2015 48th Hawaii International Conference on System Sciences

Estimating Centrality Statistics for Complete and Sampled Networks: Some Approaches and Complications

Ju-Sung Lee
School of Computer Science
Carnegie Mellon University
and
University of Twente (CSTM)
Email: jusung@cs.cmu.edu

Juergen Pfeffer
School of Computer Science
Carnegie Mellon University
Email: pfeffer@cs.cmu.edu

Abstract

The study of large, "big data" networks is becoming increasingly common and relevant to our understanding of human systems. Many of the studied networks are drawn from social media and other web-based sources. As such, in-depth analysis of these dynamic structures e.g. in the context of cybersecurity, remains especially challenging. Due to the time and resources incurred in computing network measures for large networks, it is practical to approximate these whenever possible. We present some approximation techniques exploiting any tractable relationship between the measures and network characteristics such as size and density. We find there exist distinct functional relationships between network statistics of complex "slow" measures and "fast" measures, such as the linkage between betweenness centrality and network density. We also track how these relationships scale with network size. Specifically, we explore the efficacy of both linear modeling (i.e., correlations and least squares regression) and non-linear modeling in estimating the network measures of interest. We find that sparse, but not severely sparse, networks which admit sufficient entropy incur the most variance in the network statistics and, hence, more error in the estimation. We review our approaches with three prominent network topologies: random (aka Erdős-Rényi), Watts-Strogatz small-world, and scale-free networks. Finally, we assess how well the estimation approaches perform for sub-sampled networks.

1. Introduction

Increasing computing resources have facilitated the collection of larger and larger networks from on-line venues. This evolution has progressed in tandem with the wider availability of such data particularly from social media outlets. However, the capacity to collect data appears to have outstripped the research capabilities to analyze the data in their entirety. Important and revealing network measures, such as betweenness centrality, have such complexity that it can be prohibitively time and resource consuming to calculate them for extremely large networks.

Aside from the obvious solution of increasing one's computing capabilities (and incurring further costs), primary strategies for tackling large network analysis are a) improving the measurement algorithms thereby reducing calculation time and memory usage; b) reducing the amount of data used in the measurement algorithms (yielding estimations rather than exact measures); and c) sampling from the whole network and performing standard network analysis on the sampled network. The matter is further complicated with the increasing need for the analysis of dynamic networks which entails the collection and analysis of multiple network samples across time.

Advancements in centrality measurement have included adjusting extant network calculations in an incremental fashion for changing networks [1] or exploiting characteristics of key (or otherwise) nodes in optimizing computation [2]. Other work that addressed estimation in networks include approximation using local structures from the ego-network [3] [4] and beyond [5] and also approximation for specific structures (e.g., community structures [6]) within desired errors bounds [7].

The alternative and widely-used data reduction strategy is to sample from the data which also lessens the burden imposed on resources by the data. Sampling is particularly attractive for large scale network analysis as many social network analysis (SNA) algorithms have time complexities of $O(n^2)$ or worse. Of course, sampling, especially low sampling, can introduce additional error into the analysis. Understanding the structure of the error may prove valuable in assessing not only the fidelity of results but also in finding a potential mapping between results drawn from the sample and the results that would arise from the true, complete data.

Additional complications arise when the entire data itself is not readily accessible. Hard to reach or otherwise marginal populations are often the topic of study and require the employment of additional network

1530-1605/15 $31.00 © 2015 IEEE
DOI 10.1109/HICSS.2015.203

IEEE computer society

sampling methodologies to address biases in the network samples [8]. In the area of cybersecurity, social network analysis of social media data offers improved assessments of threats by analyzing underground social media activities [9], dynamics between cyber-criminals [10], and topologies of dark networks [11]. These hidden networks (partial or complete) are less easily obtainable than others reported in the literature.

This paper explores some alternative approaches for addressing the above concerns (namely, estimation with and without sampling) by exploiting associations among network measures. We focus on the estimation of measures that are not readily calculable for large networks, i.e., "slow" (or slowly computed) measures such closeness and betweenness centralities, using "fast" measures (e.g. density and degree centrality).

