OFFER: A Motif Dimensional Framework for Network Representation Learning

Shuo Yu
Dalian University of Technology
Dalian, China
y_shuo@outlook.com

Feng Xia
Federation University Australia
Ballarat, Australia
f.xia@ieee.org

Jin Xu
Dalian University of Technology
Dalian, China
xujin0909@hotmail.com

Zhikui Chen
Dalian University of Technology
Dalian, China
zkchen@dlut.edu.cn

Ivan Lee
University of South Australia
Australia
ivan.lee@unisa.edu.au

ABSTRACT

Aiming at better representing multivariate relationships, this paper investigates a motif dimensional framework for higher-order graph learning. The graph learning effectiveness can be improved through OFFER. The proposed framework mainly aims at accelerating and improving higher-order graph learning results. We apply the acceleration procedure from the dimensional of network motifs. Specifically, the refined degree for nodes and edges are conducted in two stages: (1) employ motif degree of nodes to refine the adjacency matrix of the network; and (2) employ motif degree of edges to refine the transition probability matrix in the learning process. In order to assess the efficiency of the proposed framework, four popular network representation algorithms are modified and examined. By evaluating the performance of OFFER, both link prediction results and clustering results demonstrate that the graph representation learning algorithms enhanced with OFFER consistently outperform the original algorithms with higher efficiency.

CCS CONCEPTS

• **Mathematics of computing**; • **Computing methodologies** → *Knowledge representation and reasoning*; *Machine learning*;

KEYWORDS

Network representation learning, network motif, multivariate relationship, link prediction.

1 INTRODUCTION

Network representation learning has been widely applied. The general idea of network representation learning is to convert nodes into vectors. By generating the structural information as well as node attributes together, network representation learning methods traverse the network structure to lower-dimensional vector space [4]. The mechanism behind this is that directly connected node pairs are generally similar in the corresponding vector space, which leads to pairwise relationships that can be well represented in the corresponding low-dimensional vector space [10].

One of the most typical characteristics of network structure is that multivariate relationships can be profiled by higher-order structures [7, 11]. However, in the corresponding low-dimensional vector space, such multivariate relationships are barely represented due to the limitation of vectors. Real-world networks consist of various complex relationships, in which most entities are in multivariate relationships with others [2, 5]. Comparing with pairwise relationships, multivariate relationships are even more significant but complicated. For the previous methods, weights of edges are applied to evaluate the strength of pairwise relationships, but lack the ability to evaluate multivariate ones. Based on the network motif structures, multivariate relationships can be formulated and measured. Some previous studies realize that motif structure can help with network representation learning [8], but the underlying reasons are not studied. Meanwhile, the feasibility of motifs is not discussed as well.

Driven by the significance of multivariate relationships, we propose OFFER (mOtiF dimensional Framework for nEtwork Representation learning) to enhance the abilities of network representation learning algorithms. This framework can be customized in choosing proper motifs when representing the network. We define two metrics based on the custom motifs: *motif degree* and *motif edge degree*. We refine the network adjacency matrix by motif degree of nodes, to help reduce the differences of node degrees that are introduced by higher-order structures. Then, we refine the learning process by employing the motif degree of edges, to strengthen the multivariate relationships in the network representation. The proposed method is examined in terms of link prediction and clustering. In summary, the contributions of this paper include:

- **The design of motif dimensional framework for network representation:** We propose a motif dimensional

framework to enhance the current network representation algorithms in formulating multivariate relationships. Enhanced by the proposed framework, the entities with multivariate relationships can be represented.

- **The discovery of the correlation between motif structures and multivariate relationships:** The mechanism of OFFER makes it possible to reinforce strong relationships as well as multivariate relationships, thus it is universally applicable for such kind of network representation methods.
- **The efficiency of OFFER:** The experimental results show that the enhanced methods generally lead to superior network representation performance. All enhanced methods outperform their original network representation algorithms with higher accuracy in both prediction and clustering capabilities.

2 MODEL DESIGN

To characterize the frequent multivariate relationship structure in the network, we use the most common three-order triangle motif.

Definition 2.1. **Network Motif.** For the graph $G = (V, E)$, the network motif M is a special subgraph structure in G that satisfies three characteristics: low-order nature, high-frequency nature, and real mapping nature.

