Evidential Link Prediction in Uncertain Social Networks Based on Node Attributes

Sabrine Mallek[1,2], Imen Boukhris[1], Zied Elouedi[1], and Eric Lefevre[2]

[1] LARODEC, Institut Supérieur de Gestion de Tunis, Université de Tunis, Tunisia
sabrinemallek@yahoo.fr, imen.boukhris@hotmail.com, zied.elouedi@gmx.fr

[2] Univ. Artois, EA 3926, Laboratoire de Génie Informatique et d'Automatique de l'Artois (LGI2A), F-62400 Béthune, France eric.lefevre@univ-artois.fr

Abstract. The design of an efficient link prediction method is still an open hot issue that has been addressed mostly through topological properties in recent years. Yet, other relevant information such as the node attributes may inform the link prediction task and enhance performances. This paper presents a novel framework for link prediction that combines node attributes and structural properties. Furthermore, the proposed method handles uncertainty that characterizes social network noisy and missing data by embracing the general framework of the belief function theory. An experimental evaluation on real world social network data shows that attribute information improves further the prediction results.

Keywords: social network analysis, link prediction, uncertain social network, belief function theory, node attributes, structural properties

1 INTRODUCTION

The link prediction problem (LP) is gaining considerable attention in various fields, such as sociology, bioinformatics, and computer science. The aim is to infer missing links from partially observed networks, such as uncovering criminals and terrorists, or potential new links like suggesting future friendships or collaborations. Typically, the most basic assumption for link prediction is that the more similar two nodes are, the more likely they connect. That is, the main concern is how to evaluate and compute similarities accurately. Typical methods use topological properties based on local and global graph information. Yet, other information such as group affiliation or node attributes add semantics to connections between the nodes. Frequently, a great deal of information is available at the nodes level. Several real-world social networks (SN) handle nodes with valued attributes. For example, Facebook users have profiles with attributes containing contact information, work, education, family and relationships. This information boosts LP by adding a meaningful semantic to the nodes characteristics.

On the other hand, most link prediction methods lack functionality to properly manipulate and deal with noisy and imperfect SN data, whereas, these latter are very often biased and exposed to errors [3]. More importantly, anonymization techniques, experimental settings and social networks sampling induce high level uncertainty. In that regard, there is increasable interest to uncertain SN [10] where uncertainty relates to

nodes and/or links existence. We outlined, in previous works [4, 5], the importance of handling uncertainty when analyzing SN. We proposed a graph-based social network model that encapsulates uncertainty at the edges level and we designed methods for LP with uncertainty-handling capabilities. However, none of them take into account semantic similarity, they only consider topological information.

In this work, we develop a novel framework for LP in social networks that combines both network topology and node attributes information in a unified scheme. Furthermore, the novel approach handles uncertain networks and operates fully under uncertainty thanks to the belief function theory (BFT) [1, 8] which is a general framework for reasoning under uncertainty that provides convenient tools to deal with imperfect and missing data problems. This paper is organized as follows. In Section 2, a brief survey on link prediction related work is presented. In section 3, fundamental concepts of the BFT are defined. Section 4 exposes our proposals: a framework for link prediction in uncertain networks. Section 5 gives the experiments and Section 6 concludes the paper.

2 LINK PREDICTION RELATED WORK

Let $G(V, E)$ be a social network graph where V is the set of nodes and E is the set of edges. For two unlinked nodes x and y, most methods compute a similarity score s_{xy} on the basis of their common features. The scores are sorted in a decreasing order, and the pairs with the highest scores are most likely to exist. Typically, the scores are computed from network topology based on local or global information. Local information uses node neighborhoods where popular metrics include Common Neighbors, the Jaccard Coefficient, Preferential Attachment and Adamic/Adar. For example, the common neighbors metric, denoted by CN, counts the common neighbors between (x, y). In contrast, global information methods use topological properties of the global network. Popular scores include Hitting time, SimRank and the shortest path. Indeed, structural metrics present some advantages as they are generic and simple. Yet, they present inconveniences as local methods favor the nodes with the highest degrees whereas global methods suffer from high complexity. Most importantly, the two types of methods only relate to network topology and do not take semantic similarity into account reducing the performances of LP.