Estimation is not necessarily an end unto itself. It may also be instrumental in determining whether or not more resources should be devoted to the additional collection and analysis of data. For example, estimation might offer an indication of important structural shifts in one or more regions of a network sub-sample or in time snapshots.

2. Analysis

The goal in this paper is to assess the extent to which "fast" centrality measures may predict "slow" measures. Instead of focusing on empirical networks, we survey three prominent topologies which are often the object of studies that endeavor to uncover properties of different kinds of networks, e.g., [12]. Some empirical on-line and social networks have been shown to reflect these topologies.

We will also examine how well our predictive strategies perform for "sub-sampled" networks (i.e., smaller samples drawn from our simulated, whole, topological networks).

2.1. Network Measures

The measures we review are network size n (count of nodes), density d (count of edges divided maximum possible edge count), degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, and the local clustering coefficient. Degree centrality (C_{D}) is simply the sum of edges connecting a node or vertex to its direct neighbors and can indicate the activity level or popularity of a node [13]. Betweenness centrality (C_{B}) is the sum of the proportions of the shortest paths a node lies on for every pair of nodes (out of all shortest paths for each pair). The measure captures the extent to which a node acts as a bridge between distant sets of nodes or clusters of nodes [13]. Closeness centrality (C_{C}) of a node is typically the inverse of the sum of all shortest paths between the node and every other reachable node [13] [14]. However, due to the complications in the measure for disconnected networks (i.e., networks with more than one component), we employ the harmonic sum variant which sums the inverse of shortest paths while assigning a shortest path length of ∞ for unreachable nodes [15] [16]. We find that its behavior is less erratic than the traditional closeness measure. Eigenvector centrality (C_{E}) is derived from the first eigenvector of the network as an adjacency matrix and reflects the extent to which a node is connected to others with high eigenvector centrality scores (i.e., popular others) [17]–[20]. The local clustering coefficient (C_{Δ}) measures the local triangle density [21], or the density of sub-network defined by a node's direct connections and connections among those directly linked nodes.

The four centrality measures are particularly prominent in network research and offer much explanatory power. The first three receive excellent analytical treatment by Freeman [13], and the use of the dominant eigenvector as a centrality measure was elaborated in a social network context by Bonacich [19] [20]. The clustering coefficient is included as it is a node-level measure that is highly relevant to one of the examined topologies. In this paper, we will focus on degree, betweenness, and the clustering coefficient.

2.2. Network Generation

We generate test networks for three prominent topologies: Erdős-Rényi (ER), Watts-Strogatz small-world (WS), and scale-free (SF); our simulated networks are undirected and unweighted. For generating each Erdős-Rényi (ER) graph sample, we randomly draw a tie probability from the uniform distribution, $p \sim \mathrm{Unif}(0, 1)$, which in turn is applied as a Bernoulli probability for independently generating edges between pairs of nodes [22]. The ER stylized graph appears rarely in the real-world; it is largely used for comparison purposes and studied for its interesting properties [23]–[25].

Watts-Strogatz (WS) small world networks are characterized by 1) a high average clustering coefficient and 2) a linear relationship between the average shortest path length, $\mu(L)$, and the logarithm of the network size, n [21]. This topology has been documented to exist in nature, both in social and biological systems, e.g. [26]–[28]. Our WS network samples are initially generated using a circle lattice structure having some fixed $2k$ edges to neighbors where k represents the number of neighbors on each side. For our particular WS network samples, we employ ten half-neighborhood sizes: $k \in \{1, \ldots, 10\}$. Next, we "rewire" (i.e., randomly reassign to two other nodes lacking an edge) the edges with some probability $r \sim \mathrm{Unif}(0, 1)$ drawn for each graph sample. Analysis (from an omitted appendix) shows that the linearity constraint for WS graphs is satisfied at a low re-wiring probability, < 0.07.

