Characterization and graph embedding of weighted social networks through Diffusion Wavelets

Zhiliang Chen12, Junfeng Wu12, Huakang Li134, Guozi Sun13
1Nanjing University of Posts and Telecommunications, Nanjing, China
2New York Institute of Technology, New York, USA
3Collaborative Innovation Center for Economics crime investigation and prevention technology, Nanchang, China
4Suzhou Privacy Information Technology Co., Ltd. China
huakanglee@njupt.edu.cn

Abstract—More and more graph embedding algorithms have been proposed, which makes the similarity judgment of graph structure more and more accurate. While exploring the similarity of neighborhood structures, the existence of weights should also be taken into account, so as to reflect the relational social network graph in the real world. We use Graphwave, a kind of algorithms for graph embedding with diffusion wavelets, to incorporate weight into numerical value to calculate, and to process the returned probability distribution parameters, so that we can get some analysis about the actual complex network. Our analysis can overcome the priori misjudgment problem based on the topological structure, and then obtain the actual similarity of the network structure from the results of graph embedding.

Index Terms—Weighted relationship, graph embedding, graphwave, social network

I. INTRODUCTION

Graph embedding works ask people to find the topologically similar network neighborhoods. The early work was devoted to finding similarities of nodes on a small scale. Some new studies could span large scale networks and return some similar nodes in different areas of a graph. Those nodes with similar structural roles have a certain degree of similarity in the network [5-8,10-13]. And the results can help people to divide the networks into different levels with similar characteristics, such as distinguishing hot airports from airline networks, and dividing many different areas into several groups according to the different power demands in the power networks.

Using Laplacian transformation as an advanced mathematical mean becoming popular gradually. Also, combining it with many physical concepts [1, 25] improves the effect of graph embedding. Graphwave [1] is an advanced theory in the field of graph embedding with its utility of diffusion wavelet in the graph and calculate the probability distribution of energy in each node to learn the structure of the network. On the classification task of graphs with symmetry, it shows better effect than most other theories. Also, its computational process is proved that overall complexity is linear in the number of edges, that means it can contain large sparse networks.

Our first work is to verify the classification effect of graphwave in the weighted graph and check the selection of Clustering Method for Characteristic Functions. We first convert the data into a triple that containing two nodes information and the weight between this two nodes. Then, the unit energy is multiplied by the weight according to the original algorithm.

After that, we embed synthetic networks into graphs and obtain the practical meaning of point location by comparing artificial analysis and experimental results with clustering.

What’s more, we will show the importance of weight for social network distinction by comparing the classifications between weighted models and unweighted models of the same structure.

The main contribution of this paper is to verify the effectiveness of graph embedding algorithm in practical weighted networks. Through various experiments, we can:

- Acquire structural prominence nodes to extract key persons
- Display the overall structure similarity distribution from a split relational network
- Search for similar contents based on known information.

The whole paper will be discussed in the following order. We firstly bring a brief introduction to the basic content of Graphwave algorithm. Afterwards we give a basic explanation for principal component analysis. Later, small and large data sets are applied to evaluate the correctness of our theory. Finally, we will give a conclusion of our work.

II. GRAPHWAVE ALGORITHM

A. Definition of Graph

In this section, we consider that the definition of a graph with \( N \) nodes is an ordered triple \([2, 3]\) \( g = (\nu, \varepsilon, \phi) \) including:

1. \( \nu(g) = \{a_1, a_2, ..., a_N\} \), \( \nu(g) \neq \phi \)
2. \( \varepsilon(g) \) is a set of edges, and \( \nu(g) \cap \varepsilon(g) = \phi \)
3. \( \phi(g) \) is incidence function

If each node has one or more edges with other nodes. We can simplify the definition as:

\[
g = (\nu, \varepsilon)
\]

(1)

Then we get adjacency matrix \( A \):

\[
A = A(g) = (a_{ij})_{n \times n}
\]

(2)

For a weighted graph, the adjacency matrix \([3, 4]\) is:
\[ A[i,j] = \begin{cases} w_{ij}(a_i,a_j), (a_j,a_i) \in \varepsilon(g) \\ 0 \text{ otherwise} \end{cases} \]  

where \( w_{ij} \) is the weight between nodes \( i \) and \( j \). Equation (3) shows the adjacency matrix \( A \).

We also get a degree matrix:

\[ D_{ij} = \Sigma_j A_{ij} \]  

Equation (4) is the degree matrix \( D \).

**B. Diffusion wavelets theory**

In this section, we provide the previous work on the graphwave model [1], which is the spectral graph wavelet-based model. We omit the process of proving the correctness of the theory and focus on the definition of the model itself. The analysis of graphwave has been given in the previous work.

