The Link Prediction Problem Under a Belief Function Framework

Sabrine Mallek, Imen Boukhris, Zied Elouedi, Eric Lefevre

The Link Prediction Problem Under a Belief Function Framework

Sabrine Mallek, Imen Boukhris and Zied Elouedi
LARODEC, Institut Supérieur de Gestion de Tunis
Université de Tunis
Tunis, Tunisia
Email: sabrinemallek@yahoo.fr, Email: imen.boukhris@hotmail.com
Email: zied.elouedi@gmx.fr

Eric Lefèvre
UArtois EA 3926 LGI2A
Univ. Lille Nord de France
Arras, France
Email: eric.lefevre@univ-artois.fr

***Abstract*—Link prediction is a key research area in social network analysis that enables to understand how social networks evolve over time. It involves predicting the links that may appear in the future based on a snapshot of the social network. Various techniques addressing this problem exist but most of them deal with it under a certain framework. Yet, complete information about the social network of interest is frequently not available as knowledge about the nodes and edges may be partial and incomplete, hence any analysis approach must handle uncertainty in the prediction task. In this paper, we examine the link prediction problem in uncertain social networks by adopting the theory of belief functions. Firstly, a new graph-based model for social networks that encapsulates the uncertainties in the links' structures is proposed. Secondly, we use the assets of the belief function theory for combining pieces of evidence induced from different sources and decision making to propose a novel approach for predicting future links through information fusion of the neighboring nodes. The performance of the new method is validated on a real world social network graph of Facebook friendships.**

***Keywords*-link prediction; social network analysis; belief function theory; uncertain social network;**

I. Introduction

Social networks are very large systems that depict social interactions between millions of individuals. They are usually modeled as graphs, where nodes correspond to persons and edges represent associations between them. Social network analysis (SNA) has arisen as a tool to monitor and analyze such networks. It is a collection of specifically designed methods oriented towards an investigation of the relational aspects of the social structures [1].

Yet, social networks are very dynamic structures since new nodes and edges are added continuously. For instance, one of the interesting problems treated in social network analysis is the understanding of the dynamics that drive social networks evolution such as predicting the possibility of a future linkage between a non connected pair of nodes which is known as the link prediction problem. However, social network techniques are designed to deal with social networks under a certain framework. In fact, frequently used methods assume links with binary relationships, either 1 (exist) or 0 ($\neg$ exist). Nevertheless, the structure of such networks critically relies on the precise nature of the data, this follows to inaccurate results when applying SNA techniques. As discussed in [2], [3], datasets of social networks are prone to observation errors, and are frequently missing and affected by noise (e.g., nodes and/or edges are missing from the data) which may be due to the imperfect nature of the sources used for building the networks (human intelligence, open source intelligence, etc.) [4]. Thus, we will be compelled to face two major problems. The first one is to consider all edges and nodes and risk the possibility of adding mistakenly false ones into the network. The second one consists at removing all uncertain edges and nodes and risk the problem of missing edges and nodes [3]. For this reason, we suggest to handle uncertainty into the social network graph.

Actually, descriptions of how to cast uncertainty into social network analysis have not been addressed in the literature from sociology and other fields. Mainly, it is related to the analyzed data. In fact, data were collected manually by sociologists through direct observation or by questionnaires. Hence, data and the people it was collected from are well known. Besides, the data sets were frequently small. On the other hand, data that we encounter nowadays are very large because of the emergence of online networking applications and scalable databases which made many researchers from various fields interested in studying characteristics of large-scale social networks, however, little interest have been dedicated to the investigation of the uncertain aspects of such networks [5]. That is, uncertainty is a feature produced from lack of information regarding the world for deciding whether a statement is true or false.

Typically, most of the existing works are devoted to study weighted networks, where the weights take integer values. However, another way to represent an uncertain network is to weight the links with values in $[0, 1]$ to encode the degrees of uncertainty [4]. In fact, one of the inherent properties of various real-world networks is that they are characterized by different degrees of uncertainty especially the large-scale ones, as pointed in [2]. Thus, incorporating uncertainty into social networks can be argued to be even more important since these later are expected to be very large. Yet, even smaller-scale social networks are vulnerable to uncertainty and prone to errors because of the inherent unreliability, bias

of human informants or issues related to dodged responses in network surveys [4].

Accordingly, in this paper we adopt the belief function theory [6], [7] as a theory for reasoning under uncertainty to deal with imprecision found in data and the uncertainty that characterizes social networks. In fact, the interest of the use of the belief function theory to handle uncertainty in networks is discussed more thoroughly in [3]. The main advantage is that it is a general framework that permits to quantify imperfect knowledge and expresses the degree of ignorance in the problem. It also permits to combine several pieces of evidence induced from different sources of information to make decisions. We propose to use the tools provided by the belief function theory to define a new graph-based model for social networks that incorporates uncertainty at the edges level.

Additionally, we develop a novel link prediction approach that takes into account the uncertain aspects describing the social network. Our approach is different from the state of the art methods for link prediction, in that it operates merely with the belief function tools. The uncertainty degrees on the links existence are extended and combined as independent sources of information. A matching and fusion procedure is subsequently applied to get an insight on the existence of a new link. Moreover, a method for generating uncertain social networks is presented in order to test the performance of the proposed link prediction method.

