A Literature Review of Recent Graph Embedding Techniques for Biomedical Data

Yankai Chen[1], Yaozu Wu[2], Shicheng Ma[2], and Irwin King[1]

[1] Department of Computer Science and Engineering,
The Chinese University of Hong Kong
{ykchen,king}@cse.cuhk.edu.hk
[2] KEEP, The Chinese University of Hong Kong
yaozu279@gmail.com, shicheng@keep.edu.hk

Abstract. With the rapid development of biomedical software and hardware, a large amount of relational data interlinking genes, proteins, chemical components, drugs, diseases, and symptoms has been collected for modern biomedical research. Many graph-based learning methods have been proposed to analyze such type of data, giving a deeper insight into the topology and knowledge behind the biomedical data. However, the main difficulty is how to handle high dimensionality and sparsity of the data. Recently, graph embedding methods provide an effective and efficient way to address the above issues. It converts graph-based data into a low dimensional vector space where the graph structural properties and knowledge information are well preserved. In this paper, we conduct a literature review of recent graph embedding techniques for biomedical data. We also introduce important applications and tasks in the biomedical domain as well as associated public biomedical datasets.

Keywords: Graph embedding · Biomedical data · Biomedical informatics.

1 Introduction

With the recent advances in biomedical technology, a large number of relational data interlinking biomedical components including proteins, drugs, diseases, and symptoms, etc. has gained much attention in biomedical academic research. Relational data, also known as the graph, which captures the interactions (i.e., edges) between entities (i.e., nodes), now plays a key role in the modern machine learning domain. Analyzing these graphs provides users a deeper understanding of topology information and knowledge behind these graphs, and thus greatly benefits many biomedical applications such as biological graph analysis [2], network medicine [4], clinical phenotyping and diagnosis [30], etc.

Although graph analytics is of great importance, most existing graph analytics methods suffer the computational cost drawn by high dimensionality and sparsity of the graphs. Furthermore, owing to the heterogeneity of biomedical graphs, i.e., containing multiple types of nodes and edges, traditional analyses

over biomedical graphs remain challenging. Recently, graph embedding methods, aiming at learning a mapping that embeds nodes into a low dimensional vector space $\mathbb{R}^d$, now provide an effective and efficient way to address the problems. Specifically, the goal is to optimize this mapping so that the node representation in the embedding space can well preserve information and properties of the original graphs. After optimization of such representation learning, the learned embedding can then be used as feature inputs for many machine learning downstream tasks, which hence introduces enormous opportunities for biomedical data science. Efforts of applying graph embedding over biomedical data are recently made but still not thoroughly explored; capabilities of graph embedding for biomedical data are also not extensively evaluated. In addition, the biomedical graphs are usually sparse, incomplete, and heterogeneous, making graph embedding more complicated than other application domains. To address these issues, it is strongly motivated to understand and compare the state-of-the-art graph embedding techniques, and further study how these techniques can be adapted and applied to biomedical data science. Thus in this review, we investigate recent graph embedding techniques for biomedical data, which give us better insights into future directions. In this article, we introduce the general models related to biomedical data and omit the complete technical details. For a more comprehensive overview of graph embedding techniques and applications, we refer readers to previous well-summarized papers [7,14,33].

2 Homogeneous Graph Embedding Models

In the literature, homogeneous graphs refer to the graphs with only one type of nodes and edges. There are three main types of homogeneous graph embedding methods, i.e., *matrix factorization-based methods*, *random walk-based methods* and *deep learning-based methods*.

Matrix factorization-based methods. Matrix factorization-based methods, inspired by classic techniques for dimensionality reduction, use the form of a matrix to represent the graph properties, e.g., node pairwise similarity. Generally, there are two types of matrix factorization to compute the node embedding, i.e., *node proximity matrix* and *graph Laplacian eigenmaps*.

For node proximity matrix factorization methods, they usually approximate node proximity into a low dimension. Actually, there are many other solutions to approximate this loss function, such as low rank matrix factorization, regularized Gaussian matrix factorization, etc. For graph Laplacian eigenmaps factorization methods, the assumption is that the graph property can be interpreted as the similarity of pairwise nodes. Thus, to obtain a good representation, the normal operation is that a larger penalty will be given if two nodes with higher similarity are far embedded. There are many works using graph Laplacian-based methods and they mainly differ from how they calculate the pairwise node similarity. For example, BANE [44] defines a new Weisfeiler-Lehman proximity matrix to capture data dependence between edges and attributes; then based on this matrix, BANE learns the node embeddings by formulating a new Weisfiler-Lehman ma-

trix factorization. Recently, NetMF [28] unifies state-of-the-art approaches into a matrix factorization framework with close forms.