Meanwhile, other methods consider different information sources. Yu et al. [12] integrate network proximity and node attributes for LP in weighted SN by boosting the links connecting nodes with similar attributes. However, the method does not scale to large networks. Hasan et al. [2] test node attributes to predict co-authorships. Yet, some features appear to be not suitable to evaluate similarities between the authors. O'Madadhain et al. [7] combine network-based and entity-based features to predict co-participation in events over time under supervised learning. Although semantic attributes enhance LP, they present some limitations like unavailability. Frequently, SN policies do not impose the users to put all their information and content i.e, Facebook allows users to control their privacy settings. Generally, information is hidden because of privacy, legal, ethical or operational concerns. In addition, users information may not be reliable as they can put false and misleading information or even create total fake

profiles. To deal with this ambiguity, we incorporate uncertainty in LP by embracing the BFT as it allows to manage and deal with uncertainty.

3 BASIC DEFINITIONS OF THE BELIEF FUNCTION THEORY

In the belief function theory [1, 8], a problem is drawn through a finite set of exhaustive and mutually exclusive events denoted by Θ and called the frame of discernment. 2^Θ be its power set. A mass function $m : 2^\Theta \rightarrow [0, 1]$, called basic belief assignment (bba), on 2^Θ satisfies:

$$\sum_{A \subseteq \Theta} m(A) = 1. \tag{1}$$

To combine two mass functions m_1 and m_2 derived from reliable and distinct sources of evidence, the conjunctive rule [9] denoted by $\textcircled{\cap}$ is applied:

$$m_1 \textcircled{\cap} m_2(A) = \sum_{B,C \subseteq \Theta : B \cap C = A} m_1(B) \cdot m_2(C). \tag{2}$$

Evidence may be ambiguous or incomplete. Thus, it may not be equally trustworthy. For that, a discounting operation [8] is applied to m to get the discounted bba denoted by ${}^\alpha m$. Given the reliability degree of the source evaluated by a discounting rate $\alpha \in [0, 1]$, the discounting operator is defined by:

$$\begin{cases} {}^\alpha m(A) = (1 - \alpha) \cdot m(A), \forall A \subset \Theta \\ {}^\alpha m(\Theta) = \alpha + (1 - \alpha) \cdot m(\Theta). \end{cases} \tag{3}$$

To fuse two bba's m_1 and m_2 defined on two disjoint frames Θ and Ω, a vacuous extension, denoted by $\uparrow$, is applied. The bba's are extended to the product space of the frame of discernment $\Theta \times \Omega = \{(\theta_i, \omega_k), \forall i \in \{1, \ldots, |\Theta|\}, \forall k \in \{1, \ldots, |\Omega|\}\}$. The vacuous extension operation is defined by:

$$m^{\Theta \uparrow \Theta \times \Omega}(C) = \begin{cases} m^\Theta(A) & \text{if } C = A \times \Omega, A \subseteq \Theta, C \subseteq \Theta \times \Omega \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

The relation between two disjoint frames Θ and Ω can be fulfilled using a multi-valued mapping mechanism [1] denoted by τ. The function τ ascribes the subsets $B \subseteq \Omega$ that may fit a subset $A \subseteq \Theta$:

$$m_\tau(A) = \sum_{\tau(B) = A} m(B). \tag{5}$$

To make decision in the belief function framework, the pignistic probability denoted by $BetP$ is computed from the bba m. It is defined by [9]:

$$BetP(A) = \sum_{B \subseteq \Theta} \frac{|A \cap B|}{|B|} \frac{m(B)}{(1 - m(\emptyset))}, \forall A \in \Theta. \tag{6}$$

