STRUCTURAL NODE EMBEDDINGS IN GRAPHS VIA ANONYMOUS WALKS

Alfredo De la Fuente
Department of Statistical Learning Theory
Skolkovo Institute of Science and Technology
alfredo.delafuente@skoltech.ru

Maxim Panov
Assistant Professor
Skolkovo Institute of Science and Technology
m.panov@skoltech.ru

ABSTRACT

Graph embedding constitutes the challenging task of efficiently capturing the structural information of graph data in a lower dimensional space to be subsequently used as input for downstream tasks. We analyze the problem of network representation learning under an structural perspective; i.e., to encapsulate in our node embeddings the local structure of the subgraph surrounding each node, which in practice is helpful to describe real-world network datasets without side information and structural node identities. One of the main difficulties faced by different proposed models in the literature, specially the family of random walks based models, lies in producing embeddings based on homophily rather than structure, as well as their lack of robustness for generalizing under perturbations.

We propose and analyze a unsupervised embedding approach that builds upon anonymous walks statistics as a theoretically-justified mechanism to encapsulate the structure of the graph. We successfully implement and benchmark our proposed model with state-of-the-art random walk based models using well-known airline networks for node classification task. The experiments show that our approach under certain hyperparameter configurations outperforms the other models by a considerable margin. Additionally, we include an adversarial vulnerability analysis for the produced embeddings.

1 INTRODUCTION

Graph embedding aims to represent highly structured data into a low dimensional vector, which lies in a space that preserves structural information. Many efforts have been devoted over the last years to study and improve graph embedding approaches (Goyal & Ferrara, 2017; Cai et al., 2018) mostly due to an increasing amount of applications whose benefit from graph data across a wide variety of machine learning areas; such as natural language processing, bioinformatics (De Cao & Kipf, 2018), social network analysis (Handcock et al., 2007; Jacob et al., 2014; Li et al., 2018), and recently as building blocks of reinforcement learning algorithms (Sanchez-Gonzalez et al., 2018; Hamrick et al., 2018; Qu et al., 2018).

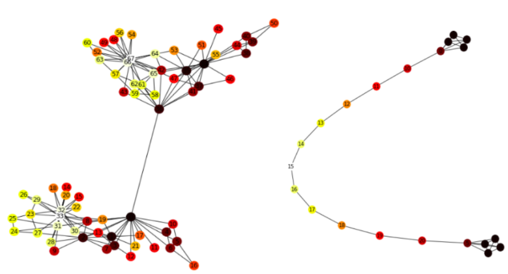

Figure 1: Mirrored Karate Club (left) and Barbel Graph (right).

Within network representation learning, we focus on the task of unsupervised node embedding, where we aim to reconstruct the structural role of a node solely based on its surrounding topology, for further node classification tasks (see example at Figure 1).

Main Contribution. We propose a scalable unsupervised node embedding algorithm that by integrating statistics derived from anonymous walks allow us to encapsulate structural local information. Our key contribution is defining an scheme that allows to efficiently learn new embeddings for node classification task in an inductive manner, so that for any new node added to the graph we will only require a set of statistics derived from its anonymous walks, and not explicitly training the full model with the new adjacency matrix. The method in addition, enforces the structural criteria over homophily at learning embeddings, which is particularly advantageous for many real datasets in comparison with random walks or graph convolutional networks based models.

1.1 Node Embeddings

The research area of node embedding have recently experienced a increasing development particularly in three directions: matrix factorization approaches, random walks based embeddings and deep learning graph convolutional networks.

Within matrix factorization graph-based embedding, we represent the node similarity as a matrix and factorize it to produce the node embeddings. We can distinguish two well-known categorizations: graph Laplacian Eigenmaps (Belkin & Niyogi, 2003); and node proximity matrix factorization methods, which include, for example, HOPE by Ou et al. (2016) or GraRep by Cao et al. (2015)). Despite both approaches consider global statistics of node proximity, one serious weakness they present, however, is that they require large computation time and memory resources, which can be inefficient for huge graphs.