Random walk-based methods. Random walk-based methods have been widely used to approximate many properties in the graph including node centrality and similarity. They are more useful when the graph can only partially be observed, or the graph is too large to measure. Two widely recognized random walk-based methods have been proposed, i.e., DeepWalk [27] and node2vec [15]. Concretely, DeepWalk considers the paths as sentences and implements an NLP model to learn node embeddings. Compared to DeepWalk, node2vec introduces a trade-off strategy using breadth-first and depth-first search to perform a biased random walk. In recent years, there are still many random walk-based papers working on improving performance. For example, AWE [19] uses a recently developed method called *anonymous walks*, i.e., an anonymized version of the random walk-based method providing characteristic graph traits and are capable to exactly reconstruct network proximity of a node. AttentionWalk [1] uses the softmax to learn a free-form context distribution in a random walk; then the learned attention parameters guide the random walk, by allowing it to focus more on short or long term dependencies when optimizing an upstream objective. BiNE [13] proposes methods for bipartite graph embedding by performing biased random walks. Then they generate vertex sequences that can well preserve the long-tail distribution of vertices in original bipartite graphs.

Deep learning-based methods. Deep learning has shown outstanding performance in a wide variety of research fields. SDNE [37] applies a deep autoencoder to model non-linearity in the graph structure. DNGR [8] learns deep low-dimensional vertex representations, by using the stacked denoising autoencoders on the high-dimensional matrix representations. Furthermore, Graph Convolutional Network (GCN) [20] introduces a well-behaved layer-wise propagation rule for the neural network model. Another important work is Graph Attention Network (GAT) [36], which leverages masked self-attentional layers to address the shortcomings of prior graph convolution-based methods. GAT computes normialized coefficients using the softmax function across different neighborhoods by a byproduct of an attentional mechanism across node pairs. To stabilize the learning process of self-attention, GAT uses multi-head attention to replicate K times of learning phases, and outputs are feature-wise aggregated, typically by concatenating or adding.

3 Heterogeneous Graph Embedding Models

Heterogeneous graphs mean that there are more than one type of nodes or edges within. The heterogeneity in both graph structures and node attributes makes it challenging for the graph embedding task to encode their diverse and rich information. In this section, we will introduce *translational distance methods* and *semantic matching methods*, which try to address the above issue by constructing different energy functions. Furthermore, we will introduce *meta-path-based methods* that use different strategies to capture graph heterogeneity.

Translational distance methods. The first work of translation distance models is TransE [6]. The basic idea of the translational distance models is, for each observed fact (h, r, t) representing head entity h having a relation r with tail entity t, to learn a good graph representation such that h and t are closely connected by relation r in low dimensional embedding space, i.e., $\mathbf{h} + \mathbf{r} \approx \mathbf{t}$, using geometric notations. Here $\mathbf{h}$, $\mathbf{r}$ and $\mathbf{t}$ are embedding vectors for entities h, t and relation r, respectively. To further improve the TransE model and address its inadequacies, many recent works have been developed. For example, RotatE [34] defines each relation as a rotation from the source entity to the target entity in the complex vector space. QuatE [45] computes node embedding vectors in the hypercomplex space with three imaginary components, as opposed to the standard complex space with a single real component and imaginary component. MuRP [3] is a hyperbolic embedding method that embeds multi-relational data in the Poincaré ball model of hyperbolic space, which can well perform in hierarchical and scale-free graphs.

Semantic matching methods. Semantic matching models exploit similarity-based scoring functions. They measure plausibility of facts by matching latent semantics of entities and relations embodied in their representations. Targetting the observed fact (h, r, t), RESCAL [26] embeds each entity with a vector to capture its latent semantics and each relation with a matrix to model pairwise interactions between latent factors. HolE [25] deals with directed graphs and composes head entity and tail entity by their circular correlation, which achieves a better performance than RESCAL. There are other works trying to extend or simplify RESCAL, e.g., DistMult [43], ComplEx [35]. Another direction of semantic matching methods is to fuse neural network architecture by considering embedding as the input layer and energy function as the output layer. For instance, SME model [5] first inputs embeddings of entities and relations in the input layer. The relation r is then combined with the head entity h to get $g_{left}(h, r) = M_1\mathbf{h} + M_2\mathbf{r} + \mathbf{b}_h$, and with the tail entity t to get $g_{right}(t, r) = M_3\mathbf{t} + M_4\mathbf{r} + \mathbf{b}_t$ in the hidden layer. The score function is defined as $f_r(h, t) = g_{left}(h, r)^T \cdot g_{right}(t, r)$. There are other semantic matching methods using neural network architecture, e.g., NTN [31], MLP [10].

