STATISTICAL CLASSIFICATION OF SOCIAL NETWORKS

Tian Wang[†]*, Hamid Krim*[⋆]

North Carolina State University
[†]Department of Physics
[⋆]Department of Electrical and Computer Engineering

ABSTRACT

This paper proposes a new social network classification method by comparing statistics of their centralities and clustering coefficients. Specifically, the proposed method uses the statistics of Degree Centralities and clustering coefficients of networks as a classification criterion. A theoretical justification to this method is also given. In relation to the widely held belief that a social network graph is solely defined by its degree distribution, the novelty of this paper consists in revealing the strong dependence of social networks on Degree Centralities and clustering coefficients, and in using them as minimal information to classify social networks. In addition, experimental classification demonstrates a very good performance of the proposed method on real social network data, and validates the hypothesis that Degree Centralities and clustering coefficients are the only two viable independent properties of a social network.

Index Terms— Social Network, Network Classification, Pattern Recognition

1. INTRODUCTION

Social networks have been of research interest for a long time. The previous research has mainly focused on individuals and the relationships in a given network[1][2][3][4]. Few works have, however, studied the relationships between different networks.

To date, different properties of networks have been measured and studied. The main ones includes: Degree Centrality, Betweenness Centrality, Closeness Centrality, Eigenvector Centrality and Clustering Coefficients. Each of them is a measure of importance of a node within a network. It was discovered that: their distributions are not arbitrary but rather obey certain laws. For instance, a degree sequence in a social network always follows power law distribution[3]. But the statistical difference between the distributions of centralities and clustering coefficients has yet to be paid close attention to.

In fact, although the degree sequences for most social networks obey a power law distribution, the statistics, i.e. variance, skewness and kurtosis of the distributions vary for different networks. Naturally, one would think that the difference between higher order statistics of the distribution of centralities of different networks is actually distinguishing characteristics between them. One can hence conclude that using the statistical moments of the distribution of centralities and clustering coefficients of the nodes in a network as a classification criterion is both reasonable and effective, as it only requires partial information of the networks.

Meanwhile, not all of the properties of social networks are equally important for their classification. Under the assumption that, in social networks, if two relationships which do not share a common node, are conditionally independent of each other, one can show that the Degree Centralities and the number of triads are the crucial properties to determine its characteristics[5]. Thus, to classify different types of social networks, we only have to compare their Degree Centralities and their number of triads. Given that clustering coefficient is the only property that is closely related to the number of triads in a network, our proposed method is based on the Degree centralities and clustering coefficients to classify social networks.

The classification of different types of networks is of importance in many applications. In criminal networks, the potential use of this technique would be to detect whether different criminal networks belong to a larger network using similar rules to control its members. In terrorist networks, this method could, for instance, be used to detect whether a terrorist network is led by the same leader that has a history of organizing terrorist plots, and hence help identify the leader.

In this paper, we first introduce some fundamental characteristics in social networks along with their definitions. The theoretical analysis about the dependence of social networks on certain properties is then illustrated. We also experimentally establish the relationship between different centralities of social networks. By combining the theoretical analysis and the experiment results together, we formulate a hypothesis which states that information included in Degree Centrality and in clustering coefficients is amply sufficient to fully characterize a social network. We proceed to test this hypothesis by constructing a classification model, which we experimentally apply to real network data for validation. And we finally conclude with some remarks and future planned work at the end of the paper.

2. BACKGROUND

In all previous studies on social networks[3][1][2][4], different types of centralities and coefficients were proposed in order to measure the important characteristics of a network. These include: Degree Centrality, Betweenness Centrality, Closeness Centrality, Eigenvector Centrality and clustering coefficients. We have thus far focused on networks with neutral simple links, i.e. network graph is undirected and the edges are un-weighted. The adjacency matrix of a network is defined as:

$$M_{ij} = \begin{cases} 1, \ if \ nodes \ i \ and \ j \ are \ connected \\ 0, \ if \ nodes \ i \ and \ j \ are \ not \ connected \end{cases} \quad (1)$$

All the network information is contained in the adjacency matrix M.