4 EVIDENTIAL LINK PREDICTION IN UNCERTAIN SOCIAL NETWORKS

We develop a framework for link prediction that handles uncertainty and combines structural properties and node attributes. We consider in this paper, a SN graph model that handles uncertainty at the edges level. This structure was introduced in [4, 5]. The uncertainty degree about the existence of a link xy is quantified using a basic belief assignment denoted by m^{xy} defined on $\Theta^{xy} = \{E_{xy}, \neg E_{xy}\}$, where E_{xy} is the event supporting the existence of the link xy and $\neg E_{xy}$ endorses its absence. Based on the intuition of local methods, in this paper, our proposed approach called evidential link prediction (ELP), considers the neighborhood of the nodes and compares their features. For instance, given two unlinked nodes x and y, we compare the features of the neighbors of y to those of x and conversely. The idea is that when an individual is similar to another one's connections then they are likely to connect. Subsequently, the most similar nodes are retained and considered as sources of evidence regarding xy existence. In fact, we did not limit our analysis to only common neighbors in order to overcome the limitation of favoring nodes with highest degrees. However, when similar nodes are also common neighbors they are considered more reliable sources of evidence. Therefore, our method takes into account both node attributes and structural properties. The information gathered from the most similar node to x and that to y is pooled to get an overall evidence about xy. To this end, the steps of our novel framework for predicting a new link xy in an evidential SN are detailed below.

Similarity assessment First, the sets of neighbors $\tau(x)$ and $\tau(y)$ of x and y are computed and the attributes of each node and the neighbors of the second are compared such that x is compared to the neighbors $y_n \in \tau(y)$ of y and vice versa. The attributes of the nodes are assumed to have categorical values with no missing values. The similarity is evaluated such that:

$$S_{node_1,node_2} = \frac{\#matched\ attributes}{\#total\ attributes} \tag{7}$$

Next, the most similar node to x denoted by y_s and that to y denoted by x_s are considered. Note that, when there are more than two most similar nodes, the common neighbor with the highest mass on the event "exist" is chosen, otherwise it is chosen randomly.

Reliability evaluation Upon getting the similar nodes, we evaluate their reliability through a discounting operation (Equation 3). When the most similar node is not a common neighbor, it is not considered as a total reliable source. A discount coefficient denoted by $\beta = 1 - S_{node_1,node_2}$ is built according to the similarity score. Actually, when the nodes have all the attributes in common i.e., $S_{node_1,node_2} = 1$, then the most similar node is fully reliable i.e., $\beta = 0$. Hence, m^{xx_s} gives the discounted mass ${}^{\beta}m^{xx_s}$ according to Equation 3.

Fusion and prediction To fuse and propagate information to the query links xy, a vacuous extension (Equation 4) on the product space $\Theta^{xx_s} \times \Theta^{yy_s}$ is applied first to make the bba's bear on the same referential. The induced bba's are fused using the conjunctive

rule to get the global mass $m^{\mathcal{PS}}_{\textcircled{\cap}}$ such that:

$$m^{\mathcal{PS}}_{\textcircled{\cap}} = m^{xx_s\uparrow\mathcal{PS}} \textcircled{\cap} m^{yy_s\uparrow\mathcal{PS}} \tag{8}$$

Upon combination, the obtained bba's are transferred to the frame Θ^{xy} of the query link by a multi-valued mapping (Equation 5) according to the method presented in [5]. Finally for links selection, pignistic probabilities $BetP^{xy}$ are ranked in a decreasing order of confidence where the top ranked links are the ones having the highest scores on the event "exists".

5 EXPERIMENTAL EVALUATION

We examine the validity and soundness of our proposals using a component of 1060 nodes and 10K edges of a real world SN of $Facebook$ friendships [6]. It includes node attributes (such as education, language, school, location, work) that have been anonymized. We proceed by constructing the evidential SN by simulating bba's according to the procedure presented in [4]. A comparative study is made with the classical Common Neighbor (CN) method and the approach from [5], called belief link prediction (BLP), which is inspired from the Common Neighbors method and uses solely structural properties.

