

Complex Networks Clustering and Edges Correlation

Igor Kanovsky

Max Stern Academic College of Emek Yezreel

Emek Yezreel, 19300,

Israel

972-4623-493

igork@yvc.ac.il

ABSTRACT

In this paper, we describe a new method for cluster recognition in complex networks. A typical real world network has Small World properties or/and some other edges correlation. The method invokes edge weighting based on the edge participants in local edges correlation. Inter-cluster edges are perceived as not correlated. We propose an extension of the popular Small World model of Watts and Strogatz and use it to test our approach. The limitations and advantages of the method are investigated in this paper.

Categories and Subject Descriptors

G.2.2 [Discrete Mathematics]: Graph Theory – *Network problems*; G.3 [Mathematics of Computing]: Probability and Statistics – *stochastic processes*; I.6.5 [Simulation and Modeling]: Model Development.

General Terms

Algorithms, Measurement, Theory.

Keywords

Complex networks, Small World, Extended Small World model, clustering, clustering coefficient, link weighting, and community structure.

1. INTRODUCTION

A lot of complex systems may be represented as a Small World (SW) graph. For example the Web, Internet, personal contacts, citation graph, protein interaction networks, etc. [1]. These systems are inhomogeneous in the perception of existence of sub-graphs with relative bigger density of edges, so called “communities” or clusters. From a practical point of view it is important to recognize clusters within real world networks. The problem of cluster recognition is a well-studied field of data mining. In [2, 3], new approaches for network clusters structure identifying in complex networks were introduced. The algorithms are based on the concept of betweenness centrality of a node or link, of which utilization may not be effective for big size graphs or when only partial graph data is available.

In this paper an alternative idea for complex network clustering is proposed. Real world networks have to provide some functionality. This is expressed by the SW effect and other link-link correlations. Thus it is expected that inside a cluster the link correlation has to be significantly stronger than for inter-cluster links. For example, the clustering coefficient [1] calculated for a

cluster has to be bigger than one calculated for its entire network. We propose link weighting based on its near neighbors’ link correlation, which represents the intro-cluster nature of the link. As usual, clusters are recognized as a set of nodes connected with links having weights bigger than some threshold value.

2. CLUSTERING ALGORITHM

2.1 Link Weighting

We define that the weight of an edge in SW graph is proportional to the number of the common neighbors in its neighborhood. For the edge $e \in E$ connecting vertices $(v_1, v_2) \in E$ we define the edge weight β as:

$$\beta(v_1, v_2) = \frac{|N(v_1) \cap N(v_2)|}{|N(v_1) \cup N(v_2)|}$$

where $N(v)$ is the neighborhood of the vertex v . Observe that $0 \leq \beta \leq 1$.

The weighting formula was inspired by the SW property which requires a big number of triangles. Actually the weight is the degree of the edge contribution in the SW formation. For other kinds of networks, edge weights have to be chosen as a measure of edge involvements in the formation of the small sub-graphs that provide network functionality. Sometimes such sub-graphs are called motifs [4].

2.2 Iterated Cluster Recognition Algorithm

Two adjacent vertices v_1, v_2 belong to the same cluster if the weight of their connecting edge is bigger than some threshold value which is a parameter of the algorithm, $(v_1, v_2) \in E$

$$\beta(v_1, v_2) > \alpha$$

where α is the level of the cluster separation for a graph. Notice that non-adjacent vertices v_1, v_2 belong to the same cluster C if there exists a path L connecting v_1 to v_2 in which each edge in L has a bigger weight than the threshold value. The simplest way for clustering is to remove all the edges with weight below α . But we propose another approach, which we consider as more applicable.

Iterated Cluster Recognition Algorithm (ICRA):

Input: graph $G=(V,E)$, level of cluster separation α ;

Output: clustering $\{V_1, V_2, \dots, V_k\}$;

Start:

$i=0$

Loop while V is not empty

$i++$

Find an arbitrary cluster C in the graph $G[V]$ induced by V .