Meta-path-based methods. Generally, a meta-path is an ordered path that consists of node types and connects via edge types defined on the graph schema, e.g., $A_1 \xrightarrow{R_1} A_2 \cdots \xrightarrow{R_{l-1}} A_l$, which describes a composite relation between node types $A_1, A_2, \cdots, A_l$ and edge types $R_1, \cdots, R_{l-1}$. Thus, meta-paths can be viewed as high-order proximity between two nodes with specific semantics. A set of recent works have been proposed. Metapath2vec [11] computes node embeddings by feeding metapath-guided random walks to a skip-gram [24] model. HAN [41] learns meta-path-oriented node embeddings from different meta-path-based graphs converted from the original heterogeneous graph and leverages the attention mechanism to combine them into one vector representation for each node. HERec [29] learns node embeddings by applying DeepWalk [27] to the meta-path-based homogeneous graphs for recommendation. MAGNN [12] comprehensively considers three main components to achieve the state-of-the-art

performance. Concretely, MAGNN [12] fuses the node content transformation to encapsulate node attributes, the intra-metapath aggregation to incorporate intermediate semantic nodes, and the inter-metapath aggregation to combine messages from multiple metapaths.

Other methods. LANE [18] constructs proximity matrices by incorporating label information, graph topology, and learns embeddings while preserving their correlations based on Laplacian matrix. EOE [42] aims to embed the graph coupled by two non-attribute graphs. In EOE, latent features encode not only intra-network edges, but also inter-network ones. To tackle the challenge of heterogeneity of two graphs, the EOE incorporates a harmonious embedding matrix to further embed the embeddings. Inspired by generative adversarial network models, HeGAN [16] is designed to be relation-aware in order to capture the rich semantics on heterogeneous graphs and further trains a discriminator and a generator in a minimax game to generate robust graph embeddings.

4 Applications and Tasks in Biomedical Domain

In recent years, graph embedding methods have been applied in biomedical data science. In this section, we will introduce some main biomedical applications of applying graph embedding techniques, including *pharmaceutical data analysis*, *multi-omics data analysis* and *clinical data analysis*.

Pharmaceutical data analysis. Generally, there are two main types of applications, i.e., (i) *drug repositioning* and (ii) *adverse drug reaction analysis*. Drug repositioning usually aims to predict unknown drug-target or drug-disease interactions. Recently, DTINet [23] generates drug and target-protein embedding by performing random walk with restart on heterogeneous biomedical graphs to make predictions based on geometric proximity. Other studies over drug repositioning focused on predicting drug disease associations. For instance, Wang et al. [39] propose to detect unknown drug-disease interactions from the medical literature by fusing NLP and graph embedding techniques. An adverse drug reaction (ADR) is defined as any undesirable drug effect out of its desired therapeutic effects that occur at a usual dosage, which now is the center of drug development before a drug is launched on the clinical trial. Recently, inspired by translational distance models, Stanovsky et al. [32] propose a deep learning model to recognize ADR mentions in social media by infusing DBpedia.

Multi-omics data analysis. The main aim of multi-omics is to study structures, functions, and dynamics of organism molecules. Fortunately, graph embedding now becomes a valuable tool to analyze relational data in omics. Concretely, the computation tasks included in multi-omics data analysis are mainly about (i) *genomics*, (ii) *proteomics* and (iii) *transcriptomics*. Works of graph embedding used in genomics data analysis usually try to decipher biology from genome sequences and related data. For example, based on gene-gene interaction data, a recent work [22] addresses representation learning for single cell RNA-seq data, which outperforms traditional dimensional reduction methods according to the experimental results. As we have introduced before, PPIs play key roles in most cell functions. Graph embedding has also been introduced to PPI graphs for proteomics data analysis, such as assessing and predicting PPIs or predicting protein

functions, etc. Recently, ProSNet [40] has been proposed for protein function prediction. In this model, they introduce DCA to a heterogeneous molecular graph and further use the meta-path-based methods to modify DCA for preserving heterogeneous structural information. As for transcriptomics study, the focus is to analyze an organism's transcriptome. For instance, Identifying miRNA-disease associations now becomes an important topic of pathogenicity; while graph embedding now provides a useful tool to involve in transcriptomics for prediction of miRNA-disease associations. Li et al. [21] propose a method by using DeepWalk to embed the bipartite miRNA-disease network to make association prediction for miRNA-disease graphs.

