A STABLE BETWEENNESS CENTRALITY MEASURE IN NETWORKS

Santiago Segarra *Alejandro Ribeiro*

Department of Electrical and Systems Engineering, University of Pennsylvania

ABSTRACT

This paper presents a formal definition of stability for node centrality measures in networks and shows that the well-known betweenness centrality is not stable with respect to that metric. An alternative definition that preserves the same centrality notion while satisfying this stability criterion is then introduced. The practical implications of stability are explored by studying the behavior of the traditional as well as the alternative stable betweenness centrality in both, synthetic random networks, and the network of interactions between sectors of the United States economy.

Index Terms— Networks, betweenness centrality, stability.

1. INTRODUCTION

Within a network, the influence that a node or agent can exert over the others is not constant across agents and depends on the network topology. Identifying the most influential nodes in a network helps in explaining the network dynamics, e.g. the distribution of power in exchange networks [1], as well as in designing optimal ways to externally influence the network, e.g. attack vulnerability of networks [2]. Node centrality measures are tools designed to identify such influential agents. Although several centrality measures can be found in the literature, the most common being degree, closeness [3,4], eigenvector [5], stress [6] and betweenness [7] centrality, the latter has been extensively used in the study of both technological [8] and social [9] networks. In betweenness centrality, the centrality of a node is given by the frequency of this node belonging to the shortest path between other two nodes in the network. Although originally presented for undirected and unweighted graphs, variations of betweenness centrality have been developed, including one suitable for application in the more general framework of weighted and directed networks [10].

Stability refers to the ability of a centrality measure to be robust to noise in the network data. From a practical perspective, a stable measure is more reliable in the presence of imperfect measurements or quantization errors. In the past decade, stability has been used as a parameter to compare the performance of different centrality measures [11–13]. In these papers, an empirical approach was followed by comparing stability indicators measured in both random and real world networks for different centrality measures. However, no formal theory was developed explaining the different behaviors among measures. Our first contribution is a formal definition of a stable centrality measure (Section 3.1). In order to build such definition, we need to rely on a metric on the space of weighted and directed networks with a common node set (Section 3). After showing that the traditional betweenness centrality measure is not stable, our second contribution is the definition of an alternative measure which captures the same betweenness centrality notion and satisfies the stability definition (Section 4). Finally, we illustrate how our formal definition of stability is correlated with practical stability indicators by analyzing the behavior of the traditional and the alternative betweenness centrality measures in random networks (Section 5.1) and a real-world network (Section 5.2).

Work supporte by NSF CCF-1217963.

2. PRELIMINARIES

In the present paper we consider weighted and directed networks. Formally, a network $N = (X, A_X)$ is defined as a finite set of n nodes X endowed with a real valued dissimilarity function $A_X : X \times X \to \mathbb{R}_+$ defined for all pairs of nodes $x, x' \in X$. Dissimilarities $A_X(x, x')$ are nonnegative for all $x, x' \in X$, and null if and only if $x = x'$, but need not satisfy the triangle inequality and may be asymmetric, i.e. $A_X(x, x') \neq A_X(x', x)$ for some $x, x' \in X$. For any given X, denote by $\mathcal{N}_X$ the space of networks with X as node set.

In the definition of betweenness centrality, the concepts of *path* and *path length* are important. Given a network (X, A_X) and $x, x' \in X$, a path $P(x, x')$ is an *ordered* sequence of nodes in X,

$$P(x, x') = [x = x_0, x_1, \ldots, x_{l-1}, x_l = x'], \tag{1}$$

which starts at x and finishes at x'. We say that $P(x, x')$ links or connects x to x'. The *links* of a path are the edges connecting consecutive nodes of the path in the direction given by the path. We define the *length* of a given path $P(x, x') = [x = x_0, \ldots, x_l = x']$ as the sum of the dissimilarities $\sum_{i=0}^{l-1} A_X(x_i, x_{i+1})$ encountered when traversing its links in order. Dissimilarities are also referred to as edge or link weights. Given the network $N = (X, A_X)$, we define the shortest path function $s_N : X \times X \to \mathbb{R}_+$ where the shortest path length $s_N(x, x')$ between nodes $x, x' \in X$ is defined as

