First publ. in: International Journal of Bifurcation and Chaos 17 (2007), 7, pp. 2303-2318

Centrality Estimation in Large Networks*

Ulrik Brandes Christian Pich

Department of Computer and Information Science
University of Konstanz

Abstract

Centrality indices are an essential concept in network analysis. For those based on shortest-path distances the computation is at least quadratic in the number of nodes, since it usually involves solving the single-source shortest-paths (SSSP) problem from every node. Therefore, exact computation is infeasible for many large networks of interest today. Centrality scores can be estimated, however, from a limited number of SSSP computations. We present results from an experimental study of the quality of such estimates under various selection strategies for the source vertices.

*Research supported in part by DFG under grant Br 2158/2-3

1 Introduction

An essential tool in the analysis of complex networks are centrality indices defined on the vertices or edges of the underlying graph [Koschützki et al. 2005a,b]. Depending on the type of network studied, they are proxies for the structural importance of an element for the overall functioning of the network. Many popular centrality indices are based on shortest paths, measuring, e.g., the average distance from other vertices, or the ratio of shortest paths a vertex lies on. Our impression is that the majority of network-analytic studies relies at least in part on an evaluation of such indices.

With the rapidly increasing amount of data gathered and made available in electronic form, there is a likewise increasing demand for the computation of centrality indices on networks that are orders of magnitude larger than before. Although exact centrality index computation is *tractable* in the conventional sense that there exist polynomial time and space algorithms, these are not practical.

It is therefore of considerable interest to evaluate the practical performance of methods for estimating centrality indices. For most feedback-based indices defined via systems of linear equations there is a natural method of approximation inherent in iterative solvers for linear equations and eigenproblems. For the discrete concepts of centrality based on shortest paths, these are not applicable. In fact, approximation of betweenness centrality (defined below) is stated as an important open problem, e.g., in [Carpenter et al. 2002].

We here present an experimental study of estimators for the two most commonly used shortest-path centralities, closeness and betweenness. The estimates are based on a restricted number of single-source shortest-paths computations from a set of selected pivots. For doing so, we generalize an approach of Eppstein and Wang [2004] in a number of ways (explained in Sec. 3), and test it experimentally.

This paper is organized as follows. The basic concepts needed are defined in Sec. 2, and algorithms for estimating shortest-path centralities using pivots are given in Sec. 3. In Sec. 4 we introduce several pivot selection strategies. The results of our experimental study are presented in Sec. 5, and we conclude in Sec. 6.

2 Shortest-Path Centralities

Indices for measuring the structural importance of nodes in a network abound (see [Brandes and Erlebach 2005] for an overview). Two of the indices most commonly used in the social sciences are closeness centrality [Beauchamp 1965, Sabidussi 1966] and betweenness centrality [Anthonisse 1971, Freeman 1977]. Both are based on shortest-path distances, but while a node has high closeness centrality if its total (and therefore also average) distance to all other vertices is small, a high betweenness centrality score indicates that a node is contained in relatively many shortest paths connecting pairs of others.

2.1 Definition

Throughout this paper the topology of a networks will be represented by a *graph* $G = (V, E)$, where V is a set of *vertices*, and $E \subseteq \binom{V}{2}$ is a set of *edges*, i.e. unordered pairs of vertices. In particular, we do not allow directions, self-loops, multiple edges between the same pair of vertices, or weights on the edges; i.e. our graphs are simple, undirected, and unweighted. If not stated otherwise, $n = |V|$ denotes the number of vertices and $m = |E|$ the number of edges. A vertex $v \in V$ is called *incident* to an edge $e \in E$, if $v \in e$, and two vertices are called *adjacent*, if they are incident to a common edge.

A *path* is an alternating sequence of vertices and edges, such that edges in the sequence appear between their two incident vertices. The *length* of a path is simply its number of edges. Two vertices $s, t \in V$ are connected, if their exists a path starting at one and ending at the other; such a path is also called an *st-path*. A graph is called connected, if every pair of vertices is connected.

We restrict ourselves to connected graphs (otherwise the connected components can be treated individually).

The *distance* $d(s, t)$ between two vertices $s, t \in V$ is the length the shortest path connecting them. In particular, $d(s, t) = d(t, s)$, since the reversal of an st-path yields a ts-path, and $d(s, s) = 0$, since the path s is an alternating sequence with no edges. The largest distance between any two vertices of a graph is called the *diameter* of G, $diam(G)$.

Closeness centrality [Beauchamp 1965, Sabidussi 1966] measures how close a vertex is to all other vertices in the graph. To obtain large values for small sums of distances, it is defined as the inverse of the total distance,

$$c_C(v) = \frac{1}{\sum_{t \in V} d(v, t)} \ . \tag{1}$$

Thus, the distance from a vertex of high closeness centrality to any other vertex is short on average. These vertices are considered to be structurally important, because they can easily reach or be reached by others.

An alternative concept of centrality is based on the idea of control over the connections between other pairs of vertices. Denote by $\sigma(s, t)$ the number of different shortest st-paths, and by $\sigma(s, t|v)$ the number of shortest st-paths that contain v as an *inner* vertex, i.e. $v \neq s, t$ or $\sigma(s, t|s) = 0 = \sigma(s, t|t)$. *Betweenness centrality* [Anthonisse 1971, Freeman 1977] measures the degree to which a vertex is needed by others when connecting along shortest paths,

$$c_B(v) = \sum_{s \neq v \neq t} \frac{\sigma(s, t|v)}{\sigma_(s, t)} \ . \tag{2}$$

There are many other structural indices that are based on similar notions of importance. For instance, we can replace the sum of distances in closeness centrality by the maximum distance to any other vertex [Harary and Hage 1995], or subtract each distance from an upper bound rather than taking the inverse [Botagfogo et al. 1992, Valente et al. 1998]. Variants of betweenness count all shortest paths equally [Shimbel 1953] or use maximum network flow instead of shortest paths [Freeman et al. 1991]. Natural variants of closeness and betweenness are also obtained by replacing spread along shortest paths with current flow [Newman 2005, Brandes and Fleischer 2005]. A different class of measures is based on feedback, i.e. the centrality of a vertex directly influences that of its neighbors. Well-known members of this class are eigenvector centrality [Bonacich 1972], Google's PageRank [Brin and Page 1998], and hubs & authorities [Kleinberg 1999].

For most of these measures, generalizations have been proposed for directed, non-simple, weighted, and unconnected graphs, and there is a similar range of indices that value the importance of edges rather than vertices. We refer to [Brandes and Erlebach 2005] for a comprehensive survey.

In this paper, we focus on shortest-path closeness and betweenness for vertices in simple, undirected, connected graphs without weights as defined by Eqs. (1) and (2). Note, however, that our results also apply to more general settings.