The rest of this paper is organized as follows: In Section 2, we provide an overview of the link prediction problem. In Section 3, we describe some basics of the belief function theory. Section 4 presents the new belief link-based graph model for an uncertain social network representation. Section 5 exposes the proposed approach for link prediction under an uncertain framework. Section 6 illustrates our new proposed method. Section 7 reports the performed experiments together with the acquired results and Section 8 concludes the paper.

II. THE LINK PREDICTION PROBLEM

Link prediction has a broad applicability in a variety of domains such as link analysis, bioinformatics and information retrieval [8]. For instance, one could predict friendships or professional ties when analyzing social networks or predict co-authorships in collaboration networks [9].

Formally, the link prediction task can be formulated as follows [10]: Given a social network $G(V, E)$ where V is the set of nodes which may be of various types (e.g., people, organizations, firms) and E is the set of edges linking them via a type of interdependency (e.g., friendship, financial exchange, physical proximity, knowledge, relationships of beliefs). An edge between a pair of nodes $(v_i, v_j) \in V$ represents an association that took place at a particular time t. The task is to predict the set of potential links to be formed at time $t + 1$.

The link prediction problem may also be treated as the problem of inferring missing links in a network. In fact, in many scenarios, one builds a network given observable data and then attempts to derive extra links that are not visible but are likely to exist [11]. The main difference with the prediction of new links is that it does not address the dynamics of the network, it considers its static state instead. Furthermore, specific properties of the graph nodes are taken into account when inferring missing links rather than the structure of the graph [12].

Most methods proposed to address the link prediction problem build upon a group of similarities between the nodes. They may be classified into two types: node neighborhood-based methods and path-based methods [8].

A. Node neighborhood based approaches

Some of the popular measures applied in previous works include the "Common Neighbors", denoted by $CN(v_i, v_j)$, which depicts the number of shared neighbors of a pair of nodes (v_i, v_j) in the social network. Let $\tau(v_i)$ denote the set of neighbors of the node v_i, then $CN(v_i, v_j) = |\tau(v_i) \cap \tau(v_j)|$. Newman [9] has used this measure in the context of collaboration networks, assuming a correlation between the number of common neighbors, and the likelihood that they will collaborate in a future work. Analysis of a large-scale social network in [13] has shown that two students that share many mutual friends may be friends in the future. The Jaccard's Coefficient takes all the neighbors of the pair (v_i, v_j). It is computed as: $JC(v_i, v_j) = \frac{|\tau(v_i) \cap \tau(v_j)|}{|\tau(v_i) \cup \tau(v_j)|}$. The Adamic/Adar Measure, denoted by $AA(v_i, v_j)$, weights each common neighbor v_k by $\frac{1}{log|\tau(v_k)|}$ to measure v_k's contribution, it is defined as $AA(v_i, v_j) = \sum_{v_k \in (\tau(v_i) \cap \tau(v_i))} \frac{1}{log|\tau(v_k)|}$.

B. Path based approaches

They include the shortest path distance, Average Commute Time (ACT), SimRank index etc. For instance, the shortest path distance is based on the fact that the shorter the distance between two nodes is, the higher the chance that they will be connected. The SimRank index relies on the assumption that two nodes are related if they are linked to similar nodes. The ACT comes from random walks on a graph, it computes the average number of steps $m(v_i, v_j)$ made by a random walker by starting from v_i to reach v_j.

The path based approaches inquire for the topological information of the whole network, although they perform better than the node neighborhood based-measures, they have two drawbacks: the first one is that computing a global index is time consuming. The second disadvantage is that global topological information is not usually accessible [14]. Therefore, we propose in this paper an approach for link prediction based on the intuition of the node neighborhood-based methods under an uncertain framework.

III. BELIEF FUNCTION FRAMEWORK

The belief function theory, also called the Dempster-Shafer theory of evidence [6], [7], 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. We recall in this section notations and formal definitions of the belief function framework used to implement the proposed method.

Let $\Theta = \{\theta_1, \theta_2, \cdots, \theta_n\}$ be an exhaustive and finite set of mutually exclusive events associated to a given problem. It is called the frame of discernment. The power set of Θ denoted by 2^Θ is defined as: $2^\Theta = \{A : A \subseteq \Theta\} = \{\emptyset, \{\theta_1\}, \cdots, \{\theta_n\}, \{\theta_1, \theta_2\}, \ldots, \Theta\}$. It includes the empty set $\emptyset$ which matches the impossible proposition or the conflict. A basic belief assignment (bba), denoted by m, represents the influence of a piece of evidence on subsets of the frame of discernment Θ. It is defined as follows:

$$\begin{aligned} &m : 2^\Theta \rightarrow [0,1] \\ &\sum_{A \subseteq \Theta} m(A) = 1 \end{aligned} \tag{1}$$

A is called a focal element if $m(A) > 0$.

