A simulator to study the stability of network centrality measures

Orsolya Kardos
University of Szeged Institute of Informatics
P.O. Box 652
H-6701 Szeged, Hungary
kardoso@inf.u-szeged.hu

András London*
University of Szeged Institute of Informatics
P.O. Box 652
H-6701 Szeged, Hungary
Poznań University of Economics, Department of Operations Research
london@inf.u-szeged.hu

Tamás Vinkó
University of Szeged Institute of Informatics
P.O. Box 652
H-6701 Szeged, Hungary
tvinko@inf.u-szeged.hu

ABSTRACT

Measuring nodes' importance in a network and ranking them accordingly is a relevant task regarding many applications. Generally, this measurement is done by a real-valued function that evaluates the nodes, called node centrality measure. Nodes with the largest values by a centrality measure usually give the highest contribution in explaining some structural and functional behavior of the network. The stability of centrality measures against perturbations in the network is of high practical importance, especially in the analysis of real network data that often contains some amount of noise. In this paper, by utilizing a simulator we implemented in R, a formal definition of stability introduced in [13] and various perturbation methods are used to experimentally analyze the stability of some commonly used node centrality measures.

Keywords

Network science, Centrality measures, Stability, R language

1. INTRODUCTION

In a complex network, being social (e.g. Facebook friendship), economical (e.g. international trade), biological (e.g. protein-protein interaction) or technological (e.g. transportation) network, the position of the nodes in the topology of the underlying graph is of central importance. Central nodes in this graph topology often have major impact, whereas peripheral nodes usually have limited effect on the structure and functioning of the network. Thus, identifying the central and most important nodes helps in better understanding the networks from many different perspectives. Node centrality measures are metrics designed to identify these important nodes. However, the importance of a node can be interpreted in many different ways, therefore, depending on the applications, many centrality measures have been developed and effectively applied in various domains [7]. The most commonly used centrality measures are degree [11, 14], closeness [1, 12], eigenvector [2], betweenness [8], PageRank [4] and HITS [10]. Degree centrality measures the importance of a node simply by the number of its neighbors. Closeness centrality shows the average shortest path length from the node to every other node in the network. Eigenvector centrality, and similarly PageRank, of a node is computed (iteratively) as a function of the importance of its neighbors. Betweenness centrality measures the relative number of shortest paths in the network that go through a node.

The stability of centrality measures has often been investigated in an empirical way by comparing the network with one obtained by modifying the original one according to some randomisation procedure [3, 6, 15]. Recently Segarra and Ribeiro gave a formal definition for the stability of centrality measures and proved that degree, closeness and eigenvector centrality are stable whereas betweenness centrality is not [13]. In this work we experimentally investigate the stability of degree and eigenvector centrality measures on various data sets, and under two different perturbation processes. By doing so we introduce our simulation environment which is implemented in R and available online as an interactive tool.

This paper is organized as follows. In Section 2 we will briefly discuss the definition of stability for centrality measures and introduce the main notations used in the paper. In Section 3 we will present a simulation environment written in R and describe the data sets used in our experiments. In Section 4 we will describe the two perturbation processes, discuss our results and draw some succinct conclusions.

2. NODE CENTRALITY AND STABILITY

Let us consider a network represented by a graph $G = (V, E)$, where V is the set of nodes and E is the set of edges (i.e. pairs of nodes) of the network. Centrality measure is a real-valued function $C^G : V_G \to \mathbb{R}_{\geq 0}$, that assigns a non-negative number to each node of network G. Here we will

*The author was partially supported by the National Research, Development and Innovation Office - NKFIH, SNN-117879

not give the formal definitions of the investigated centrality measures that can be found e.g. in [7]. We use the definition of stability introduced in [13] as follows. A node centrality measure C is said to be stable if

$$|C^G(x) - C^H(x)| \leq K_G \cdot d(G, H) \quad (1)$$

holds for every node $x \in V$, where G and H are two graphs over the same node set V, K_G is a constant, and $d(\cdot, \cdot)$ is a distance function between two graphs.