2.2 Computation

For sparse networks, which we loosely define as those for which $m \in \mathcal{O}(n \log n)$, i.e. in which the number of actual edges is small compared to the number of potential edges, the closeness centrality index is best computed by solving a single-source shortest-path (SSSP) problem from every vertex. In each iteration, we may sum up all distances found and invert the total to obtain the centrality score of the source. Using standard breadth-first search, the running time per source is bounded by $\mathcal{O}(n + m)$, and thus $\mathcal{O}(nm)$ in total.

For betweenness centrality, the computation is less straightforward, since we do not have to evaluate lengths, but numbers of shortest path between pairs with given intermediates. We reformulate (2) by introducing the *dependency* $\delta(s,t|v) = \frac{\sigma(s,t|v)}{\sigma_(s,t)}$ of a pair $s,t \in V$ on $v \in V$ and summing out all targets t,

$$c_B(v) = \sum_{s \neq v \neq t} \frac{\sigma(s,t|v)}{\sigma_(s,t)} = \sum_{s \neq v \neq t} \delta(s,t|v) = \sum_{s \neq v} \delta(s|v) \ ,$$

where $\delta(s|v) = \sum_{t \neq v} \delta(s,t|v)$ is the *one-sided dependency* of s on v. In [Brandes 2001] it is shown how to compute the one-sided dependencies of all $v \in V$ for a given $s \in V$ by solving an SSSP. Therefore, betweenness centrality can be computed in the same asymptotic time bounds, and in fact using essentially the same basic algorithm, as closeness centrality.

A notable feature of the above SSSP-based algorithms is that the space requirement is linear, since the quadratic distance matrix is needed only row-wise. All distance-information computed during one iteration can be discarded before starting the next.

3 Approximate Computation

For large graphs, the exact computation of centralities as described in the previous section is too costly since the running time is $\Omega(n^2)$ even for the sparsest connected graphs.

On the other hand, the computation consists of solving n single-source shortest-paths problems, one for each vertex, and each SSSP contributes one summand to the result. This contribution is the distance to the source for closeness, and the one-sided dependency of the source for betweenness. The vertices for which an SSSP is solved are called *pivots*. Based on an idea put forward by Eppstein and Wang [2004], the exact centrality value can be estimated by extrapolating the contributions obtained from just a few SSSP computations, i.e. from a small set of pivots.

The foundation of this idea is a bound on the deviation of the average of a given number of bounded random variables from its expectation. Hoeffding [1963] proves that for independent identically distributed random variables $X_1, \ldots, X_k$ with $0 \leq X_i \leq M$ $(i = 1, \ldots, k)$ and an arbitrary $\xi \geq 0$,

$$P\left(\left|\frac{X_1 + \ldots + X_k}{k} - E\left(\frac{X_1 + \ldots + X_k}{k}\right)\right| \geq \xi\right) \leq e^{-2k\left(\frac{\xi}{M}\right)^2} . \tag{3}$$

If pivots are selected at random, the contributions of different SSSP computations to the centrality of a single vertex can be considered the result of a random experiment. In the following two subsection we derive estimates for closeness and betweenness using this idea.

3.1 Closeness centrality

The contribution of an SSSP computation from pivot $p_i \in V$ to the centrality of a vertex $v \in V$ is $d(p_i, v) = d(v, p_i)$. In order to extrapolate from k such samples, let

$$X_i(v) = \frac{n}{n-1} \cdot d(v, p_i) \tag{4}$$

be the random variable associated with the random experiment of selecting pivot p_i. Let

$$\begin{aligned} M &= \frac{n}{n-1} \cdot diam(G) \\ \xi &= \varepsilon \cdot diam(G) \ . \end{aligned}$$

Since the expectation of estimate $\frac{1}{k}(X_1(v) + \ldots + X_k(v))$ is the sum of distances of all vertices from v, Hoeffding's bound (3) guarantees that its error is bounded from above by $\varepsilon \cdot diam(G)$ with probability at least $\exp\{-2k(\frac{\varepsilon(n-q)}{n})^2\}$.

Eppstein and Wang [2004] concludes that in graphs with constantly bounded diameter, $k \in \mathcal{O}(\log n)$ pivots are sufficient to estimate closeness centrality up to a constant with high probability. In the sequel of this paper, we will consider four generalizations with respect to this approach. Pivot-based estimation will also be computed

- on graphs of arbitrary diameter,
- using fewer pivots,
- using deterministic pivot-selection, and
- for betweenness centrality.

Clearly, we can trade estimator accuracy and confidence for running time by increasing or decreasing the number of pivots.

3.2 Betweenness centrality

When computing betweenness, the contribution of a pivot $p_i \in V$ to the centrality of a vertex $v \in V$ is $\delta(p_i|v)$. Again, to extrapolate from the average contribution of k pivots, we use

$$X_i(v) = \frac{n}{n-1} \cdot \delta(p_i|v) \tag{5}$$

for a single estimate. Setting

$$\begin{aligned} M &= \frac{n}{n-1} \cdot (n-2) \\ \xi &= \varepsilon(n-2) \ , \end{aligned}$$

we can again apply Hoeffding's bound as above. Note that one-sided dependencies are bounded by 0 from below and by $n-2$ from above. While the assumption of constantly bounded or at least small diameter made for closeness is reasonable for many practical examples, a one-sided dependency of $n-2$ is easily attained (simply consider a vertex with a neighbor that has degree one and is chosen for pivoting). It can thus be suspected that estimation of (non-normalized) betweenness is much more difficult and unreliable than estimation of (non-normalized) closeness.

4 Pivot Selection

To ensure that pivot contributions $X_i(v)$ are independent, pivots need to be selected at random. This appears to be a technical assumption introduced only to make sure that (3) holds in general. For practical purposes it might be advantageous to choose pivots deterministically, e.g. by spreading them uniformly over the graph. We used the following strategies in our experiments described in the next section. See also Tab. 1.

place Table 1 about here

All strategies are supposed to select k distinct pivots $p_1, \ldots, p_k \in V$, such that the results obtained by solving an SSSP from every pivot are representative for solving it from every vertex in V.

The most straightforward strategy, call it Random, is to select the pivots uniformly at random. Since high-degree vertices are likely to be hubs in many shortest paths, a potentially useful alternative is to choose pivots with a probability proportional to their degree. This strategy will be called RanDeg.

In the following, deterministic strategies, the first pivot p_1 is chosen uniformly at random from V. For $i = 0, \ldots, k$, let $P_i = \{p_1, \ldots, p_i\}$ be the first i pivots, and $V_i = V \setminus P_{i-1}$ be the set of non-pivots from which p_i may be chosen.