Scale-free (SF) networks are characterized by a degree distribution following a power law, $P(d) \sim d^{-\gamma}$ where d is some degree (centrality) score, $P(d)$ is the

proportion of nodes having that score, and γ is the scaling exponent. The World Wide Web (WWW), co-citation networks, and some social networks have been demonstrated to exhibit characteristics of scale-free networks [24], [29]–[31]. For scale-free (SF) networks, we restrict the generative $\gamma \in [2, 3]$ as it is implied that empirical networks operate in this region [24], [30], [31]. In generating these networks, we sample again from the uniform distribution for the generative target $\gamma = 2 + \text{Unif}(0, 1)$. We then sample a degree distribution using weights of $\{0^{-\gamma}, 1^{-\gamma}, \ldots, (n-1)^{-\gamma}\}$. Incidentally, networks having $\gamma \approx 3$ qualify as Barabási-Albert networks.

Unless otherwise noted, our simulated networks span ten sizes $n \in \{100, 200, \ldots, 1000\}$. Clearly, these networks sizes do not necessarily qualify as "large scale" or "big data". However, since our approaches are untested, it behooves us to experiment with smaller networks before tackling larger ones. We generate $m = 1000$ network samples for each topology for each size n.

2.3. Additional Notation and Terminology

Standard graph theoretic notation expresses a graph or network as comprising a set of vertices and a set edges: $G = \{V, E\}$. We use the terms 'graph' and 'network' interchangeably. We employ the network size $n = |V|$. The mean, standard deviation, and Pearson correlation coefficient are indicated by μ, σ, and r, respectively. Since our data is simulated, each complete network draw or instance is referred to as a "sample". Thus, we call an induced subgraph of a draw a "sub-sample". The s super-script denotes a term referring to a sub-sampled network, e.g. n^s is the size of a subsampled network.

2.4. Slow vs. Fast Measures

The time complexities for calculating degree centrality and density are equivalent and are either $O(n^2)$ or $O(|E|)$ depending on how the graph is encoded. We consider these to be our "fast" measures. In Table 1, we show ratios for computation times of key centrality measures; these were derived empirically by averaging actual run times. For example, $O(\frac{C_\mathsf{B}}{C_\Delta})$ indicates the ratio of computation times for betweenness centrality and the clustering coefficient as being proportional to $n^{\frac{3}{2}}$ for WS graphs.

Table 1. Complexity of computation

topo	$O\left(\frac{C_\mathsf{B}}{C_\Delta}\right)$	$O\left(\frac{C_\Delta}{C_\mathsf{D}}\right)$	$O\left(\frac{C_\mathsf{B}}{C_\mathsf{D}}\right)$
ER	n	$n^{\frac{1}{2}}$	$n^{\frac{3}{2}}$
WS	$n^{\frac{3}{2}}$	$n^{\frac{1}{2}}$	n^2
SF	$n^{\frac{3}{2}}$	$n^{\frac{1}{2}}$	n^2

The ratios are not entirely surprising. Betweenness centrality calculations should take considerably longer than the clustering coefficient which in turn takes longer than degree centrality (or density). While the complexity of the betweenness algorithm implemented is $O(nm)$ where m is the number of edges $|E|$ [32], the results we observe are likely due to the density ranges of our simulated data. Thus, we consider betweenness centrality to be a "slow" measure and the clustering coefficient to be a "medium" measure (i.e., neither slow nor fast).

2.5. Correlations

As a preliminary test, one might examine how correlations (r) between network measures might suffice for estimating slow measures [33]. We consider this a starting point for estimation.

