LLNL-TR-741432

Improving Estimation of Betweenness Centrality for Scale-Free Graphs *

Seth Bromberger†, Christine Klymko†, Keith Henderson†, Roger Pearce†, Geoff Sanders†
Lawrence Livermore National Laboratory

Abstract

Betweenness centrality is a graph statistic used to find vertices that are participants in a large number of shortest paths in a graph. This centrality measure is commonly used in path and network interdiction problems and its complete form requires the calculation of all-pairs shortest paths for each vertex. This leads to a time complexity of $O(|V||E|)$, which is impractical for large graphs. Estimation of betweenness centrality has focused on performing shortest-path calculations on a subset of randomly-selected vertices. This reduces the complexity of the centrality estimation to $O(|S||E|), |S| < |V|$, which can be scaled appropriately based on the computing resources available. An estimation strategy that uses random selection of vertices for seed selection is fast and simple to implement, but may not provide optimal estimation of betweenness centrality when the number of samples is constrained. Our experimentation has identified a number of alternate seed-selection strategies that provide lower error than random selection in common scale-free graphs. These strategies are discussed and experimental results are presented.

†{seth, klymko1, henderson43, pearce7, sanders29}@llnl.gov

1 Introduction

One question that often arises in the study of complex networks is that of identifying the most important vertices or edges in the network. Although the meaning of "importance" can vary from application to application, the most important network elements can often be identified through the use of *centrality measures* [3, 7]. One commonly used centrality measure is *betweenness centrality*, which counts the percentage of shortest paths in the network which traverse through any given node. This measure is often a very powerful tool for finding important nodes (or edges) but it is very expensive to calculate, since it requires the computation of the shortest paths between every pair of vertices in the network. Due to this, approximation methods for betweenness centrality, many of which make use of sampling strategies, have been developed over the past decade. However, it is still unclear what is the best sampling strategy. In this work, we investigate a variety of strategies and compare them to those found in existing literature.

The paper is organized as follows. In Section 2 we introduce notation and set up the problem. Section 3 provides an overview of current methods for sampling betweenness centrality. The methodology used in our experiments can be found in Section 4 and our results in Section 5. Other results not directly related to this specific research are detailed in Section 6. Finally, conclusions and strategies for future investigation into this problem are in Section 7.

2 Definitions and Notation

Let $G = \{V, E\}$ be a graph comprised of a set of vertices V and a set of edges $E \subseteq V \times V$ between vertices. The betweenness centrality of a vertex $v \in V$ is defined as

$$BC_v = \sum_{u \neq w \neq v} \frac{\sigma_{uw}(v)}{\sigma_{uw}} \tag{1}$$

where $\sigma_{st}(v)$ is the number of shortest paths between vertices s and t passing through v, and σ_{st} is the total number of shortest paths between s and t in G.

Because a true betweenness centrality calculation must calculate all-pairs shortest paths for G, it has a worst-case performance of $O(|V||E|)$ using Dijkstra's algorithm. The practical implication of this superlinear performance is that completion of the calculation becomes infeasible as the graph grows. As a result, estimations of betweenness centrality are used when the true betweenness centrality measures are too expensive.

The standard estimation strategy for betweenness centrality is to reduce the number of shortest paths calculated by selecting a subset of vertices (the "seed

set") $S \subset V$ at random, and using the shortest paths from each $v \in S$ to all other vertices in V to estimate the betweenness centrality for all vertices in G. In all strategies using a seed set, the betweenness centrality for each node is obtained using only partial information, but with a performance improvement of $O((|V| - |S|)|E|)$. For large real-world graphs, the seed set is measured in fractions of a percent of $|V|$ in order to produce an estimation that can be performed in reasonable time on modern computing hardware.

The (in)accuracy of betweenness centrality estimation strategies may be expressed by comparing the top k-ranked vertices between the estimation and the true betweenness centrality.[1] A common method of measuring the differences between the two sets of top k-ranked vertices is to use the Jaccard distance:

$$d_J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

This equation yields a real number between 0 and 1, where numbers closer to 0 indicate greater similarity (and therefore greater estimation accuracy), and numbers closer to 1 indicate greater dissimilarity (and therefore greater estimation error).