MaxMin To spread pivots uniformly over the entire graph, this strategy selects the next pivots to be as far away from any previous pivot as possible. It thus places a pivot in a region not covered well. Formally, p_i is chosen to be a vertex $v \in V_i$ maximizing

$$\min_{p \in P_{i-1}} d(p, p_i) \ . \qquad (6)$$

This strategy is a well-known 2-approximation (and best possible unless $\mathcal{P} = \mathcal{NP}$) for the k-center problem in facility location, in which the goal is to find a set of k vertices, the centers, such that the distance from any vertex to the closest center is minimized [Hochbaum and Shmoys 1986].

MaxSum Intuitively, the sum of distances is an even better indicator of how badly covered a vertex is by the current set of pivots. We may therefore wish to select the next pivot p_i from V_i by maximizing

$$\sum_{p \in P_{i-1}} d(p, p_i) \tag{7}$$

rather than the minimum. Note that this corresponds to selecting a vertex that is among the most peripheral with respect to the current estimates of closeness centrality.

MinSum The above strategies favor the selection of vertices in the periphery of the graph, thus creating a tendency to overestimate distances. The dual approach of is to choose new pivots to be the most central with respect to the closeness estimate among the non-pivots, i.e. by minimizing (7). Note that this strategy grows a connected set of pivots around the initial one. Since the corresponding variant of MaxMin exhibits the same behavior only with the added randomness of choosing any vertex connect to the current set of pivots, we did not include it in our experiments.

Mixed Note that it is easy to construct examples in which the deterministic strategies are significantly off for at least some vertices, even if the number of pivots is large. To balance systematic errors while hopefully maintaining the desired reduction in the number of pivots needed, we also consider a mixed strategy that combines Random, MaxMin and MinSum in a round-robin fashion.

5 Experiments

We have conducted an extensive suite of experiments on both generated and observed data to assess the quantitative and qualitative behavior of pivot-based centrality estimation. To be able to compute the exact centrality scores for baseline comparison, the experiments are restricted to networks of relatively small size (order of 1,000 vertices and 10,000 edges). See Tab. 2 for a summary.

place Table 2 about here

5.1 Data

There are numerous models for generating random graphs with specific structural characteristics [Baumann and Stiller 2005]. We have selected three of the more common ones.

Random Graphs. The basic random graph model of Gilbert [1959][1] is defined by two parameters, the number of vertices n and an edge probability $0 < p < 1$. Between each of the $binomn2$ pairs of the n vertices, an edge is created with probability p independently. Graphs generated from this model are typically very balanced, with similar vertex degrees, little clustering, and relatively short distances.

Small Worlds. Watts and Strogatz [1998] introduces a model in which a ring of n vertices, in which every vertex is connected to its $2r$ nearest neighbors, is modified by rewiring each edge, randomly and independently, with probability $0 < p < 1$. Despite its sparsity, the initial structure exhibits high local clustering, which is maintained while the average distance is reduced by rewiring.

[1] Note that this model is frequently named after Erdős and Rényi [1959], who introduced a model with essentially equal asymptotic characteristics in which a fixed number of edges is drawn uniformly at random from all pairs of vertices.

Preferential Attachment. A model for generating graphs with heavy-tailed degree distributions is described by Barabási and Albert [1999] and made rigorous by Bollobás et al. [2001]. The n vertices of a graph are added one at a time, and for each of them a fixed number of edges connecting to previously created vertices with probability proportional to their degree.

Efficient algorithms for generating graphs from these models are presented in Batagelj and Brandes [2005]. As for observed data, we selected the following three examples for their varying size, structure, and origin.

place Figure 1 about here

Protein Interaction. This data is taken from Jeong et al. [2001] and consists of proteins found in the yeast *Saccharomyces cerevisiae*. The edges represent protein-protein interactions, and it can be seen in Fig. 1 that the network has a sparse core with many dangling trees. Note that the centrality that the authors argue to be an indicator of lethality is degree centrality, i.e. simply the number of edges incident to a vertex.

place Figure 2 about here

Needle Exchange. Valente et al. [1998] study a network of intravenous drug users participating in a needle exchange program. Edges indicate that one person obtained a needle that another one returned. Even though the data gives rise to a weighted multigraph, we only use its simple undirected version. Except for a significant number of degree-one vertices, this network has much fewer biconnected components than the protein interaction network. Hence, there is a qualitative difference in distances and path numbers.

place Figure 9 about here

Ticker News. Reuters ticker news following the terrorist attacks of September 11, 2001, have been transformed into a network text representation proposed by Corman et al. [2002]. Vertices represent words appearing in noun phrases, and edges are introduced between pairs of vertices that appear in the same noun phrase, or consecutively within a sentence. By construction, these networks have very few dangling tree structures, and many locally dense subgraphs (see Fig. 9). This is the only graph for which multiple edges are used in the betweenness computations; they have no relevance for closeness.

5.2 Method

Since the speed-up obtained is directly proportional to the number of pivots, implementation details and actual running times are irrelevant for its assessment.

For each combination of six graphs and six strategies we carried out twenty repetitions of the following experiment. The vertices of the graph are ordered according to the pivot strategy, and divided in twenty intervals to produce increasingly large sets of pivots. For each of these sets, the centrality estimates are computed. In the experiments on generated graphs, a new one is generated for each run.

Since one is mostly interested in the centrality ranking of a network, the results of each experiment are scaled to sum to one. This way, we do not have to worry about systematic under- or overestimation, sample sizes, or normalization of centralities.

The normalized centrality indices obtained for the different strategies are compared to the exact centrality index using their Euclidean and also the inversion distance. The Euclidean distance is to assess the overall deviation in relative scores, and the inversion distance, i.e. the number of pairs that are in wrong rank order, is to assess the usefulness of the estimates in ordering the vertices according to their centrality. Though the numerical values may be far off, it could be that the ranking is already accurate, and vice versa.

5.3 Results

The results of the above experiments are presented in Figs. 3–8.

For random graphs (Fig. 3), the results are mostly as expected or even hoped for. All strategies yield accurate estimates already with few pivots. So most of the computation in exact algorithms is spent on minor improvements. Moreover, the deterministic strategies choosing peripheral vertices outperform random selection, if only slightly. It is no surprise that Random and RanDeg perform similarly, since the degree variance is small in random graphs.

The situation is entirely different for small worlds and preferential attachment graphs (Figs. 4 and 5). While MaxMin yields the most accurate results for small numbers of pivots in small worlds, it becomes one of the worst strategies when the number of pivots is increased. For preferential attachment graphs it is outperformed almost immediately. Random strategies, on the other hand, are surprisingly consistent on both classes of graphs. They exhibit essentially the same behavior as on random graphs. The most striking observation, however, is the performance of MinSum for betweenness on preferential attachment graphs, where the worst numerical estimates yield the best rankings. We have no convincing explanation so far.