In Table 2, we present mean correlations (averaged across the simulated samples) for each of the three topologies. For these results, the densities of Erdős-Rényi network samples are limited by the range of the densities of the other typologies to allow for reasonable comparison. The row names indicate the network topology: ER = Erdős-Rényi, WS = Watts-Strogatz small-world, and SF = scale-free. The columns indicate the pair-wise correlation between centrality scores. The larger correlations have been bold-faced. At first glance, some of the correlations appear highly promising particularly degree centrality (D) which appears predictive of several measures. When we examine the entire range of densities for Erdős-Rényi graphs (not shown), the mean correlations vary little, while some of the σ's (which appear under each correation) diminish due to the high linear associations that emerge between centralities at higher densities. Incidentally, Dekker [34] observed comparably low correlations between closeness and betweenness centralities in the empirical social networks he studied, which according to our table might typify SF topology.

The high mean correlations look promising, but these can be deceptive and ill-fitting particularly those exhibiting relatively higher s.d.'s. In the left plot of Figure 1, the density is plotted against the correlation between degree and betweenness centralities. The colors denote the topology: ER, WS, and SF. While for

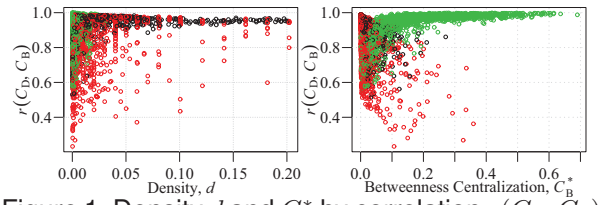


Figure 1. Density d and C^*_B by correlation $r(C_\mathsf{D}, C_\mathsf{B})$

ER, the mass of points hover near perfect correlation, the case is different for WS and SF which display poor correlations especially for networks having low density.

In the right plot of Figure 1, we examine what happens to the correlations under different betweenness centralizations. The betweenness centralization mea-

Table 2. Correlations between centralities

topo	DB	DC	DE	DΔ	BC	BE	BΔ	CE	CΔ	EΔ
ER	**0.95** (0.06)	**0.89** (0.11)	**0.90** (0.14)	0.01 (0.05)	**0.82** (0.15)	**0.90** (0.11)	−0.03 (0.05)	**0.87** (0.19)	0.01 (0.06)	0.04 (0.07)
WS	**0.88** (0.14)	**0.83** (0.17)	**0.83** (0.22)	−0.12 (0.15)	**0.83** (0.14)	0.79 (0.22)	−0.26 (0.29)	**0.80** (0.24)	−0.22 (0.28)	−0.06 (0.13)
SF	**0.93** (0.08)	0.29 (0.12)	0.78 (0.13)	0.06 (0.06)	0.27 (0.14)	0.76 (0.12)	0.03 (0.06)	0.46 (0.11)	0.16 (0.08)	0.19 (0.09)

Note: $\mathrm{D} = C_{\mathrm{D}}, \mathrm{B} = C_{\mathrm{B}}, \mathrm{C} = C_{\mathrm{C}}, \mathrm{E} = C_{\mathrm{E}}, \Delta = C_{\Delta}$, local clustering coefficient (triangle density)

sure captures the extent to which few nodes harbor higher betweenness than the rest spans the interval [0,1]. A C_{B}^* of 1.0 indicates a star-network having only one bridging node while a score of 0.0 indicates a complete clique in which each node has edges to all other nodes.

We see that the topologies differ in their sensitivities towards C_{B}^*. C_{B}^* in SF networks appears to strengthen the association between degree and betweenness.

2.6. Centrality Statistics

Correlations alone would allow us to only sample rankings of a "slow" measure using a "fast measure". Other innovations have explored faster rankings of centrality scores to find top nodes, e.g., Okamoto et al examined ranking closeness centrality [35]. For predicting a measure score itself, we require additional statistics, in particular the mean and variance or s.d.[1] We predict these network statistics using linear models. In Figure 2, we implement a least-squares (LS) regression model of a polynomial function:

$$\mu(C_{\mathrm{B}}) = \beta_0 + \beta_1 \cdot d + \beta_2 \cdot d^2 + \beta_3 \cdot d^3 + N(0, \sigma^2) \quad (2)$$