A bba that has at most one focal element A different from Θ is called a simple support function (ssf). It is defined as [15]:

$$\begin{cases} m(A) & = 1 - \omega, \forall A \subset \Theta \\ m(\Theta) & = \omega \end{cases} \tag{2}$$

Total ignorance in the belief function theory is represented by a vacuous bba. It is defined such that [7]:

$$\begin{cases} m(\Theta) & = 1 \\ m(A) & = 0, \forall A \neq \Theta \end{cases} \tag{3}$$

On the other hand, to combine evidence given by two reliable and distinct sources of information, the conjunctive rule of combination denoted by $\textcircled{\cap}$ is used. It is defined as follows [16]:

$$m_1 \textcircled{\cap} m_2(A) = \sum_{B,C \subseteq \Theta : B \cap C = A} m_1(B) \cdot m_2(C) \tag{4}$$

In order to combine two bba's m_1 and m_2 defined on two disjoint frames Θ and Ω, the vacuous extension operation is applied. For that, the bba's have to be extended to the product space $\Theta \times \Omega = \{(\theta_i, \omega_k), \forall i \in \{1, \ldots, |\Theta|\}, \forall k \in \{1, \ldots, |\Omega|\}\}$. The vacuous extension operation denoted $\uparrow$, is defined by:

$$m^{\Theta\uparrow\Theta\times\Omega}(C) = \begin{cases} m^\Theta(A) & \text{si } C = A \times \Omega, \\ & A \subseteq \Theta, C \subseteq \Theta \times \Omega \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

Given two different frames of discernment, one may use the multi-valued mapping to specify the relation between them [6]. In other words, a multi-valued mapping denoted by τ, associates to two disjoint frames of discernment Θ and Ω the subsets $S_i \subseteq \Omega$ that can possibly correspond under τ to $A_i \subseteq \Theta$:

$$m_\tau(A_i) = \sum_{\tau(S_i) = A_i} m(S_i) \tag{6}$$

Decision making within the belief function theory is ensured by different solutions depending on the interpretation. The most popular is the Transferable Belief Model (TBM) proposed by Smets [17]. Decision making within the TBM is performed at the pignistic level where beliefs are represented by pignistic measures denoted by $BetP$ [18]:

$$BetP(A) = \sum_{B \subseteq \Theta} \frac{|A \cap B|}{|B|} \frac{m(B)}{(1 - m(\emptyset))}, \text{for all } A \in \Theta \tag{7}$$

IV. BELIEF LINK-BASED SOCIAL NETWORK

A social network is usually modeled as a classical graph $G = (V, E)$ where V is the set of nodes representing the entities and E is the set of links connecting them. However, such representation does not take into account uncertainty resulting from inaccurate and incomplete data or unreliability of the tools used to construct the social network.

For instance, in [4], the authors highlighted the importance of incorporating uncertainty in social networks constructed from textual data, and proposed to code the strength of each edge between a pair of nodes using probabilities generated from a "ramp shaped membership function". But, the construction of this social network structure is only conceivable in the particular context of the proposed work. Another example worthy of consideration is the belief social network proposed in [19] in which nodes, edges and messages are weighted by $bba's$ expressing uncertainties about their types where the goal is to infer the nature of a message that flows through the network. However, our focus in this work is to treat the uncertainty regarding the existence of new links.

In this respect, we introduce our belief link-based social network for which uncertainty is encoded by the belief function theory. In fact, each edge $v_i v_j$ has assigned a basic belief assignment defined on the frame of discernment $\Theta^{v_i v_j} = \{E_{v_i v_j}, \neg E_{v_i v_j}\}$ denoted by $m^{v_i v_j}$, ($E_{v_i v_j}$ is the event describing the existence of a link between v_i and v_j and $\neg E_{v_i v_j}$ depicts its absence). Thus, an uncertain belief link-based social network graph is defined as $G(V, E)$ where: $V = \{v_1, \ldots, |V|\}$ is the set of nodes and E is the set of edges: A pair $(v_i v_j, m^{v_i v_j})$ is assigned to each edge $v_i v_j \in E$ where v_i,$v_j \in V$, $v_i \neq v_j$, and $m^{v_i v_j}$ is a bba that encodes the degree of uncertainty regarding the existence and absence of a link between v_i and v_j. An example of such a graph structure is given in Fig. 1(a) where the edges are weighted with bba's. For clarity, a link between a pair of nodes (v_i, v_j) is represented if the pignistic probability $BetP^{v_i v_j}(E_{v_i v_j}) > 0.5$. In fact, $BetP^{v_i v_j}(E_{v_i v_j}) > 0.5$ means that the likelihood that the link exists between v_i and v_j is greater than 50%.

(a) Graph $G(V, E)$ (b) Graph $G(V, E^{+1})$

Figure 1: A social network graph with bba's weighted edges at time t (a) and $t+1$ (b)

As presented in Fig. 1, instead of having links' weights taking values either 1 or 0 to express whether or not a link exists, we assign a mass function taking values in $[0, 1]$ to quantify the degree of uncertainty regarding a link existence.

V. BELIEF LINK PREDICTION

We want to be able to infer new links based on a current state of a social network. To do this, we draw on methods based on common neighbors. In fact, the simplest way to address the link prediction problem is to make use of the neighborhood-based metrics. Despite of this simplicity, algorithms based on these measures have proven to be efficient in several social network domains [9], [13]. The intuition is that the links connecting more common nodes are more likely to exist.