3 Related Work

One of the earliest papers discussing the estimation of betweenness centrality is [4]. This paper introduced the idea of sampling a small number vertices, running single source shortest path (SSSP) from these vertices, and estimating the betweenness centrality of all vertices in the network from this data. The authors investigated a number of sampling strategies, including random, maximizing the minimum distance between samples, maximizing the sum of the distances between samples, minimizing the sum of the distances between samples, and variants where a mix of these methods are used. Among these strategies, the random sampling consistently outperformed others on a variety of generated and real world networks.

Around the same time, the authors of [2] introduced an adaptive sampling technique for the approximation of betweenness centrality. They show that, once enough samples are taken, with high probability the algorithm based on this adaptive sampling technique achieves a good estimate of the betweenness centrality of high centrality nodes. Through experiments on a number of gen-

[1] In most cases, the actual betweenness centrality value is less important than the ordering of the vertices implied by the value: vertices with higher betweenness centrality values have a higher betweenness ranking.

erated and real world networks, they show that this generally holds for low centrality nodes as well.

In [5], the authors explore the use of graph coarsening for the approximation of betweenness centrality. They focus on speedup and numerical accuracy of the approximation on Erdös-Rényi graphs of 1000-2000 vertices.

All of the above papers are concerned with approximating the true values of the betweenness centrality of nodes in the network with high numerical precision. However, as was mentioned earlier, often it is not the numerical accuracy but the relative ranking of nodes in the network that is the main concern when calculating betweenness centrality. One of the few papers which addresses this issue is [8]. In this work, the authors introduce algorithms both to estimate the betweenness of all vertices in the graph with high numerical accuracy and to identify a superset containing the top k vertices in the network with high probability. Both algorithms are based on random sampling of shortest paths in the network. The number of shortest paths needed varies with the desired accuracies and various graph properties.

In [1], the authors introduce "k-path centrality," which is defined for node v as the sum over all possible source nodes s of the probability that a walker originating from s goes through node v, assuming the walker only moves along random simple paths of at most length $k + 1$. They show that a high k-path centrality is correlated with a high betweenness centrality (as well as being much faster to calculate) through a series of experiments on synthetic and real-world networks.

4 Methodology

Our objective was to determine through experimentation whether or not there exists a seed selection strategy for betweenness centrality estimation that results in better accuracy (lower error) in identifying the top k nodes than the commonly-used random seed selection strategy. We compared different strategies by using the Jaccard distance (2) across various numbers s of seeds to measure (dis)similarity among the top k-ranked vertices: that is, we sought to measure how well a given strategy identified the set of k vertices with the highest betweenness centrality.

We concerned ourselves with "real-world" graphs, each of which has scale-free properties but differs in other metrics, such as order, size, and clustering coefficient. The five graphs upon which we evaluated each strategy were all taken from the Stanford Large Network Dataset Collection [6] and are listed in Table 1.

graph name	vertices	edges
as-caida	26 475	53 381
ca-astroph	17 903	197 031
ca-condmat	21 363	91 342
ego-twitter	81 306	1 768 149
email-enron	33 696	180 811
facebook-combined	4 039	88 234
soc-slashdot0902	81 268	582 533

Table 1: List of graphs selected for evaluation

We devised several estimation strategies for our experimental testing:

- ***random selection***: seeds are selected randomly from the set of vertices in the graph.
- ***descending degree***: seeds are selected in descending order based on their degree.
- ***ascending degree***: seeds are selected in ascending order based on their degree.
- ***preferential selection 1***: randomly select an edge from the graph, and then pick the lower-degree vertex in that edge. This strategy has the effect of picking low-degree vertices that are incident to hubs (vertices of high degree).
- ***preferential selection 2***: like *preferential selection 1*, but re-rank the seeds after selection based on the difference in degree between the seed and its incident vertex.
- ***degree-seeded betweenness***: estimate betweenness centrality using the top 100 vertices chosen based on descending degree. Select seeds based on that descending estimated betweenness centrality.
- ***descending distance-based sweep***: from an initial seed selected at random, iteratively select seeds by calculating the distances between the set of selected seeds each other vertex and selecting as the next seed the vertex with the largest minimum distance from the existing seed set.

The complexity of each strategy may be calculated in terms of the number of seeds s, as shown in Table 2.

