

# VISUAL CLUSTERING OF COMPLEX NETWORK BASED ON NONLINEAR DIMENSION REDUCTION

Jiayu Li<sup>1</sup>, Shuzhong Yang<sup>2</sup>

<sup>1</sup>School of Computer and Software, Communication University of China, 100024, Beijing, China

<sup>2</sup>School of Computer and Information Technology, Beijing Jiaotong University, 100044, Beijing, China

**Abstract:** In this paper, we present a new visual clustering algorithm inspired by nonlinear dimension reduction technique: Isomap. The algorithm firstly defines a new graph distance between any two nodes in complex networks and then applies the distance matrix to Isomap and projects all nodes into a two dimensional plane. The experiments prove that the projected nodes emerge clear clustering property which is hidden in original complex networks and the distances between any two nodes reflect their close or distant relationships.

**Key words:** complex network, visual clustering, Isomap, graph distance.

## 1. INTRODUCTION

In recent years, there is a growing interest in evolving complex networks. Small world characteristic [1] and scale-free characteristic [2] are the two most important characteristics of complex networks. Many networks in real world not only have small world characteristic and scale-free characteristic, but also have community structure property. The description of the community structure of complex networks has also been one of the focuses of attention in recent years [3, 4, 5]. Visual clustering of complex networks can help us find the community structure hidden in the networks, understand the networks better and predict the behavior of networks in future. Since complex networks are usually characterized by various graphs, we can realize the visual clustering of complex networks by using the similar methods of graphs.

To realize the visual clustering of a graph, we must firstly define an effective measurement of distances between nodes. So how to define effective distances becomes a crucial issue for the visual clustering of graph. Then we need choose an appropriate projection technique to layout the result of visual clustering in a plane or a three dimensional space. Dimension reduction techniques have been widely developed as the projection techniques up to now [6, 7, 8, 9]. Since Isomap [9] can keep the global geometric property through estimating geodesic distances between pairwise nodes, we will choose it to realize the layout of visual clustering of complex works in this paper.

The rest of the paper is organized as follows. In Section 2, we show how to realize the visual clustering of complex works using nonlinear dimension reduction technique Isomap. In Section 3, some performance experiments are presented. Finally, discussions and future work are given in Section 4.

## 2. VISUAL CLUSTERING ALGORITHM

### 2.1 Quantification of Distances in Graph

A graph  $G = (V, E)$  consists of a finite set  $V$  of nodes and a finite set  $E$  of edges with  $E \in V^{(2)}$ , where  $V^{(2)}$  is the set of all subsets of  $V$  which have exactly two elements.

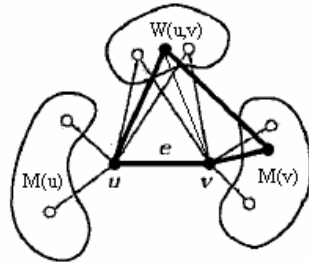


Figure 1. Dividing neighbors of edge  $(u, v)$

According to the above graph definition and edge strength definition in [10], the pairwise distances between nodes can be computed through the following steps. Given an edge  $(u, v) \in E$  (Figure.1), we can compute its strength by dividing neighbors of  $u$  or  $v$  into three distinct subsets. Firstly, denote by  $M(u)$  the set of neighbors of  $u$  that are not neighbors of  $v$  (excluding  $v$ ). Secondly, denote by  $M(v)$  the set of neighbors of  $v$  that are

not neighbors of  $u$  (excluding  $u$ ). Finally, denote by  $W(u, v)$  the set of common neighbors to  $u$  and  $v$ . Denote by  $r(A, B)$  the number of edges linking nodes in the set  $A$  to nodes in the set  $B$  such that  $s(A, B) = r(A, B) / |A||B|$  ( $|A|$  is the number of nodes in  $A$ ) defines the proportion of existing edges among all possible edges connecting nodes of  $A$  and  $B$ . By above definition, any edge connecting two of the subsets  $M(u)$ ,  $M(v)$  or  $W(u, v)$  is part of a cycle of length 4 going through the edge  $(u, v)$  (see Figure. 1, the cycle labeled by thick lines). Note that all cycles of length 4 are captured this way. Finally, we define the ratio  $|W(u, v)| / (|M(u)| + |W(u, v)| + |M(v)|)$  as a ratio related to the proportion of cycles of length 3 containing the edge  $(u, v)$ . Note that there are as many of these cycles as there are nodes in  $W(u, v)$ . Then the strength of an edge is given by computing:

$$strength(u, v) = \frac{s(M(u), W(u, v)) + s(W(u, v), M(v)) + s(W(u, v), M(u), M(v)) + |W(u, v)|}{(|M(u)| + |W(u, v)| + |M(v)|)} \quad (1)$$

(Note: we need to set  $s(A, A) = 2r(A, A) / (|A|(|A| - 1))$  when computing the proportion of edges connecting a set to itself). Then we can use the inverse of strength to define the pairwise distances between nodes. If the strength between two nodes is zero, the distance between them is infinity; otherwise the distance equals to  $1 / strength(u, v)$ .

