre-process the network using the technique presented in Section 4.2.

### 4.5.2 Link prediction process

To test the performance of our link prediction approach based on group information denoted by uLPGI, we produce three different evidential versions of the social network that we call $G_1$, $G_2$ and $G_3$. Subsequently, we compare our method to the existing approach based on social groups from literature $CNG$. We measure prediction quality using precision and recall evaluation metrics.

![Facebook Friendships](image)

**Figure 4.2**: Precision and recall results of uLPGI and $CNG$ on Facebook friendships

Figure 4.2 illustrates the obtained precision and recall results for the three experiments. As shown, the uLPGI prediction results in terms of both precision and recall are higher than those of the $CNG$ approach i.e., 55% average precision towards 54% average precision for $CNG$ and 43% average recall for uLPGI towards 43% average recall for $CNG$. That is, we validate the performance of our method empirically.
4.6 Link Prediction based on Node Attributes

In this Section, we test the validity and soundness of our framework for link prediction based on node features presented in the previous Chapter, in Section 3.5.

4.6.1 Network pre-processing

To evaluate our proposals, we use a component of 1060 nodes and 10K edges of the real-world social network of Facebook friendships (McAuley & Leskovec, 2012) presented in Section 4.5.1. Actually, this dataset includes node features such as education, language, school, location, work position and location. The dataset has been anonymized by McAuley and Leskovec (2012). The evidential version of the social network is constructed following the technique presented in Section 4.2.

4.6.2 Link prediction process

To evaluate the performance of our link prediction approach based on nodes attributes (uLPNA), we measure prediction accuracy using the precision and recall. We compare our method to the classical Common Neighbor (CN) method presented in the First Chapter, in Section 1.4.2.1 and our link prediction method based on local information (uLPLI) presented in the previous Chapter, in Section 3.3. These two methods use structural properties solely. As such, we will be able to evaluate the effect of adding semantic to the task of link prediction.

Figure 4.3 shows precision and recall values of the three methods. As illustrated, the link prediction approach using nodes attributes provides accurate predictions. As it can be seen, it gives higher performances in terms of both precision and recall than the baseline methods based on network topology solely i.e., 73% precision against 67% and 52% for respectively uLPLI and CN. Obviously, considering node features in the link prediction task allows us to improve performances. Besides, one should note that the uLPNA approach takes into account both local and semi-local structural information into. Actually, the method considers direct and indirect neighbors. This joins flexibility to link similarity evaluation as opposed to the two baseline methods. Hence, validity of the framework is approved. One can claim that node attributes information enhances link prediction results as it adds a semantic connotation to network topology.
4.7 Link Prediction based on Supervised Learning

In this Section, we evaluate the performance of our framework for link prediction based on supervised learning presented in the previous Chapter, in Section 3.6.

4.7.1 Pre-processing

To evaluate our method, we perform experiments on a component of 1K nodes and 10K edges of the Facebook dataset from McAuley and Leskovec (2012) presented in Section 4.5.1. We randomly remove a partition of the links from the network construct the testing set. Hence, we prune randomly 10% of the links. The latter set of links is used in prediction. On top of that, we add a number of false links that do not exist in the network that we generate randomly. The number of false links is the same as the number of links in the 10% pruned set. The remaining 90% of the graph is used as the source graph. Results are obtained by averaging over 10 iterations with independently random divisions of testing set and training set. To reduce computation time, we pre-process the dataset by computing local similarity scores of all links in the train and test sets.

Figure 4.3: Precision and recall of uLPNA, uLPLI and CN on the Facebook network
4.7.2 Link prediction

Link prediction performance is evaluated in terms of accuracy and precision. Since the number of true links does not vary, the recall measure would have unchanging values. Thus, we only use accuracy and precision. The required parameters of the framework are $\alpha$, $\beta$, $\gamma$ and the number of nearest neighbors $k$. As reported in (Denoeux, 1995), the parameters $\alpha$ and $\beta$ do not have a great effect on evidential $k$-NN performance. Hence, as in Denoeux (1995), $\alpha$ is set to 0.95 and $\beta$ is set to 1.

In the first phase of our experiments, we test values of $k$ ranging from 1 to 15. In addition, tests conducted to optimize the $\gamma$ parameter allowed us to fix it to 0.12. We compare our link prediction method based on the evidential $k$ nearest neighbor classification method in uniplex networks (uLPSL) with the classical $k$ nearest neighbor classification approach (KNN). In the classical method, classification of a link is made according to the majority classes of its $k$ nearest neighbors. The obtained results are shown in Figure 4.4.

![Figure 4.4: Accuracy results of uLPSL and KNN according to the number of k nearest neighbors](image)

Figure 4.4 illustrates the obtained results measured by accuracy for different nearest neighbors number. As shown, both methods provide improved performances as the number of nearest neighbors $k$ increases. Actually, when we increase $k$ we pick up more sources of information regarding the class membership of test links. However, as reported in Figure 4.4, under a certain threshold, no improvement is realized. It is reasonable to think that by boosting the number of nearest neighbors too much, we get more dissimilar ones.
Chapter 4: Link Prediction in Uniplex Social Networks: Experimental Results

As a result, the committed bba’s are close to the state of total ignorance. Hence, they have no significance in fusion and prediction. Our link prediction framework based on supervised learning gives better classification performance than the standard $k$ nearest neighbors method. Our method gives better results for some values of $k$ thanks to considering implicitly information given by the sources as opposed to the standard method which considers all nearest neighbors equally trustworthy.

In a second phase of our experiments, we conduct tests by varying the number of negative links (links that do not exist in the network) to study the behavior of our framework to class imbalance scenarios. Actually, the link prediction is characterized by imbalanced classes as the number of non existing links is bigger than the existing ones. We keep the same parameters specifications as in the first phase. The number of nearest neighbors $k$ is set to 15 since it provided maximum accuracy value (Figure 4.4). Results in terms of precision are illustrated in Figure 4.5 according to different negative links number. We use precision as evaluation measure.

![Figure 4.5: Precision results of uLPSL according to the number of negative links](image)

As illustrated in Figure 4.5, precision results decrease when the number of negative links increase. One of the reasons is that more non existing edges are falsely predicted. Yet, the plot does not fall severely but quite gently reaching 77% for 10K false edges. It maintains that our approach is able to handle class imbalance issue. In addition, it considers network topology and handles uncertainty at the same time.
4.8 Conclusion

In this Chapter, we performed experiments on real-world social data to evaluate the performance of our proposals. Experimental evaluation is made using different evaluation measures: precision, recall, as well as accuracy. We used different social network data from real-world such as Facebook. In order to construct our evidential uniplex graph structure for social networks, we proposed a new technique based on graph sampling and simulation methods. We compared our results with baseline methods.

We obtained interesting results for the tested datasets indicating quite promising performances of our proposals. Compared to other methods, our approaches give better performances which clearly demonstrate their predictive power. Furthermore, since our proposals handle uncertainty in social data, they are more suitable for real-world applications.

We compared our approaches to similar methods that perform under certainty. Our four frameworks uLPLI, uLPGI, uLPNA and uLPSL all outperformed respectively baseline methods from literature $CN$, $CNG$ and $KNN$ in terms of both precision and recall. Indeed, handling uncertainty found in social data enhances link prediction results.

In the next Chapter, we provide our proposals to networks extended to multiple links between node pairs. Obviously, in the real-world, more than one type of relationship can connect two actors, e.g., friendship, family and work relationships. Such connections can be so intertwined that it is impossible to examine them individually. Therefore, information given by multiplex structures has to be handled in link prediction approaches.
Chapter 5

Link Prediction in Multiplex Social Networks

5.1 Introduction

Traditional studies in social network analysis consider relationships structure to have single types. As a result, the shifting impacts of various interpersonal interaction types of are disregarded. Real-world social systems hold rich content about social entities, their characteristics such as habits and interests, their relationships, and so on. Unquestionably, actors connect by different relationships of multiple types. Building of a suitable analysis framework requires the consideration one of the most striking features of social network which is links multiplicity.

Most traditional algorithms consider links in a homogeneous manner and disregard multiplicity. Clearly though, treating all links in the same way limits performances. To simplify multiplex structure, other methods consider multi-relational network as multiple separate single-relational networks and then apply existing approaches for uniplex structures. Sadly, by splitting the network structure, some of the information is evidently lost. With this in mind, it is important to take into full consideration relationships variety in the network.

In this chapter, we describe our proposals to handle the link prediction problem in multi-layered social networks under the belief function theory framework. First, we present a new graph-based model for multiplex social networks with uncertainty integrated into is structure. Indeed, having a more intricated structure increases social data uncertainty especially at the links level. On top of that, we extend our approaches detailed in the
previous Chapter to cope with the complex structure of multi-relational networks. New information given by the multi-relational graph structure has to be integrated into the link prediction task accurately. Apart from this, link prediction no longer only concerns the prediction of potential existence of a link but also the prediction of its type. Each proposal is illustrated using a detailed example.

In Section 5.2, a novel graph-based representation for multiplex social networks that incorporates uncertainty at the links level is introduced. Sections 5.3, 5.4, 5.5 and 5.6 give our contributions for multi-relational link prediction based on the belief function theory. A new framework is presented in each section where different sources of information for link prediction are considered. Each method extends one of our approaches presented in Chapter 3, and integrates structural multiplex topology in the link prediction task.

5.2 The model: Evidential Multiplex Social Network (EMSN)

Social networks from real-world carry on naturally multiple relation types connecting actors. Such connections can be explicit or implicit. As presented in the first Chapter, a multiplex social network is represented by a multi-relational graph $G(V, E, L)$ where $V$ is a fixed set of nodes, $E = \{E_1, \ldots, E_\lambda\}$ is the superset of $\lambda$ link types, each set contains relationships of type $i \in \{1, \ldots, \lambda\}$, and $L = \{L_1, \ldots, L_\lambda\}$ is the set of layers (|$L$| = $\lambda$), where each layer captures a class of links.

A number of key issues arise when dealing with multiplex social network. One of the most concern aspect is their complex structure. Such convoluted structure is difficult to manage as one must pay attention to all its details and features. In particular, multi-relational networks complexity may alter their properties more skeptically. For example, it would appear that extra links are duplicates produced erroneously by social networks construction. As pointed out by Sharma et al. (2014), errors concerning multiplex social networks components are expected to be larger due to unreliable experimental adjustments and technical concerns. Most studies believe that multiplex networks are already entirely cleaned.

