

A Process for the Evaluation of Node Embedding Methods in the Context of Node Classification

Christoph Martin* Meike Riebeling

Institute of Information Systems
Leuphana University of Lüneburg, 21335 Lüneburg, Germany

June 1, 2020

Abstract

Node embedding methods find latent lower-dimensional representations which are used as features in machine learning models. In the last few years, these methods have become extremely popular as a replacement for manual feature engineering.

Since authors use various approaches for the evaluation of node embedding methods, existing studies can rarely be efficiently and accurately compared. We address this issue by developing a process for a fair and objective evaluation of node embedding procedures w.r.t. node classification. This process supports researchers and practitioners to compare new and existing methods in a reproducible way.

We apply this process to four popular node embedding methods and make valuable observations. With an appropriate combination of hyperparameters, good performance can be achieved even with embeddings of lower dimensions, which is positive for the run times of the downstream machine learning task and the embedding algorithm. Multiple hyperparameter combinations yield similar performance. Thus, no extensive, time-consuming search is required to achieve reasonable performance in most cases.

1 Introduction

Networks are used to model phenomena in various domains such as social relations, molecular graphs, biological structures, or recommender systems. Networks represent the relations (edges) between different entities (nodes). Social networks contain information about individuals or communities and the dynamics among them. This information can, for example, be used for segmentation or recommendation tasks. Networks capture not only social relationships, but also

*cmartin@uni.leuphana.de

citations, biological information, or knowledge relations [22]. Developing and experimenting with methods that leverage the information captured by these networks are important endeavors in business and research communities [30, 31].

In various fields, network data are to be used as input for machine learning models. This poses the challenge that network data must first be transformed in order to serve as features. Traditionally, handcrafted features have been created to represent the nodes. This type of feature engineering, however, has considerable weaknesses. It is very time-consuming on the one hand and, on the other hand, the handcrafted features can often not be reused [14]. Node embeddings map the nodes of a graph to a lower-dimensional vector which can subsequently be used as input for other machine learning techniques. However, due to the particular data structure of a network, the quality of network embeddings depends on preserving the structural properties of a graph while incorporating node attributes. This can be difficult as the structural similarity of nodes can either be portrayed as nodes close to each other or as nodes with similar roles in the network, node embeddings have to respect local and global node similarities together [3, 31, 11].

Node embedding methods have enormous potential, thus this area continues to be a highly active field of research. In recent years, several surveys have been published, which summarize the progress made in this area and address the comparison and categorization of node embedding methods [14, 11, 31, 3]. Due to the popularity of embedding methods, a unified way to compare them has become increasingly important. Methods proposed by existing studies can rarely be compared to each other since authors use different approaches to evaluate node embeddings.

We address this issue by developing a process (Section 3) for a fair and objective evaluation of node embedding procedures w.r.t. node classification. Building on and extending existing work [11, 31, 16, 12], we explicitly address the choice of the hyperparameters in the process presented here, under consideration of the downstream machine learning task, in this case node classification. This process supports researchers to compare new and existing methods in a reproducible way. Furthermore, end users can use this process to find the optimal method for the particular use case.

In the case study in Section 4, we apply the process to four popular node embedding methods and make valuable observations, especially for practitioners. The default hyperparameters for node embedding procedures are generally not a good choice. With an appropriate combination of hyperparameters, good performance can be achieved even with embeddings of lower dimensions, which is positive for the run times of the downstream machine learning task. Multiple hyperparameter combinations yield similar performance; hence usually there is no extensive, time-consuming search required to achieve reasonable performance.

2 Node Embeddings

Let G be a graph on N nodes with vertex set $V(G) = \{v_1, v_2, \dots, v_N\}$. Node embeddings are d -dimensional representations of the nodes in G ; usually, these are lower-dimensional (i.e., $d \ll N$). These embeddings are commonly used as input for machine learning algorithms. Node embedding methods have the objective to find such a mapping $f : V(G) \rightarrow \mathbb{R}^d$, where nodes which are “similar” to each other in the graph also “similar” to each other in the vector space. The definition of similarity differs between methods. In the literature, the terms graph embedding or network relational learning are also used for this purpose [23, 31, 14].