Supporting the motivation behind those models, the results on observed data do not resemble those on random graphs. In particular, Random appears to be the most reliable choice. The protein-protein interaction network causes the deterministic strategies to rank with irregular quality, most likely because of its many dangling trees. See Fig. 6 and also Fig. 1, which confirms that the initial pivots are placed in leaves of such trees, causing overestimation for vertices on the path to the center, and underestimation for those in the center. Figure 8 also shows that the variance over different runs is small for all strategies (recall that the first pivot is selected at random).

Again, we see that MinSum performs well in terms of inversion distance for betweenness. Given that all three observed networks have a noteworthy number of high-degree nodes, this is at least consistent with the observation for preferential attachment graphs. The reason for the counter-intuitive quality reduction for larger numbers of pivots on the ticker news text network is illustrated in Fig. 9. Observe that after filling the center with the first 1,000 pivots, MinSum continues to grow the connected set of pivots, but this extension is forced to fill a region of the graph that yields unbalanced contributions to all vertices.

place Figures from Fig. 3 about here

6 Conclusion

We have conducted a series of experiments to assess the practicality of heuristic methods for centrality computation.

Our experiments suggest that selecting pivots uniformly at random is superior to more sophisticated selection strategies, because structural imbalance present in most networks cause deterministic strategies to run into traps, even to the point that the estimates become worse when adding more pivots.

It is also important to note that, experimentally, the accuracy of random pivot selection is largely monotonic in the number of pivots used, and that the variance in quality over different runs is very small.

An alternative strategy to improve over our estimates is to use more sophisticated techniques than our simple random sampling estimators [Thompson 2002]. While we have not performed a thorough study, it seems, though, that reasonable and efficient estimators are difficult to design and subject to the same problems exhibited by skewed pivot selection strategies.

Since we can compute the exact closeness centrality of any particular vertex by solving one SSSP, a reasonable strategy to determine the k most central vertices is to estimate closeness using a sufficiently large number of pivots, followed by exact computations for those vertices ranked among the top k', $k' > k$, to determine their correct order. Note that this approach does not apply to betweenness centrality.

Acknowledgment. We thank Simon Endele for his help in running the experiments and Eric Kolaczyk for interesting comments.

References

Anthonisse, J. M. (1971). The rush in a directed graph. Technical Report BN 9/71, Stichting Mathematisch Centrum, 2e Boerhaavestraat 49 Amsterdam.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286:509–512.

Batagelj, V. and Brandes, U. (2005). Efficient generation of large random networks. *Physical Review E*, 71(036113).

Baumann, N. and Stiller, S. (2005). Network models. In Brandes and Erlebach [2005], pages 341–372.

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

Bollobás, B., Riordan, O. M., Spencer, J., and Tusnády, G. (2001). The degree sequence of a scale-free random graph process. *Randoms Structures and Algorithms*, 18:279–290.

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

Botagfogo, R. A., Rivlin, E., and Shneiderman, B. (1992). Structural analysis of hypertexts: Identifying hierarchies and useful metrics. *ACM Transactions on Information Systems*, 10(2):142–180.

Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2):163–177.

Brandes, U. and Erlebach, T., editors (2005). *Network Analysis*, volume 3418 of *Lecture Notes in Computer Science*. Springer-Verlag.

Brandes, U. and Fleischer, D. (2005). Centrality measures based on current flow. In *Proceedings of the 22nd International Symposium on Theoretical Aspects of Computer Science (STACS'05)*, volume 3404 of *Lecture Notes in Computer Science*. To appear.

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

Carpenter, T., Karakostas, G., and Shallcross, D. (2002). Practical issues and algorithms for analyzing terrorist networks. Invited paper, Western Multi-Conference (WMC 2002).

Corman, S. R., Kuhn, T., McPhee, R. D., and Dooley, K. J. (2002). Studying complex discursive systems: Centering resonance analysis of communication. *Human Communication Research*, 28(2):157–206.

Eppstein, D. and Wang, J. (2004). Fast approximation of centrality. *Journal of Graph Algorithms and Applications*, 8(1):39–45.

Erdős, P. and Rényi, A. (1959). On random graphs I. *Publicationes Mathematicae Debrecen*, 6:290–297.

Freeman, L. C. (1977). A set of measures of centrality based upon betweeness. *Sociometry*, 40:35–41.

Freeman, L. C., Borgatti, S. P., and White, D. R. (1991). Centrality in valued graphs: A measure of betweenness based on network flow. *Social Networks*, 13(2):141–154.

Gilbert, E. N. (1959). Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144.

Harary, F. and Hage, P. (1995). Eccentricity and centrality in networks. *Social Networks*, 17:57–63.

Hochbaum, D. S. and Shmoys, D. B. (1986). A unified approach to approximation algorithms for bottleneck problems. *Journal of the Association for Computing Machinery*, 33(3):533–550.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):713–721.

Jeong, H., Mason, S. P., Barabási, A.-L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411:41–42. Brief communications.

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

Koschützki, D., Lehmann, K. A., Peeters, L., Richter, S., Tenfelde-Podehl, D., and Zlotowski, O. (2005a). Centrality indices. In Brandes and Erlebach [2005], pages 16–61.

Koschützki, D., Lehmann, K. A., Tenfelde-Podehl, D., and Zlotowski, O. (2005b). Advanced centrality concepts. In Brandes and Erlebach [2005], pages 83–111.

Newman, M. E. J. (2005). A measure of betweenness centrality based on random walks. *Social Networks*, 27:39–54.

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

Shimbel, A. (1953). Structural parameters of communication networks. *Bulletin of Mathematical Biophysics*, 15:501–507.

Thompson, S. K. (2002). *Sampling*. John Wiley & Sons, 2nd edition.

Valente, T. W., Foreman, R. K., Junge, B., and Vlahov, D. (1998). Satellite exchange in the Baltimore needle exchange program. *Public Health Reports*, 113(S1):90–96.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature*, 393:440–442.

strategy	**rule**
Random	uniformly at random
RanDeg	random proportional to degree
MaxMin	non-pivot maximizing minimum distance to previous pivots
MaxSum	non-pivot maximizing sum of distances to previous pivots
MinSum	non-pivot minimizing sum of distances to previous pivots
Mixed	alternatingly MaxMin, MaxSum, and Random

Table 1: Pivot-selection strategies (first pivot selected at random)

network	**n**	**m**	**source**
random graphs	1,000	≈10,000	Gilbert [1959]
small worlds	1,000	10,000	Watts and Strogatz [1998]
preferential attachment	1,000	20,000	Barabási and Albert [1999]
protein interaction	2,114	4,480	Jeong et al. [2001]
needle exchange	4,259	61,693	courtesy of R. Foreman and T. Valente
ticker news	13,332	148,039	courtesy of S. Corman