Yet, in real-world social data, graph components are frequently noisy and ambiguous. For instance, in collaboration networks data, several authors may have the same name, in terrorist networks, we cannot be sure about the type of relationship between two persons, in social networks, the fact that two users are connected does not mean that they are friends, colleagues, family members or even if they really know each other. Even worse,
real-world data are usually missing or lack information where some links labels are absent or incorrect. As a result, one has either to delete valuable content or to consider all imperfect information in data (Kossinets, 2003). Incidentally, we must not conduct a trustworthy analysis certainly on multi-layer graphs.

In previous Chapter, we introduced an evidential uniplex graph-based model for social networks that handle uncertainty within the links. Yet, this structure only supports uni-relational homogenous relations. Given that, we extend our evidential uniplex social network graph into a multi-relational version. To this end, we define formally an evidential multiplex social network graph as $G(V, E, L)$ where $V = \{v_1, \ldots, v_{|V|}\}$ is the set of nodes, $E = \{E_1, \ldots, E_\lambda\}$ is the superset of edges with $\lambda$ being the number of link types. Each link $e_i \in E_i$ connecting a pair of nodes $(u, v)$ has assigned a bba defined on the frame of discernment $\Theta_e = \{E, \neg E\}$ denoted by $m_{e_i}$. $E$ and $\neg E$ are respectively the hypotheses speculating that $e_i$ is existing or absent in the network. The bba $m_{e_i}$ quantifies the uncertainty degree about the existence of a connection of type $i$ between $(u, v)$. $L$ is the set of layers.

![Figure 5.1: A multiplex social network graph with bba’s weighted edges](image)

Figure 5.1 presents a graph structure where social entities are connected by different connections at the same time. Links have bba’s instead of binary values (either 1 or 0). Mass distributions quantify uncertainty regarding links existence. There are three relationships types that are schematized differently. Each type depicts a particular relationship. In other terms, the network has three layers containing different links for the same number of nodes.
5.3 Multiplex Link Prediction based on Local Information (mLPLI)

Differently from traditional uni-relational approaches, where node neighborhood is organized in a homogeneous way, in multiplex networks, nodes are connected naturally via different relations. Therefore, in order to design a suitable link prediction approach that applies to multiplex social network, one has to consider links heterogeneity in the prediction task.

We presented in Chapter 3 an approach for link prediction that operates under uncertainty and draws on common neighbors method from literature. Adoption of the latter method to the multiplex version is conceivable by applying the method separately for each relation type. Yet, a meaningful feature of the whole structure of the multiplex is disregarded. In here, we extend the belief link prediction framework to handle different relationships shared between nodes (Mallek et al., 2016).

5.3.1 The method: Multi-relational Belief Link Prediction

The framework for the prediction of a link $e_i$ of type $i$ between a pair of nodes $(u, v)$ is detailed in followings. Each step is presented and illustrated at the same time.

5.3.1.1 Step 1: Information gathering and fusion

At a first stage, we consider the subgraph $G_C(V_C, E_C)$ including common neighbors of $u$ and $v$. As social entities may connect via different relationships, we consider links separately according to their types. Therefore, $G_C$ is represented by $\gamma$ graphs where each graph $G_i(V_i, E_i)$ belongs to layer $L_i \in \{1, \ldots, \gamma\}$ such that $V_C = V_i$ and $E_C = \bigcup_{i=1}^{\gamma} E_i$.

Next, in order to get information on the frame of discernment of the query link $e_i$, we apply a multivalued mapping operation (Equation 2.7). Therefore, bba’s of the links $e'_i$ shared with common neighbors are transferred to the frame of $e_i$ using the function $\tau : \Theta^{e_i} \rightarrow 2^{\Theta^{e_i}}$. As such, we get for each graph $G_i$ a mass $m_{e'_i}^{e_i}$ as follows:

- The mass $m_{e'_i}^{e_i}(\{E_{e'_i}\})$ is transferred to $m_{e_i}^{e_i}(\{E_{e_i}\})$;
- The mass $m_{e'_i}^{e_i}(\{-E_{e'_i}\})$ is transferred to $m_{e_i}^{e_i}(\{-E_{e_i}\})$;
• The mass \( m^e_i(\Theta^e_i) \) is transferred to \( m^e_{i'}(\Theta^e_{i'}) \).

Subsequently, all evidence collected from neighboring links is pooled according to the presence of common neighbors in \( G_i \). As such, we get the overall evidence regarding the existence of the link \( e_i \) denoted \( m^e_i \). However, sources of information are managed differently. For instance, when \( u \) and \( v \) are connected to all their common neighbors in the subgraph \( G_i \), evidence is combined by the conjunctive rule (Equation 2.3). As such, all sources of information are considered reliable. By contrast, when \( u \) and \( v \) are not connected to all their common neighbors in \( G_i \), i.e., there are common and uncommon neighbors, evidence is pooled using the disjunctive rule of combination (Equation 2.5). In this way, we consider that there is at least one reliable source of information. When \( u \) and \( v \) do not connect to any neighbor in \( G_i \), the latter is ignored and we get a vacuous mass function.

### 5.3.1.2 Step 2: Reliability evaluation

In this step, we are concerned with the overall reliability of the sub-graphs \( G_i \) as regards to the global graph \( G_C \). For that, the distribution \( \lambda_i \) of the common neighbors across all subgraphs \( G_i \) is computed such that \( \lambda_i = \left| \text{CN}_{uv} \right| \left| \text{CN}_{ei} \right| \). We define \( \alpha_i = 1 - \lambda_i \) as a discounting coefficient to update \( m^e_{uv} \). Therefore, we get a discounted bba \( \alpha_i m^e_i \) as follows:

\[
\begin{align*}
\alpha_i m^e_i(\{E_{e_i}\}) &= (1 - \alpha_i) \cdot m^e_i(\{E_{e_i}\}) \\
\alpha_i m^e_i(\{-E_{e_i}\}) &= (1 - \alpha_i) \cdot m^e_i(\{-E_{e_i}\}) \\
\alpha_i m^e_i(\Theta_{e_i}) &= \alpha_i + (1 - \alpha_i) \cdot m^e_i(\Theta_{e_i})
\end{align*}
\]

### 5.3.1.3 Step 3: Evidence reinforcement

The next step is to update the obtained evidence according to the structure of links in the multiplex. Consequently, we focus on the distribution of the simultaneous links in \( G(\mathcal{V}, \mathcal{E}, \mathcal{L}) \). More simply, when the nodes \( u \) and \( v \) already share a link of type \( j \neq i \). We compute the frequency of simultaneous 2-relational connections of types \( i \) and \( j \) denoted by \( S^{2}_{ij} \) towards all simultaneous relations of exactly two types in \( G \) denoted by \( S^2_{G} \).

More generally, when pair of nodes \((u, v)\) connect via \( m \leq n - 1 \) simultaneous links, we look for to the distribution \( S^{m+1}_{*j} \) where \( * = \{1, \ldots, m\} \) are the types of shared links. If \( S^{m+1}_{*j} \neq 0 \), evidence is reinforced on the hypothesis “exists” using \( \beta = \frac{S^{m}_{*j}}{S^2_{G}} \) as a reinforcement coefficient (Equation 2.9). Thus, we get the updated mass \( \beta m^{e_i} \).
5.3.1.4 Step 4: Links selection

To predict new links, pignistic probabilities $BetP_{ei}$ of query links are ranked according to the value on the element “exists”. Links with highest probabilities on the hypothesis “exists” are predicted.

Algorithm of our proposed approach is presented in Algorithm 5.1.

**Algorithm 5.1: Link prediction in evidential multiplex social networks**

**Input:** Evidential multiplex social network graph $G(V, E, L)$, Query links $Q$, $k$

**Output:** List of predicted links

```
foreach $uv_i \in Q$ do
    /* step 1 */
    get $CN_{uv}$ in $G$;
    foreach $L_i \in L$ do
        extract $G_C$;
        if $G_C$ contains at least $e'_i \in CN_{uv}$ then
            foreach $e'_i \in G_C$ do
                transfer $m^{uvw}_{ei}$ using $\tau : \Theta^{e'_i} \rightarrow 2^{\Theta^{uvw}}$;
                if all $e'_i \in CN_{uv}$ then
                    combine bba’s conjunctively to get $m^{uvw}$;
                else
                    combine bba’s disjunctively to get $m^{uvw}$;
                compute $\lambda_i = \frac{|CN_{ei}|}{|CN_{uvw}|}$;
        /* step 2 */
        revise $m^{uvw}$ using $\alpha = 1 - \lambda_i$;
        /* step 3 */
        if $(u, v)$ already share link(s) then
            reinforce $m^{uvw}$ using $\beta = \frac{S_m}{S_G}$;
        compute $BetP^{uvw}$;
    /* step 4 */
    rank all $BetP(E)$;
    return top $k$ links with highest $BetP(E)$;
```

Actually, this link prediction in evidential multiplex social networks is an extension of the uLPLI presented in Chapter 3, Section 3.3. When we have only one type relationships, when we divide the graph according the links layers, we get only one subgraph with all
presented links. Hence, bba’s are transferred and combined conjunctively as in uLPLI and we do not go through steps 2 and 3.

### 5.3.2 Illustrative example

Steps of our approach are illustrated in the following. We consider the nodes $a$ and $c$ from Figure 5.1 as query nodes. Assume that the nodes represent persons connected via three possible types of relationships, for example, friendship, co-work and connection in social media.

**Step 1** To illustrate the first step, we extract the subgraph $G_c$ including their common neighbors which are $d$, $e$ and $g$ ($b$ is not a common neighbor, thus it is not considered). We treat the subgraph $G_c$ separately in each layer of as illustrated in Figure 5.2. For example, the subgraph $G_1$ containing shared links of type 1 is represented in Layer $L_1$.