[2] c in this case is the number of top-degree vertices we select for the initial betweenness centrality calculation.

strategy	complexity
random selection	$O(s)$
descending degree	$O(\|V\| \log \|V\|)$
ascending degree	$O(\|V\| \log \|V\|)$
preferential selection 1	$O(s)$
preferential selection 2	$O(s + s \log s)$
degree-seeded betweenness[2]	$O(c\|E\| + \|V\| \log \|V\|)$
descending distance-based sweep	$O(s\|E\| + s)$

Table 2: Complexity of evaluated strategies

For each graph, we calculated the normalized "ground truth" betweenness centrality (yielding, for each vertex, a metric in the range $(0, 1]$ and ranked the vertices by descending centrality. We then applied each estimation strategy in turn across a number of seeds s and ranked the vertices by descending estimated centrality.[3]

For each set of seeds, we then calculated the Jaccard distance (2) between the ground truth ranking and the estimation ranking for the top k vertices, where $k \in \{10, 100, 1000\}$. By varying s and k, we were able to compare the changes in accuracy across different seed counts and different top-k ranks.

5 Results

The results of running each strategy for $k = 100$ and s from 1 to $0.01|V|$ are given in Figure 1.

For $k = 100$, four of the sampling strategies — *ascending degree*, *preferential selection 1*, *preferential selection 2*, and *distance-based sweep* — showed little advantage over random selection across the range of seed sets. In contrast, two strategies — *descending degree* and *degree-seeded betweenness* (shown in blue in Figures 1 through 3) — provided (sometimes significant) improvements to estimation accuracy over random seed selection, with minimal increases to complexity. These results remained largely consistent for $k = 10$ and $k = 1000$ as shown in Figures 2 and 3.

These results suggest that vertices with higher degree provide more information to the betweenness centralities of other vertices in a graph.

[3]For data collection, we selected $s <= 0.1|V|$, but our analysis focused on a much smaller seed size $s <= 0.01|V|$, which is more appropriate for the graphs we expect to estimate in real work.

6 Other Observations

In general, when s was between approximately $0.0002|V|$ and approximately $0.02|V|$, seed selection using the descending degree or degree-seeded betweenness strategies yielded results better than random seed selection for estimation of the top $k = 100$ highest-centrality vertices. When s was between approximately $0.02|V|$ and approximately $0.035|V|$, random seed selection began to outperform both descending degree and degree-seeded betweenness, and remained the optimal seed selection strategy for larger values of s. (Figure 4 shows examples of this behavior.)

7 Conclusions and Further Work

We have identified several questions as a result of this experiment that could serve as the basis for followup efforts:

- Understanding the properties of the graphs that contribute to higher estimation error (as seen in the `ca-astroph` graph as compared with the `as-caida` graph) could help improve estimation strategies.
- During the research, we discovered that as the average degree for vertices increased, the degree-based strategies' advantage over random seed selection decreased. This perhaps makes intuitive sense but having a formal explanation as to why this is the case could be beneficial.
- Finally, future efforts could focus on validating these results on larger graphs, as well as trying to determine mathematical bounds on the estimations. Additional strategies might be developed. It is expected that any new successful strategies will be derived from descending degree.

Acknowledgements

The authors would like to thank James Fairbanks of Georgia Tech Research Institute for his expertise and contribution to the methodology used in these experiments.