The definition says that a centrality measure is stable if the maximum change in node centrality is bounded by a constant times the distance of the two graphs. This constant value must be universal to any perturbed version of the initial graph. Furthermore, the constant value does not depend on the presence of normalization of centrality values. Note that the definition is similar to the definition of Lipschitz-continuity, applied in a discrete space. In order to make the above inequality meaningful a graph distance $d : G \times H \to \mathbb{R}_{\geq 0}$ should be specified. Here, the distance of two graphs with identical node set V is defined as

$$d(G, H) = \sum_{i,j} |A^G_{ij} - A^H_{ij}|,$$

where A denotes the (weighted) adjacency matrix of the network.

It is of empirical interest to study how graph H occurs from a given graph G and how it affects the constant K_G in formula (1). In Section 3 different graph perturbation methods using various input graphs and data sets in order to examine the ranges of K_G are discussed.

2.1 Theoretical values in stability concepts

Segarra and Ribeiro showed that using the stability concept (1) the degree, closeness and eigenvector centrality measures are stable, whereas betweenness centrality is not [13]. The theoretical K_G values for the three stable measures were determined. Given a directed and weighted graph G, $K_G = 1$ for degree centrality. This is because the distance of the two adjacency matrices will be at least the maximum difference of the degree centrality value. Furthermore this theoretical value for undirected weighted graphs can be reduced to $1/2$ due to the symmetry of the adjacency matrices. For closeness centrality it was proved that the theoretical bound K_G is equal to the number of nodes, hence it is not a universal constant. The eigenvector centrality is stable and the constant K_G can be computed as $4/(\lambda_1 - \lambda_2)$, where λ_1 and λ_2 are the greatest and second greatest eigenvalue of the adjacency matrix of graph G, respectively.

Although there exist some theoretical results for the constant K_G, it could still be interesting to analyze its actual value in real networks under natural perturbation scenarios. In the next section we describe our simulation environment and data sets used for experimental analysis.

3. SIMULATION ENVIRONMENT

R is an open-source programming language developed by the R Foundation and can be widely used for statistical computations and representations. The functions which are mainly used in our project for graph manipulation and related computations, generating synthetic graphs and graph visualization are part of the `igraph` package. We also use the `plotly` library which is an online analytical and data visualization tool. It can be easily integrated in various developer environments, thus combined with R can be widely used for data visualization.

With the help of these tools we designed and implemented a versatile simulation environment that we use to perform our experiments. The simulator can handle various network data structures, while the output of a simulation can be various plots, data tables, statistics depending on the user defined parameters. A version of the simulator with limited functionality that uses the data as input as discussed below is available online at:

`https://kardosorsi.shinyapps.io/stability`

The interested readers are cordially invited to visit our website and try out different experiments. The full version of the simulator is available upon request.

3.1 Data sets

We have performed a wide-range of experiments on various synthetic and real data sets using different types of perturbations. In the following two experiments are elaborated in more detail.

S&P 500

Firstly, a correlation based financial graph was used. The main motivation behind using stock data was to obtain the perturbation method directly from real-life processes. The experiments were performed using the daily closing prices of stocks of the S&P 500 in the period of 01/01/1995 – 31/12/2018, including the assets of 330 leading U.S. companies[1]. We used a time-window of 200 days to construct correlation matrices from stock return time series on that interval with starting points $T_0 = 01/01/1995$, $T_k = T_0 + k\Delta T$ with $\Delta T = 50$, $k = 1, 2, \ldots$. This way we obtained 116 consecutive networks, with the fixed set of 330 nodes and weighted edges represent the correlation coefficient of each pair of assets on the corresponding time interval. Here, the changes in edge weights between each consecutive network pairs simulates the perturbation process.