![Figure 5.2: Common neighbors' subgraphs in each layer](image)

Evidence given by neighboring links in each subgraph $G_i$ is transferred to the frame of discernment of link $ac$ that connects $a$ and $c$. Thus, we get the following bba’s for each subgraph:

\[
\begin{align*}
G_1 : m^{ac_1}_{ad_1}, m^{ac_1}_{ae_1}, m^{ac_1}_{ag_1}, m^{ac_1}_{dc_1} \\
G_2 : m^{ac_2}_{gc_2} \\
G_3 : m^{ac_3}_{ad_3}, m^{ac_3}_{ae_3}, m^{ac_3}_{ec_3}
\end{align*}
\]

The next step is to combine the transferred evidence according to common neighbors presence in the subgraphs. For example, we combine bba’s collected from $G_1$ using the
conjunctive rule of combination (Equation 2.3) because \( a \) and \( c \) connect common neighbors. That is, we get:

\[
m^{ac_1} = m^{ac_1}_{ad} \bullet m^{ac_1}_{ae} \bullet m^{ac_1}_{ag} \bullet m^{ac_1}_{gc} \bullet m^{ac_1}_{dc} \bullet m^{ac_1}_{ec}
\]

Yet, evidence collected from \( G_3 \) is combined using the disjunctive rule (Equation 2.5) as \( a \) and \( c \) have common and uncommon neighbors i.e., only \( e \) is a common neighbor, \( g \) and \( d \) are not common neighbors in this layer. Hence, evidence is pooled using the disjunctive rule:

\[
m^{ac_3} = m^{ac_3}_{ad} \bullet m^{ac_3}_{ae} \bullet m^{ac_3}_{ec}
\]

On the other hand, \( G_2 \) is not considered because \( a \) and \( c \) do not share any common neighbor in Layer \( L_2 \). Thus, we get a final bba \( (m^{ac_2}(\Theta^{ac_2}) = 1) \).

**Step 2** In Figure 5.2, \( a \) and \( c \) have three common neighbors in \( G_1 \) i.e., \( g, e \) and \( d \). Hence, \( \lambda_1 = \frac{3}{3} = 1 \). In \( G_3 \), the query nodes have only one common neighbor i.e., \( e \). Thus, \( \lambda_3 = \frac{1}{3} \). That is, reliabilities of both subgraphs \( G_1 \) and \( G_3 \) are evaluated respectively using the coefficients \( \alpha_1 = 1 - 1 = 0 \) and \( \alpha_3 = 1 - \frac{1}{3} \). That is, \( G_1 \) is fully reliable. The discounting operation gives \( \alpha_1 m^{ac_1} \) and \( \alpha_3 m^{ac_3} \).

**Step 3** As presented in Figure 5.1, \( a \) and \( c \) are not connected. Hence, we do not update the obtained bba’s after the discounting mechanism. Yet, if \( a \) and \( c \) shared one link i.e., type 2. One proceed with the reinforcement step. Thus, we consider simultaneous 2-relational associations in the multiplex. As shown in Figure 5.1, there are five 2-relational connections i.e., shared between the node pairs \( (a,b) \), \( (a,e) \), \( (a,d) \), \( (e,c) \), and \( (g,c) \). Therefore, we compute the distribution of simultaneous 2-relational connections of types 1 and 2 i.e., \( S^{12} = 2 \) and the types 2 and 3 i.e., \( S^{23} = 0 \) towards all 2-relational connections the multiplex i.e., \( S^{2} = 5 \). Consequently, the discounted masses obtained the second step \( \alpha_1 m^{ac_1}_1 \) and \( \alpha_3 m^{ac_3}_3 \) are reinforced towards the hypothesis “exists” using \( \beta_1 = \frac{S^{12}}{S^{2}} = \frac{2}{5} \) and \( \beta_3 = \frac{S^{23}}{S^{2}} = \frac{0}{5} \) respectively as reinforcement rates. Finally, we get the masses \( \beta_1 m^{ac_1}_1 \) and \( \beta_3 m^{ac_3}_3 \).
5.4 Multiplex Link Prediction based on Group Information (mLPGI)

In the previous Chapter, we presented a method to predict links in evidential uniplex social networks using group information. Now, we are interested to multi-relational social networks. In fact, the direct application of the proposed method from the uniplex case is not interesting as it inquires to carry out the same process for each relationship type separately.

Considering the multi-relational structure implies new challenges. First, information given by communities structure has to be considered carefully as it provides a relevant source of information that may enhance prediction. For example, the frequency of simultaneous links in each circle gives information about the relevance of particular interactions in the group. Besides, links structure in the whole graph is significant as it tells about the layering of relationships types. The challenge is to find a way to take advantage of both multiplex and social group information at the same time.

In the following, we present a framework for link prediction in evidential multiplex social network that takes into account social circles information. The proposed approach operates using the belief function theory tools. We assume that the social network has natural group structure as in real-world social networks, users are regrouped into social circles that connects common interests, beliefs, geographic proximity, education and so on. We consider an evidential multiplex graph $G(V, E, L, C)$ where $C$ is the set of communities.

5.4.1 The method: Evidential multi-relational Link Prediction Using Group Information

Our framework for multi-relational link prediction consists of five steps. To predict a link $e_i$ of type $i$ between a pair of nodes $(u,v)$, we consider links in the common shared groups as sources of information. Similarities are evaluated using structural measures based on node neighborhood and group information. We subsequently update evidence according the similarities rate and multiplex information. Then, global evidence is pooled to get an overall information regarding the existence of $e_i$. In the following, we detail the steps of our method and we provide an illustrative example to facilitate understanding of our proposals.
5.4.1.1 Step 1: Similarity evaluation

At a first stage, we consider the neighborhood of the nodes $u$ and $v$ in common groups. Generally, social entities that participate to social circles have common features. When two pairs of entities share the same features, they might share the same connections as well. Based on this intuition, we compute similarity between $e_i$ and the links in common groups of $u$ and $v$ using uniplex and multiplex local and group information scores for node pairs that share at least a link of type $i$. We use the Euclidean distance to evaluate similarities between $e_i$ and each neighboring link $e'_i$ in the common groups. As such, both uniplex and multiplex information is considered. It is computed as follows:

$$
\begin{align*}
    d(e_i, e'_i) &= \sqrt{\sum_{q=1}^{Q} (s^q_{e_i} - s^q_{e'_i})^2}
\end{align*}
$$

where $q$ is the index of uniplex or multiplex structural score computed in Layer $L_i$ or in $G$, $Q$ is the total number of structural scores, and $s_{e_i}$ and $s_{e'_i}$ are respectively its values for $e_i$ and $e'_i$.

Upon evaluating similarities, we consider the most similar link $e'_i$ to $e_i$ that minimizes the distance.

5.4.1.2 Step 2: Evidence discounting and transfert

Next, as we cannot consider the most similar link a completely trustworthy information. We revise evidence given by $e'_i$ using the dissimilarity measure as discount coefficient for discounting mechanism. Thus, we discount the mass $m^{e'_i}$ of the link $e'_i$ using the discounting rate $\alpha = d(e_i, e'_i)$

$$
\begin{align*}
    a m^{e'_i}(\{E_{e'_i}\}) &= (1 - \alpha) \cdot m^{e'_i}(\{E_{e'_i}\}) \\
    a m^{e'_i}(\{\neg E_{e'_i}\}) &= (1 - \alpha) \cdot m^{e'_i}(\{\neg E_{e'_i}\}) \\
    a m^{e'_i}(\Theta_{e'_i}) &= \alpha + (1 - \alpha) \cdot m^{e'_i}(\Theta_{e'_i})
\end{align*}
$$

Then, we transfer the obtained bba $m^{e'_i}$ to the frame of of the link $e_i$ using a multi-valued mapping operation.

- The discounted mass $a m^{e'_i}(\{E_{e'_i}\})$ is transferred to $m^{e'_i}(\{E_{e_i}\})$;
The discounted mass $\alpha m^e_i(\{-E_{e_i}\})$ is transferred to $m^e_i(\{-E_{e_i}\})$;

- The discounted mass $\alpha m^{\Theta e'_i}(\Theta e'_i)$ is transferred to $m^{\Theta e'_i}(\Theta e'_i)$.

### 5.4.1.3 Step 3: Multiplex information integration

In this step, we take into consideration multi-relational information in the social groups and in the multiplex graph. If $(u,v)$ already share $\Delta$ links, we compute the distribution $S_{CG}^{\Delta+1}$ of simultaneous links of types $\gamma$ shared between $u$ and $v$ in the common groups, towards the number of simultaneous links $|\Delta| + 1$ of different types in the common groups $S_{CG}^{\Delta+1}$. In addition, we compute the distribution of all simultaneous links of types $\gamma$ towards the number of simultaneous links $|\Delta| + 1$ of different types in the multiplex $S_{G}^{\Delta+1}$. Then, we perform an operation of reinforcement towards the element “exists” using $\beta = \left( \frac{S^{\Delta+1}_{CG}}{S^{\Delta+1} + \frac{S_{CG}^{|\Delta|+1}^{\Delta+1}}{S_{G}^{\Delta+1}}} \right) / 2$ as a reinforcement rate. We get the reinforced bba $\beta m^e_i$. As such, we integrate multiplex information in both common communities and the global graph in the prediction task.

### 5.4.1.4 Step 4: Evidence fusion

Next, we pool new evidence $\beta m^e_i$ regarding the existence of the link $e_i$ and the earlier bba $m^e_i$ speculating that $e_i$ does not exist in the network using the conjunctive rule of combination. We get the final mass $m^f_i$ as follows:

$$m^f_i = \beta m^e_i \bigcap m^e_i$$

\[(5.4)\]

### 5.4.1.5 Step 5: Links selection

Decision is made according to the value of the pignistic probability on the hypothesis “exists”. All non-observed links are ranked in a decreasing order according to pignistic probabilities, and the links with highest values are considered of higher existence likelihoods.

The algorithm of our proposed framework is described in Algorithm 5.2.
**Algorithm 5.2**: Link prediction in evidential multiplex social networks with social circles

**Input**: Evidential multiplex social network graph $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{L}, \mathcal{C})$, Query links $Q$, $k$

**Output**: List of predicted links

```plaintext
foreach $uv_i \in Q$ do
    get common circles $C_{uv}$ in $\mathcal{G}$;
    compute features of $uv$;
    /* step 1 */
    foreach $u'v'_i \in C_{uv}$ do
        compute features of $u'v'_i$;
        compute $d(uv_i, u'v'_i)$;
        get most similar link that minimizes $d$;
    /* step 2 */
    discount $m^{u'v'_i}$ using $\alpha = d(uv_i, u'v'_i)/d_{max}$;
    transfer bba using $\tau$: $\Theta^{u'v'_i} \rightarrow 2^{\Theta^{uv_i}}$;
    /* step 3 */
    if $(u, v)$ already share link(s) then
        $\beta = (\frac{s^{uv}_{\Delta+1}}{s_{CG}^{uv}} + \frac{s^{u'v'_i}_{\Delta+1}}{s_G^{\Delta+1}})/2$;
        reinforce $m^{uv_i}$ using $\beta$;
    /* step 4 */
    combine $m^{uv_i}$ and $m^{u'v'_i}$ conjunctively;
    /* step 5 */
    compute pignistic probability $BetP_{uv}$;
    rank all $BetP(E)$;
    return top $k$ links with highest $BetP(E)$;
```

This link prediction for evidential multiplex social networks based on social circles is extended from the uLPGI approach presented in Chapter 3, Section 3.4. When we have homogenous relationships, we do not go through step 3, thus, we get the uLPGI method.