From this point of view, we solve the link prediction problem under an uncertain framework using the belief function theory. Common neighborhood nodes are considered as independent sources of information. The uncertainty degrees are transferred and fused to infer the degree of existence of a new link. To this end, we state the problem of predicting a new link as follows: Given a snapshot of $G(V, E)$ at a time t. Predict the existence of a link $v_i v_j$ in the new set of edges E^{+1} between a pair of nodes $v_i, v_j \in V$ at $t+1$, by taking into account the relationships shared between v_i, v_j and their common neighbors at t. Note that all the links shared between each common neighbor v_k and the nodes v_i and v_j whose $m^{v_k v_i}\{(\neg E_{v_k v_i})\} \neq 1$ and $m^{v_k v_j}\{(\neg E_{v_k v_j})\} \neq 1$ are considered, even if these links are not shown schematically on the graph, i.e., $BetP^{v_k v_i}(E_{v_k v_i}) < 0.5$. To this end, we propose an overall mechanism for a method fulfilling the process of inferring a new link between a pair of nodes (v_i, v_j) composed of the four steps presented below.

A. Step 1: Information acquiring

For each common neighbor v_k, extend vacuously the frames of each link $v_k v_i$ and $v_k v_j$ to the joint frame denoted by Θ^{N_k} where $\Theta^{N_k} = \Theta^{v_k v_i} \times \Theta^{v_k v_j}$ using Equation 5. This step is essential in order to work on a unified referential. Note that the vacuous extension is a conservative process of reallocation of beliefs, thus the mass allocated to $A \subseteq \Theta^{v_k v_j}$ is reallocated to $A \times \Theta^{v_k v_i}$ after the vacuous extension operation. Hence, it minimizes any a priori on $\Theta^{v_k v_i}$ and does not favor any event $B \subseteq \Theta^{v_k v_i}$. Once the vacuous extension is applied, we combine the induced bba's using the conjunctive rule of combination to get the masses of the possible pairs included in the product space Θ^{N_k}.

B. Step 2: Information transfer

To successfully transfer the obtained bba's to the frame $\Theta^{v_i v_j}$, a multi-valued operation, denoted by τ, is used such that τ: $\Theta^{N_k} \rightarrow 2^{\Theta^{v_i v_j}}$. The τ function (6) brings together combination sets as follows:

- The masses of the pairs containing at least an element in $\{E_{v_k v_j}, E_{v_k v_i}\}$ and not in $\{\neg E_{v_k v_j}, \neg E_{v_k v_i}\}$ are transferred to $E_{v_i v_j} \subseteq \Theta^{v_i v_j}$ such that:

$$m_\tau(E_{v_i v_j}) = \sum_{\tau(S_i) = E_{v_i v_j}} m(S_i), S_i \subseteq \Theta^{N_k} \quad (8)$$

- The masses of the pairs that contain at least an element in $\{\neg E_{v_k v_j}, \neg E_{v_k v_i}\}$ and no element in $\{E_{v_k v_j}, E_{v_k v_i}\}$ are transferred to $\neg E_{v_i v_j} \subseteq \Theta^{v_i v_j}$ as:

$$m_\tau(\neg E_{v_i v_j}) = \sum_{\tau(S_i) = \neg E_{v_i v_j}} m(S_i), S_i \subseteq \Theta^{N_k} \quad (9)$$

- The masses of the pairs including at least an element in $\{E_{v_k v_j}, E_{v_k v_i}\}$ and an element in $\{\neg E_{v_k v_j}, \neg E_{v_k v_i}\}$ are transferred to $\Theta^{v_i v_j}$ such that:

$$m_\tau(\Theta^{v_i v_j}) = \sum_{\tau(S_i) = \Theta^{v_i v_j}} m(S_i), S_i \subseteq \Theta^{N_k} \quad (10)$$

C. Step 3: Pieces of evidence fusion

To get $m^{v_i v_j}$, the bba's $m^{v_i v_j}_{v_k}$ considering all the n common neighbors are combined using the conjunctive rule of combination such that:

$$m^{v_i v_j} = m^{v_i v_j}_{v_1} \textcircled{\cap} m^{v_i v_j}_{v_2} \textcircled{\cap} \ldots \textcircled{\cap} m^{v_i v_j}_{v_n} \quad (11)$$

This step is fundamental, as it permits to fuse the information provided by the neighboring nodes and treat their shared links as independent sources of evidence.

D. Step 4: Decision making

Finally, we compute the pignistic probability $BetP^{v_i v_j}(E_{v_i v_j})$ using 7 to make a decision on the existence of the link on the graph. In fact, if $BetP^{v_i v_j}(E_{v_i v_j}) > 0.5$ then the link between v_i and v_j is likely to exist with probability $> 50\%$ at $t+1$, otherwise it would be absent. In other words, the value of $BetP$ tells whether the link is more likely or not and it quantifies its degree of probability.