Cooper-Frieze graph process

Secondly, we implemented the Cooper-Frieze graph evolution process based on a general model of web graphs proposed in [5]. That is a general model of a random graph process to generate a graph of power-law degree distribution as follows. Starting from an initial graph G_0 at time $t = 0$, the process evolves randomly by the addition of new nodes and/or edges at each time step $t = 1, 2, \ldots$. The following six parameters of the process provide a high-level of freedom in graph generation. With probability $\alpha \in [0, 1]$ and $1 - \alpha$ a new node is created or an existing node generates edges, respectively. With probability $p = (p_i : i \geq 1)$ a new node generates i edges. For new nodes, with probability $\beta \in [0, 1]$ the terminal node of a new edge is made uniformly at random and with $1 - \beta$ according to degree (i.e. new edges are

[1]We selected those assets from the $S\&P$ 500 list that were complete in the considered time period.

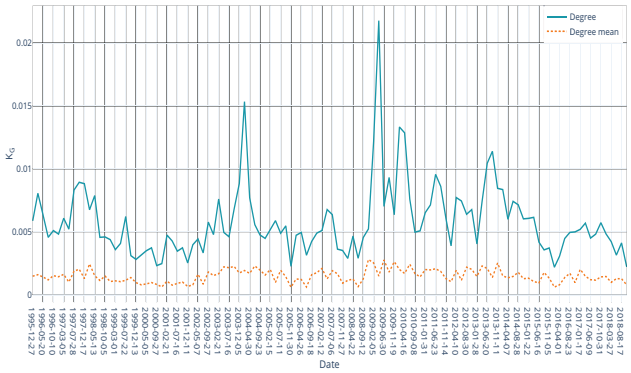

Figure 1: K_G constant values for degree centrality measure during the perturbation simulation.

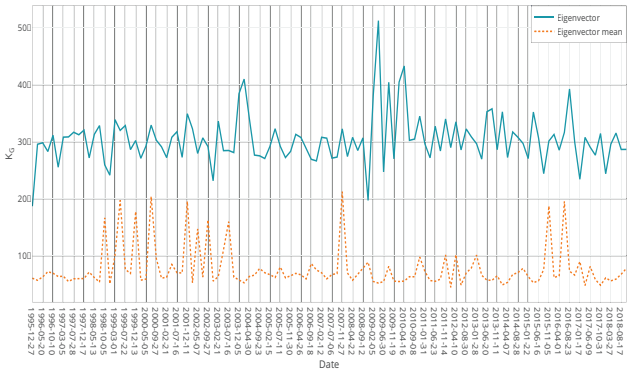

Figure 2: K_G constant values for eigenvector centrality measure during the perturbation simulation.

preferentially attached). If an already existing node generates an edge, where the number of edges given by probability $q = (q_i : i \geq 1)$, the initial node is selected uniformly with probability δ and according to the degree with $1 - \delta$. The parameter γ has similar role for existing nodes as β had in the case of new nodes. Using this process we are able to simulate a graph perturbation process. The initial graph (at time $t = 0$) can be set as an input parameter and then in every time step $t = 1, 2, \ldots$ a new (perturbed) graph is created by the evolutionary graph process.

4. RESULTS AND DISCUSSION

Two main perturbation categories are examined. The first category is the *graph structure perturbation* that can be raised from real-life data (like stock correlations) or synthetic perturbation obtained by rewiring edges selected uniformly at random. The other group is raised from *graph evolution*. Here we will present our experiments on the $S\&P$ 500 data set for the structure perturbation and on Cooper-Frieze networks for the graph evolution. Results are shown on consecutive graphs as discussed in Section 3. During our experiments reported here the degree and eigenvector centrality measures were considered[2].

Graph structure perturbation. At the global level, an interesting result is provided by the behavior of the K_G constant value regarding both degree and eigenvector centrality measures over time, see Figure 1 and Figure 2, respectively. The mean values for centrality C are calculated as

$$\frac{1}{|V|} \sum_{x \in V} |C^G(x) - C^H(x)|. \quad (2)$$

We can observe that for both centrality measures are very stable, only very slight changes in their values are observed. Interestingly, these changes happen in periods of crisis. The increases around 2004, 2007-2008 and 2010-2011 can be noticed. The 2007-2008 period can be associated with the Lehman Brothers failure, whereas the 2010-2011 may reflect the Sovereign debt crisis. It is a well-known stylized fact in finance that assets correlation increases in times of financial distress. Note that these actual K_G values are way lower than their theoretical bounds.