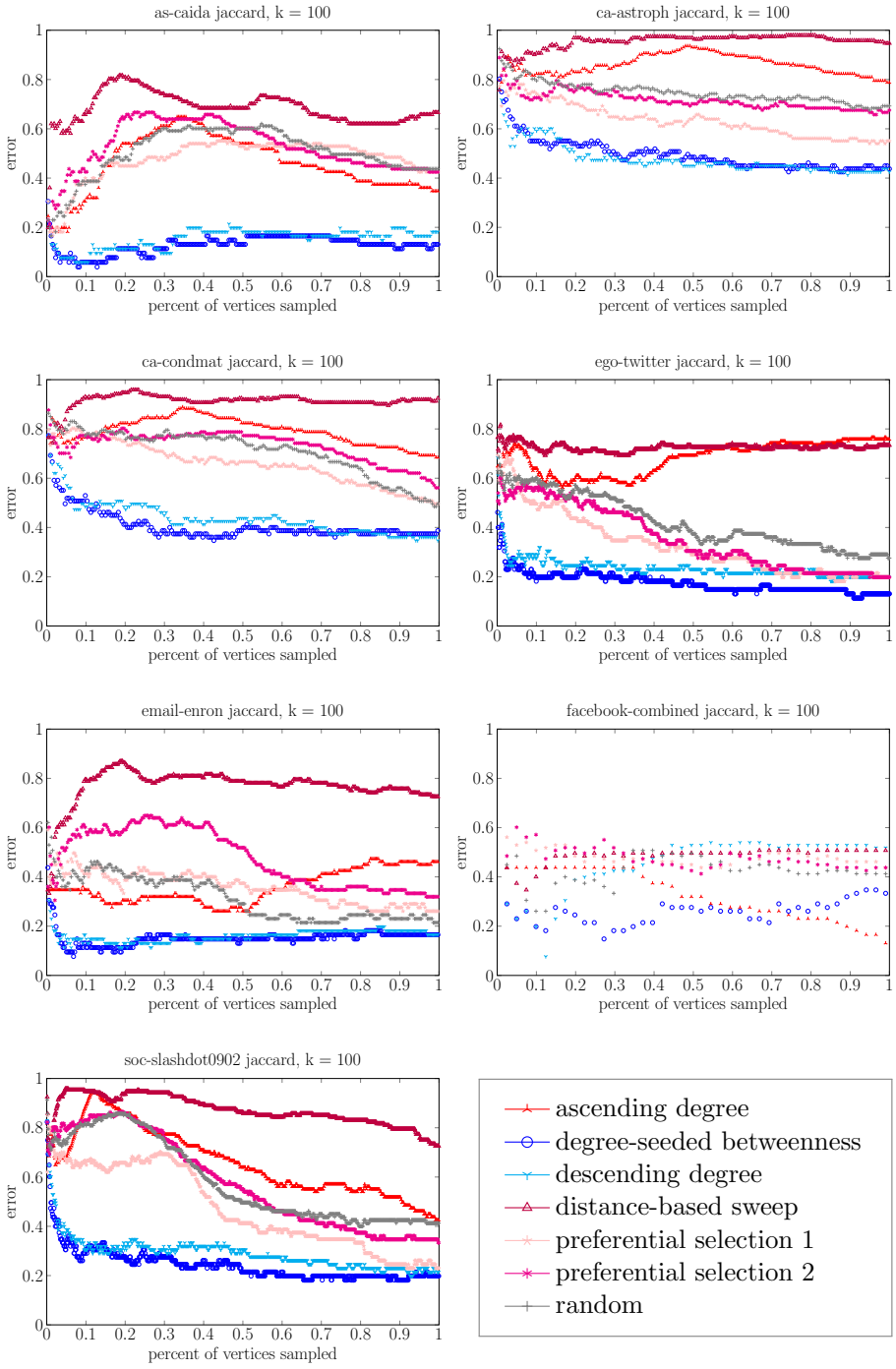

Figure 1: Estimation error of various strategies for the top $k = 100$ most central vertices (as measured by betweenness centrality) as the number of seeds used in the estimation calculation varies. *Degree-seeded betweenness* and *descending degree* strategies generally yield lower errors in all but the smallest graph.

