On Node Embedding of Uncertain Networks

Hoang H. Nguyen, Sergej Zerr
L3S Research Center
Leibniz Universität Hannover, Germany
{*ehoang,szerr*}*@l3s.de*

Tuan-Anh Hoang
Faculty of Mathematics, Mechanics, and Informatics
Hanoi University of Science, Vietnam
hoangtuananh@hus.edu.vn

***Abstract*—Node embedding has recently shown state-of-the-art performance in various network analysis tasks. However, most of the existing node embedding methods do not consider the uncertainty of the input data, which is often the case in practice. This work offers an empirical evaluation of the typical node embedding methods when applied on uncertain networks. Precisely, we examine the performance of embedding vectors obtained by these methods in a set of downstream tasks. To this end, we employ a wide range of uncertain networks and traditional prepossessing techniques for dealing with uncertainty. Our findings suggest that the existing node embedding methods perform practically well on networks with uncertainty once the network data is appropriately prepossessed.**

***Index Terms*—Node embedding, uncertain networks, network analysis**

I. Introduction

A. Motivation

The node embedding approach has shown outperforming performance in various network analysis applications [3]. However, most of the existing methods do not consider the uncertainty in the input data, which is often included through the imperfection of data collection and analyzing techniques. Therefore, the performance of existing node embedding methods on uncertain graphs has not been evaluated comprehensively and remains an open question.

B. Research Objective

In this work, we would like to address this question by examining the effectiveness of typical node embedding methods for uncertain networks by evaluating the learned embedding vectors' performance in a set of downstream tasks. We study the most typical type of uncertain networks in which the links are probabilistically observed. Precisely, there is a probability associated with each link to indicate the likelihood that the link is observed in real data.

II. Related Work

In this section, we briefly review the previous works that are closely related to ours. These works can be grouped into *node embedding methods*, *uncertain network analysis*, and *data investigation*.

A. Node Embedding

There has been an increasing number of embedding methods proposed for unsupervised learning for graphs in recent years. DeepWalk [22] and node2vec [7] are two of the most popular approaches based on random walk. These methods consider a graph as a document and employ truncated random walks on the graphs as sentences in the document, inspired by Skip-Gram model [19]. Fundamentally, node2vec is an extension of DeepWalk, with a flexible biased random walk procedure depended on second-order random walks.

LINE [23] and SDNE [24] approach graph embedding in different ways, although these methods use the given graph directly as its context graph. Unlike DeepWalk and node2vec, they focus on embedding vectors of closer nodes having either connection between them or sharing the same 1-hop neighborhood, and then concatenating the two generated vectors to form the final representation. Generally, LINE and SDNE only differ in their exact formulations of the loss functions and optimizing strategies.

Kipf et al. presented GAE [11] and VGAE [12] in 2016. Accordingly, GCN [11], a recent method for learning on graph-structured data, is utilized in these models. GAE targets to solve problems about semi-supervised learning, and VGAE tends to apply for unsupervised learning.

Notably, URGE model [8] computes some proximity matrices from original uncertain graphs and then applies matrix factorization to get embedded vectors. This approach has various advantages like the proximity matrices based on expected Jaccard similarity and probabilistic random walk with restart can capture the structure of an uncertain graph, which other approaches could not handle well.

However, most of the existing embedding models are not designed for uncertain graphs and were not compared with suitable models to give a general assessment of each model's effectiveness on these graphs.

B. Uncertain Network Analysis

There are a couple of notable studies about analyzing and mining for the networks with probabilities, i.e., structural-proximity computing [6], [8], [26], frequent subgraph mining [25], [27], clustering [13], [17], and classification [4]. Zou and Li [26] examine several structural-context similarities for uncertain graphs such as the cosine similarity, the Jaccard similarity [9], and DICE similarity [5]. Additionally, Hu et al. [8] propose the URGE model for learning a low-dimensional vector for each node on these kinds of graphs. Besides, Kollios et al. [13] present a new definition of clustering based on expected edit distance for probabilistic graphs. However, the existing studies are still fragmented and not clear systematic for evaluating the embeddings of uncertain graphs.