[2]Note that a more detailed presentation of our results will be part of a paper planned to be published later.

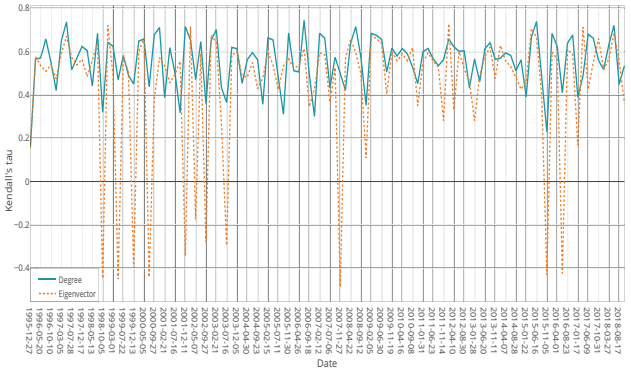

Figure 3: Kendall's tau coefficient between rank by different centrality measures during perturbation

The other interesting aspect in analyzing the stability of the different network centrality measures is the order or ranking provided by the metrics. The Kendall rank correlation coefficient [9] is used to measure the ordinal association between two measured quantities. The coefficient results in high value when observations have a similar rank (i.e. relative position label of the observations within the variable: 1st, 2nd, 3rd, etc.) between the two variables. The simulator can be parametrized in order to visualize the correlation between the order by centrality measures for the different measures respectively. Therefore it is possible to analyze the correlation between the two rank vectors during the graph perturbation procedure. On Figure 3 the Kendall correlation coefficients are reported. The degree centrality stays quite stable in the range of $0.35 - 0.7$, whereas the eigenvector centrality shows some seemingly radical changes over time. We can observe that these extreme changes in ranking shown on Figure 3 are related to the higher K_G constant values regarding the average change in centrality measures presented on Figure 2.

Graph evolution. The other aspect that we wanted to study in our experiments was the graph perturbation caused by some evolutionary process. The concept behind this was that studying the maximum of the difference in centrality measures during graph evolution can be an interesting ap-

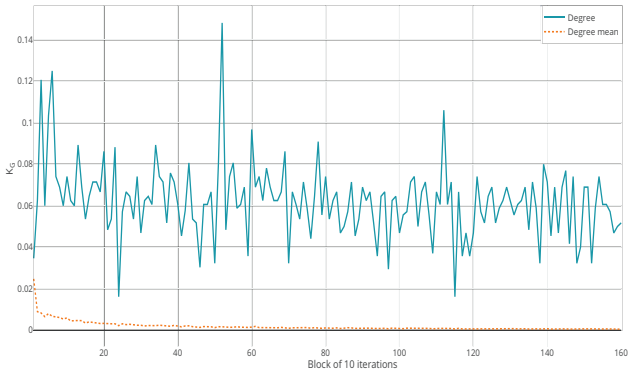

Figure 4: K_G constant values for degree centrality measure during the graph evolution process

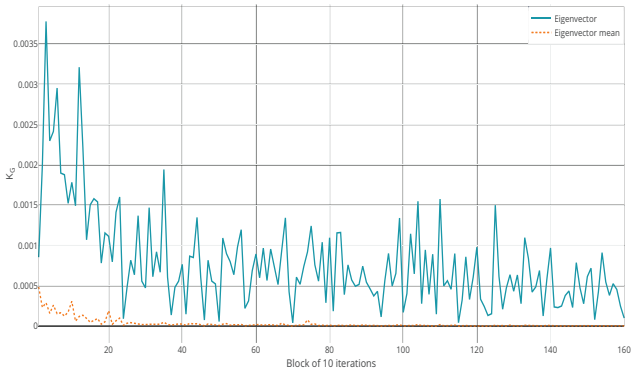

Figure 5: K_G constant values for eigenvector centrality measure during the graph evolution process

proach regarding many real-life applications. The perturbation behind evolution relies on the fact that in these networks new vertices can connect to the initial graph with one or more edges, also new edges can appear between existing nodes in the network.