In our case study (Section 4), we use the following four frequently cited and widely used node embedding methods: node2vec [13], GraRep [4], Deep Network Graph Representation (DNGR) [5], and Large-scale Information Network Embedding (LINE) [26]. We use the implementations provided by [10, 21].

3 A Process for the Comparison of Node Embedding Methods

In this section, we develop the evaluation process for node embedding methods. This process enables researchers and practitioners to perform a fair and objective evaluation of node embedding procedures. We present this process for two main reasons. The first is to compare new and existing methods in a reproducible way. Furthermore, it helps end users to find the optimal method for the particular use case. We start by arguing why the procedure for selecting hyperparameters cannot easily be transferred from previous machine learning methods to node embedding learning. Then we propose an approach and integrate it with the process.

The evaluation of algorithms and methods is an essential part of machine learning and network analysis research [6, 7]. Particularly, algorithm selection is a widely discussed topic and an essential part of the application of machine learning algorithms in practice. This is due to the fact that there is not one single method optimal for all problem settings [18, 29].

Essential components of evaluation experiments in machine learning are the data set, feature selection, feature representation, and hyperparameter settings [7]. The components of an evaluation process for node embeddings are slightly different. The data set and the hyperparameter settings can be transferred to node embeddings as essential components of the evaluation [31]. However, the feature selection process and the data representation have to be altered. Node embedding methods naturally take a network and the contained information as feature input, essentially making the step of feature selection unnecessary. The necessary representation of the network might differ between algorithms, hence the data representation is implied by choice of the embedding method.

Node embedding methods constitute an unsupervised problem setting traditionally; semi-supervised methods also exist (e.g., [17, 25]), but these are not addressed in this paper. An application task is, therefore, necessary to evaluate the quality of node embeddings and is thus an essential component of the process. In summary, the core components of the process are the network data, the application task, the evaluation metric, and the hyperparameter configuration.

Network data The choice of the network data depends on the setting in which the process is applied and the node embedding methods considered. Practitioners who are looking for the best method for their particular application should use data that is close to the production data. For the comparative evaluation of new and existing embedding methods, in the interests of reproducibility, we recommend using publicly available networks of different size and structure. These may be, for example, the data sets used in the case study in Section 4.

Application task and evaluation metric The most popular application task is node classification, which is often applied when presenting a new embedding method. Classification aims at finding class labels for each node. The vector representation serves as feature input for a classifier [3, 11]. Training a classifier requires training data, which means that labels have to be available at least for a part of the network. Common evaluation metrics in this context include F₁-score, precision, recall, or accuracy. We propose to use the F₁-score since it takes precision and recall into account, $F_1\text{-score} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$. In the case of multi-class and multi-label classification problems, we use the macro and micro variant of the F₁-score. Here the classes, respectively, the individual observations, are weighted equally [24].

For the classification task, we propose to use two popular and often used algorithms in machine learning: a logistic regression model (one-vs-rest classification for the multi-label model) and a random forest model. The regression model because of the frequent usage in the evaluation of embedding methods. The random forest is a widely used model in practical machine learning applications. Nevertheless, it is usually not applied in node embedding research. Therefore we suggest to use it in this context because it is very flexible and leads to good results on different data sets [8].

Hyperparameter configuration The selection of the best hyperparameters is a debated topic in research. The impact of different tuning parameters on each other and how they affect the performance is only poorly understood [20]. In practice, a widely used method to find a set of hyperparameters is random search, where the search space of hyperparameters is randomly explored and evaluated. Begstra and Bengio [2] showed that this type of search leads to equally good or even superior models, compared to grid search, while only a fraction of the time is needed.

In addition to the way the hyperparameter selection is performed, the data utilized for tuning is an important topic. Usually, in machine learning data