Table 2: Networks used in the experiments

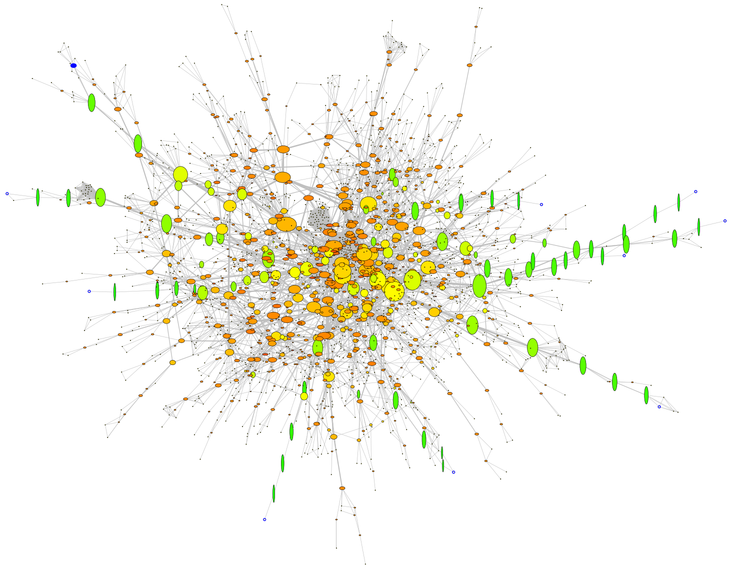


Figure 1: Protein interaction network. Node dimensions indicate exact (width) and estimated (height) closeness centrality using MaxMin for pivot (blue) selection. Other colors emphasize under- (red) and overestimation (green)

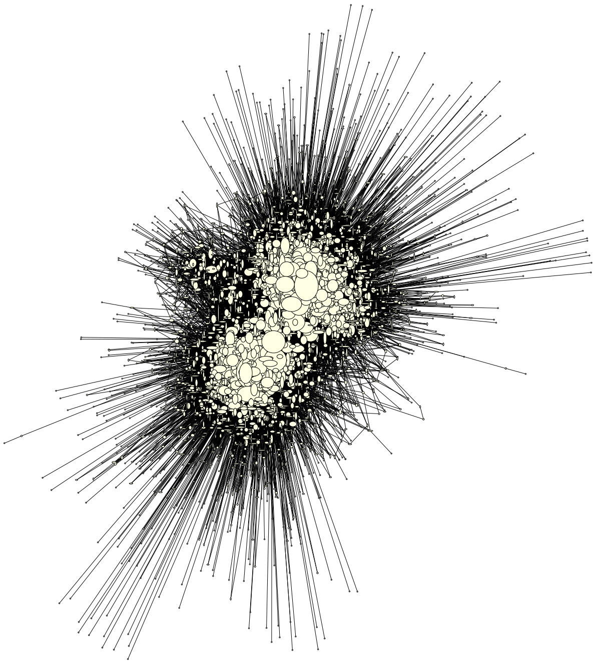


Figure 2: Needle exchange network (the apparent clustering is caused by two established and one recently opened exchange location)