Specifically, the low-order nature means that the order of M is generally not high (not over 8). The high-frequency nature means that the frequency of M in the real network is much higher than that of the corresponding random network. The real mapping nature means that M can always find the actual meaning in the real network. Because of these natures of M, especially the real mapping nature, motifs can be used to characterize specific multivariate relationship structures in the real network. Among various network motifs, triangle motifs are more popular, because they are widely present in various real networks and have a strong reality mapping.

To characterize motif features on network nodes or edges, we define the motif degree of each node and edge.

Definition 2.2. **Motif Node Degree (MND).** For a given motif M, $ND_M(i)$ represents the motif degree of node i, and $ND_M(i) = n$ means node i is involved in a number of n motifs.

Definition 2.3. **Motif Edge Degree (MED).** For a given motif M, $ED_M(i, j)$ represents the motif degree of edge (i, j), and $ED_M(i, j) = n$ means edge (i, j) is involved in a number of n motifs.

Figure 1 illustrates these two concepts. The formula on the right represents the triangle motif degree information of the corresponding nodes (green) and edges (red) on the left. Comparing with
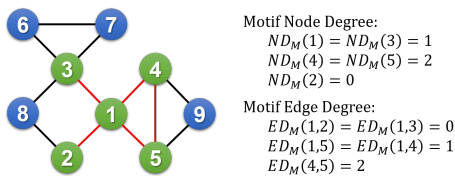

Motif Node Degree:
$ND_M(1) = ND_M(3) = 1$
$ND_M(4) = ND_M(5) = 2$
$ND_M(2) = 0$

Motif Edge Degree:
$ED_M(1,2) = ED_M(1,3) = 0$
$ED_M(1,5) = ED_M(1,4) = 1$
$ED_M(4,5) = 2$

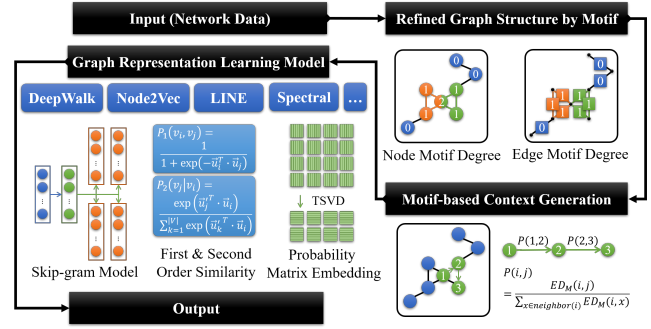**Figure 1: An example to explain triangle MND and MED.**

Figure 2: The framework of OFFER.

traditional node degree, MND and MED better reflect various network attributes such as significant network connectivity, robustness, and centrality. We refer to the analysis method using motif degree information instead of traditional degree information as the "motif-dimensional analysis method".

We define a motif-biased adjacency matrix A_M for G to refine the total network. This matrix is constructed based on MND. To ensure network connectivity, we define A_M as shown in Equation (1).

$$A_M(i, j) = \begin{cases} 1, & \text{if } (i, j) \in E, ED_M(i, j) = 0 \\ 0, & \text{if } (i, j) \notin E \\ 1 + \frac{ED_M(i,j)}{|V_M|}, & \text{if } (i, j) \in E, ED_M(i, j) \neq 0 \end{cases} \quad (1)$$

Wherein, $|V_M|$ represents the number of nodes in M.

The general framework is shown in Figure 2. For a real network, OFFER first calculates its MND and MED matrix. Then, OFFER converts them into an appropriate network representation matrix, such as A_M. Finally, OFFER deploys the improved network representation learning algorithm. As an example, for the network representation learning algorithm based on the transition probability matrix P, OFFER refine the learning process in network representation by using MED. Consequently, this will lead to a fact that edges with larger $ED_M(i, j)$ will achieve more learning weight. That is, if we denote the neighbor nodes of i as id $neighbor(i)$, then the probability from node i to node j is shown in Equation (2).

$$P_{trans}(i, j) = \frac{ED_M(i, j)}{\sum_{x \in neighbor(i)} ED_M(i, x)} \quad (2)$$

