Front. Comput. Sci., 2019, 13(2): 437–439
https://doi.org/10.1007/s11704-018-7389-1

LETTER

Learning distributed representations for community search using node embedding

Jinglian LIU[1,2], Daling WANG (✉)[1], Shi FENG[1], Yifei ZHANG[1], Weiji ZHAO[2,3]

1 School of Computer Science and Engineering, Northeastern University, Shenyang 110169, China
2 School of Information Engineering, Suihua University, Suihua 152061, China
3 School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China

1 Introduction and main contributions

Community search is a query-dependent variant of community detection problem in social network analysis [1]. Algorithms of this type usually start with a query node preknown to be in the target community, and uncover the remaining nodes in the community. The key challenge in this problem is how to find a proper way to represent network structure as a representation that can be easily exploited by downstream data mining models. Traditional approaches treat this problem as a pre-processing hand-engineered feature extraction step, which requires expert knowledge. Recently, there has been a surge of researches on node embedding which use deep learning approach to learn the representation of nodes from network structure automatically [2, 3]. The idea behind them is to learn a mapping from nodes to points in a low-dimensional vector space, which can be used as feature inputs for downstream data mining models.

In this paper, we propose to exploit node embedding technique for community search task from a new perspective. The main contributions of this work can be summarized as follows:

- We adopt edge weighting strategy in the process of random walk, and design a closest-neighbor biased random walk method. Moreover, we build a new node embedding model based on closest-neighbor biased random walk (NEMCNB for short). The model can be applied to community search task, as well as other network analysis tasks, such as link prediction and node classification.
- Our node embedding model NEMCNB provides a new similarity measure between nodes in the network. By embedding it into similarity-based community search, we design a novel community search algorithm.
- We test the proposed community search algorithm on both synthetic and real-world network datasets. The experimental results show that our algorithm is better at community search compared with related algorithms.

The technical details, proofs and evaluations can be found in the support information.

Received November 6, 2017; accepted September 12, 2018

E-mail: wangdaling@cse.neu.edu.cn

2 Problem definition of community search with a node embedding approach

Definition 1 (Network [4]) Let $G=(V, E)$ be an undirected graph, where V is the set of nodes and E is the set of edges. Let $n=|V|$ be the number of nodes in G. The set of nodes adjacent to node x is denoted by $\Gamma(x)$, $\Gamma(x)=\{y|y \in V, (x, y) \in E\}$. The degree of x is the number of nodes in $\Gamma(x)$, denoted by k_x.

Given a network G, and suppose the ground-truth commu-

nity structure of G is P, P=$\{C_1, C_2, \ldots, C_t\}$, and V=$C_1 \cup C_2 \cup \ldots \cup C_t$. Each node set C_i (i=1, 2, …, t) corresponds to a community of G. The problem definition of community search can be presented as: given a network G and a query node v ($v \in C_i$), the goal is to find out the remaining nodes in a node set C_i that contains node v. Suppose D is the node set discovered by the community search algorithm, then the algorithmic effectiveness is measured by the similarity between ground-truth community C_i and D.

In order to solve the community search problem, we address two sub-problems: node embedding and community search based on node embedding.

Problem 1 (Node embedding [3]) Given a network G, and a predefined dimensionality of the embedding dn, the problem of node embedding is to learn a mapping $f : V \rightarrow R^{n \times dn}$, in which each node is associated with a real-valued dn-dimensional vector. Equivalently, f is a matrix of size $|V| \times dn$ that contains the embedding vectors of nodes in G.

Problem 2 (Community search based on node embedding) Given a network G and a matrix f defined as the node embeddings of G, the similarity between nodes in G is measured via their embeddings. Given a query node v preknown to be in the target community, the goal of community search is to uncover the remaining nodes in the community.

3 Node embedding approach of community search

We propose a new approach for solving community search problem. We first propose a closest-neighbor biased random walk method, and then design our node embedding model NEMCNB. Moreover, the community search algorithm based on NEMCNB is given in the last part of this section.

3.1 Closest-neighbor biased random walk

Formally, given a start node u, we simulate a random walk of fixed length l. Let c_i denote the ith node in the walk, starting with $c_1 = u$. Node c_{i+1} is the next node after c_i in a node path. c_{i+1} is generated by the following distribution [3]:

$$P(c_{i+1} = x | c_i = v) = \begin{cases} \frac{w_{vx}}{z}, & (v, x) \in E, \\ 0, & (v, x) \notin E, \end{cases} \quad (1)$$

where w_{vx} is the weight of edge (v, x), and z is the normalizing constant, which is calculated as following.

$$z = \sum_{u \in \lceil(v)} w_{vu}. \quad (2)$$

During the process of random walks, both DeepWalk and node2vec neglect the closeness between c_i and c_{i+1}. However, for social networks, as we know, the tie strengths are different between friends, and the closest friends are preferred rather than others. We adopt edge weighting strategy in random walk process, and propose a closest-neighbor biased random walk method. In this work, we choose Jaccard Index as the edge weighting strategy, and w_{vx} is calculated as follows.

$$w_{vx} = \frac{|\lceil(v) \cap \lceil(x)|}{|\lceil(v) \cup \lceil(x)|}. \quad (3)$$

Suppose the current node c_i is e, $\lceil(e) = \{d, f, g, h\}$. As shown in Fig. 1, for node $x \in \lceil(e)$, $\lceil(x)$, w_{ex} and $P(x|e)$ are calculated as Table 1.

