

On Applying Meta-path for Network Embedding in Mining Heterogeneous DBLP Network

Akash Anil, Uppinder Chugh, and Sanasam Ranbir Singh

Department of Computer Science and Engineering
Indian Institute of Technology Guwahati
<a.anil.iitg, uppinderchugh@gmail.com, ranbir@iitg.ac.in

Abstract. In recent time, applications of network embedding in mining real-world information network have been widely reported in the literature. Majority of the information networks are heterogeneous in nature. Meta-path is one of the popularly used approaches for generating embedding in heterogeneous networks. As meta-path guides the models towards a specific sub-structure, it tends to lose some heterogeneous characteristics inherently present in the underlying network. In this paper, we systematically study the effects of different meta-paths using different state-of-art network embedding methods (Metapath2vec, Node2vec, and VERSE) over DBLP bibliographic network and evaluate the performance of embeddings using two applications (co-authorship prediction and authors research area classification tasks). From various experimental observations, it is evident that embedding using different meta-paths perform differently over different tasks. It shows that meta-paths are task-dependent and can not be generalized for different tasks. We further observe that embedding obtained after considering all the node and relation types in bibliographic network outperforms its meta-path based counterparts.

Keywords: Heterogeneous Network · Meta-path · Network Embedding · DBLP · Co-authorship · Classification

1 Introduction

Recently there is a surge in applying network embedding for addressing various tasks in network science such as classification, clustering, link prediction, community detection etc. [6,5,10,16]. Network embedding aims at learning low dimensional feature vector for a node which is capable of preserving its structural characteristics [4,6]. Majority of network embedding models proposed in the past focus mainly on mining homogeneous networks consisting of singular type of node and relation [5]. However, many of the real-world information networks and social networks are heterogeneous in nature consisting of different types of nodes and relations [13]. For example, an academic bibliographic network may be better represented using author, paper, venue (conference/journal) as nodes and different contextual relations such as author-writes-paper, author-publishes-at-venue, etc.

Majority of the previous studies on mining heterogeneous networks [3,12] exploit *meta-paths* [7] which is a sequence of relations between different node types as defined below. Given a heterogeneous network $G(N, E, \mathcal{V}, \mathcal{R})$ where N , E , \mathcal{V} , and \mathcal{R} are set of nodes, set of relations, set of node types, and set of relation types, a meta-path \mathcal{P} is defined as a path $\mathcal{P} : v_1 \xrightarrow{r_1} v_2 \xrightarrow{r_2} v_3 \xrightarrow{r_3} \dots$ where $v_i \in \mathcal{V}$ and $r_i \in \mathcal{R}$. For example, a meta-path *author* $\xrightarrow{\text{writes}}$ *paper* $\xrightarrow{\text{writtenBy}}$ *author* in a bibliographic network represents a co-authorship relation in a paper. While exploring a network, a meta-path defines the specific path the explorer should follow. In the past, meta-paths have been used to generate network embedding [5] and reported to obtain promising results for various applications. In this paper, we systematically analyze effectiveness of considering meta-path for generating network embedding, specifically for bibliographic network. Since, meta-path guides to explore only the partial network defined by the meta-path, it may lose some of the inherent network properties. Motivated by this, this paper attempts to understand the following two important issues while considering meta-paths for generating network embedding.

1. Does meta-path lose network information which can degrade the network embedding performance?
2. Are meta-path based embeddings independent to the end task?

To investigate the above-discussed problems, we evaluate embeddings generated using different types of meta-paths using three state-of-art embedding models namely, (i) Metapath2vec [5], (ii) Node2vec [6], and (iii) VERSE [16] on co-authorship prediction task and author’s research area classification in DBLP¹ heterogeneous bibliographic network. From various experimental observations, it is evident that embeddings generated using entire heterogeneous network outperform the embedding generated using specific meta-paths. Further, It is also observed that embedding using different meta-paths may perform differently over different tasks if not chosen carefully.

Rest of the paper is organized as follows. Section 2 reviews some of the previous works on network embedding. Section 3 describes the experimental setups and results. Paper concludes in Section 4.

2 Literature Survey

For network embedding, a majority of the initial studies attempt to map the natural graph representations like normalized adjacency or Laplacian matrix to lower dimensions by using spectral graph theory [2,9] and various non-linear dimensionality reduction techniques [11,15,1]. However, these models are not scalable to large real world networks as they exploit graph decomposition techniques at the core which requires whole matrix beforehand.

To overcome the above limitations, many network embedding models exploit a framework which first generates a neighborhood sample using a random walk

¹ <https://dblp.uni-trier.de/>

Table 1. Characteristics of different networks constructed over DBLP data

Dataset	DBLP 1968-2008						DBLP 2009-2011
	AA	APA		AVA	All		
Node Types	Author	Author	Paper	Author Venue	Author	Paper Venue	Author
# Nodes	162298	162298	155189	162298 621	162298	155189 621	18457
# Edges	461722	475828		326602	957856		29677

or proximity measure and then leverages it to learn the node embeddings using a skip-gram [8] based neural network model [6,10,14]. For example, Node2vec [6] uses a 2nd order random walk to generate the sample and learn the node embedding using skip-gram model. Further, VERSE [16] preserves the vertex-to-vertex similarity using personalized PageRank and thereby uses a single layer neural network to learn the embeddings.