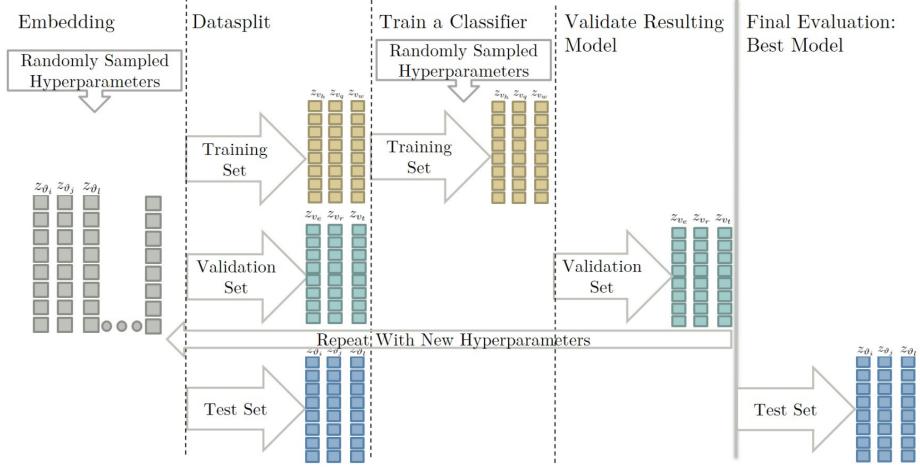


Figure 1: Setup for the hyperparameter selection The embedding algorithm is applied to the whole network resulting in a vector representation marked by gray squares on the left side which is subsequently divided into three splits: training set (yellow), validation set (turquoise) and test set (blue). The training set is used to train the classifier, the validation set is used in that last part of the hyperparameter selection to validate the performance of the combined model. The test set is not used during the hyperparameter selection. After repeating this process several times, the best hyperparameter combination is selected for the final model, which is evaluated on the test set (final step on the right separated by the bold line).

is split into a test, training, and validation set, in which the test set is only used once for the final validation. The training of the algorithm is performed on the training set with a subsequent evaluation of the performance using the validation set [15]. For network embedding procedures, this is not possible. Splitting the network data into different sub-graphs would significantly alter the results of the embedding methods as they rely on representing the whole graph mirroring the structural context information of a node and its position in the whole network. Only using part of the network for an embedding would lead to a completely different representation with important context information missing. The proposed solution for the described challenges is a combined tuning of hyperparameters of the embedding and the subsequent application algorithm. The application task serves as the basis for the performance evaluation governing the hyperparameter selection. As shown in Figure 1, the representation for the whole graph is learned, whereas only part of this data is used in the application task (for example, classification) to evaluate the hyperparameter selection. For both algorithms – the embedding algorithm and the classification algorithm – hyperparameters are selected randomly. This process is repeated several times.

Finally, the best model combination using the best hyperparameters for both algorithms is picked and evaluated on the test set.

4 Case Study for the Comparison Process

In this section, we utilize the process developed in Section 3 to compare four frequently cited and widely used node embedding methods: node2vec, GraRep, LINE, and DNGR. Especially, we are interested in the impact of the number of dimensions and the amount of training data used on the performance in the domain of node classification.

We use data sets with varying characteristics (i.e., directed and undirected as well as binary, multi-class, and multi-label classification) to get an understanding of how embedding procedures behave under different conditions. Table 1 lists basic statistics about these networks. For training and model selection, we use 50% for the training set and 25% for the validation set and test set. For the second part of the experiment, where we analyze the impact of varying amounts of training data, we use 10%, 20%, . . . , 100% of the training data. All of these values refer to the node embedding vectors.

Table 1: Summary of networks used in case study.

| Network | Nodes | Edges | Directionality | # Labels | Source |
|--------------|--------|---------|----------------|------------------|--------|
| Moreno Blogs | 1,224 | 19,025 | directed | 2 (binary) | [1] |
| CiteSeer | 3,312 | 4,660 | directed | 6 (multi-class) | [9] |
| Facebook | 4,039 | 88,234 | undirected | 4 (multi-class) | [19] |
| BlogCatalog | 10,312 | 333,983 | undirected | 39 (multi-label) | [27] |

Overall results The performance of the embedding methods w.r.t. the different classifiers and measures are listed in Table 2. The scores for the logistic regression scenarios reveal that most of the tested algorithms perform similar across the networks. The highest score for the BlogCatalog network is 0.35, which was reached by node2vec. LINE and GraRep reach equal scores of 0.34 on that network. For Facebook, the scores are even closer together, the values vary between 0.45 and 0.52. The same trend can be found in the results of the Moreno network. For the Moreno network, the score of LINE, GraRep, DNGR, and node2vec are the same with 0.95. The best scores for CiteSeer range from 0.53 to 0.57. Only the deep learning-based method yield worse results, DNGR does not work well with a score of 0.25. Overall, the results indicate that very similar scores can be reached across different methods. The observed performance of node2vec, LINE, and GraRep on the BlogCatalog data set are in line with the results reported in the literature. For GraRep and node2vec, evaluation experiments were also conducted using a one-vs-rest logistic regression [4, 13]. Moreover, in [4], LINE was included as a baseline. For all three networks, the performance was around 0.4; the slightly lower performance observed in this