### 5.4.2 Illustrative example

We illustrate our framework using a detailed example. Consider the evidential multi-relational social network graph presented in Figure 5.3. The task is to predict the existence of a link $uv_1$ between the nodes $u$ and $v$. The latter nodes have the groups $C_1$ and $C_2$ in
Step 1 First of all, one has to compute uniplex and multiplex similarity scores for node pairs connected by at least a link of type 1 in the common groups, i.e., $CN$, $AA$, $JC$, $CNG$ and $WOCG$. Table 5.1 reports structural scores values.

<table>
<thead>
<tr>
<th></th>
<th>$CN_{uni}$</th>
<th>$CN_{multi}$</th>
<th>$AA_{uni}$</th>
<th>$AA_{multi}$</th>
<th>$JC_{uni}$</th>
<th>$JC_{multi}$</th>
<th>$CNG_{uni}$</th>
<th>$CNG_{multi}$</th>
<th>$WOCG_{uni}$</th>
<th>$WOCG_{multi}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$uv$</td>
<td>3</td>
<td>5</td>
<td>8.73</td>
<td>9.5</td>
<td>0.50</td>
<td>0.83</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>$ux$</td>
<td>0</td>
<td>1</td>
<td>0.00</td>
<td>1.43</td>
<td>0.00</td>
<td>0.14</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$uy$</td>
<td>0</td>
<td>1</td>
<td>0.00</td>
<td>1.43</td>
<td>0.00</td>
<td>0.14</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$uz$</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$vz$</td>
<td>1</td>
<td>1</td>
<td>3.32</td>
<td>2.10</td>
<td>0.14</td>
<td>0.13</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$zw$</td>
<td>1</td>
<td>1</td>
<td>1.43</td>
<td>1.00</td>
<td>0.25</td>
<td>0.25</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$vw$</td>
<td>1</td>
<td>1</td>
<td>2.10</td>
<td>1.43</td>
<td>0.17</td>
<td>0.14</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$yt$</td>
<td>1</td>
<td>1</td>
<td>1.43</td>
<td>1.00</td>
<td>0.33</td>
<td>0.17</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$vx$</td>
<td>0</td>
<td>1</td>
<td>0.00</td>
<td>1.43</td>
<td>0.00</td>
<td>0.13</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$vt$</td>
<td>0</td>
<td>1</td>
<td>0.00</td>
<td>1.66</td>
<td>0.00</td>
<td>0.11</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Next, we assess similarity using the Euclidean distance where similarity scores are used as features. Table 5.2 gives the distances values.
The Euclidean distance is divided by its maximum value to get normalized values in [0,1].

<table>
<thead>
<tr>
<th>Link</th>
<th>Distance</th>
<th>$d/d_{max}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$uv - ux$</td>
<td>14.08</td>
<td>0.90</td>
</tr>
<tr>
<td>$uv - uy$</td>
<td>14.08</td>
<td>0.90</td>
</tr>
<tr>
<td>$uv - uz$</td>
<td>15.71</td>
<td>1.00</td>
</tr>
<tr>
<td>$uv - vz$</td>
<td>11.31</td>
<td>0.72</td>
</tr>
<tr>
<td>$uv - zw$</td>
<td>13.01</td>
<td>0.83</td>
</tr>
<tr>
<td>$uv - vw$</td>
<td>12.37</td>
<td>0.79</td>
</tr>
<tr>
<td>$uv - yt$</td>
<td>13.01</td>
<td>0.83</td>
</tr>
<tr>
<td>$uv - vx$</td>
<td>14.08</td>
<td>0.90</td>
</tr>
<tr>
<td>$uv - vt$</td>
<td>13.95</td>
<td>0.89</td>
</tr>
</tbody>
</table>

As shown in Table 5.2, the most similar link to $uv_1$ is the link $vz_1$ linking the pair of nodes $(v,z)$.

**Step 2** Assume that we have bba’s defined as follows:

$$
\begin{align*}
    m^{uv_1}(\{E_{uv_1}\}) &= 0.30 \\
    m^{uv_1}(\neg E_{uv_1}) &= 0.45 \\
    m^{uv_1}(\Theta) &= 0.25
\end{align*}
\quad\text{and}\quad
\begin{align*}
    m^{vz_1}(\{E_{vz_1}\}) &= 0.60 \\
    m^{vz_1}(\neg E_{vz_1}) &= 0.15 \\
    m^{vz_1}(\Theta) &= 0.25
\end{align*}
$$

Since $vz_1$ is not a fully reliable source of information i.e., $d(uv,vz) \neq 0$, we apply a discounting operation to revise evidence given the most similar link. Hence, the bba $m^{vz_1}$ is discounted using $\alpha = d(uv,vz) = 0.66$ as a discounting rate. Thus, we get:

$$
\begin{align*}
    \alpha m^{vz_1}(\{E_{vz_1}\}) &= 0.168 \\
    \alpha m^{vz_1}(\neg E_{vz_1}) &= 0.042 \\
    \alpha m^{vz_1}(\Theta) &= 0.79
\end{align*}
$$

We transfer $\alpha m^{vz_1}$ to the frame $\Theta^{uv_1}$ using a multi-valued mapping operation, we get the bba $m^{uv_1}_{vz_1}$ given the most similar link:
\[ m_{vz1}^u(\{E_{vz1}\}) = 0.168 \]
\[ m_{vz1}^u(\{-E_{vz1}\}) = 0.042 \]
\[ m_{vz1}^u(\Theta) = 0.79 \]

\section*{Step 3}
Suppose that the pair \((u, v)\) already share one link of type 3. Thus, we update our evidence thanks to this information using a reinforcement mechanism such that:

- The distribution of simultaneous links of types 3 and 1 shared between \(u\) and \(v\) in the common groups \(S_{3CG}^2 = 3\).

- The distribution of simultaneous links of types \(\gamma\) shared between \(u\) and \(v\) in the common groups, towards the number of 2-simultaneous links of different types in the common groups is \(S_{CG}^2 = 4\).

- The distribution of all simultaneous links of types 3 and 1 towards the number of 2-simultaneous links of different types in the multiplex is \(S_{31G}^2\). We assume it is equal to 14.

- The distribution of all 2-simultaneous links of different types in the multiplex is \(S_G^2\). We assume it is equal to 35.

Thus, we reinforce the bba \(m_{vz1}^u\) on the hypothesis “exists” using the reinforcement rate \(\beta = 0.575\).

\[ \beta m_{vz1}^u(\{E_{vz1}\}) = 0.65 \]
\[ \beta m_{vz1}^u(\{-E_{vz1}\}) = 0.02 \]
\[ \beta m_{vz1}^u(\Theta) = 0.33 \]

\section*{Step 4}
The next step is to combine \(m_{vz1}^u\) and \(m_{vz1}^u\) using the conjunctive rule of combination. Assume that \(m_{vz1}^u\) is distributed as follows:

\[ m_{vz1}^u(\{E_{vz1}\}) = 0.30 \]
\[ m_{vz1}^u(\{-E_{vz1}\}) = 0.45 \]
\[ m_{vz1}^u(\Theta) = 0.25 \]

Hence, we get the final bba as follows:
\[
\begin{align*}
& m_{uv}^1(\{E_{vz_1}\}) = 0.4565 \\
& m_{uv}^1(\{\neg E_{vz_1}\}) = 0.1625 \\
& m_{uv}^1(\Theta) = 0.0825 \\
& m_{uv}^1(\emptyset) = 0.2985
\end{align*}
\]

**Step 5** Finally, the pignistic probability is computed, we get \(BetP_{uv}(E) = 0.65\). Thus, there is 65% chance that the link exists in the network.

### 5.5 Multiplex Link Prediction based on Node Attributes (mLPNA)

Social networks include rich content regarding the social entities. Each entity is characterized by a set of features that might be personnel (i.e., age, gender, ethnicity), behavioral (i.e., physical activity, drinking or smoking behavior), beliefs, attitudes, or opinions (i.e., political preferences, sport). Such patterns certainly enhance link prediction because generally there is a relation between social entities features. For example, two persons that share approximately the same personnel variables and have similar behavior and interests may be connected. Even more, obviously, there is a relationship between the entities attributes and the existence of a link of a particular type. For example, two colleagues work in the same company and have lunch in the same restaurant, or two classmates study in the same university, have the same age, same behavior and interests. Hence, such information adds semantic to the prediction task. Nodes attributes may enhance the prediction process, they can give insight about the links existence and type.

We presented in the previous Chapter, a method for predicting links using node attributes. However, the proposed approach applies to uniplex structure. The idea is to extend this method to multiplex social networks and handle multi-relational information in the prediction task by taking uncertainty into account.

#### 5.5.1 The method: Evidential Multi-relational Link Prediction Using Node Attributes

Our proposals for evidential multi-relational link prediction using node attributes draw on node neighborhood based methods from literature. We extend our method for evidential link prediction using node attributes to the multiplex version. We consider the evidential
multiplex social network graph presented in the second Section of this Chapter. We take into account attributes of nodes and structural information regarding the multiplex graph. Indeed, having more than one connection layer provides additional information about the structure of the network compared to a classical uni-relational structure. We integrate such information in the link prediction task. Consequently, we consider both semantic and structural information the link task. To predict the existence of a link \( uv \) of type \( i \), we go through four steps. These steps are detailed in the following.

5.5.1.1 Step 1: Evaluating similarity

At first, we consider the sets of neighbors \( \tau(u) \) and \( \tau(v) \) connected to \( u \) and \( v \) with at least a link of type \( i \). We compare the attributes of \( u \) to those of the nodes in \( \tau(v) \), and we compare the attributes of \( v \) to those of the nodes in \( \tau(u) \). The similarity between \( \text{node}_1 \) and \( \text{node}_2 \) is evaluated as follows:

\[
S_{\text{node}_1,\text{node}_2} = \frac{\# \text{matched attributes}}{\# \text{total attributes}} \tag{5.5}
\]

Using the similarities values, we consider the most similar node \( u' \in \tau(u) \) and \( v' \in \tau(v) \) to respectively \( v \) and \( u \).