The Degree centrality describes how many direct neighbors one node has in a network. In social networks, especially a scale free network, the distribution of degree centralities obeys a power law. The definition of degree centrality of node i in a network is:

$$D(i) = \sum_{j=1,j\neq i}^{N} \frac{M_{ij}}{N-1}, \tag{2}$$

where N is the total number of nodes in a network represented as a planar graph. We also define the degree of node i as: $d(i) = \sum_{j=1,j\neq i}^{N} M_{ij}$.

The structure consisting of a node of degree d along with all its d neighbors is called a *d-star*.

Betweenness Centrality describes how important a node is when considering how much information flows through it in a network. The definition of Betweenness Centrality of node i in a network is:

$$B(i) = \sum_{j=1,j\neq k\neq i}^{N} \frac{\sigma_{jk}(i)}{\sigma_{jk}}, \tag{3}$$

where $\sigma_{jk}(i)$ is the number of geodesic paths between node j and node k that goes through node i, and σ_{jk} is the number of geodesic paths between nodes j and k.

Closeness centrality describes how a node could pass information to the other nodes. The definition of Closeness centrality of node i in a network is:

$$C(i) = \sum_{j=1,j\neq i}^{N} \frac{d_{ij}}{N-1}, \tag{4}$$

where d_{ij} is geodesic distance between nodes i and j.

Eigenvector Centrality not only counts the number of links one node has, but also counts how important is the node it connects to. The definition of Eigenvector Centrality of node i in a network is:

$$E(i) = \frac{1}{\lambda} \sum_{j=1,j\neq i}^{N} E(i)M_{ij}, \tag{5}$$

where $E(i)$ is Eigenvector Centrality for node i and λ corresponds to the largest eigenvalue of the adjacency matrix M. And $E(i)$ can be solved from the definition above.

A clustering coefficient measures how structured the neighborhood of a node is in a network. The definition of Closeness centrality of node i in a network is:

$$CC(i) = \frac{2|\{e_{jk}\}|}{d(i)(d(i)-1)}, \tag{6}$$

where set $\{e_jk\}$ contains all the links existing between the neighbors that node i immediately connects to. For each e_jk, where nodes j and k are neighbors of node i, there is a triad formed by edges e_jk,e_ik and e_ij. $|\{e_{jk}\}|$ thus represents the number of triads that includes node i.

3. REVIEW OF MARKOV GRAPH[5]

A social network described by its adjacent matrix M_{ij} can be considered as a sequence of random variables which are the elements of the adjacent matrix. Furthermore, we can also think of it as a random Markov field which has an underlying dependence structure describing the conditional dependence between adjacent matrix elements. Such dependence structure is called a dependence graph $D = \{node_D, edge_D\}$ for the social network $M = \{node_M, edge_M\}$. The nodes in graph D are all pairs of people in M: $node_D = \{\{i,j\}_{i\neq j}\}$; thus there are $N(N-1)/2$ nodes for graph D. Suppose there are four nodes in social network M: i, j, k, l. Then relationship $\{i, j\}$ and relationship $\{k, l\}$ correspond to two nodes in dependence graph D. And these two nodes in D are linked if and only if relationship $\{i, j\}$ and relationship $\{k, l\}$ conditionally depend on each other.

According to *Hammersley-Clifford theorem*, the probability of a general network to show up is:

$$P(G) = z^{-1} exp[\sum_{c\subseteq G} \alpha_c], \tag{7}$$

where z is the partition that normalizes $P(G)$ and α_c is a constant corresponding to a clique c in $\{D\}$.

It is reasonable to state that, in social networks, two relationships which do not share a common node are conditionally independent of each other. This effectively requires $edge_D$ contain no links between any two elements $\{i, j\}$,$\{k, l\}$ that are disjoint edges in M. In this case, the cliques in D only correspond to triads and stars in M and M is called a *Markov Graph*. In addition, as the social network is homogeneous, the probability $P(G)$ should be invariant to the permutation of the indices of the nodes in M. In light of the above properties, we have:

Theorem 3.1 *Any homogeneous undirected* Markov graph *has probability:*

$$P(G) = z^{-1} exp[\tau t + \sum_{k=1}^{N-1} \delta_k s_k], \tag{8}$$

where t is number of triangles in network G and s_k is the number of *k-star* in network G; τ and δ_k are arbitrary constant corresponding to them. For details of the proofs the reader is referred to *Frank* and *Strauss*[5][6].