It is important to notice that our method is incremental. Indeed, every time a new common neighbor is added, the new information can be easily combined with the previous state of the graph once the masses are transferred using the τ function. As a matter of fact, once a new common neighbor is added, steps 1 and 2 presented in Section V are applied, then the results are combined with the information already processed for the former common neighbors without going through steps 1 and 2 for these latter.

Furthermore, our belief link prediction method is generic as it does not depend to the social network domain. That is, it can be applied to various social networks e.g., collaboration networks, dark networks, citation networks, etc. Besides, other networks such as computer networks and networks that relate events to each other may also be handled.

VI. ILLUSTRATION

In this section, we illustrate the link prediction approach by a simple example. Assume that the graph in Fig. 1(a) describes a friendship network between persons constructed from data coming from manual sources. Uncertainty may arise due to implicit vagueness of informant reliability and bias. For example "I saw person A talk to person B". Yet, this does not mean that A and B are connected. On the other hand, uncertainty may occur due to partial and incomplete information i.e., the human-informant is not sure if it was A or C talking to B.

Consider Ali and Ben of Fig. 1(a) as the query nodes for whom we intend to predict the likelihood of a future connection at $t+1$. For clarity, we represent the nodes by the first letter of their labels i.e., Ali is represented by A. According to our problem statement, we should consider all the common neighbors u_k for which the masses $m^{Au_k}\{(\neg E_{Au_k})\}$ and $m^{Bu_k}\{(\neg E_{Bu_k})\}$ are $\neq 1$. Assume we have mass functions allocated as described in Table I, where $\Theta^{link} = \{E_{link}, \neg E_{link}\}$, thus the set of the common neighbors is $N = \{C, D, E\}$. That is, the direct neighboring links shared with each common neighbor are: AC, BC, AD, BD, AE and BE. The steps 1 and 2 presented in Section V are applied to all the common neighbors. The node C is considered at first. The product space $\Theta^{N_C} = \Theta^{AC} \times \Theta^{BC}$ contains the couples $\{(E_{AC}, E_{BC}), (E_{AC}, \neg E_{BC}), (\neg E_{AC}, E_{BC}), (\neg E_{AC}, \neg E_{BC})\}$.

Step 1: First, the vacuous extension of Θ^{AC} and Θ^{BC} to Θ^{N_C} is computed as described in Table II. Then, the masses on Θ^{N_C} are combined conjunctively (4) to get $m^{N_C}_{\textcircled{\cap}} = m^{AC\uparrow N_C} \textcircled{\cap} m^{BC\uparrow N_C}$ as reported in Table III.

Table I: Mass functions allocated to the links of the social network of Fig. 1(a)

Link	$m^{link}\{(E_{link})\}$	$m^{link}\{(\neg E_{link})\}$	$m^{link}(\Theta^{link})$
AC	0.6	0.1	0.3
BC	0.3	0.4	0.3
AD	0.42	0	0.58
BD	0.4	0.2	0.4
AE	0.2	0.6	0.2
BE	0.4	0.2	0.4
CE	0.65	0.2	0.15
EY	0.5	0.2	0.3
YB	0.45	0.25	0.3

Table II: Vacuous extension

$m^{AC\uparrow N_C}$	$\{E_{AC}\} \times \Theta^{BC}$	0.6
	$\{\neg E_{AC}\} \times \Theta^{BC}$	0.1
	$\Theta^{AC} \times \Theta^{BC}$	0.3
$m^{BC\uparrow N_C}$	$\Theta^{AC} \times \{E_{BC}\}$	0.3
	$\Theta^{AC} \times \{\neg E_{BC}\}$	0.4
	$\Theta^{AC} \times \Theta^{BC}$	0.3

Table III: Conjunctive combination $m^{N_C}_{\textcircled{\cap}}$

$m^{N_C}_{\textcircled{\cap}}$	$\{E_{AC}\}$	$\{\neg E_{AC}\}$	Θ^{AC}
$\{E_{BC}\}$	0.18	0.03	0.09
$\{\neg E_{BC}\}$	0.24	0.04	0.12
Θ^{BC}	0.18	0.03	0.09

Step 2: The next step is to transfer the obtained masses giving the node c using the τ function (8, 9, and 10). Hence, we get:

$m^{AB}_C(\{E_{AB}\}) = m^{N_C}(E_{AC}, E_{BC}) + m^{N_C}(E_{AC}, \Theta^{BC}) + m^{N_C}(\Theta^{AC}, E_{BC}) = 0.45$

$m^{AB}_C(\{\neg E_{AB}\}) = m^{N_C}(\neg E_{AC}, \neg E_{BC}) + m^{N_C}(\Theta^{AC}, \neg E_{BC}) + m^{N_C}(\neg E_{AC}, \Theta^{BC}) = 0.19$

$m^{AB}_C(\Theta^{AB}) = m^{N_C}(E_{AC}, \neg E_{BC}) + m^{N_C}(\Theta^{AC}, \Theta^{BC}) = 0.36$

The same process is applied to the common neighbor D. 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: Θ^{AD} and Θ^{BD} are extended vacuously to Θ^{N_D} as shown in Table IV. Application of the conjunctive rule (4) gives the masses on Θ^{N_D} described in Table V.