C. Data Investigation

Although there is a high amount of data containing the uncertainties resulting from machine learning algorithms or statistical models in, e.g., chemistry and biology, not many datasets can efficiently satisfy the examination conditions. The most notable in this context is the protein-protein interaction (PPI) networks, a fruitful dataset regarding uncertain graphs, which Nepusz et al. aggregated and presented in 2012 [20]. Accordingly, the biologists label all interactions with two proteins' probabilities if they likely interact with each other. Besides that, certain networks are utilized to generate synthetic probabilistic graphs in several previous works because of the lack of truthful datasets of uncertain graphs [8], [10], [21]. Notably, Leskovec and Krevl present various networks with ground-truth communities in their SNAP collections, including social networks (Youtube, Friendster, Orkut), collaboration networks (DBLP), hyperlinks (Wikipedia), and product network (Amazon) [14]. These certain networks with ground-truth labels of nodes or clusters can be utilized to build synthetic datasets by injecting uncertainty [1].

III. Methodology

We now describe in detail the framework used for the comprehensive assessment of the embedding methods. We start by introducing the methods to be examined. We then present techniques for dealing with the uncertainty in the input network. Next, we review the datasets used for the evaluation. Lastly, we describe the tasks and metrics for quantitative evaluation of the methods.

A. Examined Methods

We examine various typical node embedding methods from the four main categories:

- Matrix factorization: including Non-negative matrix factorization [15] and Singular value decomposition [18].
- Conventional node embedding methods: including DeepWalk [22], Node2Vec [7], and LINE [23].
- Auto-encoder methods: including SDNE [24], DNGR [2], and GAE and VGAE [12].
- Methods tailored for the uncertain networks: the URGE [8] models, which, to the best of our knowledge, the only work in this direction so far

B. Dealing With Uncertainty

For dealing with the uncertainty in the input network, we employ the Jaccard, Dice, and random walk-based similarity [8], [26] to transform the input network. These techniques compute the structural proximity between nodes in the network and convert the uncertain networks into certain ones, the suitable inputs for the examined methods.

C. Datasets

We used both real and synthetics uncertain networks for evaluating the methods. The real ones are Protein-Protein Interaction (PPI) networks [16], which were widely used in previous works on analyzing the uncertain network. We generated synthetic networks by injecting uncertainty into certain ones and employed the injection technique in [1]. We made use of a sub-network induced by top 100 communities in the Amazon product network [14]. These datasets have both nodes' labels and links' probability to enable us to evaluate the embedding methods in chosen downstream tasks.

D. Tasks and Evaluation Metrics

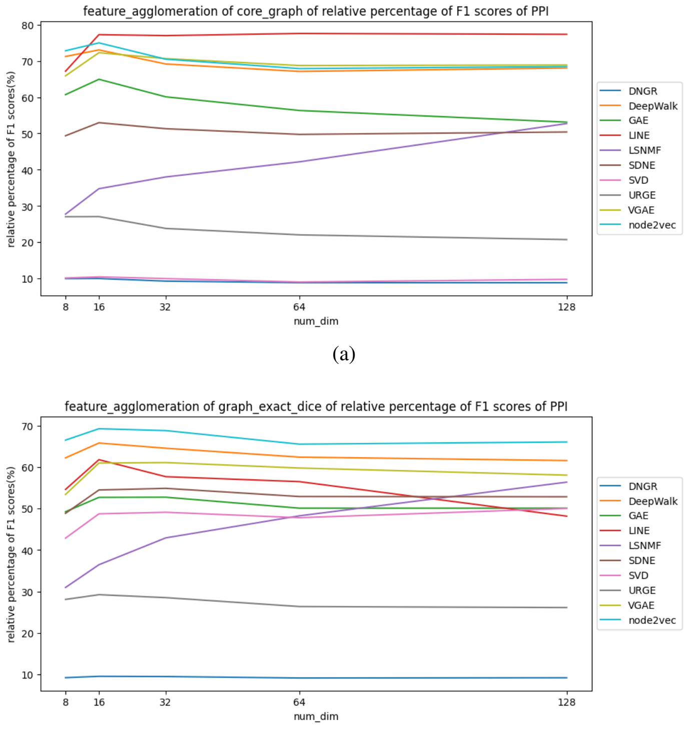
Generally, there is no known ground-truth to assess the embedding vectors directly. Hence, the downstream tasks are often employed for the indirect evaluation of the embedding vectors. In this work, we took two of such traditional downstream tasks: node clustering and link certainty regression (i.e., to estimate the certainty of a link between two nodes based on the nodes' embedding vectors). For the first task, we choose F1 scores as a quality metric in recovering the ground-truth clusters (i.e., group of the nodes having the same label) to evaluate the methods. We use the mean squared error (MSE) and R^2-coefficient as metrics for the second task.