We can also rewrite this probability as:

$$P(G) = z^{-1} exp[\tau t + \sum_{i=1}^{N} \theta_i d(i)], \tag{9}$$

where $d(i)$ is the degree for node i and $theta_i$ is the constant corresponding to it. In light of the earlier definitions of Degree Centralities and clustering coefficients, these are the only directly related characteristics of the network. Accordingly, we readily conclude that the only two crucial properties to describe a social network are degree centralities and clustering coefficients.

4. RELATIONS BETWEEN CENTRALITIES AND CLUSTERING COEFFICIENTS

As degree centralities of nodes in a social network always obey a power law distribution, we proceed to test how the other centralities behave when degree centralities are kept unchanged and obey a power law distribution. By applying the *Molloy-Reed (M-R)* algorithm[7][8], we can uniformly sample a graph which obeys a certain graphical degree sequence, coming from a power law distribution. A sequence of integers $\{d(i)\}$ is graphical if for $1 \leq k \leq N-1$:

$$\sum_{i=0}^{k} d(i) \leq k(k+1) + \sum_{i=k+1}^{N} min\{k, d(i)\}, \tag{10}$$

where *Erdös-Gallai theorem*[9] guarantees that once a sequence of integers obeys the above condition, there exists at least one graph whose degree sequence is exactly the same as the given sequence of integers. We randomly choose a graphical degree sequence that obeys power law distribution, as shown in $Figure1$.

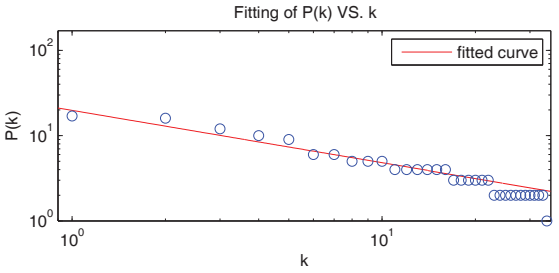

Fig. 1. A graphical degree sequence obeying power law distribution.

Then according to this degree sequence, we uniformly sample 10^5 sub-graphs that obey this degree and observe how the other centralities and clustering coefficients vary when the degree sequence is kept invariant. The result of the experiment is shown in $Figure2$.

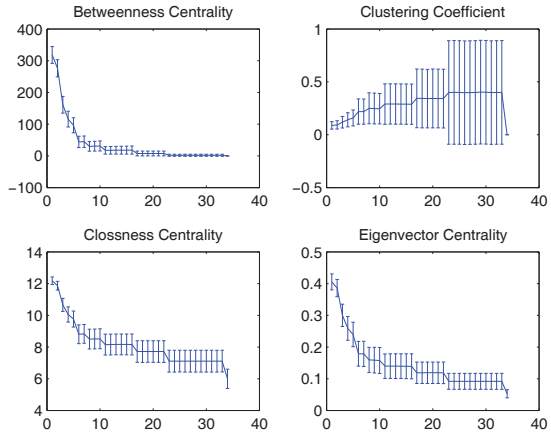

Fig. 2. Behavior of Betweenness Centrality, Closeness Centrality, Eigenvector Centrality and Clustering Coefficients.

We repeat the experiment demonstrated above 10^2 times. For each time, we randomly select a degree sequence that obeys a power law distribution. The relative average standard derivation comparing to the mean of the centralities and cluster coefficients are shown in Table 1.

Table 1. STD V.S Mean of Network Properties.

