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, Guwahati, India
a.anil.iitg@gmail.com, uppinderchugh@gmail.com, ranbir@iitg.ac.in

Abstract. Unsupervised network embedding using neural networks garnered considerable popularity in generating network features for solving various network-based problems such as link prediction, classification, clustering, etc. As majority of the information networks are heterogeneous in nature (consist of multiple types of nodes and edges), previous approaches for heterogeneous network embedding exploit predefined meta-paths. However, a meta-path guides the model towards a specific sub-structure of the underlying heterogeneous information network, it tends to lose other inherent characteristics. Further, different meta-paths capture proximities of different semantics and may affect the performance of underlying task differently. In this paper, we systematically study the effects of different meta-paths using recently proposed network embedding methods (`Metapath2vec`, `Node2vec`, and `VERSE`) over DBLP bibliographic network and evaluate the performance of embeddings on two applications, namely (i) Co-authorship prediction and (ii) Author's research area classification. From various experimental observations, it is evident that embeddings exploiting different meta-paths perform differently over different tasks. It shows that meta-path based network embedding is task-specific and can not be generalized for different tasks. We further observe that selecting particular node types in heterogeneous bibliographic network yields better quality of node embeddings in comparison to considering specific meta-path.

Keywords: Heterogeneous network · Meta-path · Heterogeneous network embedding · DBLP · Co-authorship prediction · Author 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. [5,7,12,18]. Network embedding aims at learning low dimensional feature vector for a node capable of preserving its structural characteristics [4,7]. Majority of the network embedding models proposed previously

B. Deka et al. (Eds.): PReMI 2019, LNCS 11942, pp. 249–257, 2019.
https://doi.org/10.1007/978-3-030-34872-4_28

consider homogeneous networks, i.e. network consisting of singular type of nodes and relations [7,12,16,18]. However, majority of the real-world information networks and social networks are heterogeneous in nature i.e. networks consist of multiple types of nodes and relations [15]. For example, an academic bibliographic network may be represented using Author (A), Paper (P), Venue (V) (conference/journal) as nodes and different contextual relations such as Author-`writes`-Paper (AP), Author-`publishes-at`-Venue (AV), etc.

Majority of the previous studies on mining heterogeneous networks [3,14] exploit *meta-path* [8] which is a sequence of relations between different node types. Further, symmetric meta-paths are capable of preserving heterogeneous proximity between the underlying nodes. For example, in a bibliographic network, meta-path APA gives the proximity estimate between two authors collaborating on the same paper whereas AVA represents proximity between two authors publishing at the same venue. While exploring a network, a meta-path defines a specific path the explorer should follow. Recently, meta-paths have been used to generate network embedding [5,6] and reported to obtain promising results for various applications in network mining such as node classification, link prediction, clustering, etc. In this paper, we systematically analyze the 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 embeddings.

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-the-art embedding models, namely, (i) `Metapath2vec` [5], (ii) `Node2vec` [7], and (iii) `VERSE` [18] on Co-authorship prediction task and Author's research area classification in DBLP[1] heterogeneous bibliographic network. From various experimental observations, it is evident that meta-path based network embedding cannot be generalized for graph-based problems of diverse nature. Further, selecting suitable node types in the underlying heterogeneous network seems to be more important than considering different meta-paths for heterogeneous network embedding.

Rest of the paper is organized as follows. Section 2 presents some of the previous works on network embedding. Section 3 gives a brief description for heterogeneous network, meta-path, and network embedding. Section 4 describes the experimental setups and results. Finally, Sect. 5 concludes the paper.

[1] https://dblp.uni-trier.de/.

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 [2,10] and various non-linear dimensionality reduction techniques [1,13,17]. However, these models are not scalable to large real-world networks as they exploit graph decomposition techniques at the core which requires the 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 or proximity measure and then leverages it to learn the node embeddings using a skip-gram [9] based neural network model [7,12,16]. For example, Node2vec [7] uses a second order random walk to generate the neighborhood samples and learn the node embedding using skip-gram model, VERSE [18] preserves the vertex-to-vertex similarity using Personalized PageRank [11] and then exploits a single layer neural network to learn the embeddings.

All the above graph embedding models are proposed for homogeneous network. Recently, `Metapath2vec` [5] is proposed for heterogeneous network embedding which samples the node neighborhoods using a random walk guided through a meta-path. In a similar direction, study in [6] exploits the combined effect of different meta-path of predefined length to generate node embeddings in heterogeneous network.