The performance is evaluated using the precision-recall (PR) threshold curve as it discussed [11] to provide a fair and consistent tool to analyze LP results. The plot is made by varying the threshold of the top ranked links. At each threshold, we get different decision values given by the links scores e.g., pignistic probabilities for the ELP and BLP, and CN score for the Common Neighbor method. Thus, different pairs of recall and precision are computed. Fig. 1 gives the PR curves on the three methods. As

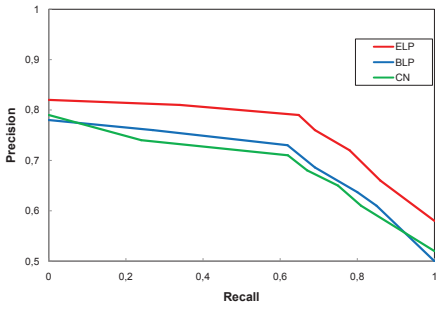

Fig. 1. PR curve of the link prediction methods

shown, the proposed novel approach gives accurate predictions. One can see that the ELP plot is above the BLP and CN curves. Interestingly, taking node attributes information into account improves performances. In addition, one should note that the ELP

method takes local and semi-local structural information into account as it considers direct and indirect neighbors. This adds flexibility to the similarity assessment unlike the traditional indices based on common neighbors. All in all, validity of the new approach is approved. Furthermore, node attributes information enhance link prediction accuracy as it adds a semantic meaning to network topology.

6 CONCLUSION

In this paper, we have developed a method for LP in uncertain SN by combining topological properties and node attributes. Uncertainty is handled thanks to the belief function theory. Similarities are evaluated by matching common attribute values where common neighbors are endorsed as more reliable sources. Evidence regarding new predicted links is estimated through a fusion and mapping procedures. Experiments confirm that semantic information given by the nodes attributes and uncertainty handling in SN enhance LP performance.

References

[1] Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. Ann. Math. Stat. 38, 325–339 (1967)

[2] Hasan, M.A., Chaoji, V., Salem, S., Zaki, M.J.: Link prediction using supervised learning. In: Proceedings of SDM'06 workshop on Link Analysis, Counterterrorism and Security (2006)

[3] Kossinets, G.: Effects of missing data in social networks. Soc. Net. 28, 247–268 (2003)

[4] Mallek, S., Boukhris, I., Elouedi, Z., Lefevre, E.: Evidential link prediction based on group information. In: Proceedings of the 3rd International Conference on Mining Intelligence and Knowledge Exploration. vol. 9468, pp. 482–492 (2015)

[5] Mallek, S., Boukhris, I., Elouedi, Z., Lefevre, E.: The link prediction problem under a belief function framework. In: Proceedings of the IEEE 27th International Conference on the Tools with Artificial Intelligence. pp. 1013–1020 (2015)

[6] McAuley, J.J., Leskovec, J.: Learning to discover social circles in ego networks. In: Proceedings of the 26th Annual Conference on Neural Information Processing Systems 2012. pp. 548–556 (2012)

[7] O'Madadhain, J., Hutchins, J., Smyth, P.: Prediction and ranking algorithms for event-based network data. SIGKDD Explor. Newsl. 7(2), 23–30 (2005)

[8] Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press (1976)

[9] Smets, P.: Application of the transferable belief model to diagnostic problems. Int. J Intell. Syst. 13(2-3), 127–157 (1998)

[10] Svenson, P.: Social network analysis of uncertain networks. In: Proceedings of the 2nd Skövde workshop on information fusion topics (2008)

[11] Yang, Y., Lichtenwalter, R.N., Chawla, N.V.: Evaluating link prediction methods. Knowl. Inf. Syst. 45(3), 751–782 (2014)

[12] Yu, Z., Kening, G., Feng, L., Ge, Y.: A new method for link prediction using various features in social networks. In: Proceedings of the 11th Web Information System and Application Conference. pp. 144–147. WISA'14 (2014)