A Comparison of Different Networks Based on the Graphlet Correlation Distance

Zhimeng Guo

Yingcai Honors College
University of Electronic Science and Technology of China
Chengdu, Sichuan, 611731, P.R.China.

Guozhimeng98@outlook.com

***Abstract* - A large number of complex systems in nature can be described by a variety of networks so that complex networks have attracted more and more attention. Nowadays, research concentrate more on the interaction between a small number of roles and some graphlet indicators have been come up with. However, a macroscopic map about different networks based on the indicators remains unclear. In this paper, we used the Graphlet Correlation Distance (GCD) to compare and classify different networks, such as road network, Facebook network and so on. The result showed that some seemingly related networks do have smaller distance, which inspires us the internal formation mechanism of similar networks can be same. This work adds significance to the literature by extending the research on network classification and help us to have a deeper understanding of the nature of the network.**

Index Terms* - *Complex networks, network classification, Graphlet Correlation Distance (GCD).

I. INTRODUCTION

We are surrounded by a variety of networks, such as social networks, neural networks, literature citation networks and so on. A typical network consists of many edges between nodes, where nodes represent different individuals in a real system and edges represent relationships between individuals. For example, in WeChat Friends Network [1] , each person is a node, and if two people are friends, there will be a link between the two nodes. Networks allow us to model complex systems. This enables us to deal with complex interaction systems and to view various physical phenomena from a more macro perspective.

Networks are too complex to undertake a complete comparison, which has been proved to be computationally intractable [2]. Therefore, former research did network statistics to mine the intrinsic attributes of the network. The simplest concept in networks is degree, which is the number of edges of a node. Some researcher think a higher degree indicates the node is more important [3]. However, a man has more friends doesn't indicate he is a important person. Based on that, some come up with the closeness and betweenness centrality to solve the problem. There are also other useful indicators, such as degree distribution and network diameter [4].

However, some evidence shows two networks with the same degree centrality and degree distribution can be extremely different, which inspire us to compare networks from single nodes to the interaction between a small number of nodes in a network. Pržulj came up with small subnetworks called graphlets to keep the interaction [5]. Begin with the work, some algorithm were designed and optimized once and once [6]. With the graphlet-base edge clustering, researcher found pathogen-interacting proteins [7]. Now the graphlet has been accepted as a important indicator in network analysis and comparison.

In this paper, we calculated the Graphlet Correlation Matrix of European aircraft route network, Minnesota Highway Network, Wikipedia vote network and Facebook network. Based on that, we compared the distance between different networks and found similar networks have a closer distance, which is in line with our intuition.

The organization of this paper is as follows: In section 2 we talk about the steps to calculate the Graphlet Degree Vector, Graphlet Correlation Matrix and Graphlet Correlation Distance of four networks. In section 3 we describe the result of comparison of the networks. The conclusion of the comparison of four networks and the method to classify more networks is discussed in section 4.

II. METHOD

We collected the data of four networks from networkrepository [8]. Then we use the Orbit Counting Algorithm [9] and Pearson correlation coefficient to get the Graphlet Correlation Matrix and Graphlet Correlation Distance [10] between of different networks, which is important indicators of network features.

A. Data

Networkrepository is a reliable network database, which collects network data in former research. We download four networks from the database. The Europe Airplane Route Network has 1174 nodes and 1417 edges at all, the nodes represent airports and edges mean that there is a route between the cities. The Facebook network is partial of the whole of the Facebook Network and it has 2981 nodes and 2888 edges . The nodes represent users and edges mean two users are friends on Facebook. The Minnesota Highway Network is a geometry-based network with 2642 nodes and 3303 edges. The Wikipedia vote network is from the vote history of Wikipedia administrators and it has 889 nodes and 2914 edges in total.

As we can see, all the four networks are unweighted and undirected and the type of networks are different. The scale of the networks are not too big so that we can calculate and validate them efficiently.

B. *Graphlet Correlation Distance*

First, we splited the network into small structure with only a few nodes. We found all subgraphs with fewer than five nodes, as shown. From this we got 30 Graphlets. In these Graphlets, we analysed the number of types of nodes. For the first Graphlet, we thought that the two nodes are symmetric, so there was only one node type. Similarly, the two nodes of the edge in the second subgraph are symmetric, and we thought that the second Graphlet has only two node types. Repeat, we can find a total of 73 node types in the graph. However, these node types are still not independent. According to the calculation of the geometric relationship, we found that there are 56 node types in the Graphlet with 5 nodes as the upper limit. Due to the amount of calculation, we can find 11 independent node types with 4 nodes as the upper limit.