## 2.2 Visual Clustering Using Isomap

The main idea behind the Isomap (isometric feature mapping) algorithm is to perform classical MDS [7] to map data points from their high-dimensional input space to low-dimensional coordinates of a nonlinear manifold. The key contribution is to compute the MDS pairwise distances not in the input Euclidean space, but in the geodesic space of the manifold. The actual geodesic distances are approximated by a sequence of short hops between neighboring sample points. Finally, MDS is applied to the geodesic distances to find a set of low-dimensional points with similar pairwise distances. We use Isomap to visualize the complex networks because it can keep the global structural properties hidden in them.

Now we list the whole visual clustering algorithm as follows:

1. Quantify the distances between any two nodes according to equation (1) and get the distance matrix  $D = \{d(i, j)\}$ ;
2. Using Isomap technique to projected all nodes into a plane. The detailed procedure is:
  - 1) Suppose the quantified distances as the distances of some points on high dimensional manifold  $M$ ;
  - 2) Determine which points are neighbors on the manifold  $M$  by identifying their  $K$  nearest neighbors, based on the distances  $d(i, j)$  between

pairs of points  $(i, j)$  in manifold  $M$ . These neighborhood relations are represented as a weighted graph,  $G$ , over the data points, with edges of weight  $d(i, j)$  between neighbors;

3) Estimate the geodesic distances  $d_M(i, j)$  between all pairs of points on the manifold,  $M$ , by computing approximations as the shortest path distances  $d_G(i, j)$  in the graph  $G$ . Then the final matrix of graph distances  $D_G = \{d_G(i, j)\}$  will contain the shortest path distance between all pairs of points in  $G$ ;

4) Apply classical MDS to the matrix  $D_G$  to construct an embedding of the data in a two dimensional plane;

5) Draw the embedding data points in a plane.

### 3. EXPERIMENTAL ANALYSIS

In the literature [11] the authors presented an energy model for visual graph clustering and proved that their LinLog model had better results than previous models. In this paper we will compare our model with LinLog model. Since lack of real world data only two typical kinds of artificial complex networks are used: structured complex network and non-structured complex network (WS small world network). The structured complex network is produced by the model presented in [12] and includes 4 communities and every community has 40 nodes. The non-structured complex network is produced by the WS small world model in [1] and includes 200 nodes. Fig. 2 and Fig. 3 list the visual clustering results of two methods, respectively.

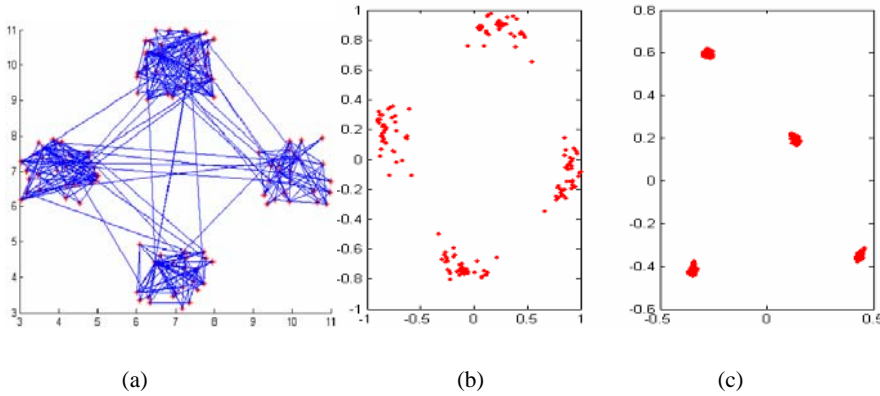


Figure 2. Visual clustering of constructed complex network. (a) Original network; (b) The result of our model; (c) The result of LinLog model