All the above graph embedding models were proposed for homogeneous network. Recently, Metapath2vec [5] first proposes embedding model for heterogeneous networks which samples the node neighborhood using a random walk guided by meta-path, and then uses skip-gram model to learn the node embedding.

3 Experimental Setups and Analysis

3.1 Experimental Dataset

This paper uses DBLP bibliographic dataset (reported in [17]) covering publication information for the period between 1968 to 2011. To generate various network embeddings using different meta-paths and evaluate the embedding performance over different applications, we further divide the dataset into two parts; (i) between 1968 to 2008 for generating network embedding, and (ii) between 2009 to 2011 for evaluating the embedding over different applications. This paper considers three types of heterogeneous entity classes namely (i) Author (A), (ii) Paper (P), (iii) Venue (V) for constructing various classes of networks defined by different meta-paths. We construct the following four different types of undirected networks from the DBLP 1968-2008 dataset.

- **AA:** It is a homogeneous unweighted co-authorship network considering only **Author** node type. Two nodes are connected if they co-author a paper.
- **APA:** It is a heterogeneous unweighted network considering **Author** and **Paper** node classes. An author is connected to a paper if he/she is one of the authors of the paper.
- **AVA:** It is a heterogeneous unweighted network considering **Author** and **venue** node classes. An author is connected to a venue if he/she published a paper in that venue. This network structure is similar to the structure considered in **Metapath2vec** [5].
- **All:** It is a heterogeneous unweighted network considering all three types of nodes (**Author**, **Paper**, and **Venue**) and corresponding relationships between them.

Table 1 shows the characteristics of these experimental networks.

3.2 Experimental Setups

As mentioned above, three popular recently proposed network embedding models namely (i) Metapath2vec [5], (ii) Node2vec [6], and (iii) VERSE [16] are considered to generate different node embeddings. For all these models, we use the same hyper-parameter values as described in the original studies cited above. All the embedding results reported in this paper consider 100 dimensional vector². To investigate the performance of different meta-paths and their associated embedding, we evaluate the embedding quality using the following two applications.

Co-authorship Prediction: Like the study [16], we also consider co-authorship prediction task as a classification problem i.e., given a node pair, classify if the node pair has a co-author relation or not. To model it as a binary classification problem, we generate feature vectors representing node pairs using Hadamard operator [6,16]. To avoid possible bias with the embedding towards the target application, we consider the DBLP 2009-2011 (non-overlapping with the embedding dataset) for generating samples for the classification task. In this sample, there are 29,677 number of co-authorship relations and 18,457 authors. We use random 80-20 split as training and test samples subjected to four different classifiers namely Gaussian Naive Bayes (NB), Random Forest (RF), Decision Tree (DT), and Logistic Regression (LR). To avoid over-fitting, above setup has been repeated 10 times.

Research Area Classification: We now investigate quality of the embeddings for predicting author’s research area. For each author in DBLP 2009-2011, we further identify (considering the `Field` attribute in [17]) the area in which the author has the maximum publication and consider it as the author’s class label. Like co-authorship prediction, we use similar random 80-20 split for all the classifiers and repeated 10 times.

3.3 Result and Discussion

From Tables 2 and 3, it is observed that LR out-performs other classifiers in 93% times for co-authorship prediction and 75% times for research area classification task. Therefore, we select LR Accuracy for further analysis.

We first investigate if meta-path based embedding loses information or not. Tables 2 and 3 present the Accuracy for co-authorship prediction and author’s research area classification using three network embedding models discussed above for all networks, i.e. `AA`, `AVA`, `APA`, and `A11`.

² While testing with different dimensions 100, 200, 300, we did not observe significant differences. We therefore consider 100 dimensional vector.

Table 2. Co-authorship Prediction by Classifiers for different Networks

Classifier	Metapath2vec				Node2vec				VERSE				Combine			
	AA	APA	AVA	All	AA	APA	AVA	All	AA	APA	AVA	All	AA	APA	AVA	All
NB	0.585	0.633	0.694	0.717	0.688	0.699	0.697	0.719	0.725	0.756	0.733	0.746	0.673	0.745	0.737	0.758
RF	0.761	0.724	0.698	0.720	0.749	0.731	0.698	0.730	0.760	0.754	0.707	0.744	0.772	0.753	0.714	0.748
DT	0.683	0.654	0.628	0.644	0.678	0.658	0.632	0.657	0.688	0.674	0.642	0.678	0.699	0.673	0.645	0.678
LR	0.736	0.739	0.738	0.766	0.773	0.766	0.75	0.777	0.788	0.784	0.764	0.796	0.799	0.795	0.778	0.806