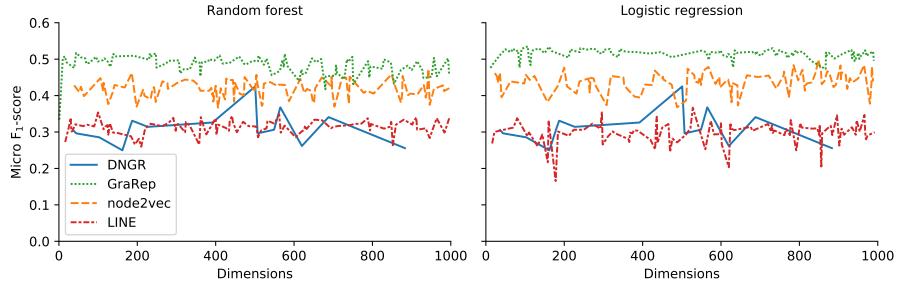
paper might be explained by the use of only 50% of the networks for training, due to the data split in training, validation and test set explained above.

Table 2: Results for the experiments in the case study.

| Score | Classifier | Network Embedding | BlogC. | CiteSeer | Facebook | Moreno |
|----------------------|-----------------|-------------------|--------|----------|----------|--------|
| Macro F ₁ | Random forest | DNGR | 0.020 | 0.180 | 0.434 | 0.941 |
| | | GraRep | 0.111 | 0.555 | 0.489 | 0.944 |
| | | LINE | 0.020 | 0.240 | 0.458 | 0.954 |
| | | node2vec | 0.032 | 0.505 | 0.456 | 0.951 |
| | Log. regression | DNGR | 0.068 | 0.153 | 0.485 | 0.954 |
| | | GraRep | 0.181 | 0.514 | 0.505 | 0.951 |
| | | LINE | 0.195 | 0.469 | 0.427 | 0.948 |
| | | node2vec | 0.212 | 0.493 | 0.412 | 0.954 |
| Micro F ₁ | Random forest | DNGR | 0.052 | 0.266 | 0.450 | 0.941 |
| | | GraRep | 0.244 | 0.607 | 0.505 | 0.944 |
| | | LINE | 0.054 | 0.273 | 0.507 | 0.954 |
| | | node2vec | 0.088 | 0.563 | 0.514 | 0.951 |
| | Log. regression | DNGR | 0.182 | 0.248 | 0.507 | 0.954 |
| | | GraRep | 0.342 | 0.574 | 0.525 | 0.951 |
| | | LINE | 0.339 | 0.536 | 0.450 | 0.948 |
| | | node2vec | 0.354 | 0.534 | 0.496 | 0.954 |
| Most frequent label | | | 0.090 | 0.212 | 0.336 | 0.520 |

Analysis of the number of dimensions The dimensionality of the embedding is the only hyperparameter shared by all node embedding methods. The performance of embedding algorithms should, intuitively, increase with an increasing number of dimension until reaching a plateau where no substantial improvement of performance happens with increasing dimensionality. Grover and Leskovec [13] observed this behavior for the node2vec algorithm. Experimenting with the number of dimensions resulted in a saturation of performance improvement at a dimension of around 100. Similar results are reported by Wang et al. [28]. They noticed a decline in performance after saturation at about 100. For some algorithms like GraRep, little influence of the dimensionality on performance was observed. The reported relation between dimension and performance is almost steady, with a slight decrease after 64 dimensions [4]. In Figure 2, the performance depending on the dimension for the case of the Facebook network is shown. These results indicate that higher dimensions do not necessarily lead to better performance. This behavior also occurs for the other networks. However, analyzing the performance with different dimensions lead to high variances. The reason might lie in the high amount of different hyperparameter combinations since the performance is not only dependent on the dimension but on the combination of parameters picked. Nonetheless, the findings suggest that in combination with the right hyperparameters, small dimen-