Table IV: Vacuous extension

$m^{AD\uparrow N_D}$	$\{E_{AD}\} \times \Theta^{BD}$	0.42
	$\Theta^{AD} \times \Theta^{BD}$	0.58
$m^{BD\uparrow N_D}$	$\Theta^{AD} \times \{E_{BD}\}$	0.4
	$\Theta^{AD} \times \{\neg E_{BD}\}$	0.2
	$\Theta^{AD} \times \Theta^{BD}$	0.4

Table V: Conjunctive combination $m^{N_D}_{\bigcap}$

$m^{N_D}_{\bigcap}$	$\{E_{AD}\}$	Θ^{AD}
$\{E_{BD}\}$	0.168	0.232
$\{\neg E_{BD}\}$	0.084	0.116
Θ^{BD}	0.168	0.232

Step 2: When we apply the τ function (8, 9 and 10) we get:

$m^{AB}_D(E_{AB}) = m^{N_D}(E_{AD}, E_{BD}) + m^{N_D}(E_{AD}, \Theta^{BD}) + m^{N_D}(\Theta^{AD}, E_{BD}) = 0.568$

$m^{AB}_D(\neg E_{AB}) = m^{N_D}(\Theta^{AD}, \neg E_{BD}) = 0.116$

$m^{AB}_D(\Theta^{AB}) = m^{N_D}(E_{AD}, \neg E_{BD}) + m^{N_D}(\Theta^{AD}, \Theta^{BD}) = 0.316$

The last common neighbor is the node 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: The vacuous extension is shown in Table VI. The conjunctive combination (4) of the obtained masses is described in Table VII.

Table VI: Vacuous extension

$m^{AE\uparrow N_E}$	$\{E_{AE}\} \times \Theta^{BE}$	0.2
	$\{\neg E_{AE}\} \times \Theta^{BE}$	0.6
	$\Theta^{AE} \times \Theta^{BE}$	0.2
$m^{BE\uparrow N_E}$	$\Theta^{AE} \times \{E_{BE}\}$	0.4
	$\Theta^{AE} \times \{\neg E_{BE}\}$	0.2
	$\Theta^{AE} \times \Theta^{BE}$	0.4

Table VII: Conjunctive combination $m^{N_E}_{\bigcap}$

$m^{N_E}_{\bigcap}$	$\{E_{AE}\}$	$\{\neg E_{AE}\}$	Θ^{AE}
$\{E_{BE}\}$	0.08	0.24	0.08
$\{\neg E_{BE}\}$	0.04	0.12	0.04
Θ^{BE}	0.08	0.24	0.08

Step 2: The application of the muti-valued mapping τ (8, 9 and 10) gives:

$m^{AB}_E(\{E_{AB}\}) = m^{N_E}(E_{AE}, E_{BE}) + m^{N_E}(E_{AE}, \Theta^{BE}) + m^{N_E}(\Theta^{AE}, E_{BE}) = 0.24$

$m^{AB}_E(\{\neg E_{AB}\}) = m^{N_E}(\neg E_{AE}, \neg E_{BE}) + m^{N_E}(\Theta^{AE}, \neg E_{BE}) + m^{N_E}(\neg E_{AE}, \Theta^{BE}) = 0.4$

$m^{AB}_E(\Theta^{AB}) = m^{N_E}(E_{AE}, \neg E_{BE}) + m^{N_E}(\Theta^{AE}, \Theta^{BE}) = 0.36$

Step 3: Once steps 1 and 2 are applied to all the common neighbors, the obtained masses given all the common neighbors m^{AB}_C, m^{AB}_D and m^{AB}_E are combined using the conjunctive rule. The results are reported in Table VIII.

Table VIII: Conjunctive combination $m^{AB}_{\bigcap}$

$m^{AB}_{\bigcap}$	$Mass$
$\{E_{AB}\}$	0.39
$\{\neg E_{AB}\}$	0.14
Θ^{AB}	0.04
$\emptyset$	0.43

Step 4: To make a decision about the existence of a new link between (A, B), we compute the pignistic probability (7) of the two events $\{E_{AB}\}$ and $\{\neg E_{AB}\}$:

$$BetP^{AB}(E_{AB}) = 0.625 \text{ and } BetP^{AB}(\neg E_{AB}) = 0.375$$

Hence, there is 62% chance that Ali and Ben will be connected in the future. That is, an edge linking them is inserted into the graph $G(V, E^{+1})$ at $t+1$ (Fig. 1(b)).

VII. EXPERIMENTS

The goal of this paper is to address the link prediction problem in uncertain social networks. Yet, data of such social networks are not available. Hence as a first phase in our experiments, we preprocessed a component of 10K nodes and 146K edges of a real-world dynamic social network of Facebook friendships obtained from [20]. The nodes represent the users and the edges between them encode friendship relations. Edges are associated with timestamps however some of them have missing values. This dataset provides a great example of the importance of incorporating uncertainty into social networks as it contains missing information. After the network pre-processing phase, we conduct the link prediction process.

A. Network pre-processing

In order to transform the obtained Facebook friendship network into a belief-link based social network, we follow two major steps: (1) we start by deriving four snapshots of the network from the data (2) then mass functions are simulated on the basis of the three first graphs to produce an uncertain belief-link based version of the social network. This latter is considered in the link prediction task.