5.5.1.2 Step 2: Reliability evaluation

Next, we revise the collected evidence using a discounting operation. The bba’s \( m^{uu'}_i \) and \( m^{vv'}_i \) are discounted respectively using \( \alpha' = 1 - S_{uv'} \) and \( \alpha'' = 1 - S_{vu'} \). When the most similar node is not a common neighbor to \( u \) and \( v \), it is not considered as a fully trustworthy source of evidence. For example, for \( m^{uu'}_i \), the discounting operation is computed as follows:

\[
\begin{align*}
\alpha' m^{uu'}_i(\{E_{uu'}\}) &= (1 - \alpha') \cdot m^{uu'}_i(\{E_{uu'}\}) \\
\alpha' m^{uu'}_i(\neg E_{uu'}) &= (1 - \alpha') \cdot m^{uu'}_i(\neg E_{uu'}) \\
\alpha' m^{uu'}_i(\Theta_{uu'}) &= \alpha' + (1 - \alpha') \cdot m^{uu'}_i(\Theta_{uu'}) \tag{5.6}
\end{align*}
\]

5.5.1.3 Step 3: Information pooling

To work on the same referential, a vacuous extension (Equation 2.6) is performed on the product space \( PS = \Theta_{uu'} \times \Theta_{vv'} \). The obtained bba’s are combined using the conjunctive
rule. We compute the global mass $m_{PS}^\cap$ as follows:

$$m_{PS}^\cap = m_{uu}^i \cap m_{vv}^i \cap$$ \hspace{0.5cm} (5.7)

Next, we transfer the obtained bba’s to the frame $\Theta_{uv}$ using a multi-valued mapping mechanism (Equation 2.7) according to the method presented in Subsection 3.3.1, in Chapter 3.

5.5.1.4 Step 4: Evidence update

If $u$ and $v$ have common neighbors, we compute the rate of common neighbors connected by the links of type $i$ and the total number of common neighbors $\delta_{uv} = \frac{|CN_{uv}|}{|CN_{tot}|}$. Then, we discount $m_{uv}^i$ using $\beta = 1 - \delta_{uv}^i$ to get $\beta m_{uv}^i$.

On the other hand, if $u$ and $v$ already share $m$ link(s) ($m \in [1, \gamma - 1]$) ($\gamma$ is the number of link types), compute the rate $S_{m+1}^i$ of simultaneous links of more than two types in the multiplex, where $* = \{1, \ldots, m\}$ are the types of the shared links. If $S_{m+1}^i \neq 0$, we reinforce the mass on the element “exists” using $\phi = \frac{S_{m+1}^i}{S_{m}^{*}}$ as a reinforcement rate to obtain the mass $\gamma m_{uv}^i$.

5.5.1.5 Step 5: Links selection

Finally, we compute the pignistic probability $BetP_{uv}^i(E)$. Links selection is made on the basis of the ranking of the highest probability values on the element “exists”.

The algorithm of our proposed framework is described in Algorithm 5.3.

This link prediction method in multiplex social networks based on node attributes is an extension of the uLPNA approach presented in Chapter 3, Section 3.5. When the social network is composed of only one type of relationships, we do not apply step 4. Hence, we get the uLPNA aproach.
Algorithm 5.3: Link prediction in multiplex social networks using node attributes

Input: Multiplex social network graph $G(\mathcal{V}, \mathcal{E}, \mathcal{L})$, Query links $Q$, Node attributes $A$, $k$

Output: List of predicted links

/* step 1 */
foreach $uv_i \in Q$ do
    get $N_u$ in $G$;
    get $N_v$ in $G$;
    foreach $x \in N_u$ do
        compute $S_{v,x}$;
        get most similar node $u'$ to $v$;
    foreach $x \in N_v$ do
        compute $S_{u,x}$;
        get most similar node $v'$ to $u$;
/* step 2 */
if $u' \notin CN_{uv}$ then
    Revise $m_{uu'}$ using $\beta = 1 - S_{v,u'}$;
if $v' \notin CN_{uv}$ then
    revise $m_{vv'}$ using $\beta = 1 - S_{u,v'}$;
/* step 3 */
apply vacuous extension to $m_{uu'}$ and $m_{vv'}$;
pool evidence using conjunctive rule to get $m_{uv}$;
transfer bba’s to the frame $\Theta_{uv}$;
/* step 4 */
if $|CN_{uv}| \neq 0$ then
    $\delta_{uv} = \frac{|CN_{uv}|}{|CN_{uv}|}$;
    discount $m_{uv}$ using $\beta = 1 - \delta$
if $(u, v)$ already share link(s) then
    $\phi = \frac{|CN_{uv}|}{|CN_{uv}|}$;
    reinforce $m_{uv}$ using $\phi$
/* step 5 */
compute $BetP_{uv}$;
rank all $BetP(E)$;
return top $k$ links with highest $BetP(E)$;
5.5.2 Illustrative example

To illustrate our approach, we consider the graph shown in Figure 5.4. We use $a$ and $b$ as the query nodes where the query link is $ab_1$.

**Step 1** We assess similarity using simple matching of the attributes as illustrated in Step 1 in Section 3.5.2. Assume that $a$ and $b$ are the query nodes and the most similar node to $a$ is $f$ and the most similar node to $b$ is $e$.

**Steps 2 and 3** Suppose we have bba’s obtained by going through Steps 1, 2 and 3 as follows (see Section 3.5.2 for details):

\[
\begin{align*}
    m^{ab_1}(\{E_{ab}\}) &= 0.63 \\
    m^{ab_1}(\{\neg E_{ab}\}) &= 0.10 \\
    m^{ab_1}(\Theta^{ab_1}) &= 0.27
\end{align*}
\]

(5.8)

**Step 4** we get $\delta^{ab_1} = \frac{|CN^{ab_1}|}{|CN^{ab_1}_{tot}|} = \frac{1}{3}$. 

Figure 5.4: Evidential multiplex social network graph
Section 5.6 – Multiplex Link Prediction based on Supervised Learning (mLPSL)

$m^{ab}$ is discounted using $\alpha = 1 - \delta^{ab}$ as discounting rate. Thus, we get the following bba:

$$
\begin{align*}
\alpha m^{ab}(\{E_{ab}\}) &= 0.21 \\
\alpha m^{ab}(\{\neg E_{ab}\}) &= 0.03 \\
\alpha m^{ab}(\Theta^{ab}) &= 0.76
\end{align*}
$$

Now, assume that $a$ and $b$ already share one link of type 2. Then, $\phi = \frac{S_{G}}{\Theta_{G}} = \frac{3}{6} = 0.5$. That is, we reinforce $\beta m^{ab}$ towards the event “exists” using $\phi$ as reinforcement rate. We get:

$$
\begin{align*}
\phi m^{ab}(\{E_{ab}\}) &= 0.605 \\
\phi m^{ab}(\{\neg E_{ab}\}) &= 0.015 \\
\phi m^{ab}(\Theta^{ab}) &= 0.38
\end{align*}
$$

Step 5 Finally, we compute the pignistic probability is computed, we get $BetP^{ab}(E) = 0.795$ and $BetP^{ab}(\neg E) = 0.205$. Therefore, the likelihood the link $ab$ exists is equal to 79.5%.

5.6 Multiplex Link Prediction based on Supervised Learning (mLPSL)

In this Section, we present a framework for link prediction in multiplex social networks using supervised learning. Indeed, we proposed in Chapter 3 a link prediction method that uses $k$-NN classification method. We extend our method to multiplex social networks to handle multiple link types. Therefore, we reformulated the classification setting as follows: given a multiplex social network graph $G(\mathcal{V}, \mathcal{E}, \mathcal{L})$, let $\Theta^{ci} = \{E_{ci}, \neg E_{ci}\}$ be the frame of discernment including the hypotheses speculating the existence and absence of a link $e_{i}$ of type $i$. The aim is to predict the class membership of the query link.

The general strategy is to extract uniplex and multiplex features from the network. Then, predict query links classes using supervised learning. For a query link, we evaluate similarities between its connecting nodes and their neighborhood using 1$^{st}$ and 2$^{nd}$ level neighbors. Most similar nodes are considered as nearest neighbors. Yet, due to uncertainty,
these sources of information cannot be considered completely trustworthy. Hence, confidence is re-evaluated and updated. Finally, we predict links according to their uncovered class labels.

5.6.1 The method: Evidential $k$-nearest neighbors for multi-relational Link Prediction

In the following, we present our framework for link prediction using evidential $k$-nearest neighbors classification in multiplex social networks. First, we construct the feature set by computing uniplex and multiplex topological measures. Then, we explore the neighborhood of query links to extract the $k$-nearest neighbors. The $k$-nearest neighbors are considered as sources of information. Their evidence is quantified and evaluated. Then, it is pooled using a fusion mechanism. Finally, links are predicted according to derived classes.

Let the link $e_i$ be the query link. The aim is to predict whether it connects the nodes $u$ and $v$ via the relation $i$. Let $1^{st}$ level neighbors of node $u$ be the direct neighbors connected to $u$ and let the $2^{nd}$ level neighbors be the neighbors of the $1^{st}$ level neighbors $u$. The node $u$ is obviously not linked to its $2^{nd}$ level neighbors. The set $N_i$ contains shared and unshared links of type $i$ with $1^{st}$ and $2^{nd}$ level neighbors of $u$ and $v$. Our framework to predict the existence of $e_i$ is detailed below.

5.6.1.1 Step 1: Feature set construction

Firstly, we construct a feature set for classification. Local topological scores are computed for each link type particularly and for all links globally. These scores are computed for each node pairs in the neighborhood of $u$ and $v$ which involves $1^{st}$ and $2^{nd}$ level neighbors. The feature set is employed to assess similarities between $(u, v)$ and their neighbors. Therefore, network topology is taken into account and the search space for detecting nearest neighbors is reduced.

5.6.1.2 Step 2: Evidential learning

The next step is to classify the query link $e_i$. For that, we detect the set of its $k$-nearest neighbors, that we denote $K_{near}$. The $k$-nearest neighbors are identified from $N_i$ by assessing similarities between $(u, v)$ and their neighborhood. The top $k$ similar neighbors are considered. We measure similarity using a distance function i.e., the Euclidean distance. Consequently, similarity between $e_i$ and a neighboring link $e'_i \in N_i$, is denoted $d(e_i, e'_i)$ and
computed as follows:

$$d(e_i, e'_i) = \sqrt{\sum_{q=1}^{Q} (s_{q e_i} - s_{q e'_i})^2} \quad (5.11)$$

where $q$ is the index of uniplex or multiplex local score computed in Layer $L_i$ or across the global graph, $Q$ is the total number of scores, and $s_{q e_i}$ and $s_{q e'_i}$ are respectively its values for $e_i$ and $e'_i$.