Figure 2: Impact of the dimensionality on the performance for the Facebook network.



sions are sufficient to reach scores, that are comparable to the performance with higher dimensions. The results highlight the influence of all hyperparameters on each other. Therefore, the optimal performance of an embedding method depends on all hyperparameters, the network, and the application task. Moreover, the results suggest that equally well results can be reached with many different hyperparameter combinations, indicating that a reasonable performance can be reached without an extensive hyperparameter search. This may also explain the difference between our results to the above. We consider the performance of the final application task (node classification) when finding embeddings.

Hyperparameter for node2vec A more detailed analysis of the hyperparameter for node2vec is listed in Table 3. The results lead to the conclusion that the best hyperparameter combination depends on the network and the application task. In the case of the BlogCatalog data set, there are also apparent differences between the two classification algorithms: The hyperparameter search leads the algorithm towards different learning strategies. The values for the sampling parameters p and q are 2 and 0.25 in the random forest case and 2 and 1 for the logistic regression. Thus for the Blog Catalog network, the random forest benefits from a depth-first sampling strategy preferring nodes further away from the source node, whereas the sampling strategy for the logistic regression is not biased towards one sampling strategy. The parameter p is 2 for both classification cases. Hence, the likelihood of revisiting a node is low.

In the paper introducing node2vec, experiments were also conducted on the BlogCatalog network. The authors described an increase in performance with small values for p and q . The results of the presented experiments suggest higher values for p and smaller values for q . Differences in the findings for the return parameter p are probably due to the variants in the remaining hyperparameters. Grover and Leskovec [13] used default values for all remaining hyperparameters and only experimented with the values for p and q . The findings of the random search suggest a strong effect of the interaction between the parameters. Even though a lower p is optimal in the case of default parameters, a higher p – leading to node sequences containing samples further away from the source – leads to

Table 3: Results for the hyperparameter analysis of node2vec.

| | Random forest | | | | Logistic regression | | | |
|-----------------------------|---------------|----------|-------|--------|---------------------|----------|-------|--------|
| | Blog | Facebook | Cite. | Moreno | Blog | Facebook | Cite. | Moreno |
| Micro F ₁ -score | 0.04 | 0.47 | 0.27 | 0.95 | 0.22 | 0.5 | 0.35 | 0.95 |
| Dimension | 74 | 943 | 103 | 733 | 197 | 848 | 245 | 600 |
| Return parameter: p | 2 | 0.5 | 0.25 | 1 | 2 | 2 | 0.75 | 2 |
| In-out parameter: q | 0.25 | 0.5 | 0.5 | 4 | 1 | 0.5 | 0.25 | 0.5 |
| Number of walks: l | 25 | 42 | 40 | 20 | 33 | 5 | 48 | 39 |
| Walk length: k | 70 | 56 | 11 | 45 | 46 | 28 | 11 | 24 |

better results, when combined with more and longer walks. These observations highlight the importance of tuning the hyperparameters of node embeddings based on the application task instead of simply using the default parameters.

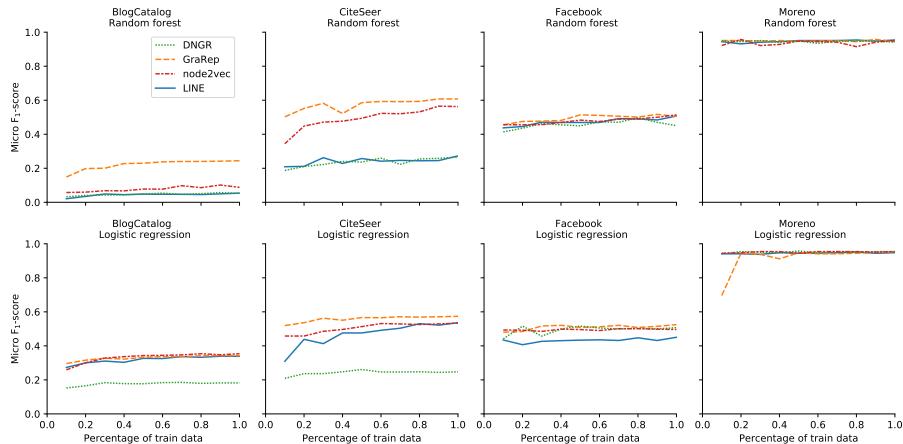


Figure 3: Results for the impact of an increasing amount of training data on the classification performance.