$$s_N(x, x') := \min_{P(x,x')} \sum_{i=0}^{l-1} A_X(x_i, x_{i+1}). \tag{2}$$

Given three arbitrary nodes $x, x', x'' \in X$, denote by $\sigma_{x'x''}$ the number of shortest paths from x' to x'', i.e. the number of paths $P(x', x'')$ of length $s_N(x', x'')$, and by $\sigma_{x'x''}(x)$ the number of these shortest paths that go through node x. For convenience, we define $\sigma_{xx} = 1$ for all $x \in X$. Notice that since A_X might be asymmetric, we can have that $\sigma_{x'x''} \neq \sigma_{x''x'}$ for some $x', x'' \in X$. The betweenness centrality $C_B(x)$ for any given node $x \in X$ is defined as [7]

$$C_B(x) := \sum_{\substack{x', x'' \in X \\ x' \neq x \neq x''}} \frac{\sigma_{x'x''}(x)}{\sigma_{x'x''}}. \tag{3}$$

In (3), we calculate the betweenness centrality value of a node $x \in X$ by sequentially looking at the shortest paths between any two nodes distinct from x and summing the proportion of shortest paths that contain node x. The higher the centrality value $C_B(x)$, the more central node x is in network N. Sometimes [7,14], (3) is normalized by the number of pairs in the network or the maximum centrality

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

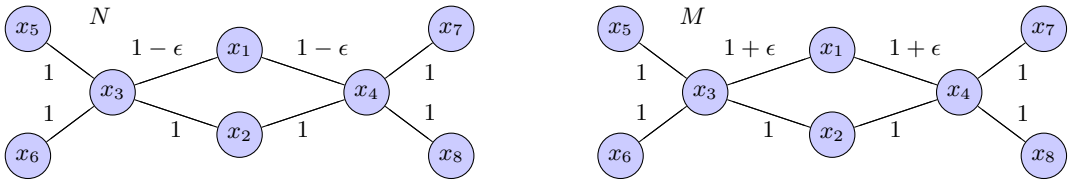

Fig. 1: Instability of betweenness centrality C_B. The distance between networks N and M vanishes with decreasing ϵ, however $C_B^N(x_1) = 18$ and $C_B^M(x_1) = 0$ for every $\epsilon > 0$.

value achievable such that $C_B(x)$ takes values in $[0, 1]$. However, we are interested in comparing centrality values between different nodes within a network and these comparisons are invariant to any normalization. Hence, we omit the normalizing constant in definition (3).

3. CENTRALITY AND STABILITY

Node centrality is a measure of the importance of nodes as intermediate steps when traversing a network. Very often, this importance relies on an underlying characteristic of the nodes. E.g., airports which are hubs for some airline have high centrality in an air transportation network. In this way, centrality detects fundamental roles played by nodes within the network. Ideally, this detection should be invariant to small perturbations in the edge weights. To formalize this, we define the metric $d_{\mathcal{N}_X} : \mathcal{N}_X \times \mathcal{N}_X \to \mathbb{R}_+$ on the space of networks $\mathcal{N}_X$ containing X as node set, as follows

$$d_{\mathcal{N}_X}(N, N') := \sum_{x,x' \in X} |A_X(x, x') - A'_X(x, x')|, \qquad (4)$$

where $N = (X, A_X)$ and $N' = (X, A'_X)$. To see that $d_{\mathcal{N}_X}$ is a well-defined metric, notice that it computes the L_1 distance between two vectors obtained by stacking the dissimilarity values in A_X and A'_X. The metric $d_{\mathcal{N}_X}$ enables the formal definition of stability presented in the following section.

