2018 International Conference on Computational, Modeling, Simulation and Mathematical Statistics (CMSMS 2018)
ISBN: 978-1-60595-562-9

Correlation Research of Centralities on Complex Network by Statistical Learning

Ying SHI, Wei WEI*, Xiang-nan FENG and Zhi-ming ZHENG

LMIB and School of Mathematics and Systems Science, Beihang University,100191, Beijing, China

*Corresponding author

Keywords: Complex network, Centrality, Correlation analysis, Statistical learning.

Abstract. In network theory and network analysis, indicators of centrality identify the most important vertices on complex networks. In this paper, we perform analysis on correlations of 13 centralities on ER random network and research how the Radial centralities interpret the Medial centralities adequately by statistical learning approaches such as linear regression, forward- and backward-stepwise selection and lasso. As a result, it is illustrated that some centralities on ER random networks with different connecting probability p always display strong correlations, and the Medial centrality can be interpreted by the Radial centralities. Furthermore, the linear regression is used to fit the relationship and retain some centralities to describe a medial centrality in our example, which will help to solve the problem that a centrality we don't have a ready algorithm and compute difficultly. The methods proposed by statistical learning provide an alternative way to obtain better understanding of the centralities and reveal the relationship among them.

Introduction

The study of random networks started with the influential work of Erdös and Rényi in the 1950s and 1960s. In the study the assumption has been made that the presence or absence of an edge between two vertices is independent with the presence or absence of any other edge, so that each edge may be considered to be present with independent probability p. If there are N vertices in a network, and each is connected to an average of z edges, it is easy to show that $p=z/(N-1)$, which for lager N is usually approximated by z/N[1]. The number of edges connected to any particular vertex is called the degree k of that vertex, and has a probability distribution p_k given by

$$p_k = \binom{N}{k} p^k (1-p)^{N-k} \approx \frac{z^k e^{-z}}{k!}, \tag{1}$$

where the second equality becomes exact in the limit of lager N.

To view a complex network, a direct way is to identify the most influential nodes and many centralities are proposed to locate important ones. Degree centrality of a node v is the fraction of nodes it is connected to. Closeness centrality of a node u is the reciprocal of the average shortest path distance to u over n-1 reachable nodes. Betweenness centrality of a node v is the sum of the fraction of all-pairs shortest paths that pass through v. Eigenvector centrality computes the centrality for a node based on the centrality of its neighbors. PageRank computes a ranking of the nodes in the network G based on the structure of the incoming links. It was originally designed as an algorithm to rank web pages. Till now, there are hundreds of ways to define the centralities.

The centrality can be classified into Radial or Medial classes according to its construction. Centralities are Radial centralities counting walks which start/end from the given vertex[2][3]. The degree and eigenvector centralities are examples of radial centralities. Medial centralities count walks which pass through the given vertex. The canonical example is betweenness centrality, counting the number of shortest paths which pass through the given vertex.

Model Description

The goal is to provide an adequate and interpretable description of how the Radial centralities affect the Medial centralities. In this paper we use linear methods of regression, including linear regression, forward- and backward-stepwise selection and the lasso to analyze the correlations between the betweenness centrality and other 12 centralities, which are degree centrality, closeness centrality, katz centrality, eigenvector centrality, pagerank, *k*-shell decomposition[4], *k*-core, eccentricity, the constraint[13], information[14], current-flowbetweenness[15] and subgraph centrality[16].

The data for the following example, shown in Figure 1, is from 100 ER random networks to study the correlation for the 13 centralities. There are 500 vertices in a network, and each is connected to an average of 20 edges (connecting probability p=0.04).

Every centrality provides a ranking which identifies the most important nodes. There are 13 centralities for every network, that is, 13 rankings. We need to describe one ranking by a distance between two ranks. Based on Spearman's rank correlation coefficient, the distance can be defined as:

$$x_{ip} = \sum_{i=1}^{n} |d_i| \times w_i,\ p = 1,2,...,12, \tag{2}$$

where n is the node number of a network, p is the different centralities. On each network, all the vertices are assigned a centrality value w_i and an order is achieved which is denoted by o_i. Sort o_i by the centrality values using the ranked values 1,2,3,...,n and create a new rank $r_i = i$ (i=1,2,…,500), and we use d_i to reflect the difference between the two ranks o_i and r_i, which can be viewed as a rearranging cost.

