Research on Motif Mining Based on Neighborhood Equivalence Class

Jian Feng, Rui Ma* and Shaojian Chen
College of Computer Science & Technology, Xi’an University of Science and Technology, Xi’an, Shaanxi, 710054, China
*Corresponding Rui Ma’s e-mail: 893289364@qq.com

Abstract. The motif is an important mesoscopic structure existing in the network, and motif mining is an important means to study the social network structure. Based on the tree traversal G-tries algorithm of common subgraphs, we propose an accurate subgraph recognition algorithm of neighborhood equivalence class Ex-Motifs to reduce the matching process of subgraph isomorphism. In addition, for the research of motif metric, we propose a motif metric index based on a common substructure, which can directly judge the significance of subgraph frequency on the original network. Experimental results show that the computational efficiency of Ex-Motifs is relatively high, and it can find a motif similar to the traditional motif metric method.

1. Introduction
Motif mining [1] is mainly divided into subgraph recognition and motif metric. At present, there are two classical methods to identify subgraphs, which are based on subgraph isomorphism matching and combination strategy. In reference [2], the recognition method based on subgraph isomorphic matching is used to expand the subgraph structure by adding the affected vertex set of the target vertex to count the matched subgraph; then, other candidate vertices in the affected vertex set are replaced until all the vertices in the vertex set are traversed. Thus, there is additional time consumption due to the backtracking process in the subgraph isomorphism process. The Acc-Motifs algorithm [3] based on the combination strategy has the problem of repeated calculation when identifying each subgraph. Both backtracking and repeated computation lead to high time complexity. In the process of motif metric, the motif is almost always determined according to the statistical results of the most primitive comparison of the subgraphs in the random network.

In view of the shortcomings of existing algorithms, we propose an accurate subgraph recognition algorithm Ex-Motifs of neighborhood equivalence class to reduce the matching process of subgraph isomorphism; the metric index of the common substructure directly judges the significance of the subgraph frequency on the original network.

2. Preliminaries
We first clarify the structure of subgraph and problem definition in the process of impacting the motif mining, and finally analyse the recognition method of G-tries method adopted in this paper.

2.1. Research object
We only study the motif graphs within three degrees, and table 1 shows the topology of all subgraphs. DG_2, DG_3, DG_4 respectively represent directed acyclic graphs with nodes 2, 3 and 4.
There are 20 different subgraph structures [4]. The white vertex of each subgraph represents the information receiver in the influence network, and the red vertex represents the information transmitter, which is also the affected vertex of the target vertex. Among them, the subgraph structure with the number of DG2-1 nodes is 2. The number DG3-1 indicates a subgraph structure composed of three nodes, and the other numbers follow.

| Number | Structure | Number | Structure | Number | Structure | Number | Structure |
|--------|-----------|--------|-----------|--------|-----------|--------|-----------|
| DG2-1  |           | DG3-1  |           | DG3-2  |           | DG3-3  |           |
| DG4-2  |           | DG4-3  |           | DG4-4  |           | DG4-5  |           |
| DG4-7  |           | DG4-8  |           | DG4-9  |           | DG4-10 |           |
| DG4-12 |           | DG4-13 |           | DG4-14 |           | DG4-15 |           |

2.2. Problem definition

**Definition 1** The directed graph DG=(V, E) represents the network, which is a network structure composed of user set V and relationship set E.

**Definition 2** In the directed graph, s is the subgraph structure, there are affected neighborhoods I(s)=|I(V(s)) - V(s)| and affected neighborhoods O(s)=|O(V(s)) - V(s)|.

**Definition 3** Neighbor equivalence class NSC refers to the set of vertices that are equivalent to the common substructure. When there is a structural equivalence vertex in the affected neighborhood of the target vertex, it will be divided into the same neighborhood equivalence class. In the process of subgraph recognition, the subgraph statistical strategy of neighborhood equivalence class is used to replace the traditional subgraph isomorphism strategy. Figure 1 illustrates the division result of the neighborhood equivalence class corresponding to the common substructure formed by the point set {3, 4} in the graph DG. There is an affected neighborhood I(DG2-1) = {0,1,2}. The "feed-forward loop" structure is formed by the point set {0,2} and DG2-1, vertex 1 and DG2-1 forms "Three chain", so I(DG2-1) is divided into two different NSC.

![Figure 1. Neighborhood equivalence class](image)

2.3. G-tries algorithm

In the process of subgraph recognition, G-tries algorithm generates a new subgraph structure by extending a new vertex on the basis of common substructure [5]. Although G-tries speeds up the recognition of subgraphs, it still uses the traditional subgraph isomorphic calculation method. This paper follows the basic idea of the G-tries algorithm, and then generalizes it to directed graphs, aiming to divide the neighborhoods of common substructures equally to perform subgraph isomorphic matching.
3. Method
The process of motif mining is divided into subgraph recognition and motif measurement. We choose different isomorphism strategies for different subgraphs in the process of subgraph recognition. In the process of motif metric, it is compared with the traditional motif metric method to find the influence motif in the social network.