More recently, by adopting the natural language model SkipGram (Mikolov et al., 2013) for graph embedding, deep learning approaches started to gain popularity in the research community by obtaining successful results on node classification tasks. This formulation considers a context associated to each instance (node) , and our goal is to minimize the log loss of predicting the context using the embedding of an instance as input. DeepWalk (Perozzi et al., 2014) uses the embedding of a node to predict its surrounding context, which is obtained by random walks. Similarly, Node2vec (Grover & Leskovec, 2016) performs random walks to define its context but incorporating a trade-off between breath-first-search(BFS) and depth-first-search(DFS) to guide the walks and produce more meaningful embeddings. We must realize that the principle behind these methods lies on the fact that measures of similarity and proximity can be encoded by random walks.

In the same spirit, other approaches make use of deep learning architectures to learn embeddings in a unsupervised setting, by the help of context generating mechanism: via diffusion wavelets (GraphWave, Donnat et al. (2017)), via anonymous walks (Ivanov & Burnaev, 2018), by incorporating structural metrics (struc2vec, (Ribeiro et al., 2017)) , similarity metrics (VERSE, Tsitsulin et al. (2018)) or using embedding propagation mechanisms (Garcia-Duran & Niepert, 2017).

Despite the remarkable results obtained by the random walk based models, these methods can be very computationally expensive specially for large and sparse graphs. For that reason, Graph Convolutional Networks (GCNs) (Kipf & Welling, 2016) were introduced to tackle this problem by defining a convolution operator which are based on spectral filters. The model iteratively adds the embeddings of neighbor nodes in order to update the new node embedding. The idea was extended by Kipf & Welling to evaluate the use of GCN architectures for Variational Graph Autoencoders in a graph embedding task. Continuing with the line of work, later Generative Adversarial Models were proposed (Dai et al., 2017; Bojchevski et al., 2018) for graph embedding tasks, which generalizes the problem to tackle a wider range of applications for structured data.

Although these models have obtained impressive results in graphs and nodes embedding tasks, most of them require to process the entire graph adjancency matrix as input for the neural network. Instead, our proposed model is able to efficiently learn node embeddings for a subset of nodes as long as enough random anonymized walks are sampled. Therefore, by similar nature, we will compare thorough this work our method against random-walk based models for benchmarking purposes.

2 THEORY

2.1 ANONYMOUS RANDOM WALKS

Intuitively, we can define anonymous random walks as a sequence of anonymized (order-based) labeled nodes where each node is independently selected from the set of previous node's neighbors. Under this particular scenario, based on theoretical provable guarantees (Micali & Zhu, 2016), anonymous random walks allows us to encapsulate and reconstruct the structure of the graph despite the lack of global node labels. Let us introduce the following definitions to formally express the above described concept.

DEFINITION Let us consider $s = (v_1, \ldots, v_m)$ to be an ordered sequence of length m, such that $v_i \in V \ \forall i$. We define the following mapping $pos(s, u_i) \mapsto p$, such that for any ordered sequence s and an element $u_i \in V$, returns a list of integers $p = (p_1, \ldots, p_k)$, such that p_i is the position where u_i appeared in sequence s.

DEFINITION Let us consider $w = (v_1, \ldots, v_m)$ a random walk of length m, we define the associated anonymous random walk given by $a = (f(v_1), \ldots f(v_m))$ such that $f(v_i) = \min_{p_j \in pos(w,v_i)} \ pos(w, v_i)$.

According to Micali & Zhu (2016), we can guarantee that for a single node v_i the distribution D_m of anonymous random walks of length m is sufficient to reconstruct the immediate topology within a certain radius around v_i. Therefore, we intuitively justify the use of anonymous walks as described by Ivanov & Burnaev (2018), to encode local topological features of the graph for each node and therefore adding consistent information about the general global structure in the embeddings.