3 EXPERIMENTS

3.1 Data Sets

Four network datasets are selected for performance evaluation of link prediction, and these datasets include **Wiki**[1], **Routers**[2], **Twitter**[3], and **Facebook**[4]. The records in these datasets can be used to generate undirected and unweighted networks. For the clustering scenario, we use **Twitter** social dataset and **Routers** dataset. Meanwhile, we also use two new datasets, namely, **Hamsterster**[5] and **Openflights**[6]. Both have higher network density to show strong

(a) Precision   (b) Recall   (c) Accuracy

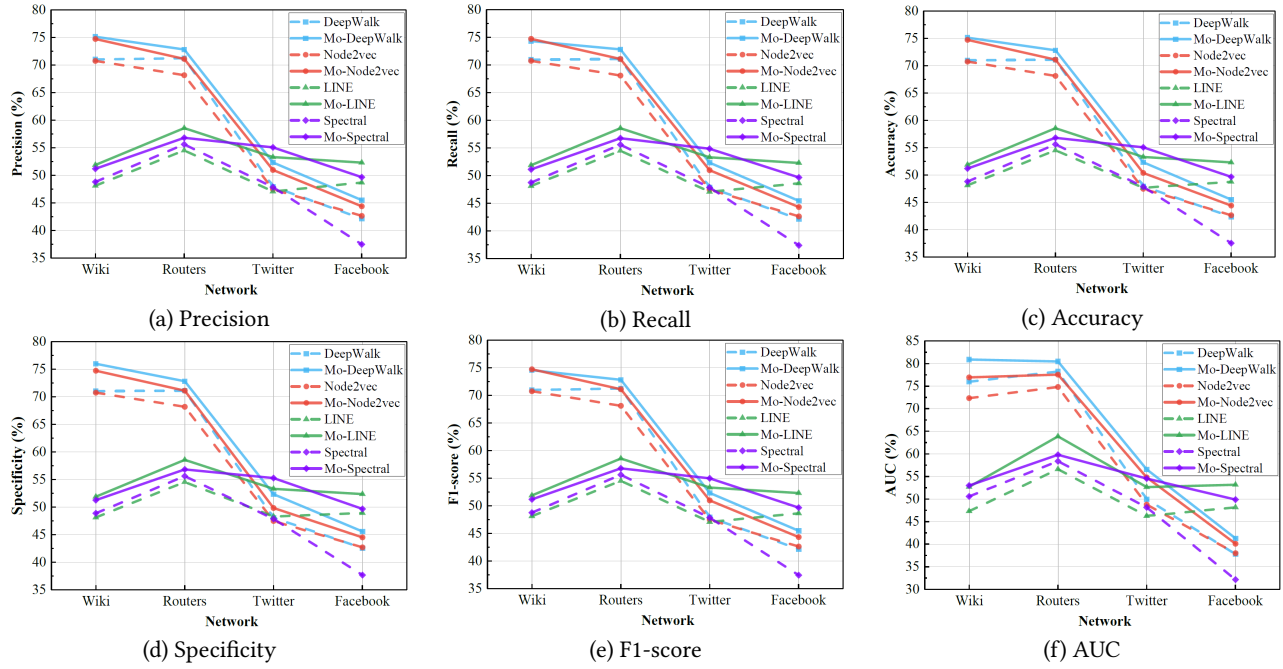(d) Specificity   (e) F1-score   (f) AUC

Figure 3: Six indicators comparison in four networks.

universality of our method in various network environments. Basic properties of these networks are shown in Table 1. Wherein, $|V|$ is the number of nodes, $|E|$ is the number of edges, $\max(d)$ is the max node degree, $\bar{d}$ is the average node degree and ρ is the density.

In these network environments, triangle motifs have different realistic mapping meanings, which also represent different multivariate relationships. In social networks (Facebook, Twitter, Hamsterster), they are employed to represent ternary closures. In Openflights, they correspond to connecting flights. As for Routers, they represent the routing structures of network devices. Due to the realistic significance, we specifically make statistics of triangle motif distributions among the above-mentioned datasets and their corresponding random networks (shown in Figure 4).

3.2 Evaluation Metrics

Link Prediction. We use cosine similarity in the link prediction based on previous work [1], wherein the cosine similarity of two vectors X, Y is calculated by the following formula: $CosSim(X, Y) = \vec{x} \cdot \vec{y}/\|x\| \cdot \|y\|$.