Each training instance $(e'_i, \Theta e'_i)$ in $N_i$ expresses a source of information about the class of $e_i$ being of type $i$. If $e_i$ is close to $e'_i$ according to $d$, they probably have the same class. The provided information solely cannot approve the class of $e_i$. Such situation is modeled by simple support functions in the belief function theory framework. Where only some parts of the belief are committed to $m_{e_i} (\{E_{e_i}\})$ or $m_{e_i} (\{-E_{e_i}\})$ and the rest is assigned to $m_{e_i} (\Theta e_i)$. As a result, each link $e'_i$ in $K_{near}$ represents a source of evidence that induces a bba $m_{e_i e'_i}$ over $\Theta e_i$ defined by:

$$m_{e_i e'_i} (\{\theta_{e_i}\}) = \sigma \phi (d(e_i, e'_i))$$
$$m_{e_i e'_i} (\Theta e_i) = 1 - \sigma \phi (d(e_i, e'_i))$$
$$m_{e_i e'_i} (A) = 0, \forall A \in 2^e \setminus \{\theta_{e_i}, \Theta e_i\}.$$  

where $\theta_{e_i}$ is the class of $e'_i$, i.e., $E_{e'_i}$ if $e'_i$ is a 1st level neighbor and $-E_{e'_i}$ if it is a 2nd level neighbor, $\sigma$ is a parameter such that $0 < \sigma < 1$ and $\phi$ is a decreasing function verifying $\phi(0) = 1$ and $\lim_{d \to \infty} \phi(d) = 0$ such that:

$$\phi (d(e_i, e'_i)) = e^{-\gamma d(e_i, e'_i)^\tau} \quad (5.12)$$

where $\gamma > 0$ and $\tau \in \{1, 2, \ldots\}$.

The $k$-nearest neighbors in $K_{near}$ give $k$ bba’s that can be pooled according to their corresponding layer using the Dempster’s rule of combination (Equation 2.4) to get the overall bba $m$ regarding the class membership of $e_i$ in Layer $L_i$:

$$m^{e_i} = m_{e_i}^{e_1} \oplus \cdots \oplus m_{e_i}^{e_k} \quad (5.13)$$

$k$ is the number of nearest neighbors in Layer $L_i$.

Next, we update the obtained bba $m^{e_i}$ according to simultaneous links in the graph $G(V, E, L)$. Since the network contains multiple associations, a links might already exist $e'_i \neq e_i$ between $u$ and $v$. Such relevant information, only handed over by multiplex
networks, has to be considered in link prediction. It stands to reason that multiple connections of particular types improve evidence regarding the existence of particular links. For instance, if many nodes are related by two simultaneous links of types \( l \) and \( k \), then it is within the bounds of possibility that a query nodes that are already connected by a relation of type \( l \) also connect by a link of type \( k \). Suppose that \( u \) and \( v \) already share exactly one link \( e_j \in L_j \), different from the query link \( e_i \in L_i \), \( i \neq j \). We seek for the distribution of simultaneous links of both types \( j \) and \( i \) denoted by \( S^2_{ij} \) towards all simultaneous links of exactly two types in \( G \), denoted by \( S^2_G \). More generally, having \( \delta \leq \lambda - 1 \) simultaneous links between \((u, v)\), the distribution \( S^{\delta+1}_{*j} \), where \( * = \{1, \ldots, \lambda\} \) are the types of simultaneous links is computed. When \( S^{\delta+1}_{*j} \neq 0 \), the bba \( m \) is reinforced towards the element “exists” using the reinforcement rate \( \beta = \frac{S^{\delta+1}_{*j}}{S^{\delta+1}_G} \) (Equation 2.9). Hence, we get the final bba \( \beta m^{e_i} \).

Finally, the obtained bba expresses the belief regarding the class membership of the query link. If evidence committed to the class “\( E_{e_i} \)” is higher than that concerning the class “\( \neg E_{e_i} \)”, the query link is predicted. Otherwise, it is not predicted.

The algorithm of our approach is given in Algorithm 5.4.

This method is an extension of the uLPSL approach presented in Chapter 3, Section 3.6. When we have uniplex social network, we only use uniplex similarity measures in the first step of feature set construction. Additionally, we do not revise evidence according the multiplex structure in step 2. Hence, we get the uniplex method uLPSL.
Algorithm 5.4: Evidential $k$-NN link prediction in multiplex social network

**Input:** Multiplex social network graph $G(V,E,L)$, Query links $Q$, $k$, $\sigma$, $\tau$

**Output:** List of predicted links

/* step 1 */
foreach $e_i \in Q$ do
  compute structural features of $e_i$;
  get the list of neighboring links $N_i$ in $G$  
/* step 2 */
foreach $e'_i \in N_i$ do
  get the structural features of $e'_i$;
  compute the Euclidean distance $d(e,e'_i)$;
  select the $k$ links that minimize $d$;
foreach $e'_i \in K_{near}$ do
  if $e'_i$ has class $E_{e_i}$ then
    $m_{e'_i}^{\Theta_{e_i}}(\{E_{e_i}\}) = \sigma \phi(d(e_i,e'_i))$
    $m_{e'_i}^{\Theta_{e_i}}(\Theta_{e_i}) = 1 - \sigma \phi(d(e_i,e'_i))$;
  else
    $m_{e'_i}^{\Theta_{e_i}}(\{\neg E_{e_i}\}) = \sigma \phi(d(e_i,e'_i))$
    $m_{e'_i}^{\Theta_{e_i}}(\Theta_{e_i}) = 1 - \sigma \phi(d(e_i,e'_i))$;
pool evidence using Dempster’s rule ;
if the end nodes of $e_i$ already share $\delta$ links of types $\chi \in \Lambda \setminus i$ then
  compute $S_{\delta+1}^{\Lambda_{\chi}}$
  compute $S_{\delta+1}^{L}$
  if $S_{\delta+1}^{L} \neq 0$ then
    reinforce $m_{e_i}^{\{E_{e_i}\}}$ using $\frac{S_{\delta+1}^{L \chi}}{S_{\delta+1}^{L}}$;  
return class of $e_i$;

5.6.2 Illustrative example

To illustrate our method, consider the graph given in Figure 5.5. Suppose that the link $ab_1$ between $a$ and $b$ is a test link. Hence, the set of neighbors of $a$ and $b$ are respectively $\tau(a) = \{c,d,e\}$ and $\tau(b) = \{f,g,e\}$. That is, the set of links shared with 1st neighbors is $L^1 = \{ac,ad,ae,bd,bg,be\}$. The sets of 2nd level neighbors of respectively $a$ and $b$ include $\{f,b\}$ and $\{a,c\}$. Therefore, $L^2 = \{af,ac\}$ and $L = \{ac,ad,ae,bd,bg,be,af,ac\}$.
Step 1 The first step is to compute structural similarity measures, evaluate similarities using a distance measure and extract the $k$ nearest neighbors (see Section 3.6.2 for details). Suppose that $k$ is equal to 3 and the 3 nearest neighbors to $ab$ that minimize the distance are the links $ae_1$, $be_1$ and $bf_1$.

Step 2 The next step is to get evidence from the $k$ nearest neighbors. Assume that after applying Dempter’s rule of combination, we get the following bba:

\[
\begin{align*}
    m^{ab_1}(\{E_{ab}\}) &= 0.98 \\
    m^{ab_1}(\Theta^{ab_1}) &= 0.02
\end{align*}
\]

If $a$ and $b$ shared already shared a link $ab_2$ of type 2. Then, we update $m^{ab_1}$ as follows: $S^2_{i_2} = 3$, and $S^2_G = 6$. Thus, $\beta = \frac{1}{2}$.

We reinforce $m_i^{ab}$ towards the event “exists” using $\beta$ as reinforcement rate, we get:

\[
\begin{align*}
    m^{ab_1}(\{E_{ab}\}) &= 0.99 \\
    m^{ab_1}(\Theta^{ab_1}) &= 0.01
\end{align*}
\]

Since $m^{ab_1}(\{E_{ab}\}) > m^{ab}(\neg E_{ab})$, then the class of $ab_1$ is “exists”. Hence, the link is predicted.
5.7 Overview and theoretical comparison

In this section, we present an overview of all the different approaches presented in this Chapter. Besides, we give the estimated theoretical complexity in Table 5.3.

It should be noted that the four presented methods in this Chapter extend our approaches presented in Chapter 3. The approaches mLPLI, mLPGI, mLPNA and mLPSL extend respectively uLPLI, uLPGI, uLPNA and uLPSL. As explained in Chapter 3, Section 3.7, each method can be applied according to the available information given by data. The major contribution of these methods is handling multiplex data in the link prediction task. We predict new links as well as their types.

Multiplex methods take over the computational costs of the uniplex ones and add another layer of complexity as results of taking the multiplex structure into account. Actually, in all algorithms, we look for the number of simultaneous links in the multiplex which costs $\mathcal{O}(N)$, where $N$ is the number of nodes. Hence, we get the theoretical complexity where $k$ is the average degree of nodes in the network illustrated in Table 5.3.

<table>
<thead>
<tr>
<th>Method</th>
<th>Input</th>
<th>Output</th>
<th>Source of information</th>
<th>Theoretical complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>mLPLI</td>
<td>EMSN</td>
<td>Predicted links and their types</td>
<td>Local information across layers</td>
<td>$\mathcal{O}(N^2.k^2)$</td>
</tr>
<tr>
<td>mLPGI</td>
<td>EMSN</td>
<td>Predicted links and their types</td>
<td>Social circles and global multiplex topological information based on local measures</td>
<td>$\mathcal{O}(N^2.k^2)$</td>
</tr>
<tr>
<td>mLPNA</td>
<td>EMSN</td>
<td>Predicted links and their types</td>
<td>Node attributes and global multiplex topological information based on local measures</td>
<td>$\mathcal{O}(N^2.k^2)$</td>
</tr>
<tr>
<td>mLPSL</td>
<td>multiplex SN</td>
<td>Classified links</td>
<td>Structural local information of uniplex and multiplex topology</td>
<td>$\mathcal{O}(N^2.k^3)$</td>
</tr>
</tbody>
</table>
5.8 Conclusion

In this Chapter, we detailed our contributions to handle the link prediction problem in multi-relational social networks. Multiplex social networks are complex structures that enable simultaneous relationships of different types between actors. We introduced a new graph-based model for multiplex social networks characterized by uncertainty regarding edges existence. Uncertainty is quantified and represented using the belief function theory.