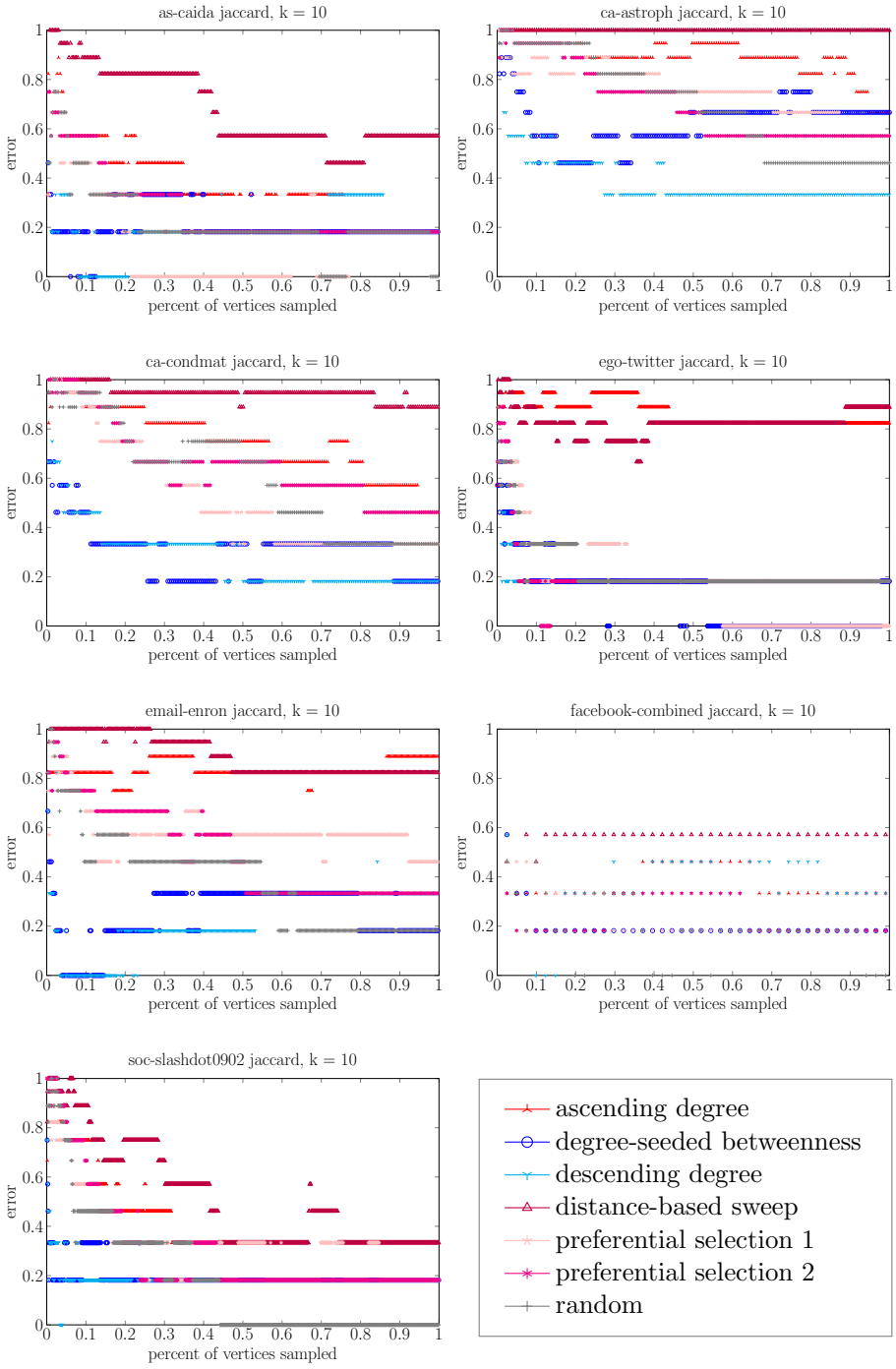

Figure 2: Estimation error of various strategies for the top $k = 10$ most central vertices (as measured by betweenness centrality) as the number of seeds used in the estimation calculation varies. *Degree-seeded betweenness* and *descending degree* strategies generally yield lower errors in all but the smallest graph.