3.1. Subgraph recognition
In the process of isomorphism of subgraphs, different subgraph isomorphism strategies are selected for different subgraphs.

3.1.1 Create structural growth graph DG-Sexts. First, construct the structural growth graph DG-Sexts according to the order of newly added vertices, find the instances of nodes in DG-Sexts in the actual network, and count the isomorphism. Figure 2. shows the growth process of subgraphs using common substructures.

![Figure 2. Directed structural growth graph DG-Sexts](image)

In subgraph recognition, the subgraph can be accurately counted only if the subgraph structure formed by the set of points is uniquely identified. In the subgraph matrix, 0 represents the absence of edges and 1 represents the presence of edges, namely, the adjacency matrix of the subgraph is constructed.

![Figure 3. Graph G matching DG_{3,3} broken symmetry example](image)
3.1.2 Add a unique ID to avoid automorphism. The number of isomorphisms will be calculated repeatedly when calculating these subgraphs. To avoid the automorphism of the subgraphs, symmetry breaking conditions need to be added. This article solves the problem of automorphism by adding an identity Id. As shown in figure 3, \( N_{3,3} = 3 \). In the path DG\(_{2,1}\) composed of the point set \{0,2\}, because the mark \( \text{Id}(1) < \text{Id}(2) \) of the vertex 1 is newly added, the point set is no longer calculated \{0,2,1\} constitutes DG\(_{3,3}\), which is similar to DG\(_{2,1}\) composed of point set \{0,3\}.

It is worth noting that the subgraph of the point sets \{0,1,2\} and \{0,1,3\} is isomorphic on the path DG\(_{2,1}\). During the expansion of the G-tries subgraph, it is necessary to trace the DG\(_{3,3}\) constituted by the point set \{0,1,2\} to the DG\(_{2,1}\) constituted by the point set \{0,1\}, then expand the candidate vertex 3 to form DG\(_{3,3}\), and finally judge that the two subgraphs are isomorphic subgraphs. Obviously, there are a lot of backtracking processes in the practical application of this algorithm, so this paper introduces the neighborhood equivalent class NSC to avoid the backtracking process in the subgraph isomorphism.

3.1.3 Introducing neighborhood equivalent class NSC to achieve subgraph isomorphism. For the common substructure DG\(_{2,1}\), The vertices in I(DG\(_{2,1}\)) = \{2,3\} is isomorphic, so the set of vertices in the affected neighborhood is equivalent, that is, NSC(DG\(_{2,1}\)) = \{2,3\}. Figure 4. shows the subgraph isomorphism process after the introduction of neighborhood equivalence classes. The numbers in parentheses in the nodes represent the number of subgraph isomorphisms. The isomorphism subgraph can be counted only once in each path, so the backtracking problem is effectively avoided.

![Figure 4. Example of subgraph isomorphism of neighborhood equivalence class](image)

3.1.4. Equivalent formula. In the process of recognition of subgraphs, the equivalent calculation formulas of three subgraphs are extracted. The isomorphism of DG\(_{2,1}\), DG\(_{3,3}\) and DG\(_{4,16}\) can be calculated by using the equivalent formula.

\[
N_{2,1} = \text{ID}(u) \quad (1)
\]

\[
N_{3,3} = \binom{2}{N_{2,1}} - N_{3,1} \quad (2)
\]

\[
N_{4,16} = \binom{3}{N_{2,1}} - N_{3,1} \times (N_{2,1} - 2) + (N_{4,2} + N_{4,9} + N_{4,11} + 2 \times N_{4,4}) \quad (3)
\]

Among them, \( \text{ID}(u) \) in formula (1) refers to the degree of the vertex \( u \), and \( N_{2,1} \) and \( N_{3,1} \) in formula (2) are Refers to the corresponding frequency at the target vertex.

3.2. motif metric

In order to save the extra consumption caused by the random network, we propose a motif metric method based on the evolution process of common substructure on the original network, which is not compared with the random network. Therefore, to measure the saliency of each subgraph, it is necessary to obtain the frequency of all subgraphs that have a common substructure with the subgraph,
and also to obtain the saliency of the common substructure. The metric value of the motif is calculated by formula (4).

\[
M_S = \frac{N_{2-1}}{N_S \sum N_m \times 100\%} \times \frac{N_{4-16}}{\sum N_i \times 100\%}
\]  

In the formula (4), the \( M_S \) refers to the metric value of the pattern graph \( S \), \( N_S \) refers to the frequency of the pattern graph \( S \), \( \sum N_m \) refers to the sum of the frequency of \( n \) subgraph with the same common substructure as the subgraph.

4. Experiment
This section compares the accurate motif recognition algorithm Ex-Motifs proposed in this paper with the currently representative ESU, G-tries and Acc-Motifs algorithms, and analyses the calculation accuracy and time on different datasets.