3.1. Stability

We say that a centrality measure $C : X \to \mathbb{R}$ is *stable* if, for every node set X and any two networks $N, M \in \mathcal{N}_X$, there exists a finite constant K_X such that

$$|C^N(x) - C^M(x)| \leq K_X \; d_{\mathcal{N}_X}(N, M), \qquad (5)$$

for every node $x \in X$, where $C^N(x)$ is the centrality value of node x in network N and similarly for M. The above definition states that a centrality measure is stable if the difference in centrality values for a given node in two different networks is bounded by a constant K_X times the distance between these networks. K_X is a normalizing constant that depends on the cardinality of the node set X and not on the dissimilarities of the networks. The inclusion of K_X in (5) ensures that the stability of a centrality measure does not depend on the appearance of a normalization term in the definition of the measure. E.g., the betweenness centrality measure as defined in (3) is stable if and only if the normalized version [7, 14] is stable. In fact, betweenness centrality is an unstable measure as we show next.

Proposition 1 *The betweenness centrality measure C_B defined in* (3) *is not stable as defined in* (5)*.*

Proof: Consider the two networks $N = (X, A_X)$ and $M = (X, B_X)$ depicted in Fig. 1 where undirected edges represent symmetric dissimilarities. Moreover, undrawn edges are identical in both networks and have weights large enough, e.g. larger than 5, such that they do not belong to a shortest path between any pair of nodes. Notice that $N, M \in \mathcal{N}_X$ where $X = \{x_1, x_2, \ldots, x_8\}$ and, thus, we can compute $d_{\mathcal{N}_X}(N, M) = 4\epsilon$ from (4).

For any $\epsilon > 0$, according to (3) we have that $C_B^N(x_1) = 18$ since the node x_1 is part of the unique shortest path from any node in $\{x_3, x_5, x_6\}$ to any node in $\{x_4, x_7, x_8\}$ and vice versa. However, for that same ϵ, $C_B^M(x_1) = 0$ since x_1 is not an intermediate node in any shortest path in network M. This implies that,

$$\frac{|C_B^N(x_1) - C_B^M(x_1)|}{d_{\mathcal{N}_X}(N, M)} = \frac{18}{4\epsilon}. \qquad (6)$$

Note that for any candidate constant K_X in (5), there exists a small enough $\epsilon > 0$ such that the above ratio is greater than the proposed K_X. Thus, such constant cannot exist and C_B is not stable. ■

The instability of the traditional betweenness centrality measure motivates an alternative definition presented in the following section.

4. A STABLE CENTRALITY MEASURE

Given an arbitrary network $N = (X, A_X)$ and a node $x \in X$, define a new network $N^x = (X^x, A_X^x)$ with $X^x = X \backslash \{x\}$ and $A_X^x = A_X|_{X^x \times X^x}$. I.e., the network N^x is constructed by eliminating from N the node x and every edge directed to or from it. Define the *stable betweenness centrality* $C_{SB}(x)$ of any node $x \in X$ as

$$C_{SB}(x) := \sum_{\substack{x', x'' \in X \\ x' \neq x \neq x''}} s_{N^x}(x', x'') - s_N(x', x''). \qquad (7)$$

Note that every term in the above summation is nonnegative since shortest paths in the network N^x cannot be shorter than the corresponding paths in N. Measure C_{SB} quantifies the centrality of a given node x by the change in the length of shortest paths once this node is removed. Intuitively, if a node is part of many shortest paths, when we remove this node the corresponding paths will increase in length and result in a high centrality value. In this sense, measure C_{SB} is similar to the original betweenness centrality measure C_B. However, how critical a given node is in connecting the network depends on the best alternative path if this node fails. In opposition to the traditional centrality measure, C_{SB} is stable as shown next.

Proposition 2 *The stable betweenness centrality measure C_{SB} defined in* (7) *is stable as defined in* (5)*, with $K_X = 2(n-1)(n-2)$ where $|X| = n$.*

Proof: See [15]. ■

For a stable centrality measure, the difference in the centrality value for a given node in two networks is bounded by the distance between these networks. In particular, if given a network we generate a new network by perturbing the original one, these two networks

must be close to each other. Thus, the stable centrality measure C_{SB} ensures that the change in centrality for every node due to the perturbation is bounded. This generates a robust measure in the presence of noise as illustrated through examples in Section 5.