Network Properties	STD/Mean
Betweenness Centrality	0.9101
Closeness Centrality	0.4705
Eigenvector Centrality	0.9178
Clustering Coefficients	2.6701

According to these results, we can see that for the same degree sequence, and for a large number of generated graphs, the variance of Betweenness Centrality, Closeness Centrality and Eigenvector Centrality between graphs remain small compared to that of clustering coefficients. This indicates that when the distribution of Degree Centralities obeys a power law, the other centralities are largely determined. So they cannot provide any more information about the characteristics of the network than degree centrality itself. However, clustering coefficients vary widely from graph to graph, indicating that different information is provided by degree centrality and the related centralities.

5. HYPOTHESIS AND CLASSIFICATION MODEL

Based on the theoretical analysis in *Section 2* and the experimental results in *Section 3*, we propose the followings:

Hypothesis 1:

The characteristics of a network are determined by its clustering coefficient and its degree centrality.

In the course of testing this hypothesis, we build a model that can be used to classify different types of networks. The model is as follows:

The characteristics of a network are described by a set of indexed features:

$$A_1, A_2, A_3 \ldots,$$

where A_i's are functions of the set of centralities and clustering coefficients of the network:

$$D, B, C, E, CC.$$

According to *hypothesis 1*, $\{A_i\}$ may be chosen to just be a function of Degree Centralities and clustering coefficients of a social network. And due to the fact that social networks are most likely scale-free, i.e. the degree sequence obeys a power law distribution, we can just use the statistics of a degree sequence to represent the information containing in degree centralities of the network. Although the distribution of clustering coefficients is unknown to us, using its statistics up to fourth order should also be sufficient to present most of the statistical information of the clustering coefficients. To that end, the set $\{A_i\}$ is chosen to be:
$\{A_i\} = \{mean(D), var(D), skewness(D), kurtosis(D), mean(CC), var(CC), skewness(CC), knotisos(CC)\}$

By comparing the different sets $\{A_i\}$ of different social networks, we can easily establish the class similarity of different networks.

6. NETWORK CLASSIFICATION

The data we use in our experiment is 800 sub-networks sampled from two giant networks: 1. a snap shot of the Internet at the level of autonomous systems measured by Mark Newman from data in July 22, 2006[10]. 2. a weighted network of co-authorships between scientists posting preprints on High-Energy Theory E-Print Archive between Jan 1, 1995 and December 31, 1999[10]. When the sub-networks are sampled from the above two giant networks, all links between nodes are set to be un-weighted and natural, and the size of the sub-network is set to be from 50 nodes to 200 nodes. There are 400 sub-networks for each type of the networks, 60% of which are treated as a learning data base, and 40% are used to test the classification model. The mean of vector $\{A_i\}$ is calculated for the data base, which is called $\{\bar{A}_i\}_{INT}$ for the sub-networks sampled from internet snap shot and $\{\bar{A}_i\}_{HEP}$ for the sub-networks sampled from the network of co-authorships among the scientists posting on High-Energy Theory E-Print Archive. Then, the normalized distance from the $\{A_i\}$ of the tested sub-network to $\{\bar{A}_i\}_j$ is calculated as:

$$ND_j = \sum_i \frac{|A_i - \bar{A}_{ij}|}{\bar{A}_{ij}}, \tag{11}$$

where $j = INT$ or HEP.

In order to show that the performance is best when $\{A_i\}$ is the statistics of Degree Centralities and clustering coefficients, $\{A_i\}$ is chosen to be in five cases:

1. Statistics of Degree Centralities and Clustering coefficients.
2. Statistics of Degree Centralities.
3. Statistics of clustering coefficients.
4. Statistics of Betweenness Centralities.
5. Statistics of Closeness Centralities.
6. Statistics of Eigenvector Centralities.

Experimental results of cases 1, 2 and 3 are shown in *Figure 3* and experimental results of cases 4, 5 and 6 are shown in *Figure 4*.