$V_i=C; \quad V=V-C;$

$k=i$

End.

Algorithm for finding an arbitrary cluster:

Input: graph $G=(V,E)$;

Output: a cluster $C \subset G$.

Start:

1. Put arbitrary vertex $v \in V$ into queue Q ;

2. Loop while Q is not empty

a. get vertex u from Q ; add u to C ;

b. Loop for each $w \in N(u)$, $w \notin C, Q$

If $\beta(u, w) > \alpha$

put w into Q

End If

End.

The average computational complexity of ICRA is $z^2|E|$, where z is average vertex degree.

3. APPLICATIONS OF THE METHOD

The proposed algorithm was applied for some networks for method ability testing.

3.1 Extended Watts and Strogatz model

The algorithm was developed for SW networks only. In order to test its ability to recognize clusters we need a simple SW model with a known cluster structure. Such model may be obtained by extension of the first and simplest SW model, proposed by Watts and Strogatz [5]. Let us consider a set of M one-dimensional lattices. By redirecting the p_{in} fraction of edges uniformly at random inside the lattice and the p_{out} fraction of edges between the lattices, the extended SW model is obtained. There are exactly M clusters in the model which disappear with p_{out} increasing. The SW property of the model disappears with p_{out} or p_{in} increasing. By simulation, we tested the fraction of vertices $p_{correct}$ which were correctly classified by our method as belonging to clusters. The $p_{correct}$ value is near 100% for p_{out} or p_{in} less than some critical value p_c and it sharply decreases with ongoing randomization. The p_c value is the same for p_{out} and p_{in} , and it increases from approximately 38% to 70% with increasing z from 6 to 16. Other parameters of the model show a weak effect on p_c . It is important to notice that the p_c is the critical value for the clustering coefficient as well, thus the simulations show the strong ability of the method for cluster recognition in the case of SW graph only. Even if there are clusters which may be

recognized by other methods, absence of SW property makes ISRA inapplicable.

3.2 Real world graphs.

Our method was applied for some real world networks as well. In the case of Zachary's Karate Club Study [see 2 for details], the method demonstrated the ability to predict the structure of two known communities. The method was applied for the Web sub-graph. We probed the sites of big organizations (universities, corporations) and found that it detects significant clusters. The method seems to be applicable for data mining. Other networks testing is currently being carried out.

4. CONCLUSION

The proposed method involves a new approach for graph clustering based on the assumption that inside a cluster there exist some local small structures of the edges, which do not exist for inter-cluster edges. A similar approach was used in a recent publication [6], where clusters were detected by percolation of small k -clique graph. This indicates that for real-world networks local link-link correlations have significant influence on cluster formation.

The proposed approach has a set of advantages. Efficiency: has a polynomial complexity $O(|E|)$. Locality: no need to know all the data of the graph or number of clusters for local cluster recognition. Applicability: applicable for graphs of a different nature ("big" and "small", power-law, etc.). It helps to recognize Small World sub-graphs within a non-Small World graph.

The method may be extended for different networks by adopting the weighting mechanism. The problem of the clusters overlapping is connected to the choice of the correct thresholds. A big number of existing vertices not belonging to any clusters is a problem as well. Additional efforts in this direction will lead to the developing of a set of methods to solve different practical needs.

5. REFERENCES

- [1] Newman, M. E. J., The structure and function of complex networks, SIAM Review 45, 167-256 (2003).
- [2] Girvan M. and Newman, M. E. J., Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99, 7821-7826 (2002).
- [3] Wilkinson, D. and Huberman, B.A., A method for finding communities of related genes, Preprint, Stanford University (2002).
- [4] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon U., Network motifs: Simple building blocs of complex networks, Science 298,824-827 (2002).
- [5] Watts, D.J. and Strogatz, S.H., Collective dynamics of 'small-world' networks, Nature, 393,440-442 (1998).
- [6] Palla G., Derenyi I., Farkas I. and Vicsek T., Uncovering the overlapping community structure of complex networks in nature and society, Nature, 435, 814-818 (2005).