Impact of % training data on the performance Figure 3 shows the impact of increasing the amount of training data on the performance, the overall impact is small. The behavior of the curves, however, shows that with small ratios, an increase in the amount of training data has a high impact on the performance. At some point, an increase in the data only leads to a small improvement. As an example, for both classification methods, the performance for the Moreno network reaches a peak in performance increase at around 20% of the training data. After that point, the impact on performance is relatively low. Similarly, the score for the embedding methods on the BlogCatalog network are increasing until a ratio of 0.2 to 0.3, the performance score of node2vec in the logistic regression scenario is 0.26 with 10% of the data. However, with

30% of the data, the score is already 0.33. There is only little improvement thereafter as the best score is 0.35. This is consistent with previous studies that showed that the performance of node2vec shows large improvements until 30% after that, the increase in performance is small [11]. For CiteSeer differences between the random forest and the logistic regression scenario can be observed. In the case of the random forest combined with GraRep and node2vec there is a substantial increase in performance. The starting value of node2vec, for example, is 0.34, whereas the best performance is 0.56. However, in the logistic regression scenario, the difference is only 0.05, which is consistent with a similar experiment conducted by [31], who compared the results using 5% and 50% of the whole network data and found an increase of 0.08 points for CiteSeer. The reason for these differences in the two application scenarios is not apparent. However, it might be because the random forest needs more labeled observations to separate them efficiently. The CiteSeer network has many labels with only a few observations. Therefore, a small amount of data might lead to an underrepresentation of training data for some labels.

5 Conclusions

Recently, node embeddings became popular as an alternative to handcrafted feature engineering [14]. In this paper, we proposed a process for the comparison of node embedding methods w.r.t. node classification. This process enables researchers and practitioners to perform a fair and objective evaluation of node embedding procedures and helps end users to find the optimal method for the particular use case.

Moreover, in a case study, we applied this process to four popular node embedding methods. These experiments showed that the introduced process provides a foundation for a standardized evaluation of node embedding methods. Additionally, we made valuable observations, especially for practitioners: The default parameters for node embedding procedures are generally not a good choice. We analyzed this in detail for node2vec. Analyzing the impact of the dimensionality of the embeddings, we noticed that the appropriate combination of hyperparameters yields good performance with a lower number of dimensions, which is positive for the run times of the downstream machine learning task and the embedding algorithm. We also observed that multiple hyperparameter combinations yield similar performance. Hence there no extensive, time-consuming search required to achieve reasonable performance.

Although the proposed process provides a robust foundation for the comparison of node embedding methods, there are some aspects which should be addressed by future research. For example, the application task link prediction. It would be particularly interesting to understand how the procedure has to be adjusted differently for missing and future link prediction. A comprehensive comparison of semi-supervised methods would also be of interest.

References

- [1] ADAMIC, L. A., AND GLANCE, N. The political blogosphere and the 2004 US election: Divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery* (2005), ACM, pp. 36–43.
- [2] BERGSTRA, J., AND BENGIO, Y. Random search for hyper-parameter optimization. *Journal of Machine Learning Research* 13, Feb (2012), 281–305.
- [3] CAI, H. Y., ZHENG, V. W., AND CHANG, K. A Comprehensive Survey of Graph Embedding: Problems, Techniques and Applications. *IEEE Transactions on Knowledge and Data Engineering* 30, 9 (2018), 1616–1637.
- [4] CAO, S., LU, W., AND XU, Q. GraRep: Learning Graph Representations with Global Structural Information. *Proceedings of the 24th ACM International Conference on Information and Knowledge Management - CIKM '15* (2015), 891–900.
- [5] CAO, S., LU, W., AND XU, Q. Deep neural networks for learning graph representations. In *AAAI* (2016), pp. 1145–1152.
- [6] CARUANA, R., AND NICULESCU-MIZIL, A. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd International Conference on Machine Learning* (2006), ACM, pp. 161–168.
- [7] DAELEMANS, W., AND HOSTE, V. Evaluation of machine learning methods for natural language processing tasks. In *3rd International Conference on Language Resources and Evaluation (LREC 2002)* (2002), European Language Resources Association (ELRA).
- [8] FERNÁNDEZ-DELGADO, M., CERNADAS, E., BARRO, S., AND AMORIM, D. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research* 15 (2014), 3133–3181.
- [9] GETTOOR, L. Link-based classification. In *Advanced Methods for Knowledge Discovery from Complex Data*. Springer, 2005, pp. 189–207.
- [10] GOYAL, P., AND FERRARA, E. GEM: A Python package for graph embedding methods. *Journal of Open Source Software* 3, 29 (2018).
- [11] GOYAL, P., AND FERRARA, E. Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems* 151 (2018), 78–94.
- [12] GOYAL, P., HUANG, D., GOSWAMI, A., CHHETRI, S. R., CANEDO, A., AND FERRARA, E. Benchmarks for graph embedding evaluation. *CoRR abs/1908.06543* (2019).