1) Graphs generation: Firstly, four graphs are extracted from the data each belonging to a time interval according to the edges' timestamps. Thus, we get four snapshots that we call $G(t-2)$, $G(t-1)$, $G(t)$ and $G(t+1)$. Each graph contains edges with timestamps included in its time interval along with the edges that belong to earlier time. For example, $G(t)$ incorporates the edges that are present in $G(t-2)$ and $G(t-1)$ occurred in $[t-2, t]$. Subsequently, edges with missing timestamps are divided into three joint sets and are randomly added to the first three graphs $G(t-2)$, $G(t-1)$ and $G(t)$. As to the fourth graph, the entire set of edges with missing timestamps is added to its set of links.

In fact, our proposed evaluation method for graphs generation is stimulated from a widely used technique in link prediction literature. Most of the existing methods prune randomly a number of edges of the graph according to some nodes characteristics (e.g., nodes degrees) and try to predict the missing links in the prediction process [21], [22]. In other words, they consider the link prediction problem as the problem of inferring missing links and discard thereby the dynamic aspect of the network. However, our proposed sampling procedure of the network snapshots permits to take into account the dynamics of the social network by considering the edges timestamps. Furthermore, as pointed out in [5], combining sampling techniques and simulation-based methods is a straight-forward way to model and analyze social networks that contain uncertain data.

Table IX provides a description of the four generated graphs $G(t-2)$, $G(t-1)$, $G(t)$ and $G(t+1)$. Note that $G(t+1)$ is the whole considered component of the Facebook friendships network.

Table IX: Graphs description

Graph	#edges with timestamps	#edges with missing timestamps
$G(t-2)$	10,250	26,250
$G(t-1)$	20,500	26,250
$G(t)$	30,750	26,250
$G(t+1)$	41,000	105,000

2) Mass functions simulation: The graphs $G(t-2)$, $G(t-1)$ and $G(t)$ are used to generate our belief link-based social network by weighting the links of $G(t)$ with simulated bba's regarding the links existence as follows:

- If a link v_iv_j exists in the three graphs $G(t-2)$, $G(t-1)$ and $G(t)$ then a simple support function $m^{v_iv_j}$ is assigned such that $m^{v_iv_j}(\{E_{v_iv_j}\}) \in [2/3, 1]$;
- If a link v_iv_j exists in $G(t-2)$ and $G(t)$ or $G(t-1)$ and $G(t)$ then a mass $m^{v_iv_j}$ is generated such that $m^{v_iv_j}(\{E_{v_iv_j}\}) \in [1/3, 2/3[$, $m^{v_iv_j}(\{\neg E_{v_iv_j}\}) \in]0, 1/3]$ and the rest is assigned to $m^{v_iv_j}(\Theta^{v_iv_j})$;
- If a link v_iv_j exists only in $G(t)$ then a mass function $m^{v_iv_j}$ is assigned such that $m^{v_iv_j}(\{E_{v_iv_j}\}) \in]0, 1/3]$, $m^{v_iv_j}(\{\neg E_{v_iv_j}\}) \in [1/3, 2/3]$ and the rest is ascribed to $m^{v_iv_j}(\Theta^{v_iv_j})$;
- If a link v_iv_j does not exist in $G(t)$ and exists in $G(t-2)$ and $G(t-1)$ then a simple support function $m^{v_iv_j}$ is assigned such that $m^{v_iv_j}(\{\neg E_{v_iv_j}\}) \in]1/3, 2/3]$;
- If a link v_iv_j exists only in $G(t-2)$ or in $G(t-1)$ then a simple support function $m^{v_iv_j}$ is assigned such that $m^{v_iv_j}(\{\neg E_{v_iv_j}\}) \in]0, 1/3]$.

Once the simulation phase is achieved, we get an uncertain version of the graph $G(t)$ with $62K$ bba's weighted edges which, according the corresponding $BetP$ values, it has 57K existing edges and 5K non existing ones.

B. Link prediction process

In the experimental phase, we apply the proposed belief link prediction method to $G(t)$. The masses of the edges without a priori knowledge (e.g., edges with no assigned bba's) are determined on the basis of the common neighbors as described in Section V. They are subsequently used to compute pignistic probabilities to make decisions about the links existence in $t+1$. Finally, the results are compared with respect to $G(t+1)$. To test the accuracy of our link prediction algorithm, we use the precision as an evaluation measure which is defined as follows:

$$precision = \frac{n_c}{n} \tag{12}$$

It expresses the number of correctly predicted existent links n_c with respect to the set of analyzed links n.

C. Results

In order to test the performance of our proposed method, we conducted three experiments with three different values of n: 100K, 150K and 170K. Figure 2 shows the precision values according to the number of correctly predicted edges and the analyzed ones relative each experiment.