Correlation Analysis

Denote X as the $N\times(p+1)$ matrix with each row an input vector, and similarly y be the N-vector of outputs. Given a vector inputs $X^T=(x_1,x_2,...,x_n)$, the aim is to bulid the relations between the inputs X and the output y. Each $x_i=(x_{i1},x_{i2},...,x_{ip})^T$ is a vector of 12 centralities for the i_{th} sample/network. In the following example, the betweenness centrality distance is considered to be the output y.

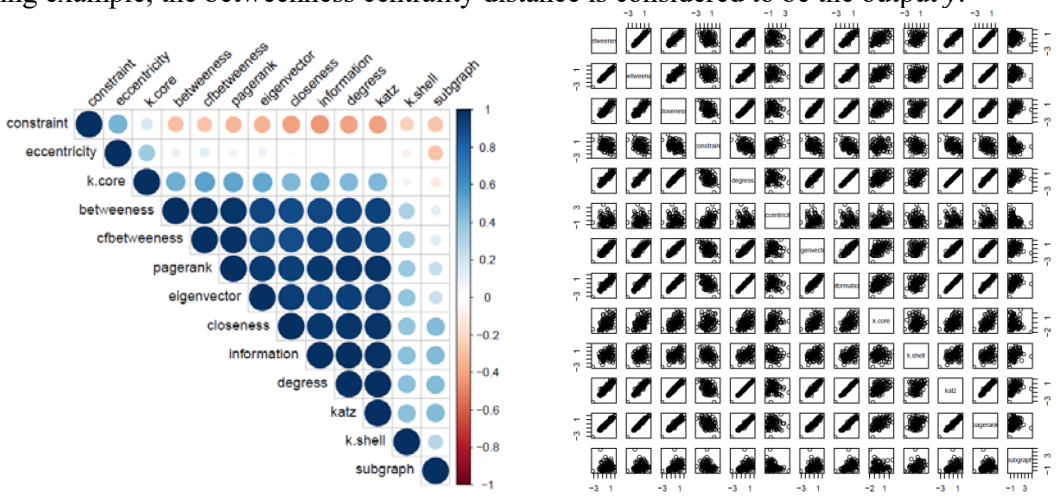

Figure 1. (Above) A correlation plot matrix of the 13 centralities data. The deeper the blue color of the circle is, the stronger a relationship about a pair of centralities is. (below) A scatter plot matrix of the 13 centralities data. Each plot shows a pair of centralities. The 100 network samples are ER ones with average degree 4 and the node number n is equal to 500.

The correlation matrix of the inputs given in Figure 1 (above) shows that there are many strong correlations between different centralities. Figure 1 (below) is a scatter plot matrix showing every pairwise plot between the variables. We see, for example, that the centralities betweenness, cfbetweenness, pagerank, eigenvector, closeness, information, degree and katz have strong

relationship with each other (Figure 1 (above)), and degree and katz show a strong linear relationship (Figure 1 (below))! But a good predictive model is difficult to construct.