- [13] GROVER, A., AND LESKOVEC, J. Node2vec: Scalable Feature Learning for Networks. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16* (2016), 855–864.
- [14] HAMILTON, W. L., YING, R., AND LESKOVEC, J. Representation Learning on Graphs: Methods and Applications. *IEEE Data Eng. Bull.* 40, 3 (2017), 52–74.
- [15] JAMES, G., WITTEN, D., HASTIE, T., AND TIBSHIRANI, R. *Introduction to Statistical Learning*, vol. 112. Springer, 2013.
- [16] KHOSLA, M., ANAND, A., AND SETTY, V. A comprehensive comparison of unsupervised network representation learning methods. *CoRR abs/1903.07902* (2019).
- [17] KIPF, T. N., AND WELLING, M. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings* (2017).
- [18] KOU, G., LU, Y., PENG, Y., AND SHI, Y. Evaluation of classification algorithms using MCDM and rank correlation. *International Journal of Information Technology & Decision Making* 11, 01 (2012), 197–225.
- [19] LESKOVEC, J., AND MCAULEY, J. J. Learning to discover social circles in ego networks. In *Advances in Neural Information Processing Systems* (2012), pp. 539–547.
- [20] LI, L., JAMIESON, K. G., DESALVO, G., ROSTAMIZADEH, A., AND TALWALKAR, A. Efficient Hyperparameter Optimization and Infinitely Many Armed Bandits. *CoRR abs/1603.06560* (2016).
- [21] NATURAL LANGUAGE PROCESSING LAB AT TSINGHUA UNIVERSITY. OpenNE: An open source toolkit for Network Embedding. <https://github.com/thunlp/OpenNE>. [accessed on 20-May-2019].
- [22] NEWMAN, M. The Structure and Function of Complex Networks. *SIAM Review* 45, 2 (2003), 167–256.
- [23] PEROZZI, B., AL-RFOU, R., AND SKIENA, S. DeepWalk. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '14* (New York, New York, USA, 2014), ACM Press, pp. 701–710.
- [24] POWERS, D. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies* 2 (Jan. 2008).

- [25] SHCHUR, O., MUMME, M., BOJCHEVSKI, A., AND GÜNNEMANN, S. Pitfalls of graph neural network evaluation. *CoRR abs/1811.05868* (2018).
- [26] TANG, J., QU, M., WANG, M., ZHANG, M., YAN, J., AND MEI, Q. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web* (2015), International World Wide Web Conferences Steering Committee, pp. 1067–1077.
- [27] TANG, L., AND LIU, H. Relational learning via latent social dimensions. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2009), ACM, pp. 817–826.
- [28] WANG, D., CUI, P., AND ZHU, W. Structural deep network embedding. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), ACM, pp. 1225–1234.
- [29] WOLPERT, D. H., AND MACREADY, W. G. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation* 1, 1 (1997), 67–82.
- [30] YANG, C., LIU, Z., ZHAO, D., SUN, M., AND CHANG, E. Y. Network Representation Learning with Rich Text Information. In *IJCAI* (2015), pp. 2111–2117.
- [31] ZHANG, D., YIN, J., ZHU, X., AND ZHANG, C. Network representation learning: A survey. *IEEE Transactions on Big Data* (2018).