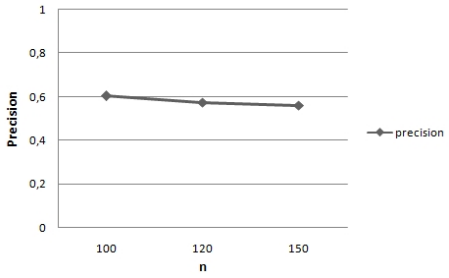

Figure 2: Precision values of the three experiments

As it can be seen, our novel link prediction method gave good precision performance. The prediction quality measured by the precision measure gives values higher

than 50% reaching a maximum performance of 60% when $n = 100$K. That is, performance and validity of the proposed algorithm is empirically confirmed.

Unfortunately, a comparative study cannot be performed at this stage since, to the best of our knowledge, there is no existing method that addresses the link prediction problem under an uncertain framework.

VIII. CONCLUSION

In this paper, we have proposed a new graph-based model for social networks that handles uncertainty at the edge level. We have developed a new approach for predicting social links under an uncertain framework. The belief function theory is used since it enables to represent both the belief regarding the link existence and the uncertainty using mass functions. Our proposed method allows to deal with uncertain relations, through the incorporation of *bba*'s weighted edges. Using the belief function theory, information from the neighboring nodes is transferred and combined to predict the existence of a new association between non linked nodes. We have evaluated our proposals on a real world online social network of Facebook friendhip with missing data.

In order to get our belief link-based graph model, we have proposed a new technique that takes into account the dynamics of the network and uncertainties in data. In summary, experiments have showed that our proposals have contributed to the link prediction problem by considering uncertainty into the social network graph structure. Our method have given good precision results, it is effective and can be applied on social networks from real world data. Extension to the case of several types of simultaneous relationships between the nodes would be considered in future work to predict the links in a more functional manner by inferring their types. We also plan to compare the performance of our method with the existing link prediction methods.

REFERENCES

[1] J. Scott, *Social Network Analysis, A Handbook*. Sage Publications., 1991.

[2] E. Adar and C. Ré, "Managing uncertainty in social networks," *Data Engineering Bulletin*, vol. 30, no. 2, pp. 23–31, 2007.

[3] J. Dahlin and P. Svenson, "A method for community detection in uncertain networks," in *Intelligence and Security Informatics Conference*, 2011, pp. 155–162.

[4] F. Johansson and P. Svenson, "Constructing and analyzing uncertain social networks from unstructured textual data," in *Mining Social Networks and Security Informatics*, 2013, pp. 41–61.

[5] P. Svenson, "Social network analysis of uncertain networks," in *Proceedings of the 2nd Skövde workshop on information fusion topics*, 2008.

[6] A. P. Dempster, "Upper and lower probabilities induced by a multivalued mapping," *Annals of Mathematical Statistics*, vol. 38, pp. 325–339, 1967.

[7] G. Shafer, *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[8] M. A. Hasan and M. J. Zaki, "A survey of link prediction in social networks," in *Social Network Data Analytics*. Springer, 2011, pp. 243–275.

[9] M. E. J. Newman, "Clustering and preferential attachment in growing networks," *Phys. Rev. E*, 2001.

[10] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *J. Am. Soc. Inf. Sci. Technol.*, vol. 58, no. 7, pp. 1019–1031, 2007.

[11] B. Taskar, M. Wong, P. Abbeel, and D. Koller, "Link prediction in relational data," in *Neural Information Processing Systems*, 2003.

[12] C. J. Rhodes and P. Jones, "Inferring missing links in partially observed social networks," *JORS*, vol. 60, no. 10, pp. 1373–1383, 2009.

[13] G. Kossinets and D. Watts, "Empirical analysis of an evolving social network," *Science*, vol. 311, no. 5757, pp. 88–90, 2006.

[14] L. Lu and T. Zhou, "Link prediction in complex networks: A survey," *Physica A*, vol. 390, no. 6, pp. 1150–1170, 2011.

[15] P. Smets, "The canonical decomposition of a weighted belief," in *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 95*, vol. 14, 1995, pp. 1896–1901.

[16] S. P, "Application of the transferable belief model to diagnostic problems," *International Journal of Intelligent Systems*, vol. 13, no. 2-3, pp. 127–157, 1998.

[17] P. Smets and R. Kennes, "The transferable belief model," *Artif. Intell.*, vol. 66, no. 2, pp. 191–234, 1994.

[18] P. Smets, "The transferable belief model for quantified belief representation," in *Handbook of Defeasible Reasoning and Uncertainty Management Systems*, vol. 1, 1988, pp. 267–301.

[19] S. Ben Dhaou, M. Kharoune, A. Martin, and B. Ben Yaghlane, "Belief approach for social networks," in *Belief Functions: Theory and Applications*, vol. 8764, 2014, pp. 115–123.

[20] R. A. Rossi and N. K. Ahmed, "The network data repository with interactive graph analytics and visualization," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015. [Online]. Available: http://networkrepository.com

[21] D. Yin, L. Hong, and B. D. Davison, "Structural link analysis and prediction in microblogs," in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*. ACM, 2011, pp. 1163–1168.

[22] Q.-M. Zhang, L. Lü, W.-Q. Wang, Y.-X. Zhu, and T. Zhou, "Potential Theory for Directed Networks," *PLoS ONE*, vol. 8, no. 2, p. e55437, 2013.