The first step is to accomplish the Laplacian Matrix of the graph:

\[ L = D - A = U \Lambda U^T \]  

Equation (5) is the Laplacian matrix \( L \), where \( U \) is the eigenvector decomposition of the unnormalized graph and the eigenvalues of this Laplacian Matrix is shown as:

\[ \lambda_1 < \lambda_2 \leq \cdots \leq \lambda_N (\Lambda = \text{Diag}(\lambda_1, \lambda_2, ... , \lambda_N)) \]  

Equation (6) is the eigenvalues of the graphwave.

Then, we should find a filter kernel to apply in a specific model, in the previous work, a common selection is the heat kernel \( k_s(\lambda = e^{-\lambda s}) \) with scaling parameter \( s \). A brief introduction of finding an appropriate value of will be given in the next subsection, and the complete search process has been confirmed in previous studies.

Since we need to transfer the graph matrix and the heat kernel into spectral graph wavelet in signal field, graphwave theory introduces Dirac Delta Signal to activate unit energy at the labeled node. In their work [15], the definition is as:

\[ \phi_x(t) = \mathbb{E}[e^{itX}], t \in \mathbb{R} \]  

Equation (9) is the characteristic function as a probability distribution.

And the same equation for the wavelet coefficient is like:

\[ \phi_a(t) = \frac{1}{N} \sum_{m=1}^{N} e^{it\Psi_{ma}} \]  

Equation (10) is the wavelet coefficient.

So, we can divide the probability distribution with sampling into projection of complex space at \( d \) evenly spaced points from 1 to \( d \), that helps us obtain 2 different factors of the wavelet coefficient. The structural embedding of node \( a \) is defined as:

\[ \chi_a(t) = \{Re(\phi_a(t_i)), Im(\phi_a(t_i))\}_{t_1,t_2,...,t_d} \]  

Equation (11) is the structural embedding.

**D. Selection of appropriate range for scaling parameter \( s \)**

The scaling parameter \([9, 16]\) brings multi-scale version on previous subsection, so we do need to find a correct appropriate range of that. Luckily, the scholars who put forward the theory of graphwave have solved this question.

In their work [1], two propositions are proved:

1. **PROPOSITION 1.** The variance of off-diagonal coefficients in heat diffusion wavelet \( \Psi_{ma}^{(s)} \) is proportional to:

\[ \text{Var}[(\Psi_{ma}^{(s)})] \propto \Delta^{(1)}(2s) - (\Delta^{(s)})^2 \]  

Equation (12) is the variance of off-diagonal coefficients.

2. **PROPOSITION 2.** The convergence of heat diffusion wavelet coefficient \( \Psi_{ma}^{(s)} \) is bounded by:

\[ e^{-\lambda_N[s]} \Delta^{(s)}(s) \leq \Delta^{(s)}(s) \leq e^{-\lambda_2[s]} \Delta^{(s)}(s) \]  

Equation (13) is the convergence of heat diffusion wavelet coefficient.

To cover an appropriate range of scale, we suggest setting \( \eta = 0.85 \) and \( \gamma = 0.95 \).

**Algorithm 1 PCA by using SVD method.**

1: Input: Matrix \( A \equiv [\chi_1, \chi_2, ..., \chi_{N}] \) 
2: Output: Matrix \( A' \equiv [\chi_1, \chi_2, ..., \chi_{N}] \) 
3: For each \( \chi_a \) and its eigenvalue, apply Mean Removal Method and return \( A^{MR} \) 
4: Compute \( S \) of \( A^{MR} \) (Eq.1) 
5: Sort all \( h \) eigenvalues of \( S \) with descending order 
6: Chaos top \( h \) values of that order 
7: \( \chi_{trans} = [\xi_1, \xi_2, ..., \xi_N] \) 
8: Transfer the elements in origin matrix into this new eigenvectors-matrix space

where \( \Delta^{(s)}(s) = |\Psi_{ma}^{(s)} - \frac{1}{N}| \) decreases monotonically with \( s \).

Equation (14) is the convergence of heat diffusion wavelet coefficient.
III. PCA OF WEIGHTED GRAPH

When it comes to the analysis of specific networks, clustering algorithm will act on the final manifestation of wavelet coefficient [20] in section II. But the coefficient is difficult to represent the actual parametric significance, even though we can find the similar nodes by using it. PCA (principal component analysis) provides us a method to switch the origin components of coefficient into weight-dependent factors [17,18,19,21]. In this section, we explain the analysis process and physical meaning of new components. It is of great significance to distinguish the characteristics of nodes from classification results.