Our proposals for uni-relational evidential link prediction are extended to multiplex networks. Four frameworks are presented. Each framework handles uncertainty in social networks using the belief function theory tools. One has to note that each approach for link prediction in multiplex networks generalizes one of our proposed methods for link prediction in uniplex networks. In the particular case, when all links have the same types, we fall back on the special case of uniplex methods. Therefore, we get the same results. It is clear that methods proposed for uniplex networks can be applied to multi-relational networks \( n \) times according to the number of link types. Still, global information about the multiplex is not taken into account which is an unfortunate loss of valuable knowledge.

In the next Chapter, we report experimental analysis of our proposals for multiplex structures. In order to evaluate the performance of our methods, tests on real-world social network data are performed. Furthermore, we conduct comparative studies with baseline methods.
Chapter 6

Link Prediction in Multiplex Social Networks: Experimental Results

6.1 Introduction

Multiplex social networks provide a more detailed picture of real life associations. Most traditional methods for link prediction are dedicated to uniplex networks assuming the links to have the same types. They can be adjusted to the multiplex case but this requires the application of the same process several times according to the number of the layers which is time consuming. Besides, global properties of the multiplex are neglected.

We presented, in the previous Chapter, methods for link prediction in multiplex social networks under the belief function theory framework. In order to evaluate our proposals, we conduct experiments on real-world social data. Since our proposals apply to evidential multiplex networks, we apply graph generation technique presented in Chapter 4 in Section 4.2.

Different data of social networks from real world with multiple types of links are used. We report results obtained by applying our proposals for link prediction on evidential multiplex social networks in order to prove the effectiveness of our approaches under the belief function theory framework. We compare our approaches to existing methods and we measure results using different evaluation criteria such as precision and recall. Sections 6.2, 6.3, 6.4 and 6.5 report the experimental results of each of our multiplex link prediction frameworks.
6.2 Multiplex Link Prediction based on Local Information

In this section, we present experimental tests performed to evaluate our method for multi-relational link prediction based on local information.

6.2.1 Network pre-processing

We test our approach on two real-word social networks. The first dataset is a network of 185 students cooperation (Fire et al., 2012) collected during a “Computer and Network Security” course. Nodes represent students connected by 362 shared relations. Edges can be of three possible types: partnership, same computer, same session. There are 241 partners links, 23 same computer links and 98 same session links. The second dataset is a relationship network of 21K links connecting 84 people according to 5 link types collected from the social evolution dataset (Madan et al., 2012).

The networks are pre-processed to get evidential multiplex social networks. To do so, the technique presented in Chapter 4, in Section 4.2 is applied. This technique is based on a procedure of graph sampling broadly used in link prediction literature (Q.-M. Zhang et al., 2013).

6.2.2 Experimental study

For the two datasets, we compare our multi-relational link prediction method based on local information (mLPLI) with our method proposed for the uniplex case (uLPLI) in Chapter 3, in Section 3.3. Indeed, the two link prediction methods perform under the belief function theory framework. Besides, both methods are based on the common neighbors approach from literature. However, only new links of single types are predicted by the uLPLI method. For that reason, we applied it separately for each set of links of a particular type. Precision and recall are used as performance measures.

Figure 6.1 reports results in terms of precision and recall obtained for the two social networks datasets. The obtained results of precision and recall given by the uLPLI method are averaged in each layer in the networks. One can notice that the proposed framework outperforms the uLPLI method for the two datasets in terms of both precision and recall. Actually, mLPLI gives us higher prediction quality.
For the students cooperation dataset, precision results given by mLPLI reach 80% compared to 46% for the uLPLI. For the relationships network: 72% precision performance is obtained by the mLPLI compared to 53% for the uLPLI. Therefore, our method can predict efficiently both link existence and types. The same observation is made for recall. The mLPLI gives higher values for the students cooperation than the uLPLI method i.e., 65% compared to 35%.

The homogeneous relationships are scarce and this is a possible reason for having such results by the uLPLI method. This indicates that general information about the network is partly accountable for the prediction task. On the other hand, our method is more suitable because it does not apply as many times as there exist layers. Yet, results obtained by the precision are higher than those of the recall measure for the two methods. This indicates that we have more incorrect existing links than incorrect non-existing links. However, the mLPLI approach has demonstrated empirically its validity and performance.
6.3 Multiplex Link Prediction based on Group Information

In this section, we present the experimental study conducted to evaluate our framework for multi-relational link prediction in multiplex social networks based on group information. We denote our method mLPGI (multi-relational link prediction group information). At first, we present social network data used for experiments. Then, we report experimental tests and results.

6.3.1 Network pre-processing

We test our approach on the dataset presented in Section 6.2. It is a relationship network connecting 84 undergraduate dormitory residents via 21K links according to 5 link types collected from the social evolution dataset (Madan et al., 2012). Data include node attributes collected through sociometric surveys for relationships such as political opinions, favorite music genres, attitudes towards diet, smoking behavior and dormitory floor. We extracted social circles from node attributes to obtain a natural structure of groups. Actually, social circles regroup actors with common interests and beliefs. Therefore, actors with common attribute values may form a common social circles. For example, using political interests attribute with values such as democratic and republican. We can get two social circles of democratic and republicans. The network is pre-processed to get an evidential multiplex social networks. We apply our evidential network generation technique presented in Chapter 4, in Section 4.2.

6.3.2 Experimental study

In the experimental tests, we compare our multiplex link prediction method based on group information (mLPGI) with the link prediction method based on group information uLPGI proposed for the uniplex case in Chapter 3, in Section 3.4.

mLPGI and uLPGI approaches use the belief function theory tools to predict links in social networks. They apply to evidential social networks. Further, they are all based on structural local information methods for link prediction from literature. uLPGI only operates on uniplex networks. Hence, it is applied separately for each link type. What we are mainly concerned about here, is to find out the impact of considering social circles information on link prediction in multiplex networks. We use precision and recall as
evaluation measures.

Figure 6.2 illustrates the performance results in terms of precision and recall obtained for the two approaches. At first glance, one can notice that the mLPGI method outperforms the uLPGI approach. It reaches 79% precision and 61% recall compared to 74% precision and 54% recall for uLPGI. Bearing this in mind, we can state that considering group information in multiplex social network enhances link prediction results. The mLPGI is able to predict both links and their types.

Precison results given by both link prediction methods based on social circles mLPGI and uLPGI are quite interesting i.e., 79% and 74% for respectively mLPGI and uLPGI. This indicates that social circles information improves prediction. Therefore, adding more meaningful information such as social circles into the prediction task is actually a good solution. Yet, mLPGI outperforms uLPGI as a result of taking multiplex information into account in the prediction task.

6.4 Multiplex Link Prediction based on Node Attributes

In this section, we report the experimental evaluation of our framework for multiplex link prediction based on node attributes denoted mLPN. Firstly, social network data employed for experiments are presented. Then, experimental results are reported.
6.4.1 Network pre-processing

To evaluate the performance behavior of our approach, we use a social network dataset from real-world. The network described in the previous Section 6.3 is considered. These social data include semantic information regarding the actors. It is pre-processed to get an evidential multiplex version by applying our technique presented in Chapter 4, in Section 4.2.

6.4.2 Experimental study

The multi-relational link prediction framework based on node attributes information, mLPNA, is compared to the uLPNA link prediction method based on node attributes denoted proposed for the uniplex case in Chapter 3, in Section 3.5.

For uLPNA, tests are performed for layer separately then performance is averaged over layers. The two methods apply to evidential networks. What we are mainly concerned about here, is to find out the impact of considering multiplex structure in the prediction task. Figure 6.3 reports experimental results in terms of precision and recall obtained for the two methods.

![Figure 6.3: Precision and recall results of mLPNA and uLPNA](image)

From Figure 6.3, we can see that mLPNA and uLPNA are well-matched on precision. They both reach high precision results i.e., 86% and 81% precision for respectively mLPNA
and uLPNA. On the other hand, mLPNA has higher recall results than uLPNA i.e., 64% and 59% respectively. As shown in Figure 6.3, we can find that performance results in terms of precision and recall reach their maximum when node attribute information and multiplex structure are considered in the prediction task.

The results are consistent with our intuition. Node attributes information combined to network topology add semantic the link prediction. That is to say, the mLPNA framework is effective. It is able to predict new links across layers in multiplex social network.

### 6.5 Multiplex Link Prediction based on Supervised Learning

This section reports the experimental study conducted to test our framework for multi-relational link prediction in multiplex social networks that denoted by mLPSL. First, we present the used datasets in Subsection 6.5.1. Subsequently, we give the performed experimental tests and discussions in Subsection 6.5.2.

#### 6.5.1 Data description

We conduct experiments on two real-world social networks data. The first dataset is the students cooperation network (SC) (Fire et al., 2012) presented in Section 6.2.1. The second dataset is a network of 61 employees with 1240 links from a University department, called AUCS (Kim & Lee, 2015), connected via five relationships: co-work, Facebook connection, having lunch, co-authorship and leisure.

#### 6.5.2 Experimental evaluation

In the experiments phase, the required parameters are $\sigma$, $\tau$, $\gamma$ and the number of nearest neighbors $k$. We set $\sigma$ is to 0.95 and $\tau$ to 1 as in (Denoeux, 1995), these two parameters do not have significant impact on performances (Denoeux, 1995). Tests for the optimization of the parameter $\gamma$ allowed as to it set to 0.11. We measure recall on 10-fold cross validation. Ten tests are conducted by partitioning randomly the set of links into ten equal subsets. In each test, a single subset composed of 10% of the total set of links, is used as test set. The remaining set of links (90%) is employed as training set. Results are obtained by computing average recall over all tests.
At first, we evaluate the effect of the number of $k$ nearest neighbors according to values ranging from 1 to 15. Figure 6.4 shows the performances in terms of recall.

As shown in Figure 6.4, the mLPSL framework gives satisfying performances when the number of nearest neighbors is equal to 3 for the student cooperation data. The best recall results are obtained when $k$ is equal to 7 for the AUCS dataset. It can be said that performances are correlated to the sparsity of the network. Actually, the student cooperation network sparser than the AUCS network, where the average degree of nodes is smaller. Hence, boosting the number of $k$ nearest neighbors covers more 2nd level neighbors. These latter have classes “¬E”. Therefore, test links are more likely to be predicted nonexistent inducing miss classifications.