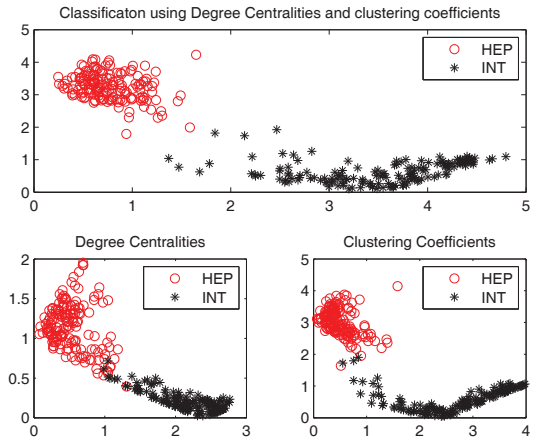

Fig. 3. Classification result of case 1, 2 and 3. y axis represents ND_{HEP}, and x axis represents ND_{INT}.

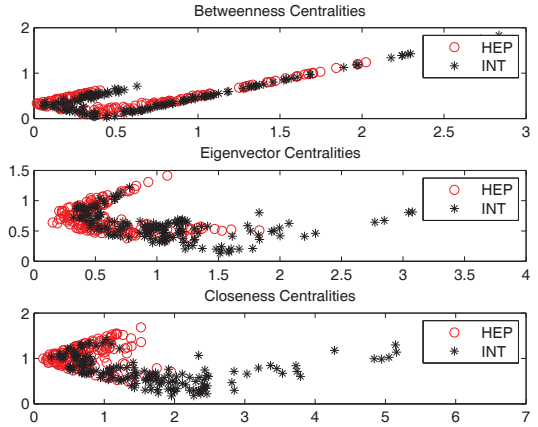

Fig. 4. Classification result of case 4, 5 and 6. y axis represents ND_{HEP}, and x axis represents ND_{INT}

From the experimental results, we can see that by using the statistics of Degree Centralities and clustering coefficients, the performance of classification model is very good. This is in contrast to only using degree centralities or clustering coefficients, where the results are difficult to evaluate. In fact, when we use any of Betweenness Centralities, Eigenvector Centralities or Closeness Centralities along, the performance of the classification model dramatically drops.

7. CONCLUSION AND FUTURE WORK

In this paper, we have theoretically shown, with experimental validations, that Degree Centralities and clustering coefficients are two fundamental statistics for classifying social networks. Based on this hypothesis, we further proposed a social network model which can classify different types of social networks. Both the hypothesis and the classification model are supported by experimental results. In the future work, weighted links and directed network will be considered in order to improve the performance of the classification. Temporal change of the centralities and clustering coefficients will also be added to the model.

8. REFERENCES

[1] Stanley Wasserman and Philippa Patrlson, "Logit models and logistic regressions for social networks," *PSYCHOMETRIKA*, vol. 61, pp. 401–425, September 1996.

[2] Stanley Wasserman and Katherine Faust, *SOCIAL NETWORK ANALYSIS: METHOD AND APPLICATIONS*, Cambridge University Press, 1994.

[3] Rka Albert and Albert Barabsi, "Statistical mechanics of complex networks," *REVIEWS OF MODERN PHYSICS*, vol. 74, JANUARY 2002.

[4] Juyong Park and M. E. J. Newman, "Statistical mechanics of networks," *PHYSICAL REVIEW E*, vol. 70, 2004.

[5] Ove Frank and David Strauss, "Markov graphs," *Journal of the American Statistical Association*, vol. 81, pp. 832–842, Sep 1986.

[6] Tian Wang and Hamid Krim, "Application of markov graph on social network classification," in preparation.

[7] Michael Molloy and Bruce Reed, "A critical point for random graphs with a given degree sequence," *Random Structures and Algorithms*, vol. 6, pp. 161–179, 1995.

[8] Milena Mihail and Nisheeth K. Vishnoi, "On generating graphs with prescribed vertex degrees for complex network modeling," *ARACNE*, pp. 1–11, 2002.

[9] S Choudum, "A simple proof of the erdos-gallai theorem on graph sequences," *BULL. AUSTRAL. MATH. SOC.*, vol. 33, pp. 67–70, 1986.

[10] M. E. J. Newman, "Proc. natl. acad. sci. usa 98, 404-409," http://www-personal.umich.edu/ mejn/netdata/, 2001.