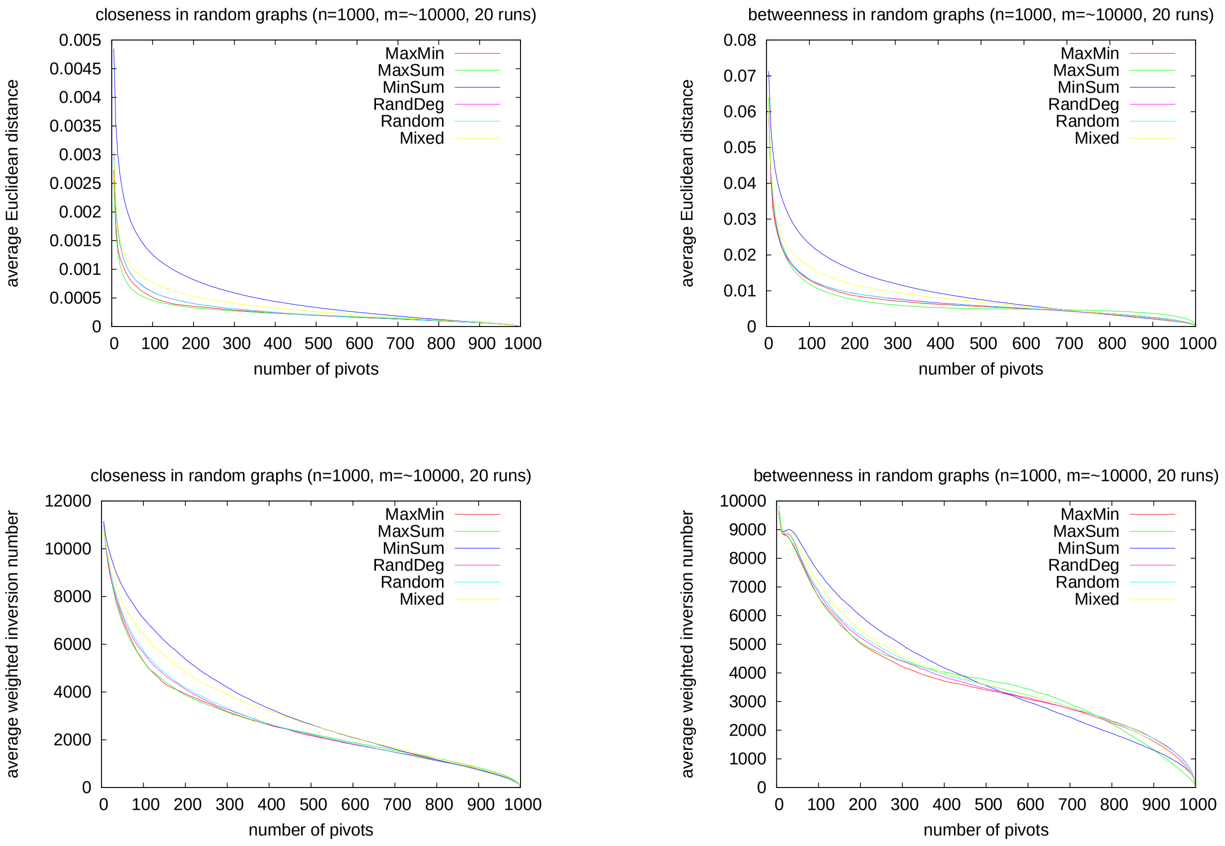


Figure 3: Centrality estimation in random graphs

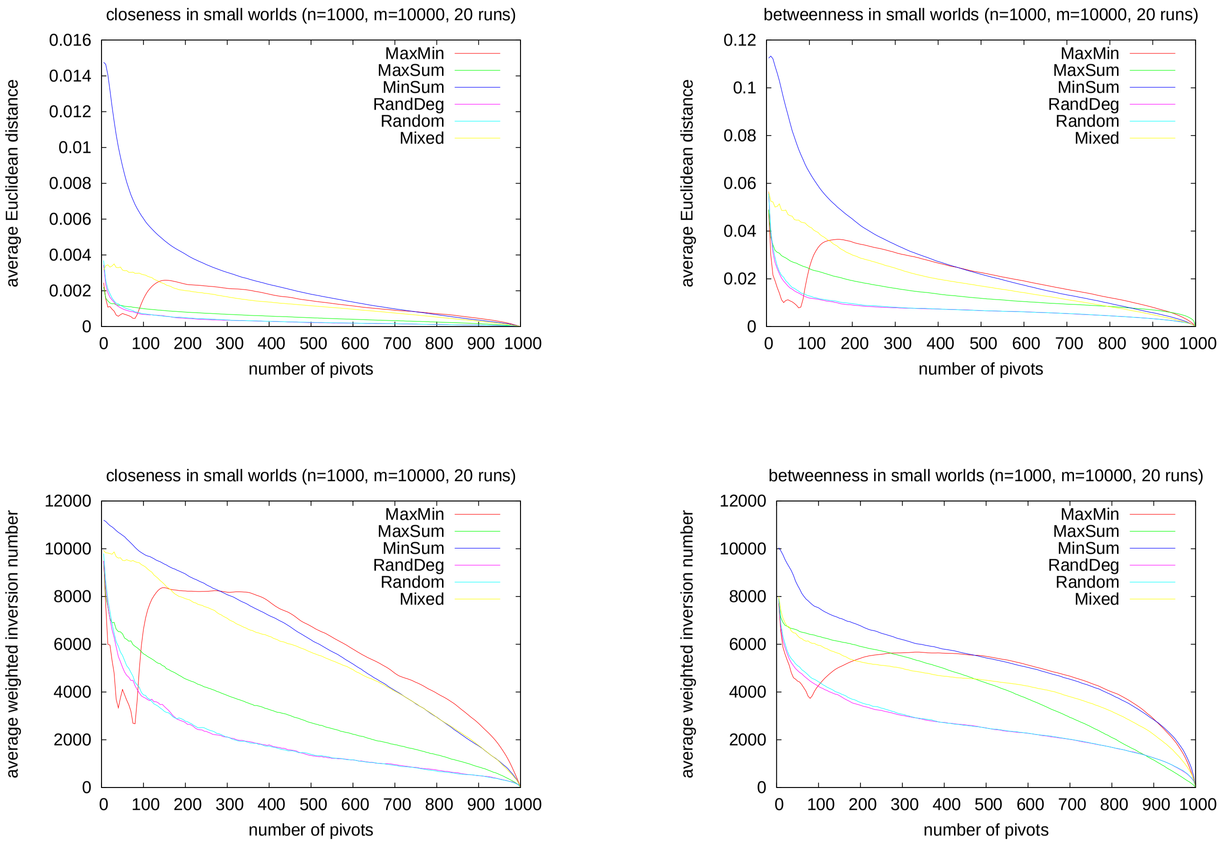


Figure 4: Centrality estimation in small world graphs

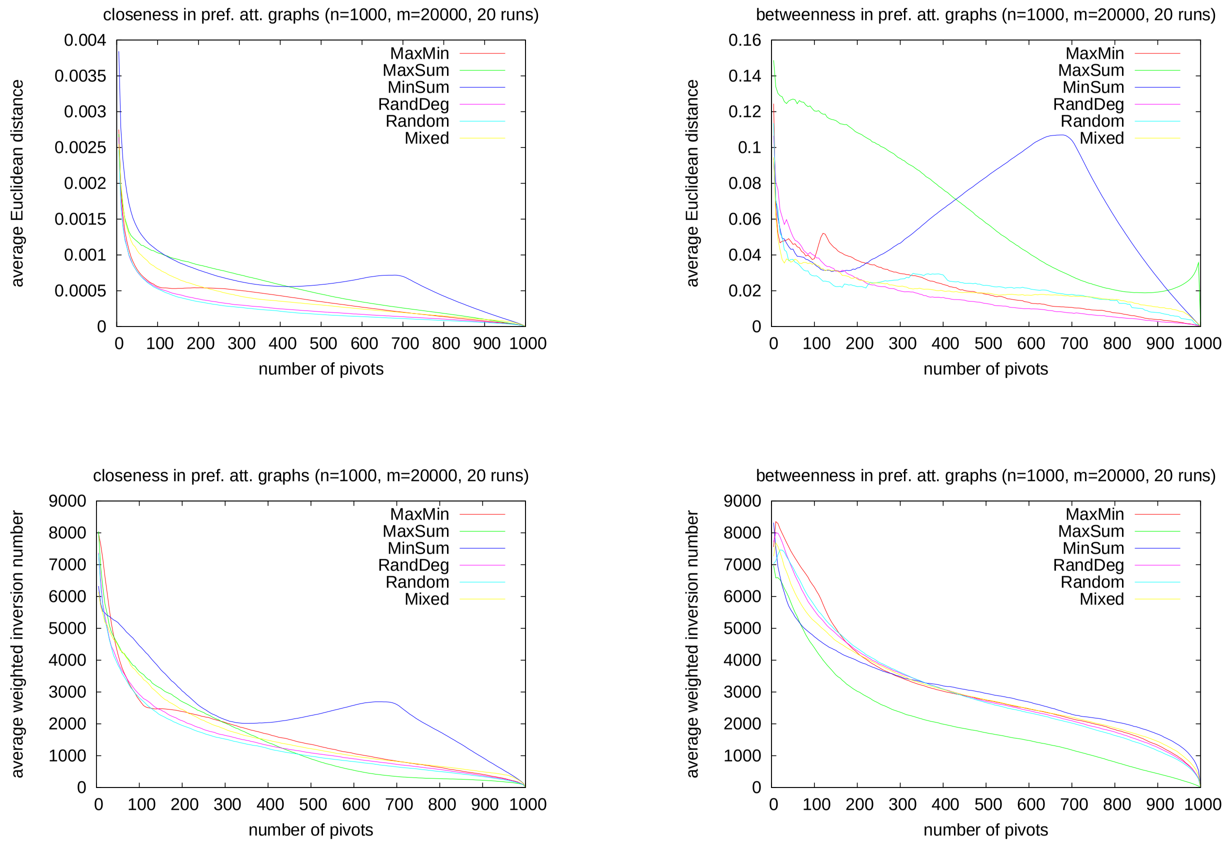


Figure 5: Centrality estimation in preferential attachment graphs