6.5.2.1 Comparison with baseline methods

In this evaluation step, we compare mLPSL method with ten baseline link prediction approaches from literature based on the local similarity measures: Common neighbors (CN), Adamic/Adar (AA), Jaccard Coefficient (JC), Resource Allocation (RA) and Preferential Attachment (PA). Local methods are tested according to two versions: uni-relational and multi-relationnal. In the uni-relational configuration, measures based on uniplex local information are computed to each test link of a particular type $i$ in a Layer $i$. Thus, the global
structure of the multiplex is ignored. Uni-relational approaches are: $CN_{uniplex}$, $CN_{uniplex}$, $AA_{uniplex}$, $JC_{uniplex}$, $RA_{uniplex}$, $PA_{uniplex}$. In the multi-relational setup, multiplex local scores are measured to each test link across all layers. Thus, multi-relational methods are: $CN_{multiplex}$, $CN_{multiplex}$, $AA_{multiplex}$, $JC_{multiplex}$, $RA_{multiplex}$, $PA_{multiplex}$. Figure 6.5 reports the obtained results measured by recall for the different methods. One should note that $k$ is set to 3 for the student cooperation dataset and it is set to 7 for AUCS dataset.

![Figure 6.5: Recall results of mLPSL and baseline uniplex and multiplex similarity scores on SC and AUCS networks](image)

As shown in Figure 6.5, it clearly sticks out that mLPSL is capable to accurately predict links in multi-relational networks. Actually, the evidential $k$ nearest neighbor classifier leverages structural information encoded in each relation from data naturally and it incorporates new links prediction. The mLPSL method pays attention to the importance of relation types in the network. As illustrated, the mLPSL algorithm outperforms baseline methods. Traditional approaches that only use uniplex information give poor results. This is due to neglecting the global information of the whole multiplex. Hence, only a partial structure is treated. Although multiplex information based methods give better results than the uniplex approaches, performance is still not satisfying. In general, we can claim validity and effectiveness of mLPSL method.

**6.5.2.2 Class imbalance test**

We test the performance of our framework against class imbalance scenarios. To do that, we increase the number of negative instances (non-existing links). The same parameters requirements are set as the former experiments. Figure 6.6 and Figure 6.7 illustrate precision results of our approach and baseline method obtained for different numbers of negative links on the two networks.
Figure 6.6: Class imbalance test of mLPSL and baseline methods on the SC network in terms of precision

Figure 6.7: Class imbalance test of mLPSL and baseline methods on the AUCS network in terms of precision
As reported in Figure 6.6 and Figure 6.7, results measured in terms of precisions decrease moderately for all methods for both datasets as more non-existing edges are predicted. However, the plots decrease slowly without falling below 82% for the AUCS dataset for 500 false edges, and 67% for the SC dataset according to 400 false links. Besides, precision results obtained by the mLPSL method are significantly higher than the those of baseline approaches for the two datasets. We can conjecture that our mLPSL method shows consistent performance to larger analyzed data which permits to handle the class imbalance problem.

**6.6 Conclusion**

In this Chapter, we have evaluated our methods for links prediction in multiplex social networks under the belief function theory framework. Various experiments have been conducted on different real-world datasets with multiple links information such as students cooperation, social evolution and employees relationships.

To create our evidential multiplex social network graph structure, we used our technique based on graph sampling and simulation methods presented in Chapter 4. Performance has been measured using popular evaluation scores such precision and recall. We compared our approaches to baseline method in order to demonstrate the efficiency of our proposals. The obtained results clearly confirm the effectiveness of our frameworks.

The mLPLI, mLPGI and mLPNA methods were compared respectively to their uniplex versions uLPLI, uLPGI and uLPNA approaches. All multiplex methods gave better results in terms of precision and recall than the uniplex ones. This is due to integrating multiplex information into the link prediction task. The mLPSL method was compared to uniplex and multiplex local information similarity measures from literature. Our approach proved its efficiency in terms of both precision and recall. It also handled the class imbalance scenarios effectively.
Conclusion

This research sought to alleviate an essential problem related to the understanding of the patterning of social ties in networks. We have conducted an extensive literature review concerning the major topics tackled through this thesis, i.e. social network analysis, link prediction and uncertainty in social data. Although several mining tasks can be applied to understand the relational aspects of social structures, in this thesis we were concerned with the link prediction problem as it typically discloses the processes that carry out social networks and the different phenomena that lead to the creation of new relationships among actors.

The main concern was to investigate methods for analyzing social ties emerging from the diverse interactions of social entities. Building on top of the belief function theory tools and classical link prediction approaches, we proposed novel methods to take benefits of their assets and conduct an effective analysis with the goal of identifying potential new associations in social networks. This conclusion summarizes the methodological and practical contributions of this thesis, as well as indications of future research.

Our contributions for the literature of social networks analysis and link prediction are listed in the following:

(1) review the current state of research on link prediction,

(2) overcome uncertainty issues due to social data imperfections and provide social network representations that handle such uncertainty,

(3) investigate methods that address the link prediction problem effectively while handling uncertainty, a concern that has not been tackled in the literature.

Despite published surveys on link prediction, most of them did not detail the problem statement or did not display a clear classification of existing approaches. Therefore, we
have extensively recalled all the aspects of the link prediction problem. First, we positioned link prediction as one of the key tasks in social network analysis and link mining. Then, we detailed the different categories of link prediction methods.

Social networks are generally modeled by mathematical structures representing pairwise relations between social entities. These graph models have different structures established according to the architecture of the social networks. Graph representations are obviously very sensitive to the reliability of social data. Yet, we demonstrated that such data can be unreliable due to noise, false or missing information. We proposed to address these issues by quantifying and integrating data imperfection in the graph structures.

Since most common social network graphs are uniplex graphs. We readjusted such structures to handle uncertainty at the edges level. Each link is weighted by an uncertainty degree regarding its existence. Such uncertainty is quantified and represented using mass functions given by the belief function framework. As such, whether the information given by social data is imperfect, partial or even missing, evidence is managed efficiently thanks to the BFT tools. Next, we were interested to a more complex structure of social networks which accommodates multiple link types between node pairs, as such structure is very close to real-world networks where entities connect simultaneously according to different interactions. We managed to extend our uniplex uncertain social network graph to the multiplex case. Indeed, uncertainty appears at the edges level. Thus, noisy and imperfect social data are represented perfectly across the social network graph structure.

Despite the large number of research published to handle link prediction and understand social entities behavior, in this thesis, we were concerned on different features few discussed of this context. While most social data are still restricted to linkage information, i.e., existing nodes and links, real world social networks contain richer information concerning the social substructures. Bearing this in mind, we introduced novel link prediction frameworks that use different information about the network. Mostly, because particular information are not usually available in data.

Firstly, we proposed two link predicators for evidential uniplex and multiplex social networks that use solely basic graph structure. We have presented strategies to manipulate topological information based on node neighborhoods. Based on this strategy, we have showed that the use of local information combined to uncertainty handling in the link prediction context may indeed contribute to accurate predictions. We defined in this context two link prediction frameworks under the BFT for uniplex and multiplex evidential social networks. Secondly, we analyzed the feasibility of predicting links using social circles information in networks. Accordingly, we proposed two link prediction approaches for both uniplex and multiplex networks based on social circles and structural information. Next, in order to add more semantic to link prediction, we specified two frameworks based on node
attributes for uniplex and multiplex evidential networks. We proved that node attribute information besides of contributing to enhance link prediction, also brings relevant clues to better explain linkage patterns. Finally, we adapted the evidential $k$-nearest neighbor classification approach to handle the link prediction in uniplex and multiplex social networks. Our frameworks are inspired from state-of-the-art similarity-based methods. They are simple and not costly in terms of computational complexity.

All methods operate exclusively using the belief function theory tools. We showed how to adequately collect information from social entities, how to fuse it to get an overall view thanks to combination rules given by the BFT, and even how to revise it by evaluating the degree of trustworthiness of these social entities and how such information could be exploited to predict potential links. Further, we demonstrated how to effectively make use of natural information of social networks to consider their semantics and enhance link prediction. Our proposals are relevant for real world applications since they significantly reduce wrong predictions and boost the right ones.

A part from this, we developed a novel technique for evidential social networks generation since such models do not exist in the literature. Obtained results on real world networks, i.e, Facebook, of all evidential link predictors showed that they performed greater than baseline methods in terms of precision and recall. This led us to prove the feasibility of our proposals and their validity in the link prediction context in social networks. In addition, this confirmed that the use of different social information sources can greatly contribute to better predictions.

Interesting avenues for future works have to be mentioned. To start, it would be interesting to investigate other uncertainty aspects in social networks. For example, at the nodes level. Actually, social networks are very dynamic structures, nodes are added or even deleted all the time. One cannot be certain about the true existence of a particular node. Further, we also suggest to handle uncertainty at node attributes. Such information is imperfect in real-world data. For example, users insert false and misleading information, or even create fake profiles. Such imperfection can be also represented by the belief function theory due to its assets.

One step further, we can extend this research to other networks types. In particular, we intend to test our link prediction framework based on group information on biological networks. This intention is supported by the fact that some biological networks, i.e., disease interaction networks, have natural overlapping groups. Also, we aim to investigate the usage of our contributions to build a general framework which combines many sources of information i.e., network topology, social circles and node attributes at the same time. Probably, the more the link predicator is informed the more prediction quality is enhanced.
Besides, other social contents can be explored such as temporal and location information.

Furthermore, there remain a number of research opportunities such as exploring other social content to predict potential links. For example, temporal information regarding creation and/or deletion of links can be used to dynamically predict links in social networks. Actually, most link prediction approaches are designed to predict links from one static snapshot of the networks. Yet, some social data often carry extra temporal information. Hence, data can give as a picture regarding the evolution of the network. The link prediction task can be applied in the sequential snapshot setting by integrating links’ temporal information into the algorithm. On the other hand, in real world social networks, entities and links may appear and disappear dynamically over time. For example, a user may delete his profile in Facebook, or two people may cut off their friendship. It would be interesting to study the different reasons of links disappearance and build an overall mechanism to predict new links existence and old links deletion.
Published contributions in the context of this Thesis


- Mallek S., Boukhris I., Elouedi Z., Lefèvre E.: Evidential Link Prediction Based on Group Information. The Third International Conference on Mining Intelligence and Knowledge Exploration, MIKE 2015, 482-492, 9th-11th December 2015, Hyderabad, India, Springer.

- Mallek S., Boukhris I., Elouedi Z., Lefèvre E.: The Link Prediction Problem Under a Belief Function Framework. 27th International Conference on Tools with Artificial Intelligence, ICTAI 2015, 9th-11th November 2015, Vietri Sul Mare, Italy, IEEE.


REFERENCES


REFERENCES


Simmel, G. (1902). The number of members as determining the sociological form of the
REFERENCES