To illustrate the stable betweenness centrality measure C_{SB}, consider the network N in Fig. 1. The centrality of node x_1 is limited by the existence of a comparable path through node x_2. More precisely, if node x_1 is deleted from N, the 18 shortest paths [cf. proof of Proposition 1] of which x_1 was originally a part of, have their length increased by 2ϵ. Consequently, $C_{SB}^{N}(x_1) = 36\,\epsilon$. If the length of the alternative path through x_2 increases, the centrality of node x_1 increases as well, reflecting that a failure of x_1 has a greater impact on the connectivity of the network. This phenomenon is not detected by the traditional measure C_B.

Computing the betweenness centrality C_B value for every node in a network with n nodes requires $O(n^3)$ computations [16]. For C_{SB}, we can use the Floyd-Warshall algorithm [17, 18] to compute all-pairs shortest paths in a network in $O(n^3)$ time. In a naive computation of C_{SB} for every node in the network, we can compute the shortest paths for all pairs of nodes in the original network and in every network generated when deleting one node at a time. This requires $n+1$ implementations of Floyd-Warshall with a total complexity of $O(n^4)$. A faster algorithm could exist since, when a node is deleted from the network, only the shortest paths originally containing this node need to be recomputed.

5. NUMERICAL EXPERIMENTS

We illustrate the practical implications of the stability of C_{SB} in contrast with the instability of C_B through the analysis of both random networks and real-world networks.

5.1. Random networks

We define a random network of n nodes as one where every dissimilarity is randomly chosen from a uniform distribution in $[0, 1]$. For the following experiment, we generate 100 random networks of n nodes, where n varies from 10 to 200 in multiples of 10. We then generate perturbed versions of each of these networks by multiplying every dissimilarity by a random number uniformly distributed in $[0.99, 1.01]$. For every network, we generate a centrality ranking of the nodes, i.e. we sort the nodes in decreasing order of centrality value, and compare it with the centrality ranking of the perturbed version of that network. We perform this comparison for the rankings output by C_B and C_{SB}.

Two stability indicators are analyzed and depicted in Fig. 2a-2d. The first indicator is the magnitude of the changes in the rankings when the networks are perturbed. Two consecutive nodes in the ranking switching positions is a smaller change that a node drastically varying its position in the ranking. Thus, for every network we record the maximum variation in ranking position experienced by a node when perturbing the network. In Fig. 2a we plot the mean of this indicator among the networks analyzed as a function of the network size. For example, for a network with 100 nodes, the perturbation generates a maximum change of 1.9 positions on average for the C_{SB} ranking and 4.6 positions for the C_B ranking. Both measures experience an approximately linear increase of the maximum change with the size of the network, but the increasing rate is faster for the C_B case, generating big performance differences between the measures for larger networks. Apart from computing the average maximum change across networks, we are interested in the distributions of this maximum change for the two centrality measures. Thus, in Fig. 2b we plot the probability that the maximum change in the ranking generated by a perturbation is greater than d positions for $d \in \{1, 3, 5, 10\}$ as a function of the network size. E.g., when $d = 1$, the solid lines in the figure inform the probability that a perturbed network suffers a change of more than 1 position, i.e. a change greater than a switch between consecutive nodes. For a network of 50 nodes, this probability is 0.27 for the C_{SB} measure and 0.70 for the C_B measure. If we look at $d = 3$, for networks of 200 nodes, we see that almost every network (97%) suffers a change of more than three positions in the ranking generated by C_B while only 26% of the networks present this phenomenon for C_{SB}. For $d = 5$ and $d = 10$, there are almost no changes greater than d for the C_{SB} case, while 65% of the networks with 200 nodes present changes greater than 5 positions and 12% changes greater than 10 positions for the C_B ranking. To facilitate the understanding of figures 2a and 2b, in Fig. 2c we present the histogram of the maximum change found in the rankings when perturbing a network for the particular case of networks with 100 nodes. The mean of the blue histogram in Fig. 2c corresponds to the blue circle in Fig. 2a for networks with 100 nodes and similarly for the red histogram. To relate the histogram with Fig. 2b, notice that the support of the red histogram starts at 2, thus every observed change is greater than $d = 1$ and, consequently, the solid red line in Fig. 2b reaches probability 1 for networks with 100 nodes. Similarly, the blue histogram has a weight of 25 on changes of 1 position, thus the other 75 out of the 100 networks analyzed presented changes of 2 or more. Consequently, the blue solid line in Fig. 2b reaches a probability of 0.75 for networks of 100 nodes. Notice that for one of the studied networks, the C_B ranking presented a change of 17 positions when the perturbation was introduced. This is an empirical example of instability as shown in Proposition 1.

