# NetGloVe: Learning Node Representations for Community Detection

Kumaran Gunasekaran, Jeyavaishnavi Muralikumar, Srinivasa Sudarshan, Balasubramaniam Srinivasan, Fragkiskos Malliaros

# NetGloVe: Learning Node Representations for Community Detection

Kumaran Gunasekaran[1*], Jeyavaishnavi Muralikumar[1*], Sudarshan Srinivasa Ramanujam[1*], Balasubramaniam Srinivasan[1*], and Fragkiskos D. Malliaros[1,2]

[1] Department of Computer Science and Engineering, UC San Diego, USA
[2] Center for Visual Computing, CentraleSupélec and Inria, France
{kugunase, jmuralik, sus046, bsriniva, fmalliaros}@eng.ucsd.edu

## 1 Introduction and Problem Statement

Community detection is a fundamental task in network analysis, with plenty of applications in social networking, biology and neuroscience. In the related literature, a variety of algorithms and methodologies have been proposed to identify the community structure of networks, including graph partitioning methods, hierarchical graph clustering, modularity optimization and spectral techniques (such as spectral clustering and modularity optimization).

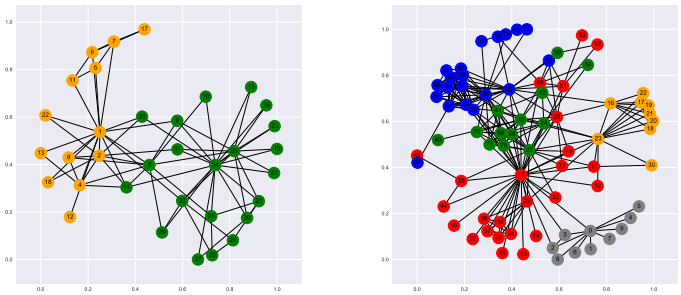
The recent advances in *representation learning* techniques, have allowed us to represent graphs (or nodes) as vectors in a lower dimensional space, that can further be used in graph mining and learning tasks. That way, instead of "manually" extracting features that can be utilized by a graph learning algorithm, we can *learn* informative and discriminative feature representations by solving an optimization problem that takes into account the structural properties of the graph. To this direction, several network feature learning algorithms have been proposed, including node2vec [1] and LINE [4].

The goal of this work is to propose NetGloVe, a new representation learning method for graphs inspired from the domain of Natural Language Processing (NLP), and to examine its application to the task of community detection.

## 2 Feature Learning with NetGloVe

In this paper, we propose NetGloVe, a node representation learning method inspired by GloVe (Global Vectors for Word Representation) [3], a word embeddings technique in NLP. GloVe uses a log bilinear model to derive vector representations of words, taking into consideration both the word co-occurrence statistics as well as the words context. GloVe is comparable, if not superior, to the Skip-gram model, which considers only the words local context to derive the word representation. Our goal here is to extend GloVe to the context of graphs, by finding a suitable analogy for the word-word co-occurrence matrix used by GloVe. Our intuition is that, nodes that belong to the same communities would have similar embeddings, and thus would be clustered together. That way, we have employed the inverse of the shortest path distance between individual pairs of

*Equal contribution; the authors are listed in alphabetical order.



(a) Zachary's karate club (b) Les Misérables

**Fig. 1.** Community detection results using the NetGloVe embeddings for (a) Zachary's karate club and (b) Les Misérables networks.

nodes in the graph, to populate GloVe's co-occurrence matrix – as we wanted to weight higher nodes that are at close proximity. Thus, the loss function of the node embedding scheme is derived as follows:

$$J = \sum_{i} \sum_{j|d_{ij}<k} f\left(\frac{1}{d_{ij}}\right)\left(w_i^T w_j - \frac{1}{d_{ij}}\right)^2, \tag{1}$$

where $f$ corresponds to a weighting function, $w_i$ and $w_j$ correspond to node vectors and $d_{ij}$ corresponds to the distance between nodes $i$ and $j$.

## 3 Experimental Results and Discussion

In our preliminary empirical analysis, we evaluate NetGloVe in the task of community. In particular, we apply NetGloVe to learn feature vectors (i.e., node embeddings) for the nodes of a network, and then we perform $k$-means clustering to extract the underlying communities. For demonstration purposes, we have applied NetGloVe to the well-known *Zachary's karate club* and *Les Misérables* networks, and the results are depicted in Fig. 1.

Our main experimental evaluation results for NetGloVe are based on artificial networks produced by the LFR benchmark generator [2], where we observe the performance of different methods for a range of mixing parameters. We have used several state-of-the-art baseline methods, including popular representation learning methods (node2vec [1] and LINE [4]), as well as more traditional community detection algorithms (Louvain modularity optimization and spectral clustering). The performance is measured using the normalized mutual information (NMI) criterion between the ground-truth community structure given by the benchmarks and the actual clustering results produced by NetGloVe and the rest baseline methods.

Figure 2 depicts the NMI score obtained by NetGloVe and the baseline methods, on artificial networks with $1,000$ nodes, average degree equal to 20 and maximum degree

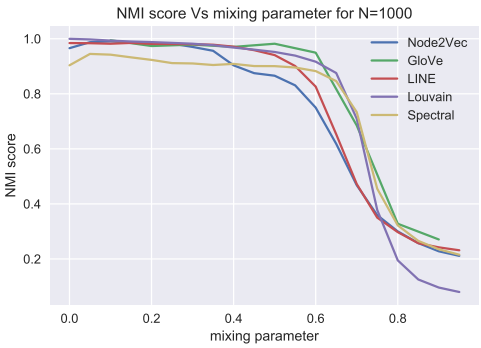



**Fig. 2.** Comparison of community detection methods for LFR graphs with average degree equal to 20 and maximum degree 50.

equal to 50 (average over multiple realizations of the benchmark graphs). For all the representation learning methods (NetGloVe, node2vec and LINE), we set the number of dimensions to $d = 64$ (i.e., each node is represented as a vector $v \in \mathbb{R}^d$). Lastly, the mixing parameter $\mu$ of the generator (x-axis), signifies the ratio of external to internal edges in the graph, with respect to community structure. We observe that NMI declines steeply for all methods for values of $\mu$ greater than 0.6, as the community structure becomes less well-defined. Moreover, we can see the NetGloVe performs as good as the rest representation learning methods (Node2vec and LINE) for lower values of $\mu$, and better for higher values of mixing parameter $\mu$ – which makes it suitable for community detection in graphs with not well-defined community structure.

*Future Work.* As future work, we plan to evaluate the performance of NetGloVe on real-world networks with ground-truth community structure. Another future research direction of particular importance is to examine the performance of the features produced by NetGloVe to *supervised* learning tasks over graphs, including the ones of link prediction and node classification.

## References

1. Grover, A., Leskovec, J.: node2vec: Scalable feature learning for networks. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 855–864 (2016)
2. Lancichinetti, A., Fortunato, S., Radicchi, F.: Benchmark graphs for testing community detection algorithms. Phys. Rev. E, 78(4):046110 (2008)
3. Pennington, J., Socher, R., Manning, C. D.: Glove: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), volume 14, pages 1532–1543 (2014)
4. Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., Mei, Q.: Line: Large-scale information network embedding. Proceedings of the 24th International Conference on World Wide Web (WWW), pages 1067–1077 (2015)