To get robust results, for each dataset, each embedding method, and each number of embedding dimension K (i.e., the dimension of embedding vectors), we run the method on the dataset 10 times independently, each with a different random seed. We then measure the performance of the obtained embedding vectors from each run in the downstream tasks and take the average performance across the runs as the method's performance on the dataset.

IV. Results & Discussion

We conduct the evaluation described above with a different number of embedding dimensions $K = 8, 16, 32, 64, 128$. For each value of K, to aggregate the performance of the examined methods across datasets, we first normalize the methods' performance on each dataset by dividing them by the best performance obtained on the dataset. In the next step, the performances of other methods are measured by their relative percentage to the respective best. Finally, we take the average of each method's runs across PPI datasets and Amazon-based synthetic datasets as their performance on the two types of datasets correspondingly.

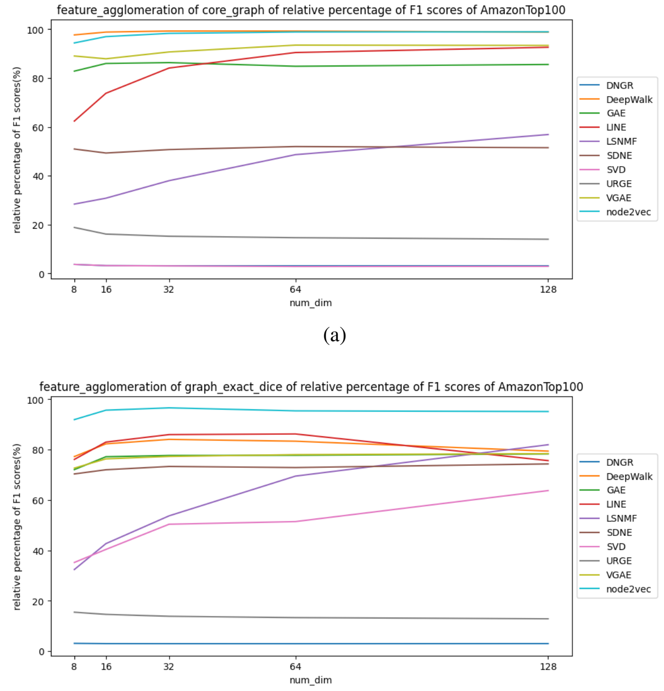
Figure 1(a) depicts the aggregated performance of the examined methods on PPI datasets where we do not employ any pre-processing techniques for dealing with the uncertainty (i.e., the links' probability are considered as their weight). Figure 1(b) shows the performance when we apply Dice similarity computation for preprocessing the input networks. Similarly, Figure 2 shows the performance obtained from Amazon synthetic datasets. The figures suggest that: (1) generally, the existing node embedding methods perform practically well on uncertain networks only by considering links' uncertainty as their weight, (2) the conventional methods (DeepWalk, Node2vec, and LINE) outperform in node clustering tasks. We obtained qualitatively similar results to those shown in the above figures, which imply that the matrix factorization - based methods are better in link uncertainty regression tasks. This



(a)

(b)

Fig. 1: Performance of the examined methods in clustering task on (a) original PPI networks, and (b) on the PPI networks with Dice similarity



(a)

(b)

Fig. 2: Performance of the examined methods in clustering task on (a) original Amazon synthetic networks, and (b) on the Amazon synthetic networks with Dice similarity

finding is reasonable as these methods are directly optimized for estimating the weight of the edges.