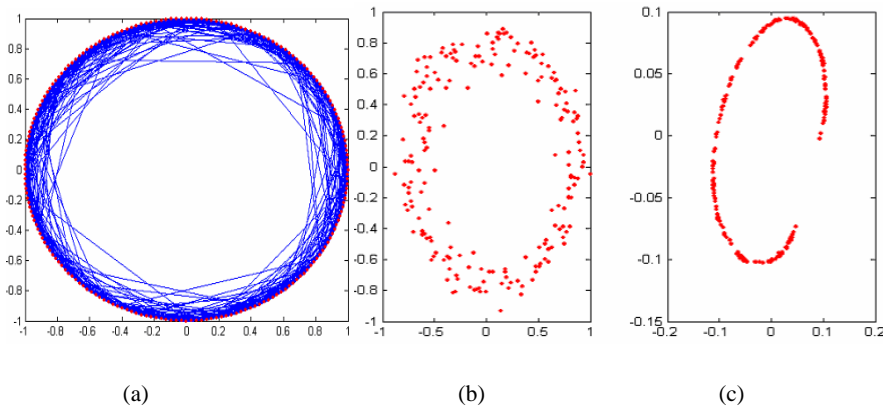


Figure. 3. Visual clustering of WS small world network. (a) Original network; (b) The result of our model; (c) The result of LinLog model

In Fig.2 (a) original network consists of four communities. Our model and LinLog model both visualize the network into four communities (Fig.2 (b) and (c)). But it is easy to see that our model can make clustering structure of complex network clearer and project the nodes more uniformly. In contrast to our model, LinLog model overlaps the nodes in the same community after projection so that we can't distinguish them. However, when applied to WS small world networks (Fig. 3), our method emerges distinct advantages compared with LinLog model: our model can keep both the local and global clustering property of small world networks but LinLog model can not. That is to say that the visual clustering result of our model not only preserves the adjacency between nodes but also preserves the original circular distribution. Compared to our model, LinLog model can not keep the circular distribution because its visual clustering result emerges split. The phenomenon can be explained as: the distance of LinLog model between two nodes is only proportional to the coupling, but our distance also considers the neighborhood of two nodes and the following estimation of geodesic distance extends the Euclidean distance. So when the clustering structure of complex networks is less distinct, our visual clustering model has more advantage than LinLog model.

In addition, regular and random network based on ER model are also used to compare the performance between our model and LinLog model. Our model also has better performance in node distribution after projection than LinLog model. But the advantage is less distinctive because of degree homogeneity of network nodes.

## 4. DISCUSSIONS

In this paper we present a new visual clustering method. Our contribution includes two aspects: introducing a new measurement of distances between nodes in a graph and using nonlinear dimension reduction technique Isomap to realize the layout of complex networks. The experimental analysis shows that our method has better performance than other visual clustering methods and the resultant nodes in graph have more clear clustering property.

## ACKNOWLEDGEMENTS

The research is supported by the National Natural Science Foundation of China under Grant Nos. 60373029 and the National Research Foundation for the Doctoral Program of Higher Education of China under Grant Nos. 20050004001.

## REFERENCES

1. Watts, D., and Strogatz, S.: Collective dynamics of 'small-world' networks. *Nature* 393 (1998) 440-442.
2. Albert-Laszlo Barabasi, Reka Albert.: Emergence of scaling in random networks, *Science* 286 (1999) 509-512.
3. Girvan, M. and Newman, M. E. J.: Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*. 99(12) (2002) 7821-7826.
4. MEJ Newman and M. Girvan.: Finding and Evaluating Community Structure in Networks. *Phys. Rev. E* 69, 026113 (2004).
5. MEJ Newman.: Detecting Community Structure in Networks. *Eur. Phys. J. B* 38, (2004) 321-330.
6. J. Karhunen, J. Joutsensalo.: Generalizations of principal component analysis, optimization problems, and neural networks. *Neural Networks*, Vol. 8, No. 4, (1995) 549-562.
7. T.F. Cox, M.A.A. Cox.: *Multidimensional Scaling*. Chapman & Hall, (2000).
8. S.T. Roweis, L.K. Saul.: "Nonlinear Dimensionality Reduction by Locally Linear Embedding". *Science* 290 (2000) 2323-2326.
9. J.B. Tenenbaum, V.d.Silva, J.C. Langford.: "A Global Geometric Framework for Nonlinear Dimensionality Reduction". *Science* 290 (2000) 2319-2323.
10. Auber, D., Chiricota, Y., Jourdan, F., and Melancon, G.: Multiscale visualization of small world networks. In *Proceedings of the 2003 IEEE Symposium on Information Visualization* (2003) 75-81.
11. Noack, A.: An energy model for visual graph clustering. In *Proc. 11th International Symposium on Graph Drawing, LNCS 2912* (2004) 425-436.
12. Chunguang Li and Philip K. Maini.: An evolving network model with community structure. *Journal of Physics A: Mathematical and General*. 38 (2005) 9741-9749.