A. Singular value decomposition (SVD)

In this paper, we choose SVD method [22] to find a decomposition matrix. It is a method that can be applied to any matrix. There is always an SVD for any matrix $A$:

$$A = U\Sigma V^T$$

with:

1. $U$ is the unitized eigenvectors of $AA^T$
2. $V$ is the unitized eigenvectors of $A^TA$
3. $\Sigma$ is the square root of eigenvalue of $AA^T$.

Then, we describe PCA of structural embedding (Algorithm 1). For a natural network, loose numerical distribution may cause some problems of calculation. So, the first step is using Mean Removal Method to eliminate potential risks (Line 3). SVD requires eigenvalues and eigenvectors of scatter matrix(Line 4), it is defined as:

$$S = \sum_{k=1}^{n} (x_k - m)(x_k - m)^T$$

where $m$ is the mean vector: $m = \frac{1}{n}\sum_{k=1}^{n}x_k$.

After that, we use SVD to find and sort the eigenvalues. For $n$ eigenvalues of the matrix, we choose the top $k$ from peak to bottom and use these new eigenvalues ($\xi$) rank to build new eigenvectors matrix, then we can transfer the origin data into this new matrix space (Line5-8).

B. Representation of Component Reduction via trace method

In this section, we try to explain the components after the PCA. Although the network of a graph could be very complex, its feature can be simplified because of the basic definition. The neighborhood-influence of a node only involves a few aspects. When we talk about a weighted graph (unidirectional), this scope is limited to the following factors: degree, total-weight and other things derived from the above factors (transmission of the parameters originates from degree matrix and adjacency matrix).

To find out which parameters have the greatest impact on a graph, we fix other scales invariant and only control $h$ as the argument. For the implementation of actual code, it is easy to control this variable. In section IV.A.(1), we will apply algorithms to synthetic graphics with different structures and find a reasonable explanation.

IV. EXPERIMENTAL EVALUATION

In what follows we examine Graphwave in different scenarios in order to illustrate classification quality and the meaning of principal component. As a matter of fact, graph algorithm has achieved excellent measurement level in some common simple classification problems.

A. Barbell Graph

Let’s consider constructing a barbell graph in a special case. Since the traditional unweighted barbell graph has symmetry, in the process of adding weight, we intend to distribute weights symmetrically. Also, in this composition, we reduce the weight at both ends of the barbell. Finally, we constructed the model shown in Fig. 1.

Fig. 1. We choose B(4, 4) as the base structure. The barbell graph has 6 distinct classes of structurally equivalent node as indicated by color; the same distinction is applied in the equivalent edges. Legend is intended to illustrate the mapping pattern of weight and color.

According to previous theories [1, 23], our synthetic graph gives each node with two strong features. They can obtain great weights from adjacent nodes but sacrifice distance for all else nodes, or vice versa.

When other situations are fixed, we change $h$ (Section III.A) from 2 (the least value of 2D-PCA) to 5. The results of these classifications are compared, and it means that when principal components are more than two, it is next to nothing for the newly increased components to classification work (Fig. IV-A). At this point, we conclude that no new component can be obtained by manipulating the input polynomial [24] in PCA process and its basis vectors are not more than two types, which is obviously about weight and degree.

What’s more, under weighted conditions, graphwave continues the previous excellent performance [1]. All points in symmetrical positions are overlapped in figures.

B. Complete Graph

A complete graph has some specific properties that can help us find out the shortcomings of the algorithm. It can
also represent the relationship network of a closed loop world and is very important to analyze the characteristics of local network. As is shown in Fig. 3, all points should have the same properties when they are weighted uniformly on paper. That assumption is proved as following (Fig. 3. b). What’s more, when we try to distinguish different areas in the same structure (Fig. 3. c), and separation has also been successfully achieved.

V. APPLICATION FOR LARGE DATA SET

A. Flight network

Aviation data sets, which are commonly used in graph embedding, are based on airline distribution. Their structures cannot support weighted analysis. Meanwhile, the flight data set can give weight to each side. This time, the data pre-treatment is very tedious. Each data can only support adding one weight to a single edge, we need to use external data processing software to convert existing tables.

We selected 2016 China’s domestic aviation data from China Civil Aviation Administration. We also find official ranking of 2016 domestic airports to be accuracy index of experimental result. The total number of flights in this data set is 15074 and corresponds to 186 airports in China.

The existence of weights makes recurrent airports have local prominent characteristics. Similar cases are also found in classification results.