Table 1: The statistic information of experimental datasets

Network	$|V|$	$|E|$	$\max(d)$	$\bar{d}$	ρ
Wiki	889	2914	102	6.5556	0.007383
Routers	2113	6632	109	6.2773	0.002972
Twitter	761	1029	37	2.7043	0.003558
Facebook	2888	2981	769	2.0644	0.000715
Hamsterster	2426	16630	273	13.7098	0.005654
Openflights	2939	15677	242	10.6682	0.003631

Six metrics (AUC, Accuracy, Precision, Recall, Specificity, and F1-score) are employed to verify the performance of link prediction: **AUC** reflects the overall performance of the proposed model, which refers to the area under ROC (Receiver Operating Characteristic) curve. **Accuracy** reflects the right prediction results ratio of the total samples, which is calculated by $Accuracy = (TP + TN)/(TP + TN + FP + FN)$. **Precision** reflects how many of the prediction positive results are positive samples, defined as $Precision = TP/(TP + FP)$. **Recall** evaluates the recognition ratio of the positive sample, defined as $Recall = TP/(TP + FN)$. **Specificity** evaluates the recognition ratio of the negative sample, defined as $Specificity = TN/(TN + FP)$. **F1-score** is calculated based on Precision and Recall, defined as: $F1\text{-}score = (2 \times Precision \times Recall)/(Precision + Recall)$.

Clustering. Herein, we use SC (Silhouette Coefficient) to evaluate the clustering results. SC is proposed with the combination
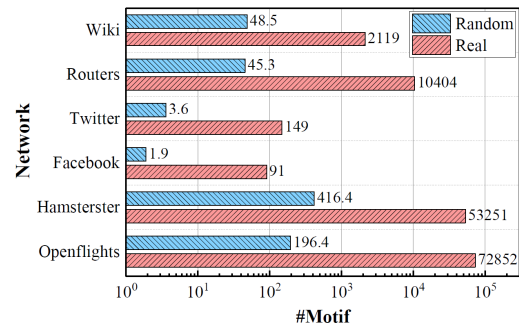

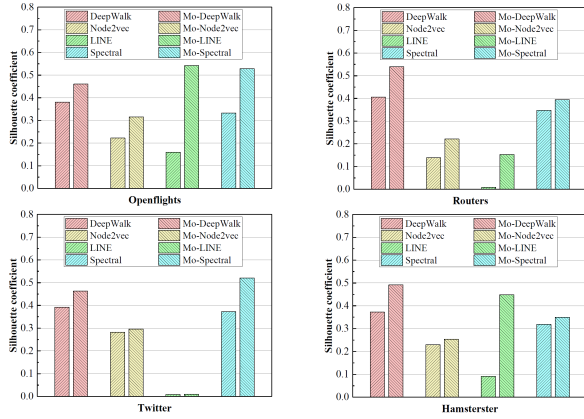**Figure 4: The triangle motif numbers in 6 datasets.**

Figure 5: Silhouette coefficient of the four networks.

of cohesion and separation. A higher SC value reflects better clustering results. Specifically, SC equals $\text{avg}\{s(i)\}$, wherein, $s(i)$ is the SC of sample i and it can be calculated by $s(i) = [b(i) - a(i)]/\max\{a(i), b(i)\}$. Among them, $a(i)$ refers to the average distance of sample i to other samples, which is considered as the evaluation metric for cohesion degree within the cluster. Correspondingly, $b(i)$ refers to the average distance of sample i and all the nodes in another cluster that i does not belong to. It is widely used to evaluate the separation degree between clusters.

3.3 Results and Discussions

Figure 3 shows the experimental results of link prediction in the four datasets. We used DeepWalk [6], Node2vec [3], LINE [9], Spectral [12], which are popular algorithms in research and engineering to optimize. The corresponding optimization algorithm is named Mo-DeepWalk, Mo-Node2vec, Mo-LINE, Mo-Spectral.

Figure 3 (a)-(d) shows experimental results of four first-level indicators (**Precision**, **Recall**, **Accuracy** and **Specificity**). It has been observed that enhanced algorithms consistently yield superior performance. Meanwhile, the performance of these four metrics is close to each other. For example, in Wiki, Mo-DeepWalk outperforms DeepWalk by about 4% for all four metrics. As for three other enhanced algorithms, more than 3% performance gain is found.