2.2 PROPOSED GRAPH EMBEDDING

In contrast with Ivanov & Burnaev (2018) model for graph embeddings, we define a memory efficient way to represent the anonymous walks distribution, by defining an sparse tensor $T^k_{i,j}$ which stores the frequency of walks that contain a transition from a node i to a node j during step k for each particular anonymous walk up to a certain length. This simple reparametrization allow us to reduce the encoding space of each anonymous walk from an exponential to a polynomial (at most cube) number of parameters. Thus, by normalizing the probabilities appropriately, we obtain the following expression to encode the set of anonymous walks from node t,

$$\text{enc}(AW_t) = \mathbf{vec}(\hat{T}^k_{i,j}) \ , \hat{T}^k_{i,j} = \frac{T^k_{i,j}}{\sum_j T^k_{i,j}} \tag{1}$$

Based o the theoretical guarantees mentioned above, we proposed a model that uses our AW representation, which capture the local structure statistics, and encodes those sets of walks using *Set2Vec* proposed by Vinyals et al. (2015), which uses attention mechanism together with LSTM in order to find an order invariant representation.

$$
\begin{aligned}
s_t &= \mathrm{LSTM}(s_{t-1}) \\
e_{it} &= f(m_i, q_t) \\
a_{it} &= \frac{\exp(e_{it})}{\sum_j \exp(e_{it})} \\
r_t &= \sum_i a_{it} m_i \\
s_t^* &= [s_t, m_t]
\end{aligned}
\tag{2}
$$

3 RESULTS

We evaluate our AWE based model on toy graphs (Barbel and Mirrored Karate) as well as real network data (Europe and Brazil Airport datasets) following the same experimental setting as described by Ribeiro et al. (2017). For the toy examples, we display in Fig. 2 the 2d projection of the embedding vectors, using MultiDimensional Scaling (MDS) for reducing dimensionality, in order to visually inspect the structural role component within the learned embeddings.

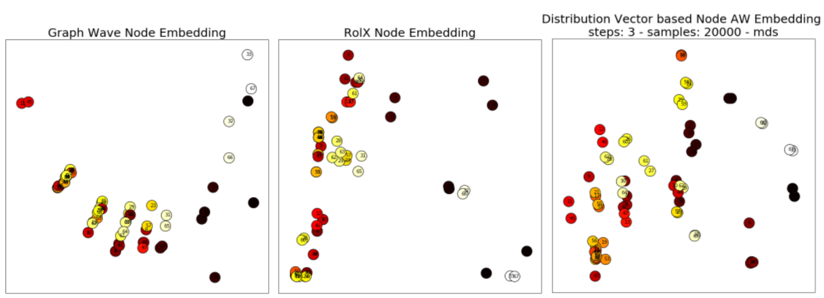

Figure 2: Embeddings produced by *GraphWave*, *RolX* and our model for Mirrored-Karate graph.

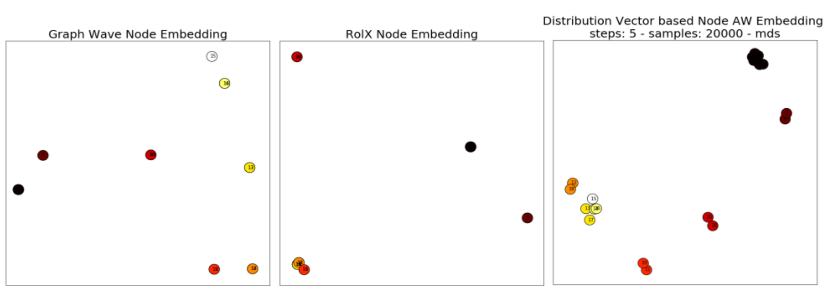

Figure 3: Embeddings produced by *GraphWave*, *RolX* and our model for Babel Graph.

For the real airport datasets, we evaluated the accuracy at one vs rest classification for 4 classes. We compared results by different methods *RolX*, *GraphWave*, *Node2Vec*, *Struc2Vec*, *DeepWalk*, Laplacian EigenMap and *Graph Factorization*. Our proposed model is an embedding which leverages *Struc2Vec* with our Set2Vec model which uses our AW representations up to length 7.