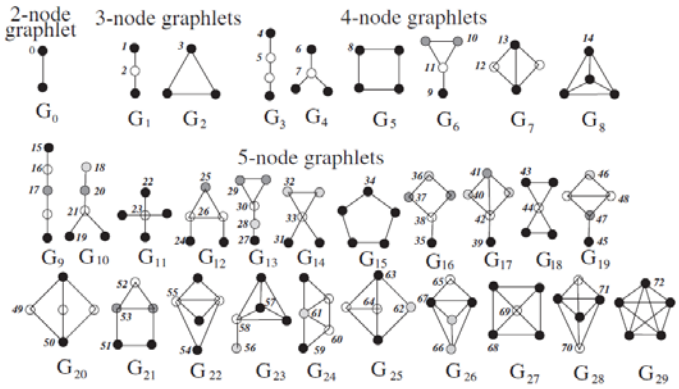

Fig. 1 Graphlets up to 5 nodes and node types [5]

As we know, one node in a network will work in different Graphlets. So for each node, we can calculate its frequency of occurrence in each of the 11 node types. A node corresponds to an 11-dimensional row vector, which is called the Graphlet Degree Vector. Repeat the calculation, we can convert any network into a matrix, in which the number of rows is the number of nodes, the number of columns is always 11. This matrix is called Graghlet Degree Distribution Agreement [5].

Then we can calculate the Pearson correlation coefficient for the 11 columns of the matrix. The resulting matrix is called the Graphlet Correlation Matrix (GCM), which is an 11th-order symmetric matrix with values between -1 and 1. We believe that a GCM implies the essential attributes of the network. So, we calculated the Euclidean distance between GCMs of different networks and we got the Graphlet Correlation Distance [5].

II. Result

Figures 1-4 shows the Graphlet Correlation Matrix (GCM) of Wikipedia vote network, Facebook network, Minnesota Highway Network and European aircraft route network. The GCMs 11th-order symmetric matrixes with values from -1 to 1. We can see four matrixes have different patterns, which is the feature of their intrinsic property. However, the patterns of Fig. 4 and Fig. 5 seem similar, which indicates they have Similar formation mechanism.

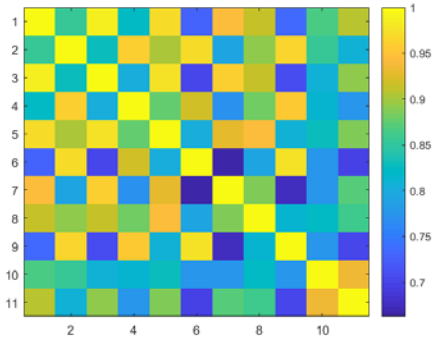

Fig. 2 GCM of Wikipedia vote network

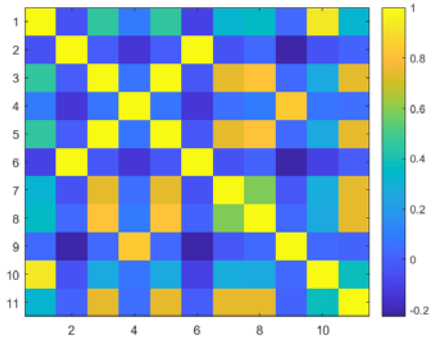

Fig. 3 GCM of Facebook network

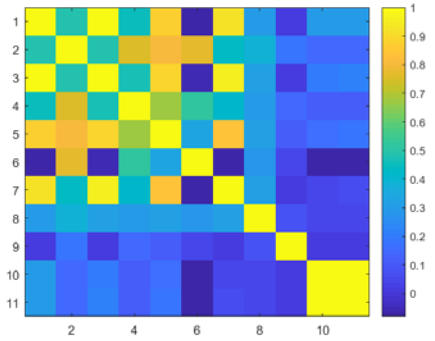

Fig. 4 GCM of Minnesota Highway Network.