Table 3. Author’s Research Area Prediction by Classifiers for different Networks

Classifier	Metapath2vec				Node2vec				VERSE				Combine			
	AA	APA	AVA	All	AA	APA	AVA	All	AA	APA	AVA	All	AA	APA	AVA	All
NB	0.392	0.476	0.503	0.499	0.500	0.582	0.497	0.488	0.492	0.557	0.550	0.552	0.429	0.58	0.529	0.522
RF	0.484	0.486	0.491	0.482	0.488	0.536	0.518	0.509	0.495	0.499	0.530	0.545	0.499	0.529	0.527	0.53
DT	0.442	0.439	0.439	0.428	0.436	0.481	0.472	0.449	0.445	0.440	0.476	0.490	0.456	0.471	0.474	0.495
LR	0.504	0.539	0.565	0.566	0.486	0.544	0.559	0.555	0.536	0.531	0.605	0.624	0.552	0.592	0.612	0.625

It is evident from Tables 2 and 3 that almost all the models perform best by exploiting All and show poor performance with AA, APA and AVA networks for both tasks, i.e. co-authorship prediction and area classification. Thus, it can be inferred that meta-path alone may be a weak representation for the network because it does not incorporate the impacts of other relational properties while capturing node neighborhood.

Secondly, we intent to investigate if same embedding responds coherently to different problems. From Tables 2 and 3, it is clearly visible that APA performs better than AVA for co-authorship prediction whereas AVA performs better than APA for classifying author’s research area. This observation is true for all the embedding techniques used in this study. Therefore, meta-path based approaches may fail in capturing heterogeneous characteristics of the underlying heterogeneous network if chosen independent to the end task.

Among all the embedding models, VERSE consistently outperforms others for almost all the networks and classifiers for both co-authorship prediction and research area classification tasks.

We further investigate combining all the three embeddings (Metapath2vec, Node2vec, VERSE) by concatenating the feature vectors. From Tables 2 and 3, it is observed that combined embeddings always out-performs individual embedding for co-authorship prediction and research area classification over all the four networks.

4 Conclusion

In this paper, we investigate the applicability of meta-paths in network embedding for co-authorship prediction and author’s research area classification problems in heterogeneous DBLP database. From various experimental results, we observe that by using the entire network majority of the embedding methods out-perform their counter-parts exploiting meta-path based network for both of the above-discussed tasks. Further, it is also evident that exploiting past

co-authorship relation or APA meta-path yield better co-author prediction in comparison to AVA meta-path which exploits author’s publication venue. On the other hand AVA meta-path contributes positively for author’s research area classification problem and have superior performance than APA meta-path. Thus, for heterogeneous network embedding one should carefully choose the node types, relation types and meta-paths which can capture better the network characteristics to address the underlying problem.

References

1. Ahmed, A., Shervashidze, N., Narayanamurthy, S., Josifovski, V., Smola, A.J.: Distributed large-scale natural graph factorization. In: WWW. pp. 37–48 (2013)
2. Belkin, M., Niyogi, P.: Laplacian eigenmaps and spectral techniques for embedding and clustering. In: NIPS. pp. 585–591 (2002)
3. Cao, B., Kong, X., Philip, S.Y.: Collective prediction of multiple types of links in heterogeneous information networks. In: ICDM. pp. 50–59 (2014)
4. Cao, S., Lu, W., Xu, Q.: Grarep: Learning graph representations with global structural information. In: CIKM. pp. 891–900 (2015)
5. Dong, Y., Chawla, N.V., Swami, A.: metapath2vec: Scalable representation learning for heterogeneous networks. In: SIGKDD. pp. 135–144 (2017)
6. Grover, A., Leskovec, J.: Node2vec: Scalable feature learning for networks. In: SIGKDD. pp. 855–864. ACM (2016)
7. Kong, X., Yu, P.S., Ding, Y., Wild, D.J.: Meta path-based collective classification in heterogeneous information networks. In: CIKM. pp. 1567–1571 (2012)
8. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS. pp. 3111–3119 (2013)
9. Ou, M., Cui, P., Pei, J., Zhang, Z., Zhu, W.: Asymmetric transitivity preserving graph embedding. In: SIGKDD. pp. 1105–1114 (2016)
10. Perozzi, B., Al-Rfou, R., Skiena, S.: Deepwalk: Online learning of social representations. In: SIGKDD. pp. 701–710 (2014)
11. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. *science* **290**(5500), 2323–2326 (2000)
12. Sun, Y., Barber, R., Gupta, M., Aggarwal, C.C., Han, J.: Co-author relationship prediction in heterogeneous bibliographic networks. In: ASONAM. pp. 121–128. IEEE (2011)
13. Sun, Y., Han, J.: Mining heterogeneous information networks: principles and methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery* **3**(2), 1–159 (2012)
14. Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., Mei, Q.: Line: Large-scale information network embedding. In: WWW. pp. 1067–1077 (2015)
15. Tenenbaum, J.B., De Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. *science* **290**(5500), 2319–2323 (2000)
16. Tsitsulin, A., Mottin, D., Karras, P., Müller, E.: Verse: Versatile graph embeddings from similarity measures. In: WWW. pp. 539–548 (2018)
17. Yang, D., Xiao, Y., Xu, B., Tong, H., Wang, W., Huang, S.: Which topic will you follow? In: ECML PKDD. pp. 597–612 (2012)