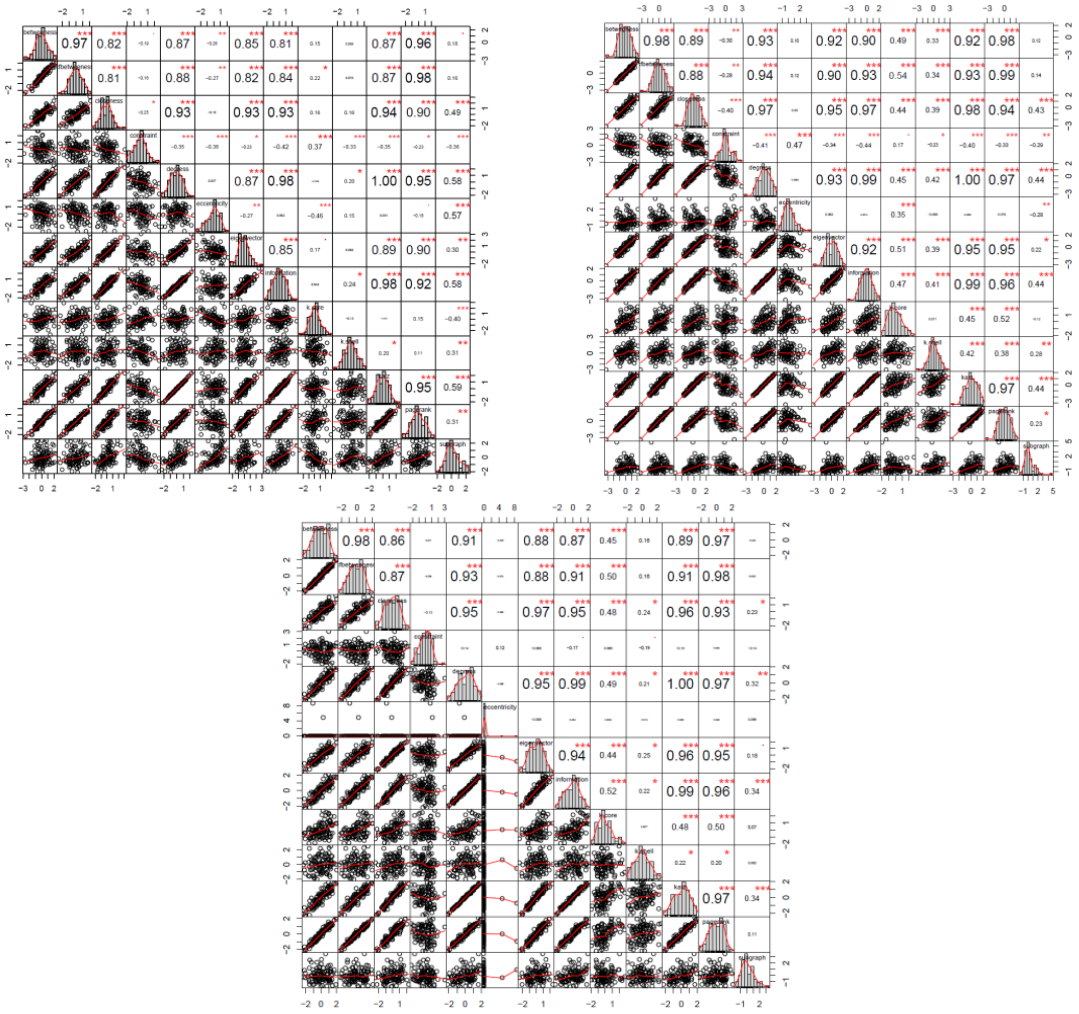

Figure 2. A correlation plot matrix of the 13 centralities data, coming from 100 ER random networks for three different connecting probability p (p=0.03,0.04,0.05) from top panel to bottom panel. Included are the correlation coefficient of centralities with each other. The diagonal shows the centrality values distribution.

For the three different values of probability p, the networks display similar correlations of the 13 centralities but a little differences:

***p*=0.03**: (1) The correlation coefficient of degree and katz centrality is 1, so there is a linear relationship. There are three centralities information, pagerank and betweenness which respectively have the same correlation coefficient (0.98,0.95 and 0.87) with degree and katz. (2) The pagerank, katz, betweenness, cfbetweeness, degree, information have a strong correlation with the correlation coefficient larger than 0.90. The closeness has the same correlation coefficient (0.93) with degree, eigenvector and information. (3) In contrast, the five centralities constraint, eccentricity, *k*-core, *k*-shell and subgraph have a weak relationship with other centralities.

***p*=0.04**: (1) The correlation coefficient of degree and katz centrality is also 1, so there is still a linear relationship. There are two centralities information and pagerank which respectively have the same high correlation coefficient (0.99 and 0.97) with degree and katz. (2) The pagerank, katz, betweenness, cfbetweeness, degree, information, eigenvector, closeness have a strong correlation with the correlation coefficient not less than 0.90. The closeness has the same correlation coefficient (0.97) with degree and information. The (3) is the same as p=0.03.

***p*=0.05**: There are three similar properties as p=0.04. Only the correlation coefficients have ±0.03 differences.

The above three networks with different connecting probability p draw the common conclusion about the correlation of the 13 centralities : (1) The correlation coefficient of degree and katz centrality is always 1, so there should be a linear relationship between them. (2) The eight centralities pagerank, katz, betweenness, cfbetweeness, degree, information, eigenvector, closeness have a strong correlation. (3) In contrast, the five centralities constraint, eccentricity, *k*-core, *k*-shell, subgraph have a weak relationship with other centralities. We need to fit the effects jointly to untangle the relationships between the inputs and the outputs.

Statistical Learning Approaches on Centralities

We fit a linear model to the betweenness, in which the inputs are scaled to have unit variance. The least square estimation is applied to the dataset, and the estimation values, standard errors and Z-score are shown in Table 1. The Z-scores measure the effect of dropping the corresponding variable from the model. A Z-score greater than 2 in absolute value is approximately significant at the 5% level[6]. The output is the betweenness centrality, which belongs to the Medial centralities[7]. The linear regression model has the form:

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j . \tag{3}$$

Table 1. Linear model fit to the centralities data. Roughly a Z-score lager than two in absolute value is significantly nonzero at the 0.05 level.