B. International Football Exchange Network

We might as well analogize countries and regions to people on social networks. Communication between countries has the same characteristics as interpersonal communication. Football has developed into one of the most important sports to promote exchanges between countries. The data set we chose was all international events between the first formal international match in 1872 and the end of the 2018 World Cup in Russia.

We apply our algorithm to the data set, and then observe the clustering model generated after the graph is embedded. The result can reflect the distribution of this football network graph. The topological structure of the network can make us judge it to a certain extent, but weight cannot be clearly reflected in the structure graph. Although we think that there are regional clustering in football from the Figure 6, we get some different
results after the actual operation with the weighted algorithm. By calculation, the Dispersion Coefficient of this data set is 0.9338. The results show that the dispersion of the entire international football network is much lower than expected: the more frequent exchanges between countries with strong football culture make their structures very similar and weak countries have the same result. That means that football has a mature and complete interactive system in the world.

When some data sets present the distribution like this football dataset, it is difficult to find some research values other than searching jobs. However, if the embedding result is uniformly distributed with its long-period statistics, it will be able to find new embedding methods from another level. In this section, the data set has lasted for more than a century. Therefore, if the sample is split into uniform timestamps, the content obtained is quite different from the result of integration. The method of embedding networks of different time layers into a vector space is still undetermined in this paper. A possible scenario is posted in Fig. 8. Our team will make a concrete analysis and answer to this question in the future.

C. Enron Email Network

Enron email data set is one of the most classic mail sets. In fact, email reflects people’s communication activities. Constructing the network of sending mail between people also constructs the social network of the whole company. We regard the number of mail exchanges as a weight. In this way, the structural characteristics and differences among different people in the whole company network are analyzed.

In the process of data processing, we delete all the external chains. We pick up the e-mail exchanges of 150 staffs under the investigation of Enron case and then use social relationships between suspects to analyze key people. After the screening is completed, Tana, Jones and Stephanie Panus is excluded because there is no internal network relationship. After that, we embed the remaining 148 nodes into a graph.
these two men, there may have the same suspicions. Dutch. Quigley, Kevin. Ruscitti and Errol. Mclaughlin are screened out under this term. Another way is to find the great liaisons in this network in order to obtain important information through questioning these people. Higher weighted nodes have been separated to the right of the graph by the algorithm. We can get some important people from it, like Phillip. Allen and John. Lavorato.

D. Discussion on the Advantage of Adding Weights

Data analysis in the previous sections has made us aware of the application of weighted graphs in real network relations. In this section, we will compare the unweighted and weighted embedding of the same network structure, to find the advantage of weighted algorithm in application.

We choose the flight network as the basic structure. Since we can obtain the rank of those airports, tagging is also within our reach. Then two different kinds of graph embeddings applications are applied to this structure (Fig. 12). By calculating clustering index of pre-marked points, we can clearly see that the weighted algorithm has outstanding classification effect in the application of large data sets.

![Fig. 10. Embedding results of Enron emails data set.](image)

![Fig. 11. Different applications of the enron network](image)

![Fig. 12. PCA of two embedding algorithms extreme distributions of network embedding.](image)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Homogeneity</th>
<th>Completeness</th>
<th>Silhouette</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unweighted</td>
<td>54.09%</td>
<td>54.10%</td>
<td>24.12%</td>
</tr>
<tr>
<td>Weighted</td>
<td>61.18%</td>
<td>61.07%</td>
<td>52.36%</td>
</tr>
</tbody>
</table>

VI. Conclusions

We validate the Graphwave theory in the weighted graph and prove that it also has good learning effect in the weighted graph. In addition, we also test the classification results, and find the basis vectors of the classification through a series of weighted graphs with manual-construction.
Weighted graphs are of great significance in practical scenarios. When we search for isomorphic things in a real network, we need to measure the relationship between different points inevitably. Interpersonal relationships in social networks and transmission relations in Electric Power Network all have this characteristic. Acceptable algorithm complexity makes this theory valuable for application, and it is more in line with the physical meaning of point-to-point interaction.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (Grant No.61906098, No. 61906099), China Postdoctoral Science Foundation (Grant No.2016M600434, 2016M591840), Natural Science Foundation of Jiangsu Province (Grant No. BK20150856), and Postdoctoral Science Foundation of Jiangsu Province(Grant No. 1601128B, 1501054B), Opening Project of Collaborative Innovation Center for Economics crime investigation and prevention technology(Grant No. JXJZXTCX-015), Angel Project of Suzhou City science and technology(Grant No. CYTS2018233).

REFERENCES