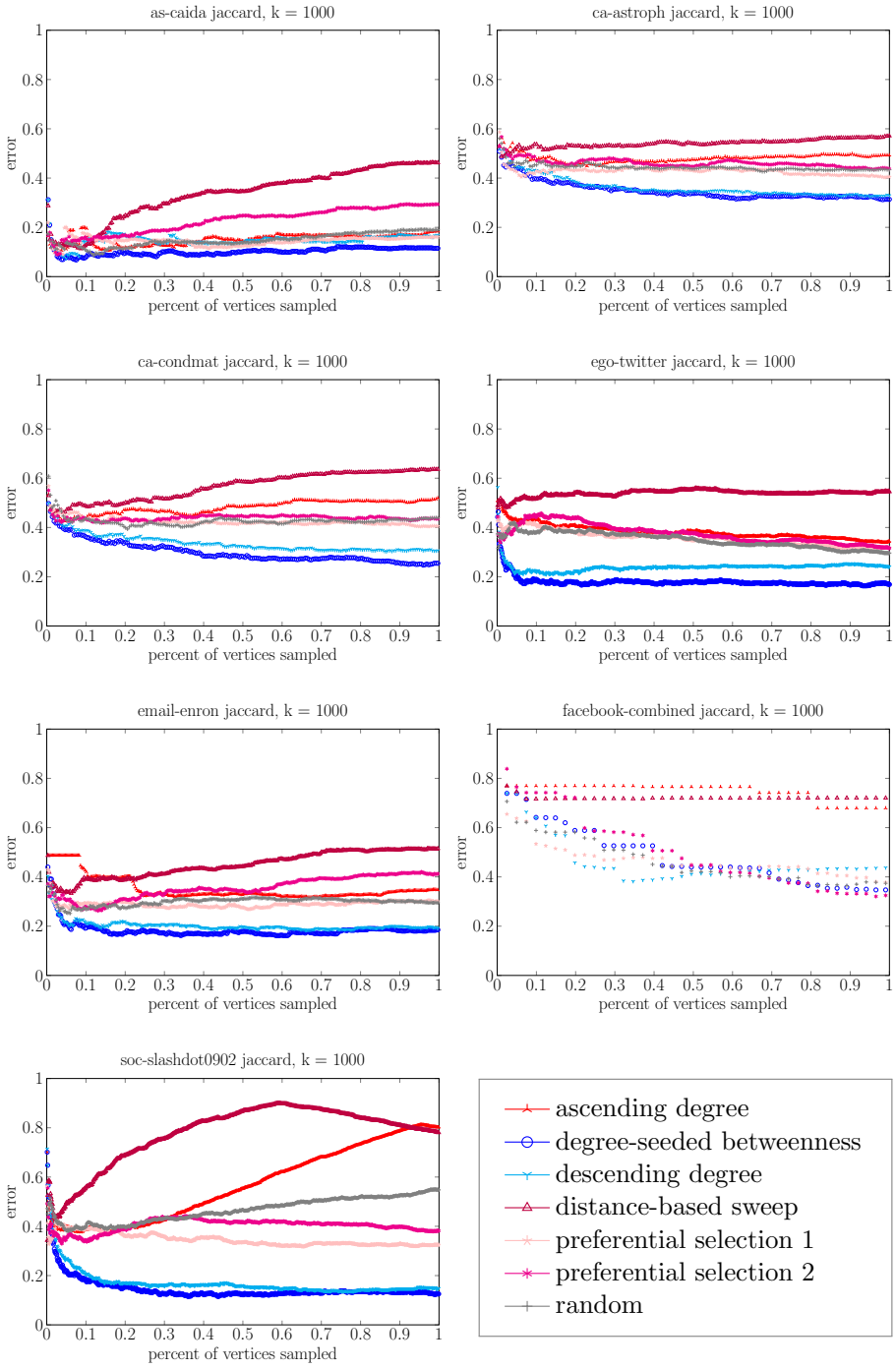

Figure 3: Estimation error of various strategies for the top $k = 1000$ most central vertices (as measured by betweenness centrality) as the number of seeds used in the estimation calculation varies. *Degree-seeded betweenness* and *descending degree* strategies generally yield lower errors in all but the smallest graph.