Figure 3 (e)-(f) shows experimental results of two second-level indicators: **F1-score** and **AUC**. **F1-score** can explore the balance of **Precision** and **Recall**. **AUC** avoids converting prediction probabilities into categories, and is not sensitive to whether the sample categories are balanced. The figure shows that the proposed motif dimensional network representation learning has consistently improved all of the original algorithms. Such finding also indicates that the motifs represented the multivariate relationship can improve network representation. The improvement is significant on Wiki and Twitter. The observation demonstrates that OFFER effectively improves the performance of the link prediction.

Figure 5 shows the SC value comparison of the four enhanced algorithms against the original algorithms. It can be seen that in the four networks, the enhanced algorithms outperform with higher SCs. The mechanism of OFFER is that it can capture multivariate relationships. Among them, the clustering results of DeepWalk,

Node2vec, and Spectral generally perform better than LINE does. This may be due to 2 labels used in the experiments, whereas LINE is usually applied for scenarios with more labels.

4 CONCLUSION AND FUTURE WORK

In this work, we propose a framework called OFFER to enhance the performance of network representation algorithms. We take four network representation algorithms as examples for optimization, including DeepWalk, Node2vec, LINE, and Spectral Clustering. And link prediction is employed to verify the performance of the enhanced network representation. Experimental results show that all enhanced algorithms outperform the original ones. Limitation of the OFFER framework includes its only focus on the network structure characteristics and its refinement of the learning process. Besides, only the triangle motif has been applied to represent multivariate relationships in the experiments, while higher-order motifs can capture relationships among more entities. But the price is a more complicated calculation and choosing a more suitable motif.

Multivariate relationships with higher-order embedding mechanism proposed in this work can be applied to graph topology inference is significant and interesting. Moreover, identifying a matching motif for a certain network also worth further investigation. Therefore, enhancing network representation learning of multivariate relationships is foreseeable future work.

5 ACKNOWLEDGMENTS

REFERENCES

[1] Caterina De Bacco, Eleanor A Power, Daniel B Larremore, and Cristopher Moore. 2017. Community detection, link prediction, and layer interdependence in multilayer networks. *Physical Review E* 95, 4 (2017), 042317.
[2] Asim K Dey, Yulia R Gel, and H Vincent Poor. 2019. What network motifs tell us about resilience and reliability of complex networks. *Proceedings of the National Academy of Sciences* 116, 39 (2019), 19368–19373.
[3] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 855–864.
[4] Anuraj Mohan and KV Pramod. 2019. Network representation learning: models, methods and applications. *SN Applied Sciences* 1, 9 (2019), 1014.
[5] Sabyasachi Patra and Anjali Mohapatra. 2018. Clustering of proteins in interaction networks based on motif features. In *2018 International Conference on Bioinformatics and Systems Biology (BSB)*. IEEE, 141–146.
[6] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*. ACM, 701–710.
[7] Ryan Rossi, Nesreen Ahmed, and Eunyee Koh. 2018. Interactive Higher-Order Network Analysis. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 1441–1446.
[8] Ryan A Rossi, Nesreen K Ahmed, and Eunyee Koh. 2018. Higher-order network representation learning. In *Companion Proceedings of the The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 3–4.
[9] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *Proceedings of the 24th International conference on World Wide Web*. 1067–1077.
[10] Wei Wang, Ming Zhu, Xuewen Zeng, Xiaozhou Ye, and Yiqiang Sheng. 2017. Malware traffic classification using convolutional neural network for representation learning. In *2017 International Conference on Information Networking (ICOIN)*. IEEE, 712–717.
[11] Jin Xu, Shuo Yu, Ke Sun, Jing Ren, Ivan Lee, Shirui Pan, and Feng Xia. 2020. Multivariate Relations Aggregation Learning in Social Networks. In *2020 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE.
[12] Jie Zhang, Yuxiao Dong, Yan Wang, Jie Tang, and Ming Ding. 2019. ProNE: fast and scalable network representation learning. In *Proc. 28th Int. Joint Conf. Artif. Intell., IJCAI*. 4278–4284.