The second indicator we analyze is the position where the change in the ranking occurs. A change towards the last positions of the ranking is irrelevant whereas a change where the most central node is modified carries important implications. In Fig. 2d, we plot the probability that the first change occurs after position p for $p \in \{5, 10, 20\}$ as a function of the size of the network. Observe that the probabilities remain approximately constant for network sizes past a certain threshold. In this way, we can state that for around 90% of the networks there is no change in the top 5 centrality ranking computed with C_{SB} independently of the network size, whereas this number is around 80% for C_B. If we are interested in the top 10 ranking remaining unperturbed, this occurs for above 70% of the networks with C_{SB} and around 50% for C_B.

In conclusion, this experiment shows that in practice the stability property of C_{SB} shown in Proposition 2 entails centrality rankings with variations which are less important (Fig. 2d) and smaller in magnitude (Fig. 2a - Fig. 2c) than those obtained with the traditional centrality measure C_B.

5.2. Real-world network: interaction of economic sectors

We also test both centrality measures in a real-world network that records interactions between sectors of the economy. The Bureau of Economic Analysis of the U.S. Department of Commerce publishes a yearly table of inputs and outputs organized by economic sectors [19]. More precisely, we are given a set I of 61 industrial sectors as defined by the North American Industry Classification System (NAICS) and a function $U : I \times I \to \mathbb{R}_+$ where $U(i, i')$ for all $i, i' \in I$ represents how much of the production of sector i, expressed in dollars, is used as an input of sector i'. We define the