	Coefficient	Std.Error	Z score
(Intercept)	0.0002308	0.0065953	0.035
cfbetweeness	0.9277653	0.1994278	4.652***
closeness	0.3377421	0.0468863	7.203***
constraint	-0.1086307	0.0129228	-8.406***
degree	-0.8591810	0.6927479	-1.240
eccentricity	0.0074802	0.0087399	0.856
eigenvector	-0.3208954	0.1049362	-3.058**
information	-1.3244697	0.0689821	-19.200***
k-core	0.0196397	0.0104180	1.885.
k-shell	-0.0063838	0.0078514	-0.813
katz	1.8661590	0.7071784	2.639**
pagerank	0.3214727	0.3638269	0.884
subgraph	-0.0519436	0.0273320	-1.900.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The inputs cfbetweeness, closeness and information show the strongest effect, with eigenvector and katz also strong. The R-squared of this linear model fit is 0.9962, which means that the inputs can interpret the output by 99.62% and the fit effect of the formula is wonderful.

Rather than search through all possible subsets, we can seek a good path to detect the key variables (centralities) affecting the output, which is forward- and backward-stepwise selection[11]. Some software package implements hybrid stepwise-selection strategies that consider both forward and backward stepwise moves at each step, and select the "best" of the two[9]. For example in the R package the step function uses the AIC criterion for weighing the choices, which takes proper account of the number of parameters fit; the sum of square, RSS and AIC are shown in TABLE 2. At each step an adding or dropping will be performed that minimizes the AIC score[5].

Hybrid stepwise-selection chooses to use the 8 inputs cfbetweeness, closeness, constraint, eigenvector, information, *k*-core, katz and subgraph. The AIC score is minimized to -535.03.

Table 2. Hybrid stepwise-selection uses the AIC criterion for weighing the inputs.

	Sum of Sq	RSS	AIC
<none>		0.39649	-535.03
+ degress	0.00737	0.38912	-534.90
+ *k*-shell	0.00642	0.39007	-534.66
+ eccentricity	0.00309	0.39340	-533.81
+ pagerank	0.00281	0.39369	-533.74
- *k*-core	0.02054	0.41703	-531.98
- subgraph	0.02989	0.42638	-529.76
- eigenvector	0.08878	0.48527	-516.82
- closeness	0.37116	0.76765	-470.96
- constraint	0.37880	0.77529	-469.97
- katz	0.43915	0.83564	-462.47
- information	1.87503	2.27152	-362.47
- cfbetweeness	1.88719	2.28368	-361.94

The lasso works by constraining the sum of the absolute values of standardized estimated coefficients to some constant *t*, in math: $\sum_{j=1}^{p} |\beta_j| \le t$.

If *t* is chosen to be too small, the model may not capture important characteristics of the data; if *t* is chosen to be too large, the mode may over-fit the data in the sample, providing an inaccurate representation for the results[10]. Making *t* sufficiently small will cause some of the coefficients to be exactly zero[8].

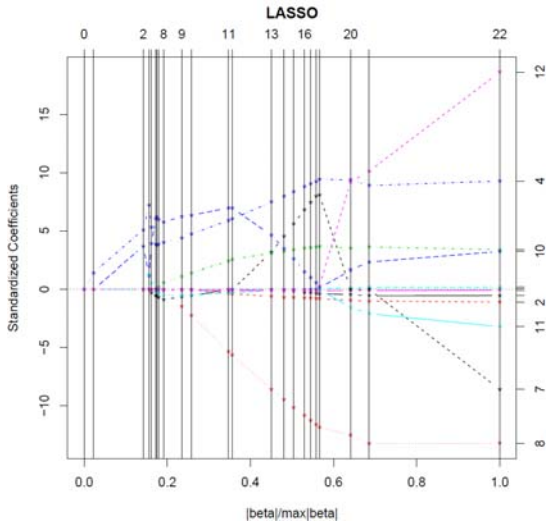

Figure 3. Profiles of lasso coefficients, as the tuning parameter is varied. Coefficients are plotted versus. The profiles are piece-wise linear, and so are computed only at the points displayed.