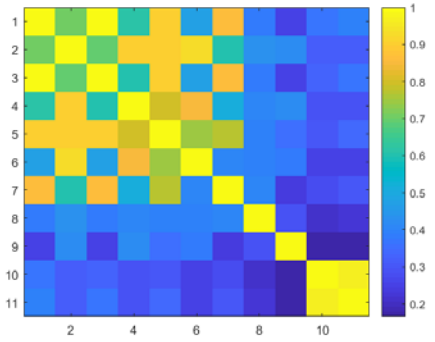

Fig. 5 GCM of European aircraft route network

Table 1 shows the distance of GCMs of four networks. If the distance is small, it seems that the networks have more similarity. We can see the distance of Minnesota Highway Network and European aircraft route network is the smallest one, which is in line with our formal expectation. As we know, both the two networks are geometry-based network. Although they stand for different places and different transportation, they show a higher similarity. This result corresponds to our intuition and reflect that Graphlet Correlation Distance is a good indicator to mine the intrinsic property of different networks.

TABLE I
DISTANCE OF DIFFERENT NETWORKS

Network Distance	European Aircraft	Wiki	Facebook	Minnesota Highway
European Aircraft	0.00	3.05	3.46	1.60
Wiki	3.05	0.00	5.06	4.36
Facebook	3.46	5.06	0.00	2.93
Minnesota Highway	1.60	4.36	2.93	0.00

Besides, the distance between European aircraft route network and Facebook network, the distance between European aircraft route network and Wiki vote network are longer than the distance between Minnesota Highway Network and European aircraft route network, which is also easily to understand. However, the distance between Facebook network and Wiki network is the longest one, which is out of our expectation. A potential explanation can be that although both the two networks are social network online, they form in very different situations. When one person votes another person as the Wikipedia administrator, there will be an edge between them in Wiki vote network. When two people are friends on Facebook, there will be an edge between them in Facebook network. Therefore, the different form of network maybe affect the real network structure, though they are both online social networks.

III. CONCLUSION

The study was designed to compare a number of networks and to classify them accurately based on the Graphlet Correlation Distance. The findings were comparison of four networks with the GCD indicator based on the data collected from networkrepository. It was found that intuition-similar networks tend to have a smaller distance.

The result gave us an accurate classification method to compare wholly different networks with scale disparity quantitatively. Meanwhile, the indicator may have also contributed to the deep understanding of network attributes.

However, the results are limited due to the short time to finish this work. We need more time to analyze and compare at least thirty networks, which will make our work more reliable. From our current result, we think the GCD is an accurate indicator to reflect the intrinsic property of different networks and we will deal with more networks based on it.

It is recommended that we can use Graphlet Degree Vector, which keeps the most of the information in a network, to compare the features of different networks. We also need to take advantage of all kinds of indicators to have a synthetical, reliable network classification method.

ACKNOWLEDGMENT

We thank Hao Wang for his comments and assistance with the work.

REFERENCES

[1] Z. Li, L. Chen, Y. Bai, K. Bian, and P. Zhou, “On diffusion-restricted social network: A measurement study of WeChat moments,” in *2016 IEEE International Conference on Communications (ICC)*, 2016, pp. 1–6.

[2] S. A. Cook, “The Complexity of Theorem-proving Procedures,” in *Proceedings of the Third Annual ACM Symposium on Theory of Computing*, New York, NY, USA, 1971, pp. 151–158.

[3] T. Opsahl, F. Agneessens, and J. Skvoretz, “Node centrality in weighted networks: Generalizing degree and shortest paths,” *Social Networks*, vol. 32, no. 3, pp. 245–251, Jul. 2010.

[4] B. Bollobás* and O. Riordan, “The Diameter of a Scale-Free RandomGraph,” *Combinatorica*, vol. 24, no. 1, pp. 5–34, Jan. 2004.

[5] N. Pržulj, “Biological network comparison using graphlet degree distribution,” *Bioinformatics*, vol. 23, pp. 177–183, 2007.

[6] D. Marcus and Y. Shavitt, “RAGE – a rapid graphlet enumerator for large networks,” *Comput. Netw.*, vol. 56, pp. 810–819, 2012.

[7] W. Hayes, K. Sun, and N. Pržulj, “Graphlet-based measures are suitable for biological network comparison,” *Bioinformatics*, vol. 29, pp. 483–491, 2013.

[8] R. Rossi and N. Ahmed, “The network data repository with interactive graph analytics and visualization,” in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[9] T. Hočevar and J. Demšar, “A combinatorial approach to graphlet counting,” *Bioinformatics*, vol. 30, no. 4, pp. 559–565, 2014.

[10] Ö. N. Yaveroğlu *et al.*, “Revealing the Hidden Language of Complex Networks,” *Scientific Reports*, vol. 4, p. 4547, Apr. 2014.