In order to evaluate the quality of the AWE embeddings, both under unsupervised and supervised metrics, we run our model for a toy graph dataset described on GraphWave paper by Donnat et al. (2017). This set of simple graphs provides a valid benchmark where we observe our proposed model performs as good as the best comparison models, however it exhibits more robustness under the presence of noise in the adjacency matrix (adversarial perturbations) as observed in Table 1.

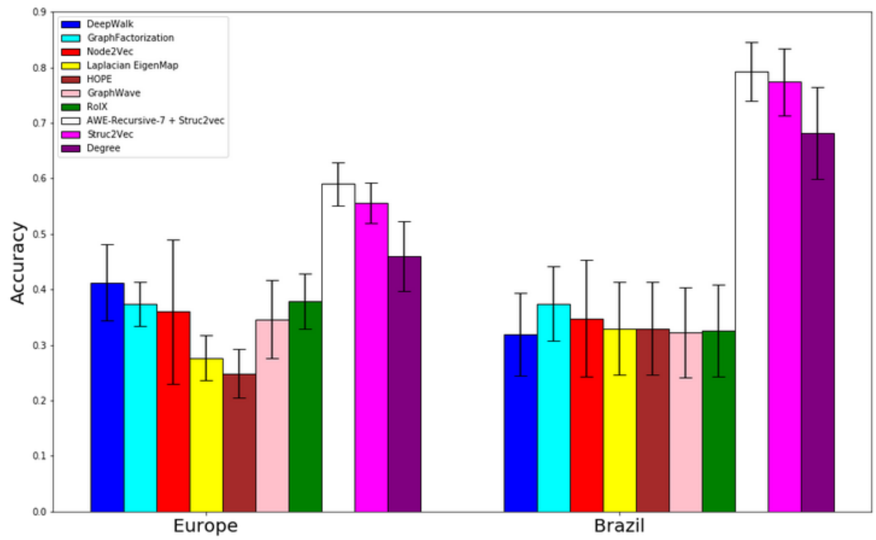

Figure 4: Node Classification Accuracy Benchmark using random walk and matrix decomposition based models, in contrast with anonymous walks proposed model stacked with *Struc2Vec*.

Graph	Method	Homogeneity	Completeness	Silhouette	F1	Accuracy
House	AWE	1.00	1.00	0.88	1.00	1.00
	GraphWave	1.00	1.00	1.00	1.00	1.00
	RolX	1.00	1.00	1.00	1.00	1.00
	DeepWalk	0.17	0.17	0.33	0.15	0.23
	Node2Vec	0.36	0.37	0.18	0.17	0.24
	Struc2vec	1.00	1.00	0.88	1.00	1.00
House Perturbed	AWE	**0.78**	**0.78**	0.40	**0.86**	**0.89**
	GraphWave	0.60	0.59	0.44	0.52	0.62
	RolX	0.68	0.69	0.48	0.74	0.80
	DeepWalk	0.26	0.26	0.35	0.19	0.29
	Node2Vec	0.31	0.33	0.17	0.27	0.36
	Struc2vec	0.61	0.66	0.55	0.58	0.68

Table 1: Benchmark results with toy structured graphs reporting unsupervised and supervised learning metrics following the experimental setting given by Donnat et al. (2017) averaging results over 10 random samples.

4 CONCLUSIONS

Anonymous Walks Embeddings constitute a valuable source of structural local information in graph data, which can substantially leverage the quality of embeddings derived from other methods. We aim to extend our current model to incorporate recent results with Variational Graph Auto-Encoders (VGAE) with the addition of Anonymous Walks samples, to obtain more meaningful latent representations and better performance.

REFERENCES

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

Aleksandar Bojchevski, Oleksandr Shchur, Daniel Zügner, and Stephan Günnemann. NetGAN: Generating Graphs via Random Walks. 2018. URL `http://arxiv.org/abs/1803.00816`.

Hong Yun Cai, Vincent W. Zheng, and Kevin Chang. A Comprehensive Survey of Graph Embedding: Problems, Techniques and Applications. *IEEE Transactions on Knowledge and Data Engineering*, XX(Xx), 2018. ISSN 10414347. doi: 10.1109/TKDE.2018.2807452.