3 Background Study

Definition 1 (Heterogeneous Network). *A Heterogeneous Network can be defined as six-tuple* $<N, E, N^\tau, E^\tau, \phi, \psi>$ *where* N *is a set of nodes,* E *is a set of edges,* N^τ *is a set of node types,* E^τ *is a set of edge types,* $\phi : N \rightarrow N^\tau$ *maps any node* $n \in N$ *to a node type* $n^\tau \in N^\tau$*, and* $\psi : E \rightarrow E^\tau$ *maps any edge* $e \in E$ *to an edge type* $e^\tau \in E^\tau$*. A homogeneous network is a special case of heterogeneous network where cardinalities of* N^τ *and* E^τ *are equal to one i.e.* $|N^\tau| = |E^\tau| = 1$.

Definition 2 (Meta-path). *Given a heterogeneous network* G *where* $N^\tau = \{n_1^\tau, n_2^\tau, \cdots, n_l^\tau\}$ *and* $E^\tau = \{e_1^\tau, e_2^\tau, \cdots, e_{l-1}^\tau\}$*, a meta-path* $\mathcal{P}_{(n_1^\tau, n_l^\tau)}$ *can be defined as an ordered sequence of edge types required to traverse for visiting a node type* n_l^τ *from node type* n_1^τ*, i.e.* $\mathcal{P}_{(n_1^\tau, n_l^\tau)} = n_1^\tau \xrightarrow{e_1^\tau} n_2^\tau \xrightarrow{e_2^\tau} \cdots \xrightarrow{e_{l-1}^\tau} n_l^\tau$.

3.1 Homogeneous Network Embedding

With the popularity of word2vec model using skip-gram proposed in [9] for generating word embedding from large sentence corpus, studies in [7,12,16] adapt skip-gram for network embedding. These network embedding frameworks exploit random walk based sampling strategy to generate node sequences capturing

node's neighborhood characteristics similar to a sentence which captures contextual relation between two words. Formally, for a given network $G(N, E)$, network embedding using skip-gram model aims at maximizing neighborhood probability for a given node:

$$argmax_\theta \sum_{n \in N} \sum_{c \in \mathcal{N}(n)} log\ p(c|n;\theta) \tag{1}$$

where $\mathcal{N}(n)$ gives the neighbors of n and $p(c|n;\theta)$ is the conditional probability of observing neighbor node c for the given node n.

3.2 Heterogeneous Network Embedding

For a given heterogeneous network $G(N, E, N^\tau, E^\tau)$, the skip-gram model defined in Eq. (1) can be transformed into heterogeneous skip-gram model as follows [5]:

$$argmax_\theta \sum_{n \in N} \sum_{\tau \in N^\tau} \sum_{c_\tau \in \mathcal{N}_\tau(n)} log\ p(c_\tau|n;\theta) \tag{2}$$

where $\mathcal{N}_\tau(n)$ gives the neighbor nodes of n from τ^{th} type. Furthermore, $p(c_\tau|n;\theta)$ is defined using softmax function, i.e. $p(c_\tau|n;\theta) = \frac{exp(X_{c_\tau} \cdot X_n)}{\sum_{u \in N} exp(X_u \cdot X_n)}$, where X_n corresponds to the embedding vector of node n.

3.3 Meta-path Based Heterogeneous Network Embedding

The meta-path based heterogeneous network embedding model exploits heterogeneous skip-gram defined in Eq. (2). Further, random walks guided through meta-paths are used to generate neighborhood samples for all the nodes. In other words, random walker traverses partial heterogeneous network specific to underlying meta-path. For example, `Metapath2vec` exploits APVPA (or AVA) meta-path while generating random walk based node sequences [5].

While `Metapath2vec` has been proposed specifically for heterogeneous network embedding, the above-discussed meta-path based network embedding framework can be easily adapted by homogeneous network embedding methods through redefining the input network with specific meta-path. Therefore, this paper further exploits two homogeneous network embedding models namely `Node2vec` [7] and `VERSE` [18] for meta-path based heterogeneous network embedding.

4 Experimental Setups and Analysis

4.1 Experimental Dataset

This paper uses DBLP bibliographic dataset (reported in [19]) covering publication information for the period between years 1968 to 2011. To generate various

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

network embeddings using different meta-paths and to evaluate the embedding performance over different applications, we further divide the dataset into two parts; (i) from 1968 to 2008 for generating network embedding, and (ii) from 2009 to 2011 for evaluating the embeddings over different applications. This paper considers three types of nodes, namely (i) Author (A), (ii) Paper (P), and (iii) Venue (V) for constructing various types of networks defined by different meta-paths. We construct the following four 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 types. 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 types. 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.