For these experiments the perturbed versions of the initial graph were provided by the Cooper–Frieze graph process. In the reported results a graph of two nodes connected with an edge as initial input graph was used. We started to measure the centrality stability values after the 100th iteration by blocks of ten iterations. Thus, at the end of an iteration block consisting of 10 time steps t, a perturbed graph is produced with new nodes and edges compared to the graph from the previous block.

As it can be seen on Figures 4 and 5 the empirical values of K_G bound are about one order magnitude higher than those for the $S\&P$ dataset. Note that they are still much lower than their theoretical values and they show only slight fluctuation. Moreover, the mean values (calculated as (2) converges to zero by the growth of the size of the network. Similar convergence can be noticed on Figure 6 regarding the Kendall's correlation which shows the evidence that even the nodes ranking remain practically unchanged during the graph evolution process.

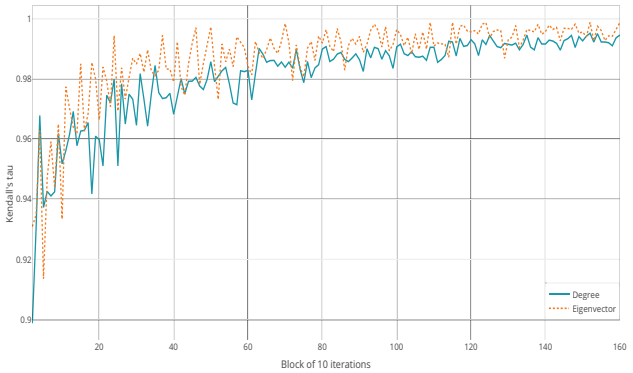

Figure 6: Kendall's tau coefficient between rank by different centrality measures during the graph evolution process

Acknowledgments

The project has been supported by the European Union, co-financed by the European Social Fund (EFOP-3.6.3-VEKOP-16-2017-00002).

5. REFERENCES

[1] M. A. Beauchamp. An improved index of centrality. *Behavioral Science*, 10(2):161–163, 1965.

[2] P. Bonacich. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2(1):113–120, 1972.

[3] S. P. Borgatti, K. M. Carley, and D. Krackhardt. On the robustness of centrality measures under conditions of imperfect data. *Social Networks*, 28(2):124–136, 2006.

[4] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998.

[5] C. Cooper and A. Frieze. A general model of web graphs. *Random Structures & Algorithms*, 22(3):311–335, 2003.

[6] E. Costenbader and T. W. Valente. The stability of centrality measures when networks are sampled. *Social Networks*, 25(4):283–307, 2003.

[7] K. Das, S. Samanta, and M. Pal. Study on centrality measures in social networks: a survey. *Social Network Analysis and Mining*, 8(1):13, 2018.

[8] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

[9] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[10] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[11] U. Nieminen. On the centrality in a directed graph. *Social Science Research*, 2(4):371–378, 1973.

[12] G. Sabidussi. The centrality index of a graph. *Psychometrika*, 31(4):581–603, 1966.

[13] S. Segarra and A. Ribeiro. Stability and continuity of centrality measures in weighted graphs. *IEEE Transactions on Signal Processing*, 64(3):543–555, 2015.

[14] M. E. Shaw. Group structure and the behavior of individuals in small groups. *The Journal of Psychology*, 38(1):139–149, 1954.

[15] B. Zemljič and V. Hlebec. Reliability of measures of centrality and prominence. *Social Networks*, 27(1):73–88, 2005.