The computational difficulty with the lasso has been solved. Least angle regression (LAR) provides an extremely efficient algorithm for computing the entire lasso path as Figure 3. The curves produced by R for each variable are in different colors. The vertical lines show each time a variable is added to the model[12]. On the far right the numbers of variables are shown[8]. Finally, the lasso retains 8 centralities to interpret the output and sets 4 centralities closeness, information, page rank and subgraph to be zero.

Summary

In this paper some correlation matrices are studied to understand the relationship between different centralities. By retaining a subset of the centralities and discarding the rest, the betweenness centrality has an interpretable description. We discuss the three approaches for picking the parsimonious inputs to interpret the output as much as possible. The least squares estimation of linear regression chooses the significant variables by measuring Z-score; the hybrid stepwise-

selection adds or drops some variables by minimizing the AIC score; the lasso sets some variables to be zero by imposing a sufficient penalty. Table 3 shows the coefficients for the three different methods. However, the further research is necessary for how to choose the best methods based on the bias-variance tradeoff, offering some hope that the theory will prove useful once more complete data becomes available.

Table 3. Coefficients for three different methods applied to the centralities data. The blank entries correspond to variables omitted.

	LS	Hybrid Stepwise	Lasso
(Intercept)	0.0002308	0.0002616	
cfbetweeness	0.9277653	1.0210889	-0.053867023
closeness	0.3377421	0.3787196	
constraint	-0.1086307	-0.0975746	-0.004068718
degree	-0.8591810		0.470646707
eccentricity	0.0074802		-0.048692188
eigenvector	-0.3208954	-0.1865454	-0.010421400
information	-1.3244697	-1.3460385	
k-core	0.0196397	0.0223194	-0.224414907
k-shell	-0.0063838		0.141985762
katz	1.8661590	1.0935594	0.633926897
pagerank	0.3214727		
subgraph	-0.0519436	-0.0628975	

Acknowledgment

This work is supported by the Fundamental Research Funds for the Central Universities, the National Natural Science Foundation of China (No.11201019), the International Cooperation Project No.2010DFR00700 and Fundamental Research of Civil Aircraft No.MJ-F-2012-04.

References

[1] M.E.J. Newman, S.H. Strongatz and D.J. Watts. Random graphs with arbitrary degree distributions and their applications. Physical Review E, Volume 64, 026118(2001).

[2] Borgatti, Stephen P. and Everett, Martin G. A Graph-theoretic perspective on centrality. Social Networks, Volume 28, pp. 466-484.

[3] Brandes, Ulrik. A faster alogoritm for betweenness centralities. Journal of Mathematical Sociology, 2001, pp. 163-177.

[4] Shai Carmi, Shlomo Havlin, Scott Kirkpatrick, Yuval Shavitt and Eran Shir. A model of Internet topology using k-shell decomposition. PNAS, 2007.

[5] T. Hastie, R. Tibshirani and J. Friedman. The Elements of Statistical Learning. Springer, New York, 2001.

[6] Bishop, C. Pattern Recognition and Machine Learning, Springer, New York, 2006.

[7] J.H. Friedman, T. Hastie and R. Tibshirani. Regularzation paths for generalized linear models via coordinate descent. Technical report, Stanford University, 2008.

[8] D. Wright and K. London. Modern Regression Techniques Using R: a practical guide for students and researchers. SAGE, 2009.

[9] M. Yuan and Y. Lin. Model selection and estimate in regression with grouped variances, Journal of Royal Statistical Society, Series B, 2007, pp. 49-67.

[10] P. Zhao and B. Yu. On model selection consistency of lasso, Journal of Machine Learning Research 7: 2541-2563.

[11] T. Hastie, J. Taylor, R. Tibshirani and G. Walther. Forward stagewise regression and the monotone lasso. Electronic Journal of Statistics 1: 1-29.

[12] R. Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58: 267-288.

[13] Burt, Ronald S. Structural holes and good ideas. American Journal of Sociology (110): 349–399, 2004.

[14] M.E.J. Newman. A measure of betweenness centrality based on random walks. Social Networks 27, 39-54, 2005.

[15] Ulrik Brandes and Daniel Fleischer. Centrality Measures Based on Current Flow. Proc. 22nd Symp. Theoretical Aspects of Computer Science. LNCS 3404, pp. 533-544. Springer-Verlag, 2005.

[16] Ernesto Estrada, Juan A. Rodriguez-Velazquez. Subgraph centrality in complex networks. Physical Review E 71, 056103, 2005.