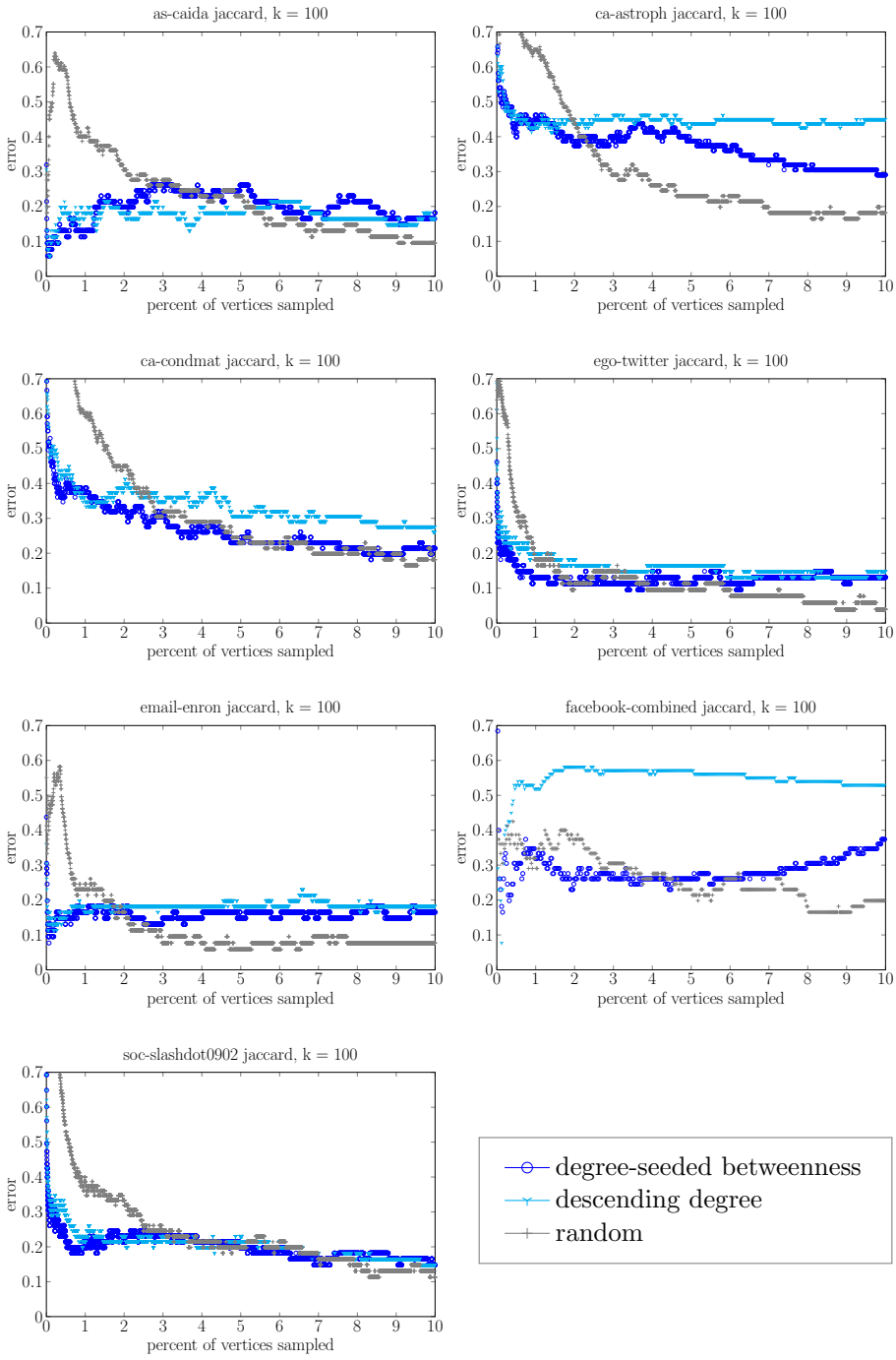

Figure 4: Estimation error of various strategies for the top $k = 100$ most central vertices (as measured by betweenness centrality) as the number of seeds used in the estimation calculation varies up to $0.1|V|$. Random seed selection performs better than *degree-seeded betweenness* and *descending degree* when the number of seeds exceeds $\approx 0.035|V|$.

References

[1] Tharaka Alahakoon, Rahul Tripathi, Nicolas Kourtellis, Ramanuja Simha, and Adriana Iamnitchi. K-path centrality: A new centrality measure in social networks. In *SNS'11*, 2011.

[2] David A. Bader, Shiva Kintali, Kamesh Madduri, and Milena Mihail. Approximating betweenness centrality. In *WAW 2007: Algorithms and Models for the Web-Graph*, pages 124–137, 2007.

[3] Albert-Lázló Barabási. *Network Science*. Cambridge University Press, 2016.

[4] Ulrik Brandes and Christian Pich. Centrality estimation in large networks. *Int. J. Bifurcation Chaos*, 17, 2007.

[5] Mikhail Chernoskutov, Yves Ineichen, and Costas Bekas. Heuristic algorithm for approximation betweenness centrality using graph coarsening. *Procedia Computer Science*, 66:83–92, 2015.

[6] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`, June 2014.

[7] M. E. J. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[8] Matteo Riondato and Evgenios M. Kornaropoulos. Fast approximation of betweenness centrality through sampling. *Data Mining and Knowledge Discovery*, 30(2):438–475, 2016.