Clinical data analysis. Graph embedding techniques have been applied to clinic data, such as electronic medical records (EMRs), electronic health records (EHRs) and medical knowledge graphs, providing useful assistance and support for clinicians in recent clinic development. To address the heterogeneity of EMRs and EHRs data, GRAM [9] learns EHR representation with the help of hierarchical information inherent to medical ontologies. ProSNet [17] constructs a biomedical knowledge graph to learn the embeddings of medical entities. The proposed method is used to visualize the Parkinson's disease data set. Conducting medical knowledge graph is of great importance and attention recently. For instance, Zhao et al. [47] define energy function by considering the relation between the symptoms of patients and diseases as a translation vector to further learn the representation of medical forum data. Then a new method is proposed to learn embeddings of medical entities in the medical knowledge graph, based on the energy functions of RESCAL and TransE [46]. Wang et al. [38] construct the objective function by using both the energy function of TransR and LINE's 2nd-order proximity measurement to learn embeddings from a heterogeneous medical knowledge graph to further recommend proper medicine to patients.

5 Conclusion

Graph embedding methods aim to learn compact and informative representations for graph analysis and thus provide a powerful opportunity to solve the traditional graph-based machine learning problems both effectively and efficiently. With the rapid development of relational data in the biomedical domain, applying graph embedding techniques now draws much attention in numerous biomedical applications. In this paper, we introduce recent developments of graph embedding methods. By summarizing biomedical applications with graph embedding methods, we provide perspectives over this emerging research domain for better improvement in human health care.

Acknowledgement

The work described in this paper was partially supported by The Chinese University of Hong Kong (CUHK 3133238, Research Sustainability of Major RGC Funding Schemes (RSFS)).

References

1. Abu-El-Haija, S., Perozzi, B., Al-Rfou, R., Alemi, A.A.: Watch your step: Learning node embeddings via graph attention. In: NeurIPS. pp. 9180–9190 (2018)
2. Albert, R.: Scale-free networks in cell biology. Journal of cell science (2005)
3. Balazevic, I., Allen, C., Hospedales, T.: Multi-relational poincaré graph embeddings. In: NeurIPS. pp. 4465–4475 (2019)
4. Barabási, A.L., Gulbahce, N., Loscalzo, J.: Network medicine: a network-based approach to human disease. Nature reviews genetics **12**(1), 56–68 (2011)
5. Bordes, A., Glorot, X., Weston, J., Bengio, Y.: A semantic matching energy function for learning with multi-relational data. ML **94**(2), 233–259 (2014)
6. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: NeurIPS. pp. 2787–2795 (2013)
7. Cai, H., Zheng, V.W., Chang, K.C.C.: A comprehensive survey of graph embedding: Problems, techniques, and applications. TKDE **30**(9), 1616–1637 (2018)
8. Cao, S., Lu, W., Xu, Q.: Deep neural networks for learning graph representations. In: AAAI (2016)
9. Choi, E., Bahadori, M.T., Song, L., Stewart, W.F., Sun, J.: Gram: graph-based attention model for healthcare representation learning. In: SIGKDD (2017)
10. Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., Zhang, W.: Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In: SIGKDD. pp. 601–610 (2014)
11. Dong, Y., Chawla, N.V., Swami, A.: metapath2vec: Scalable representation learning for heterogeneous networks. In: SIGKDD. pp. 135–144 (2017)
12. Fu, X., Zhang, J., Meng, Z., King, I.: Magnn: Metapath aggregated graph neural network for heterogeneous graph embedding. In: WWW. pp. 2331–2341 (2020)
13. Gao, M., Chen, L., He, X., Zhou, A.: Bine: Bipartite network embedding. In: SIGIR. pp. 715–724 (2018)
14. Goyal, P., Ferrara, E.: Graph embedding techniques, applications, and performance: A survey. Knowledge-Based Systems **151**, 78–94 (2018)
15. Grover, A., Leskovec, J.: node2vec: Scalable feature learning for networks. In: SIGKDD. pp. 855–864 (2016)
16. Hu, B., Fang, Y., Shi, C.: Adversarial learning on heterogeneous information networks. In: SIGKDD. pp. 120–129 (2019)
17. Huang, E.W., Wang, S., Zhai, C.: Visage: Integrating external knowledge into electronic medical record visualization. In: PSB. pp. 578–589. World Scientific (2018)
18. Huang, X., Li, J., Hu, X.: Label informed attributed network embedding. In: WSDM. pp. 731–739 (2017)
19. Ivanov, S., Burnaev, E.: Anonymous walk embeddings. arXiv:1805.11921 (2018)
20. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv:1609.02907 (2016)
21. Li, G., Luo, J., Xiao, Q., Liang, C., Ding, P., Cao, B.: Predicting microrna-disease associations using network topological similarity based on deepwalk. IEEE Access **5**, 24032–24039 (2017)
22. Li, X., Chen, W., Chen, Y., Zhang, X., Gu, J., Zhang, M.Q.: Network embedding-based representation learning for single cell rna-seq data. Nucleic acids research **45**(19), e166–e166 (2017)
23. Luo, Y., Zhao, X., Zhou, J., Yang, J., Zhang, Y., Kuang, W., Peng, J., Chen, L., Zeng, J.: A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. Nature communications **8**(1), 1–13 (2017)

24. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: ICLR (Workshop Poster) (2013)
25. Nickel, M., Rosasco, L., Poggio, T.: Holographic embeddings of knowledge graphs. In: AAAI (2016)
26. Nickel, M., Tresp, V., Kriegel, H.P.: A three-way model for collective learning on multi-relational data. In: ICML. vol. 11, pp. 809–816 (2011)
27. Perozzi, B., Al-Rfou, R., Skiena, S.: Deepwalk: Online learning of social representations. In: SIGKDD. pp. 701–710 (2014)
28. Qiu, J., Dong, Y., Ma, H., Li, J., Wang, K., Tang, J.: Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec. In: WSDM (2018)
29. Shi, C., Hu, B., Zhao, W.X., Philip, S.Y.: Heterogeneous information network embedding for recommendation. TKDE **31**(2), 357–370 (2018)
30. Shickel, B., Tighe, P.J., Bihorac, A., Rashidi, P.: Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. IEEE journal of biomedical and health informatics **22**(5), 1589–1604 (2017)
31. Socher, R., Chen, D., Manning, C.D., Ng, A.: Reasoning with neural tensor networks for knowledge base completion. In: NeurIPS. pp. 926–934 (2013)
32. Stanovsky, G., Gruhl, D., Mendes, P.: Recognizing mentions of adverse drug reaction in social media using knowledge-infused recurrent models. In: EACL (2017)
33. Su, C., Tong, J., Zhu, Y., Cui, P., Wang, F.: Network embedding in biomedical data science. Briefings in bioinformatics **21**(1), 182–197 (2020)
34. Sun, Z., Deng, Z., Nie, J., Tang, J.: Rotate: Knowledge graph embedding by relational rotation in complex space. In: ICLR (Poster). OpenReview.net (2019)
35. Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., Bouchard, G.: Complex embeddings for simple link prediction. ICML (2016)
36. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. arXiv:1710.10903 (2017)
37. Wang, D., Cui, P., Zhu, W.: Structural deep network embedding. In: SIGKDD. pp. 1225–1234 (2016)
38. Wang, M., Liu, M., Liu, J., Wang, S., Long, G., Qian, B.: Safe medicine recommendation via medical knowledge graph embedding. arXiv:1710.05980 (2017)
39. Wang, P., Hao, T., Yan, J., Jin, L.: Large-scale extraction of drug–disease pairs from the medical literature. Journal of the AIST **68**(11), 2649–2661 (2017)
40. Wang, S., Qu, M., Peng, J.: Prosnet: Integrating homology with molecular networks for protein function prediction. In: PSB. pp. 27–38. World Scientific (2017)
41. Wang, X., Ji, H., Shi, C., Wang, B., Ye, Y., Cui, P., Yu, P.S.: Heterogeneous graph attention network. In: WWW. pp. 2022–2032 (2019)
42. Xu, L., Wei, X., Cao, J., Yu, P.S.: Embedding of embedding (eoe) joint embedding for coupled heterogeneous networks. In: WSDM. pp. 741–749 (2017)
43. Yang, B., Yih, W.t., He, X., Gao, J., Deng, L.: Embedding entities and relations for learning and inference in knowledge bases. arXiv:1412.6575 (2014)
44. Yang, H., Pan, S., Zhang, P., Chen, L., Lian, D., Zhang, C.: Binarized attributed network embedding. In: ICDM. pp. 1476–1481. IEEE (2018)
45. Zhang, S., Tay, Y., Yao, L., Liu, Q.: Quaternion knowledge graph embeddings. In: NeurIPS. pp. 2731–2741 (2019)
46. Zhao, C., Jiang, J., Guan, Y., Guo, X., He, B.: EMR-based medical knowledge representation and inference via markov random fields and distributed representation learning. Artificial intelligence in medicine **87**, 49–59 (2018)
47. Zhao, S., Jiang, M., Yuan, Q., Qin, B., Liu, T., Zhai, C.: Contextcare: Incorporating contextual information networks to representation learning on medical forum data. In: IJCAI. pp. 3497–3503 (2017)