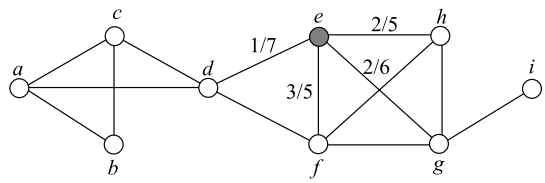

Fig. 1 Weights of edges in closest-neighbor biased random walk

Table 1 $P(x|e)$ in closest-neighbor biased random walk

x	$\lceil(x)$	w_{ex}	$P(x\|e)$
d	$\{a, c, e, f\}$	1/7	0.097
f	$\{d, e, g, h\}$	3/5	0.406
g	$\{e, f, h, i\}$	2/6	0.226
h	$\{e, f, g\}$	2/5	0.271

For all neighbor nodes of e, the probability of being c_{i+1} is directly proportional to their closenesses with e. We use alias method [3] to sample a node from $\lceil(e)$ according to their closenesses with e.

3.2 Our node embedding method NEMCNB

Based on closest-neighbor biased random walks, we propose a node embedding model NEMCNB. The input of NEMCNB is a corpus of node paths. First, we simulate closest-neighbor biased random walks on G of fixed length l starting from each node in G r times, and get a sequence of node paths. Then, by adopting CBOW (continuous bag-of-words) model [5], we take these node paths as input and produce low-dimensional vector representations for nodes in G as output.

CBOW model is one of the word2vec implementation techniques which predicts the current word based on its context words [5]. By viewing nodes as words and random paths

as sentences, we denote the context nodes of a given node u by $Context(u)$, and leverage CBOW model, i.e., maximizing objective function

$$\mathcal{L} = \sum_{u \in V} \log P(u|Context(u)), \tag{4}$$

to learn vector representations for nodes in G.

For nodes u and v, we denote their vector representations by $f[u]$ and $f[v]$, and denote the similarity based on node embedding by NES (i.e., node embedding similarity), which is the dot product of $f[u]$ and $f[v]$.

3.3 Community search based on NEMCNB

The challenge of similarity-based community search approaches is how to define a good node pairwise proximity measurement. Different from the existing hand-engineered similarity measurements, NES operates on a derived vector space which is produced by an automatical deep learning approach. Based on NEMCNB, we propose a new metric for community quality called *Closeness-Isolation*, and design our community search algorithm.

Definition 2 (Closeness-isolation metric) Given a network G, a matrix f defined as the node embeddings of network G, and a similarity measurement NES based on node vectors, for a community D with shell node set N, $N = \{x|v \in D, x \in \Gamma(v), x \notin D\}$, the *Closeness-Isolation* Metric of D, denoted by $CI(D)$, is defined as

$$CI(D) = \frac{C(D)}{1 + I(D)}, \tag{5}$$

where

$$C(D) = \sum_{u \in D, v \in D, (u,v) \in E} NES(u, v), \tag{6}$$

$$I(D) = \sum_{a \in D, b \in N, (a,b) \in E} NES(a, b). \tag{7}$$

$C(D)$ is the closeness of community D, and $I(D)$ is the isolation of D. To avoid division by zero, the denominator of $CI(D)$ is designed as one plus $I(D)$. Consequently, we expand community D by greedy adding in each step a node in shell node set N that currently has the largest similarity with nodes in D. The community expansion stops once the CI gain of every node in N that could still be added is negative.

To validate the performance of our community search algorithm, we compare it with Clauset [6], LWP [7], GMAC [4], DeepWalk [2], node2vec [3], and Spectral Clustering [8] on ten LFR benchmark networks [9] and three real-world networks. In the experiment, NEMCNB performs best, followed by DeepWalk, GMAC, and node2vec. The average F-$score$ of NEMCNB is 3.55% higher than DeepWalk, 7.95% higher than GMAC, and 11.27% higher than node2vec.

4 Conclusion

Community search is an important task in network analysis and many algorithms have been proposed. Inspired by that node embedding offering an alternative to traditional hand-engineered feature engineering, we introduce node embedding technique into community search, and design a new community search algorithm. Moreover, we propose a closest-neighbor biased random walk method, and design a new node embedding model NEMCNB. Compared with other related algorithms, NEMCNB achieves good performance on both synthetic and real-world networks.

Acknowledgements The project was supported by the National Key R&D Program of China (2018YFB1004700), and the National Natural Science Foundation of China (Grant Nos. 61772122, 61872074).

Supporting information The supporting information is available online at journal.hep.com.cn and link.springer.com.

References

1. Sozio M, Gionis A. The community-search problem and how to plan a successful cocktail party. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2010, 939–948
2. Perozzi B, Al-Rfou R, Skiena S. Deepwalk: online learning of social representations. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014, 701–710
3. Grover A, Leskovec J. node2vec: scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016, 855–864
4. Ma L, Huang H, He Q, Chiew K, Wu J, Che Y. GMAC: a seed-insensitive approach to local community detection. In: Proceedings of the 15th International Conference on Data Warehousing and Knowledge Discovery. 2013, 297–308
5. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. In: Proceedings of the 1st International Conference on Learing Representations. 2013
6. Clauset A. Finding local community structure in networks. Physical Review E, 2005, 72(2): 026132
7. Luo F, Wang J Z, Promislow E. Exploring local community structures in large networks. Web Intelligence and Agent Systems, 2008, 6(4): 387–400
8. Tang L, Liu H. Leveraging social media networks for classification. Data Mining and Knowledge Discovery, 2011, 23(3): 447–478
9. Lancichinetti A, Fortunato S, Radicchi F. Benchmark graphs for testing community detection algorithms. Physical Review E, 2008, 78(4): 046110