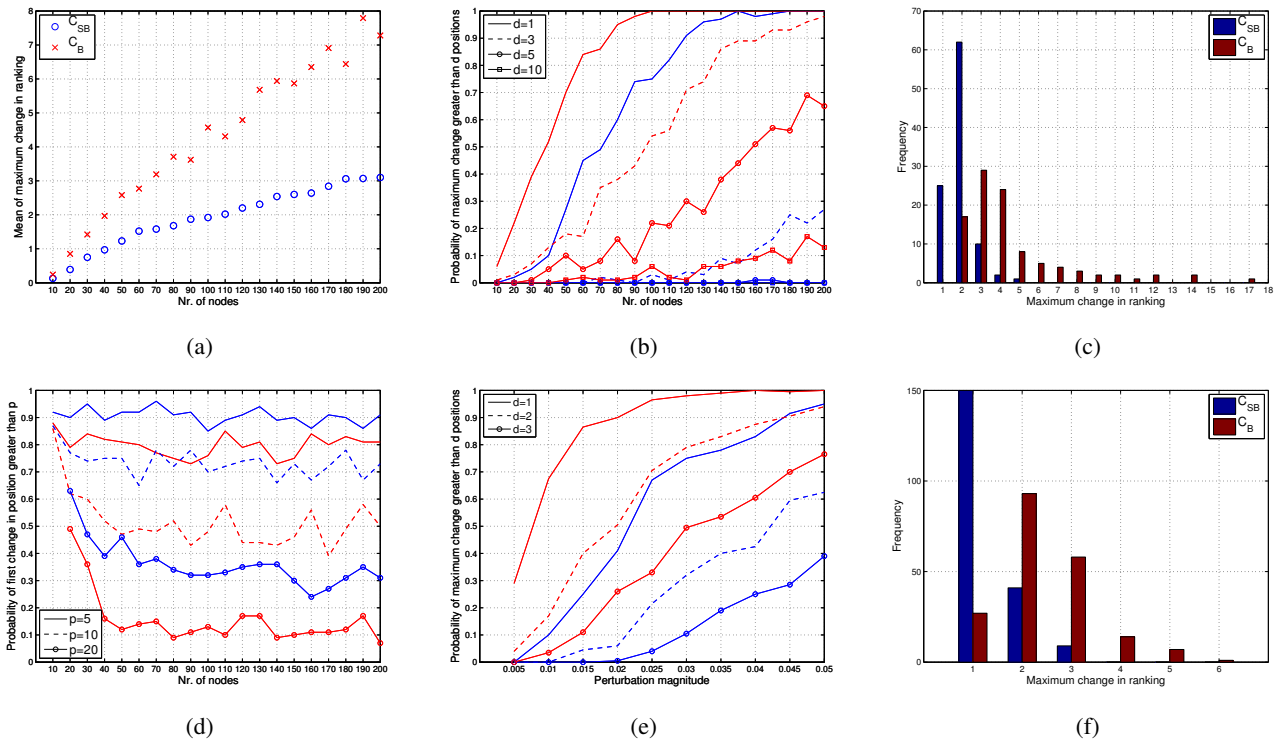

Fig. 2: Comparison of stability indicators for the betweenness centrality measure C_B (red) and the stable betweenness centrality measure C_{SB} (blue). (a) Average of the maximum change recorded when perturbing a random network as a function of network size. (b) Probability that the maximum change in the ranking when perturbing a network exceeds d positions as a function of the network size. (c) Histogram of the maximum change recorded when perturbing random networks with 100 nodes. (d) Probability that the first change position in the ranking when perturbing a network is greater than position p as a function of the network size. (e) Probability that the maximum change in the ranking when perturbing network N_I exceeds d positions as a function of the perturbation magnitude. (f) Histogram of the maximum change recorded when applying a perturbation of $\delta = 0.015$ to network N_I.

network $N_I = (I, A_I)$ where $A_I(i, i') = 1/U(i, i')$ for all $i, i' \in I$ and rank the nodes in I by their centrality computed with the traditional measure C_B and the proposed alternative C_{SB}, see Table 1. Notice that, although the rankings are different, both top 10 lists share 9 economic sectors. Moreover, the top 3 and the top 8 sectors are the same in both lists, but in different order. This verifies that, even though the definitions of C_B in (3) and of C_{SB} in (7) differ, they capture the same notion of centrality. However, the latter is more robust as we illustrate by perturbing the network N_I. In this experiment, we compute the probability of observing a change in the ranking of more than d positions for $d \in \{1, 2, 3\}$ as a function of the magnitude of the perturbation; see Fig. 2e. A perturbation magnitude of δ implies that every dissimilarity in the network is multiplied by a random number in $[1 - \delta, 1 + \delta]$. For every perturbation level, we generated 200 perturbed networks. As expected, the probability of observing a change in the network increases with the perturbation magnitude. Moreover, due to the stability property, if we fix the magnitude of perturbation, larger changes are observed in the rankings generated by C_B compared with those generated by C_{SB}. E.g., for a perturbation of 0.03, 11% of the rankings generated by C_{SB} presented a change greater than $d = 3$ whereas half the rankings generated by C_B presented this phenomenon. Finally, in Fig. 2f we present the histogram of maximum changes observed for a perturbation of $\delta = 0.015$. For example, 50 out of the 200 networks analyzed presented a maximum change greater than 1 position for the C_{SB} ranking, thus, in Fig. 2f the solid blue line reaches probability 0.25 for a perturbation magnitude of 0.015.

Table 1: Top 10 central sectors as computed with C_B and C_{SB}.