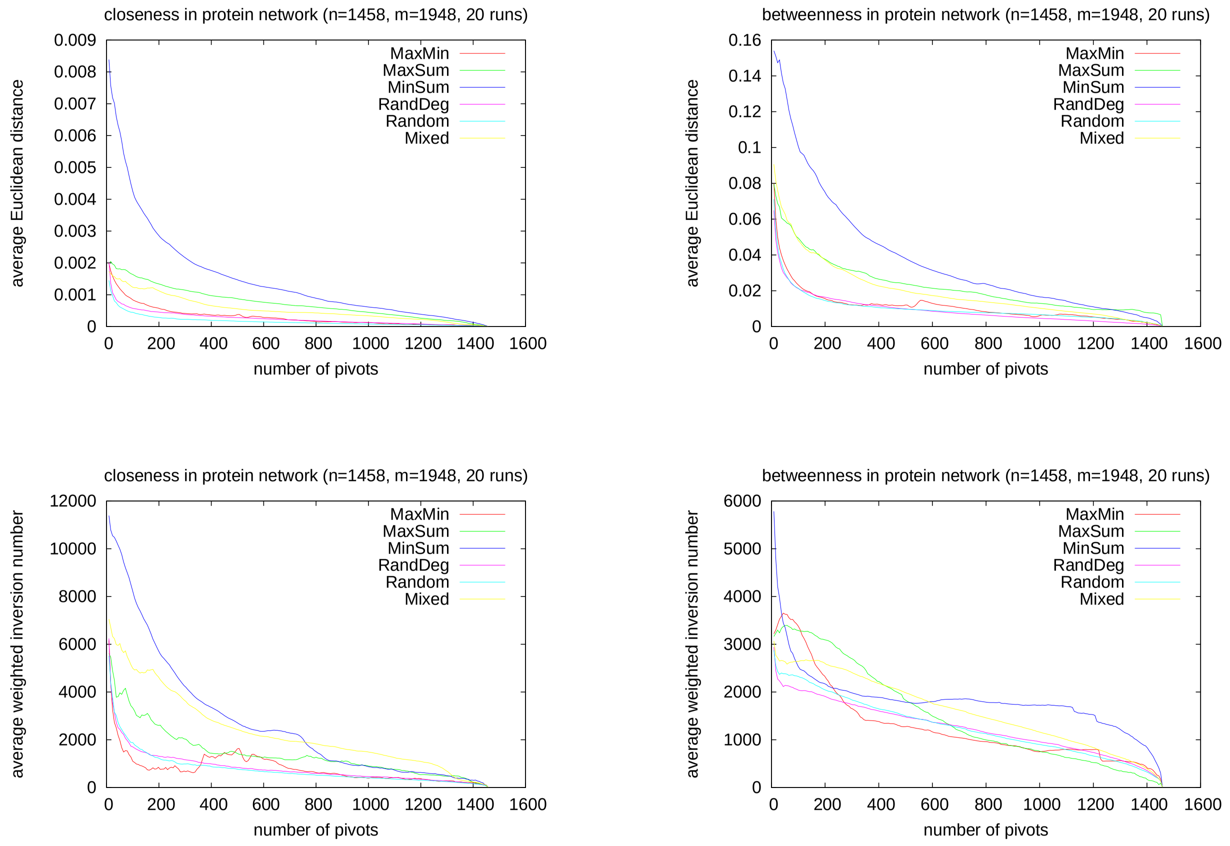


Figure 6: Centrality estimation in protein interaction network

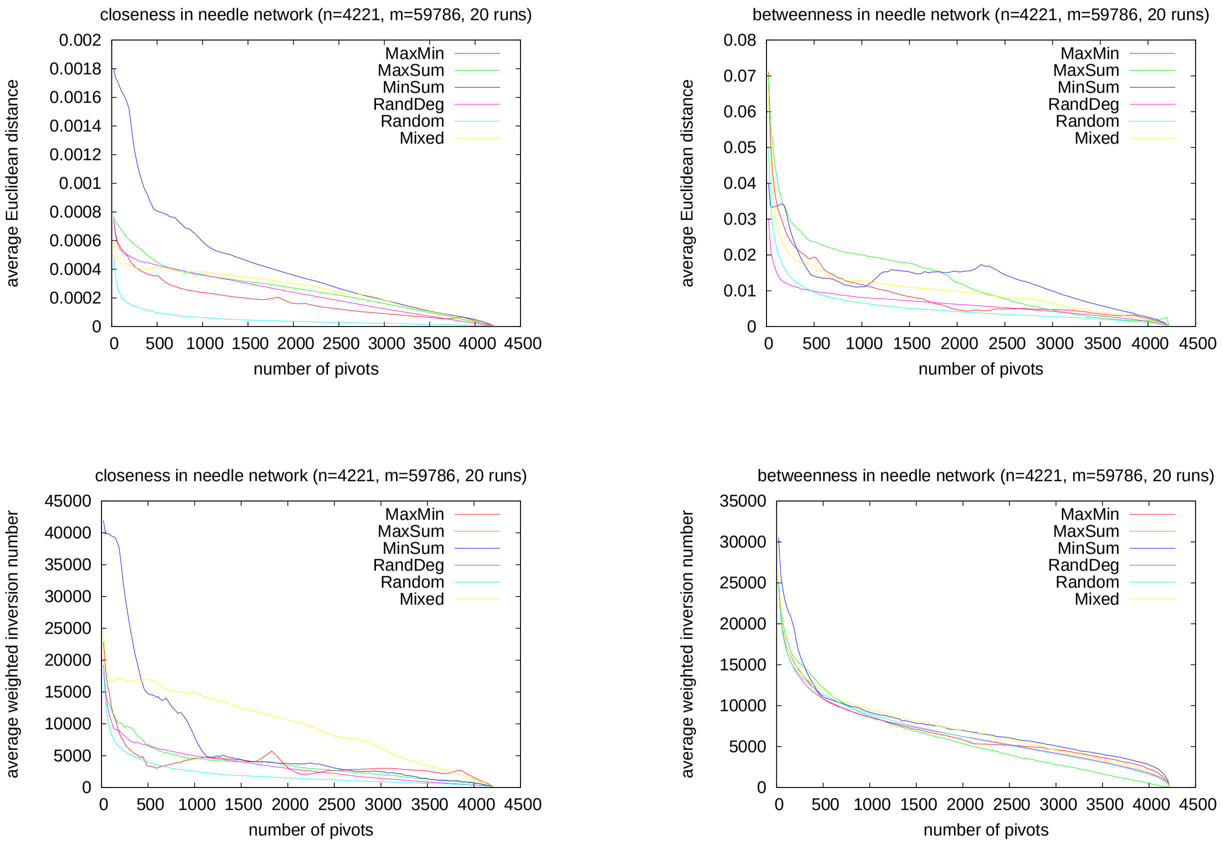


Figure 7: Centrality estimation in needle exchange network

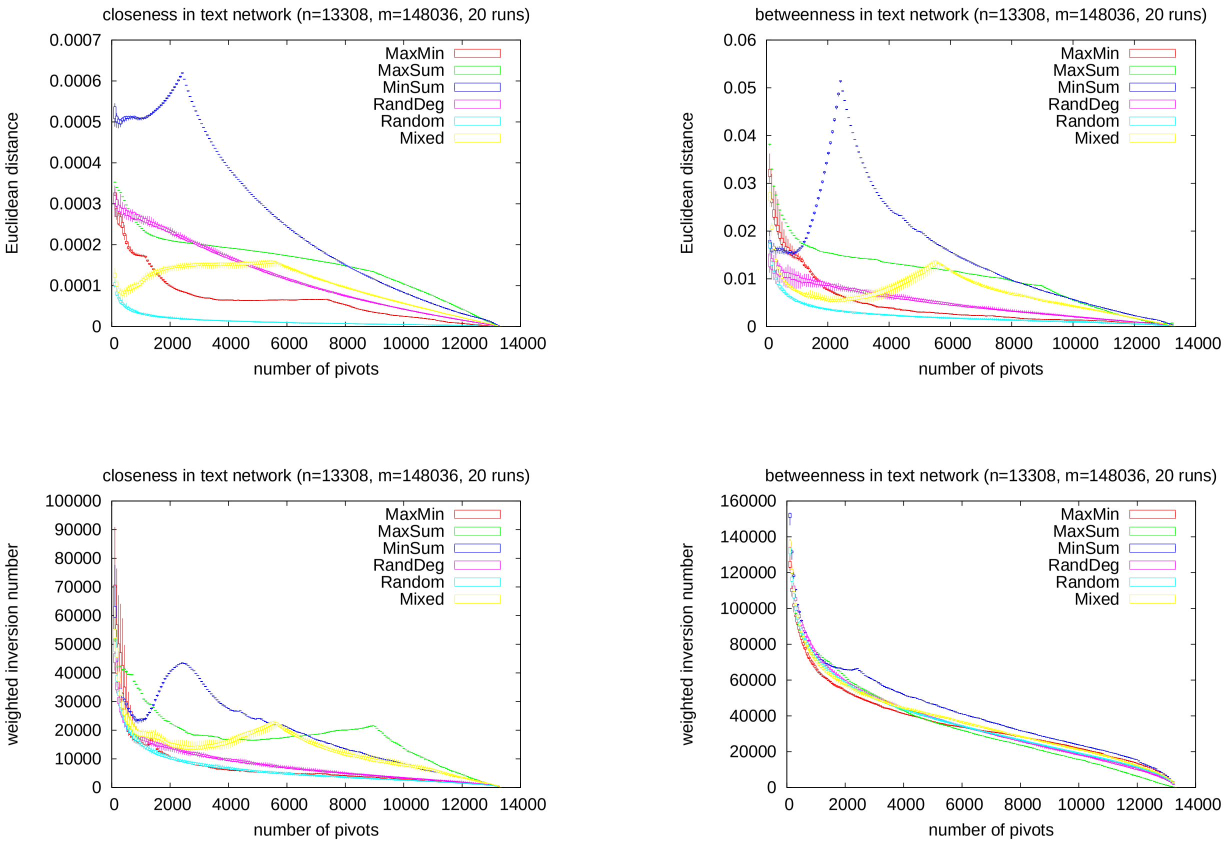