Shaosheng Cao, Wei Lu, and Qiongkai Xu. Grarep: Learning graph representations with global structural information. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pp. 891–900. ACM, 2015.

Q. Dai, Q. Li, J. Tang, and D. Wang. Adversarial Network Embedding. *ArXiv e-prints*, November 2017.

N. De Cao and T. Kipf. MolGAN: An implicit generative model for small molecular graphs. *ArXiv e-prints*, May 2018.

C. Donnat, M. Zitnik, D. Hallac, and J. Leskovec. Learning Structural Node Embeddings Via Diffusion Wavelets. *ArXiv e-prints*, October 2017.

A. Garcia-Duran and M. Niepert. Learning Graph Representations with Embedding Propagation. *ArXiv e-prints*, October 2017.

Palash Goyal and Emilio Ferrara. Graph Embedding Techniques, Applications, and Performance: A Survey. 2017. URL `http://arxiv.org/abs/1705.02801`.

Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 855–864. ACM, 2016.

J. B. Hamrick, K. R. Allen, V. Bapst, T. Zhu, K. R. McKee, J. B. Tenenbaum, and P. W. Battaglia. Relational inductive bias for physical construction in humans and machines. *ArXiv e-prints*, June 2018.

Mark S Handcock, Adrian E Raftery, and Jeremy M Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354, 2007.

S. Ivanov and E. Burnaev. Anonymous Walk Embeddings. *ArXiv e-prints*, May 2018.

Yann Jacob, Ludovic Denoyer, and Patrick Gallinari. Learning latent representations of nodes for classifying in heterogeneous social networks. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pp. 373–382. ACM, 2014.

Thomas N Kipf and Max Welling. Variational Graph Auto-Encoders. (2):1–3.

Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. pp. 1–14, 2016. ISSN 0004-6361. doi: 10.1051/0004-6361/201527329. URL `http://arxiv.org/abs/1609.02907`.

Ye Li, Chaofeng Sha, Xin Huang, and Yanchun Zhang. Community Detection in Attributed Graphs: An Embedding Approach. *Proceedingss of AAAI*, pp. 338–345, 2018.

Silvio Micali and Zeyuan Allen Zhu. Reconstructing Markov processes from independent and anonymous experiments. *Discrete Applied Mathematics*, 200:108–122, 2016. ISSN 0166218X. doi: 10.1016/j.dam.2015.06.035. URL `http://dx.doi.org/10.1016/j.dam.2015.06.035`.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

Mingdong Ou, Peng Cui, Jian Pei, Ziwei Zhang, and Wenwu Zhu. Asymmetric transitivity preserving graph embedding. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1105–1114. ACM, 2016.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. DeepWalk: Online Learning of Social Representations. 2014. ISSN 9781450329569. doi: 10.1145/2623330.2623732. URL http://arxiv.org/abs/1403.6652{%}0Ahttp://dx.doi.org/10.1145/2623330.2623732.

Meng Qu, Jian Tang, and Jiawei Han. Curriculum learning for heterogeneous star network embedding via deep reinforcement learning. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, pp. 468–476, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5581-0. doi: 10.1145/3159652.3159711. URL http://doi.acm.org/10.1145/3159652.3159711.

L. F. R. Ribeiro, P. H. P. Saverese, and D. R. Figueiredo. struc2vec: Learning Node Representations from Structural Identity. *ArXiv e-prints*, April 2017.

Alvaro Sanchez-Gonzalez, Nicolas Heess, Jost Tobias Springenberg, Josh Merel, Martin Riedmiller, Raia Hadsell, and Peter Battaglia. Graph networks as learnable physics engines for inference and control. 2018. URL http://arxiv.org/abs/1806.01242.

Anton Tsitsulin, Davide Mottin, Panagiotis Karras, and Emmanuel Müller. VERSE: Versatile Graph Embeddings from Similarity Measures. 2018. doi: 10.1145/3178876.3186120. URL http://arxiv.org/abs/1803.04742{%}0Ahttp://dx.doi.org/10.1145/3178876.3186120.

Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. Order matters: Sequence to sequence for sets. *arXiv preprint arXiv:1511.06391*, 2015.