	C_B	C_{SB}
1	Real Estate	Misc. Professional services
2	Construction	Real Estate
3	Misc. Professional services	Construction
4	Wholesale trade	Petroleum and coal prod.
5	FR banks, credit inter.	FR banks, credit inter.
6	Petroleum and coal prod.	Chemical products
7	Oil and gas extraction	Oil and gas extraction
8	Chemical products	Wholesale trade
9	Retail trade	Securities and investments
10	Securities and investments	Food and beverage

6. CONCLUSION

Stability, as a formal property of node centrality measures, was introduced. The traditional betweenness centrality measure was shown not to be stable, thus, a stable alternative definition was proposed. The stability difference between both measures was illustrated by studying indicators in both random and real-world networks.

7. REFERENCES

[1] K. S. Cook, R. M. Emerson, M. R. Gillmore, and T. Yamagishi, "The distribution of power in exchange networks: theory and experimental results," *American Journal of Sociology*, vol. 89, no. 2, pp. 275–305, 1983.

[2] P. Holme, B. J. Kim, C. N. Yoon, and S. K. Han, "Attack vulnerability of complex networks," *Phys. Rev. E*, vol. 65, pp. 056109, May 2002.

[3] G. Sabidussi, "The centrality of a graph," *Psychometrika*, vol. 31, pp. 581–603, Dec 1966.

[4] M. A. Beauchamp, "An improved index of centrality," *Behavioral Science*, vol. 10, no. 2, pp. 161–163, 1965.

[5] P. Bonacich, "Factoring and weighting approaches to clique identification," *Journal of Mathematical Sociology*, vol. 2, pp. 113–120, 1972.

[6] A. Shimbel, "Structural parameters of communication networks," *The bulletin of mathematical biophysics*, vol. 15, no. 4, pp. 501–507, 1953.

[7] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, vol. 40, no. 1, pp. 35–41, 1977.

[8] J. P. Onnela, J. Saramki, J. Hyvnen, G. Szab, D. Lazer, K. Kaski, J. Kertsz, and A. L. Barabsi, "Structure and tie strengths in mobile communication networks," *Proceedings of the National Academy of Sciences*, vol. 104, no. 18, pp. 7332–7336, 2007.

[9] M. E. J. Newman, "Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality," *Phys. Rev. E*, vol. 64, pp. 016132, Jun 2001.

[10] U. Brandes, "On variants of shortest-path betweenness centrality and their generic computation," *Social Networks*, vol. 30, no. 2, pp. 136 – 145, 2008.

[11] E. Costenbader and T. W. Valente, "The stability of centrality measures when networks are sampled," *Social Networks*, vol. 25, no. 4, pp. 283 – 307, 2003.

[12] S. P. Borgatti, K. M. Carley, and D. Krackhardt, "On the robustness of centrality measures under conditions of imperfect data," *Social Networks*, vol. 28, no. 2, pp. 124 – 136, 2006.

[13] B. Zemljič and V. Hlebec, "Reliability of measures of centrality and prominence," *Social Networks*, vol. 27, no. 1, pp. 73 – 88, 2005.

[14] D. R. White and S. P. Borgatti, "Betweenness centrality measures for directed graphs," *Social Networks*, vol. 16, no. 4, pp. 335 – 346, 1994.

[15] S. Segarra and A. Ribeiro, "A stable betweenness centrality measure in networks," *Preprint*, 2013, Available at https://fling.seas.upenn.edu/∼ssegarra/wiki/index.php?n =Research.Publications.

[16] U. Brandes, "A faster algorithm for betweenness centrality," *Journal of Mathematical Sociology*, vol. 25, pp. 163–177, 2001.

[17] R. W. Floyd, "Algorithm 97: shortest path," *Commun. ACM*, vol. 5, no. 6, pp. 345, Jun 1962.

[18] S. Warshall, "A theorem on boolean matrices," *J. ACM*, vol. 9, no. 1, pp. 11–12, Jan 1962.

[19] Bureau of Economic Analysis, "Input-output accounts: the use of commodities by industries before redefinitions," *U.S. Department of Commerce*, 2011, Available at http://www.bea.gov/iTable/index_industry.cfm.