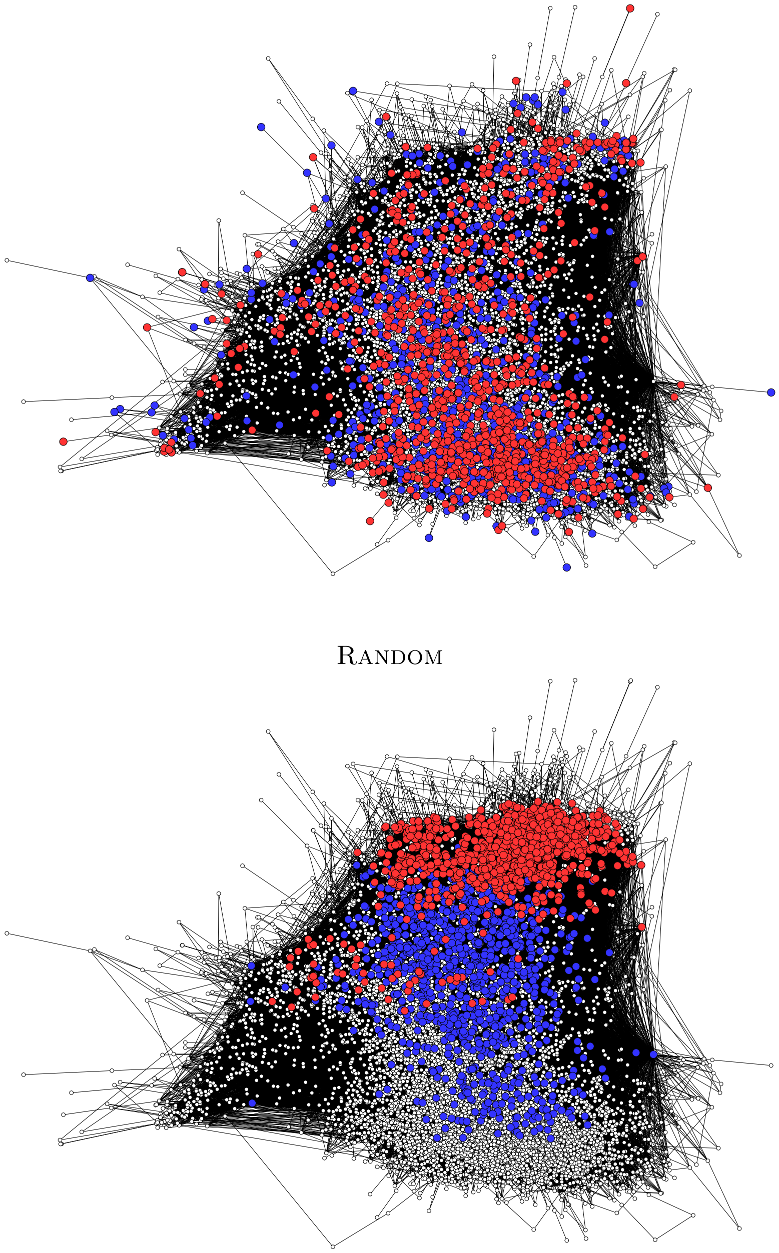


Figure 8: Centrality estimation in ticker news text network (box and whiskers)



Random

MinSum

Figure 9: MinSum fails to utilize larger number of pivots on ticker news network (white – non-pivots, blue – first 1,000 pivots, red – next 1,000 pivots)