4.1. Experimental environment and dataset
The experimental programming language is C++, and the development environment is VS2013. The dataset uses four groups of data of different orders of magnitude. Table 2. shows the size of each dataset. Virtual data is created manually; real data is the network structure formed by users forwarding microblogs.

| Source | V   | E     |
|--------|-----|-------|
| S-data | virtual | 20    | 90   |
| M-data | True  | 1000  | 15196|
| V-data | virtual | 10240 | 573440|
| B-data | True  | 1787443 | 413503687|

4.2. Evaluation index
This experiment mainly uses the two evaluation indexes of motif calculation accuracy and running time to evaluate the performance of different accurate subgraph recognition algorithms. Formula (5) shows the calculation process of accuracy. Among them, \( A_{k-i} \) stands for the accuracy of \( DG_{k-i} \), \( N_{k-i} \) stands for the frequency of \( DG_{k-i} \), \( NE_{k-i} \) represents the frequency of \( DG_{k-i} \) calculated by ESU algorithm in the original network. Because the ESU algorithm is accurately enumerated, it is reliable to compare the calculation accuracy with this result. The higher the accuracy, the more accurate the subgraph recognition algorithm.

\[
A_{k-i} = \frac{N_{k-i}}{NE_{k-i} \times 100\%}
\]  

4.3. Algorithm accuracy and time comparison
Table 3. lists the average calculation accuracy and average running time of M-data in different accurate subgraph recognition algorithms. \( AE_{k-i} \), \( AG_{k-i} \), \( AA_{k-i} \) and \( AM_{k-i} \) are the average calculation accuracy of the four algorithms respectively. The Ex-Motifs algorithm is more efficient than the ESU and G-tries algorithms; and less efficient than ACC motifs algorithm. The reason is that Ex-motifs algorithm can only count some subgraphs accurately by using equivalent formula, so it is faster than ESU and G-tries algorithm with complete isomorphism matching, and slower than Acc-motifs algorithm using combination strategy completely. Table 4. lists the running time of the four algorithms on the datasets.
Table 3. Calculation accuracy and time comparison.

| Name       | Average calculation accuracy | Average running time (s) |
|------------|------------------------------|--------------------------|
| ESU        | \( AE_{k,i} = 100.0\% \)    | 119                      |
| G-tries    | \( AG_{k,i} = 100.0\% \)    | 301                      |
| Acc-Motifs | \( AA_{k,i} = 99.0\% \)    | 4                        |
| Ex-Motifs  | \( AM_{k,i} = 100.0\% \)    | 80                       |

Table 4. Time comparison of dataset (s).

| Algorithm  | S-data | M-data | V-data | B-data |
|------------|--------|--------|--------|--------|
| ESU        | 0.036  | 119    | 6097   | 24035  |
| G-tries    | 0.102  | 301    | 18701  | 73847  |
| Acc-Motifs | 0.059  | 4      | 134    | 1618   |
| Ex-Motifs  | 0.036  | 80     | 4217   | 18554  |

4.4. Subgraph saliency analysis

After the frequency statistics of each subgraph in the dataset, the significance of each subgraph is analysed by using the motif metric index proposed in this paper. Figure 5. shows the significance of all subgraphs calculated by the four datasets using formula (4). As shown in the figure, the subgraphs in the four datasets have the same trend of saliency, all of which have the highest significance in DG_{4,10} and DG_{4,15}. Although the figure shows that DG_{4,16} has higher significance, it is more significant than DG_{4,10} and DG_{4,15} slightly lower, it means that in the case of more affected neighborhoods of the target vertex, the more dispersed structure is not necessarily more prominent.

Figure 5. Saliency of subgraph

5. Conclusion

In this paper, Ex-Motifs, an accurate subgraph recognition algorithm is proposed to compute the isomorphism of the three subgraphs with more frequent occurrence in the network, thus accelerates the computation time of the isomorphism of the traditional accurate subgraph recognition. To solve the problem of motif metric, we propose a method of motif metric based on common substructure. The social network motif can be identified without comparing with the frequency of the subgraph of the random network, which saves the time and space consumption caused by the random network. Finally, the effectiveness of the algorithm was verified on four datasets of different sizes. In summary, the accurate subgraph recognition algorithm proposed in this paper can effectively identify the influence motifs existing in the social network.

Acknowledgments

This work was financially supported by the Basic Research Project of Natural Science of Shaanxi Province (2020JM-533).
References
[1] Milo R, Shenorr S, Itzkovitz S. (2002) Network motifs: simple building blocks of complex networks. J. Science, 298(5594):824-827.
[2] Ribeiro P. (2011) Efficient and Scalable Algorithms for Network Motifs Discovery. J. Faculty of Science of the University of Porto, 110(1):605-612.
[3] Kashani Z R M, Ahrabian H, Elahi E, et al. (2009) Kavosh: a new algorithm for finding network motifs. J. Bmc Bioinformatics, 10(1):318-321.
[4] Li G, Luo J, Xiao Z. (2019) MTMO: an efficient network-centric algorithm for subtree counting and enumeration. J. Quantitative Biology. 6(2):142-154.
[5] Meira L A, Maximo V R, Fazenda A L. (2019) An Faster Network Motif Detection Tool. J. Data Structures and Algorithms, 10(4):294-307.