4.2 Experimental Setups

As mentioned above, three popular recently proposed network embedding models, namely (i) `Metapath2vec` [5], (ii) `Node2vec` [7], and (iii) `VERSE` [18] are considered to generate node embeddings. For all the 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 vector[2]. To investigate the performance of different meta-paths and their associated embedding, we evaluate the embedding quality using the following two applications.

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

Co-authorship Prediction: Like the study [18], 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 [7,18]. 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, the above setup has been repeated 10 times.

Research Area Classification: We now investigate the 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 [19]) the area in which author has 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.

4.3 Result and Discussion

Tables 2 and 3 present the Accuracy for Co-authorship prediction and Author's research area classification respectively using three network embedding models discussed above for all networks, i.e. `AA, AVA, APA`, and `All`. 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 Author's research area classification task. Therefore, we select LR Accuracy for further analysis.

Table 2. Accuracy for co-authorship prediction by classifiers for different networks, (**Combine** = **Concat**(**Metapath2vec**, **Node2vec**, **VERSE**))

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**

We first investigate if meta-path based embedding loses information or not. It is evident from Tables 2 and 3 that almost all the models perform best by exploiting `All` network and show poor performance with `AA, APA`, and `AVA` networks for both tasks, i.e. Co-authorship prediction and area classification. Thus,

Table 3. Accuracy for author's research area classification by classifiers for different networks, (**Combine** = **Concat**(**Metapath2vec**, **Node2vec**, **VERSE**))

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 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 intend to investigate if the 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. Thus, meta-path based heterogeneous network embedding cannot be generalized for the tasks of different nature.

The homogeneous network `AA` and heterogeneous network `APA`, preserve similar proximity, i.e. co-authorship between underlying pair of authors. From Table 2, it is evident that `AA` performs better than `APA` for Co-authorship prediction in majority of the cases. However, for Author's research area classification in Table 3, `APA` performs better than `AA` in almost all the scenarios. Thus, it can be inferred that meta-path based heterogeneous network embedding may perform differently (poor or better) compared to homogeneous network embedding when subjected to tasks of diverse nature.

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. It may be because unlike `Metapath2vec` and `Node2vec`, `VERSE` exploits a Personalized PageRank [11] capturing vertex-to-vertex similarity while generating the neighborhood sequences.

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 embedding always out-performs individual embedding for Co-authorship prediction and Author's research area classification over all the four networks.

5 Conclusion

In this paper, we investigate the applicability of meta-paths in heterogeneous 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 appropriate node types, majority of the embedding methods out-perform their counter-parts exploiting meta-path based

network for both of the above-mentioned tasks. Further, it is also evident that exploiting past co-authorship relation or `APA` meta-path yields 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 to 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. Fu, T.y., Lee, W.C., Lei, Z.: Hin2vec: Explore meta-paths in heterogeneous information networks for representation learning. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management CIKM 2017, pp. 1797–1806. ACM, New York (2017). https://doi.org/10.1145/3132847.3132953
7. Grover, A., Leskovec, J.: Node2vec: scalable feature learning for networks. In: SIGKDD, pp. 855–864. ACM (2016)
8. 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)
9. 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)
10. Ou, M., Cui, P., Pei, J., Zhang, Z., Zhu, W.: Asymmetric transitivity preserving graph embedding. In: SIGKDD, pp. 1105–1114 (2016)
11. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. In: Proceedings of the 7th International World Wide Web Conference,Brisbane, Australia, pp. 161–172 (1998). https://www.citeseer.nj.nec.com/page98pagerank.html
12. Perozzi, B., Al-Rfou, R., Skiena, S.: Deepwalk: online learning of social representations. In: SIGKDD, pp. 701–710 (2014)
13. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**(5500), 2323–2326 (2000)
14. 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)
15. Sun, Y., Han, J.: Mining heterogeneous information networks: principles and methodologies. Synth. Lect. Data Min. Knowl. Discovery **3**(2), 1–159 (2012)
16. Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., Mei, Q.: Line: large-scale information network embedding. In: WWW, pp. 1067–1077 (2015)

17. Tenenbaum, J.B., De Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science **290**(5500), 2319–2323 (2000)
18. Tsitsulin, A., Mottin, D., Karras, P., Müller, E.: Verse: versatile graph embeddings from similarity measures. In: WWW, pp. 539–548 (2018)
19. Yang, D., Xiao, Y., Xu, B., Tong, H., Wang, W., Huang, S.: Which topic will you follow? In: Flach, P.A., De Bie, T., Cristianini, N. (eds.) ECML PKDD 2012. LNCS (LNAI), vol. 7524, pp. 597–612. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33486-3_38