V. Acknowledgement

This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreement No. 833635 (project ROXANNE: Real-time network, text, and speaker analytics for combating organized crime, 2019-2022). This work is also supported by Vingroup Innovation Foundation (VINIF) in project code VINIF.2020.DA14.

References

[1] Paolo Boldi, Francesco Bonchi, Aristides Gionis, and Tamir Tassa. Injecting uncertainty in graphs for identity obfuscation. *VLDB*, 2012.

[2] Shaosheng Cao, Wei Lu, and Qiongkai Xu. Deep neural networks for learning graph representations. 2016.

[3] P. Cui, X. Wang, J. Pei, and W. Zhu. A survey on network embedding. *TKDE*, 31(5):833–852, 2019.

[4] Michele Dallachiesa, Charu Aggarwal, and Themis Palpanas. Node classification in uncertain graphs. In *SSDBM*, 2014.

[5] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[6] Lingxia Du, Cuiping Li, Hong Chen, Liwen Tan, and Yinglong Zhang. Probabilistic simrank computation over uncertain graphs. *Information Sciences*, 295:521–535, 2015.

[7] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *KDD*, pages 855–864, 2016.

[8] Jiafeng Hu, Reynold Cheng, Zhipeng Huang, Yixang Fang, and Siqiang Luo. On embedding uncertain graphs. In *CIKM*, pages 157–166, 2017.

[9] Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull.Soc.Vaudoise.Sci Nat*, 1901.

[10] Ruoming Jin, Lin Liu, and Charu C Aggarwal. Discovering highly reliable subgraphs in uncertain graphs. In *KDD*, pages 992–1000, 2011.

[11] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[12] Thomas N Kipf and Max Welling. Variational graph auto-encoders. *NIPS Workshop on Bayesian Deep Learning*, 2016.

[13] George Kollios, Michalis Potamias, and Evimaria Terzi. Clustering large probabilistic graphs. *TKDE*, 25(2):325–336, 2011.

[14] Jure Leskovec and Andrej Krevl. Snap datasets: Stanford large network dataset collection. http://snap.stanford.edu/data, June 2014.

[15] Chih-Jen Lin. Projected gradient methods for nonnegative matrix factorization. *Neural computation*, 19(10):2756–2779, 2007.

[16] Guimei Liu, Limsoon Wong, and Hon Nian Chua. Complex discovery from weighted ppi networks. *Bioinformatics*, 25(15):1891–1897, 2009.

[17] Lin Liu, Ruoming Jin, Charu Aggarwal, and Yelong Shen. Reliable clustering on uncertain graphs. In *ICDM*. IEEE, 2012.

[18] Rahul Mazumder and et. al. Spectral regularization algorithms for learning large incomplete matrices. *JMLR*, 2010.

[19] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

[20] Tamás Nepusz, Haiyuan Yu, and Alberto Paccanaro. Detecting overlapping protein complexes in protein-protein interaction networks. *Nature methods*, 9(5):471, 2012.

[21] Panos Parchas, Nikolaos Papailiou, Dimitris Papadias, and Francesco Bonchi. Uncertain graph sparsification. *TKDE*, 30(12):2435–2449, 2018.

[22] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *KDD*, 2014.

[23] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *WWW*, 2015.

[24] Daixin Wang, Peng Cui, and Wenwu Zhu. Structural deep network embedding. In *KDD*, pages 1225–1234, 2016.

[25] Zhaonian Zou, Hong Gao, and Jianzhong Li. Discovering frequent subgraphs over uncertain graph databases under probabilistic semantics. In *KDD*, pages 633–642, 2010.

[26] Zhaonian Zou and Jianzhong Li. Structural-context similarities for uncertain graphs. In *ICDM*, pages 1325–1330. IEEE, 2013.

[27] Zhaonian Zou, Jianzhong Li, Hong Gao, and Shuo Zhang. Mining frequent subgraph patterns from uncertain graph data. *TKDE*, 2010.