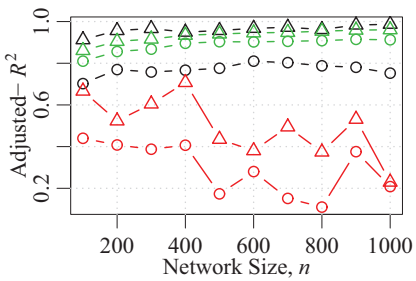
where $N(0, \sigma^2)$ indicates the error fitted to a Gaussian (i.e., the normal distribution) having an unknown variance of σ^2 which is estimated during the regression calculations. Essentially, we attempt to predict the mean betweenness centrality statistic from a cubic function of density (d), one of the more basic network measures. We report on a polynomial of order 3 as the gains in adjusted-R^2 beyond that order is marginal. The adjusted-R^2 statistic of a regression model indicates the extent to which the variation in the dependent data is accounted for by the predictors.

In the upper plot of Figure 2, we display the adj-R^2s produced at varying levels of network size n; again, the colors denote the graph typologies. The lower plot shows the adj-R^2s for the WS-specific model (which we discuss later). The mean betweenness centralities $\mu(C_{\mathrm{B}})$ of both ER and SF are readily predictable by the model defined by Eq. 2 (i.e., circle points in the figure). However, the statistic for WS is not predicted very well, with decreasing accuracy for larger networks.

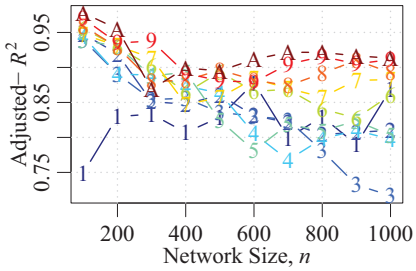
1. The Pearson's correlation (r) equation is

$$r_{X,Y} = \frac{E[(X - \mu(x))(Y - \mu(y))]}{\sigma_X \sigma_Y} \quad (1)$$

So, it is feasible to construct a distribution for a target measure Y from the correlation and relevant statistics.



Note: the circles refers to Eq. 2; triangles, Eq. 3



Note: both colors and numbers/letters denote the neighborhood size k

Figure 2. $n \times$ adjusted-R^2

We expect a superior fit when we include degree and clustering coefficient statistics:[2]

$$\mu(C_{\mathrm{B}}) = f(n, \mu(C_{\mathrm{D}}), \mu(C_{\Delta}), \sigma(C_{\mathrm{D}}), \sigma(C_{\Delta})) + N(0, \sigma^2) \quad (3)$$

where $f(\ldots)$ represents a full interaction model with non-linear predictors. An interaction model includes not only additive main effects/parameters but also the products of each combination of those parameters.

The new model yields an adj-R^2 of 0.957 for Erdős-Rényi networks of size 100; the fit is marginally improved by the inclusion of density (d). Incidentally, the prediction of $\sigma(C_{\mathrm{B}})$ (for the same network topology) and size does not fare as well, achieving an adjR^2 of 0.823 (not shown).[3] The triangle points in Figure 2 display the superior adjusted-R^2s achieved with Eq. 3. However, the enhanced model fails to sufficiently improve the prediction for WS graphs. We note that the residuals for the linear regression models are not entirely Gaussian or homoskedastic, largely due to the constriction of centrality statistics at densities where the pattern becomes strictly linear (see Fig. 3). However, the residuals are most normal and homoskedastic for the models of Eq. 3 and WS networks. For the latter set of networks, the outlying residuals appear

2. Incidentally, density and $\mu(C_{\mathrm{D}})$ are perfectly correlated within any network of any of the three topologies.

3. "Prediction" in this paper refers to statistical prediction and is, for all intents and purposes, identical to "estimation".

structural, in that they appear exacerbated by structures that lack WS features, e.g., not enough bridging links, particularly for the models generated from Eq. 4.

The WS graphs were constructed with fixed neighborhood sizes (hence fixed densities). Given that $\mu(C_{\mathrm{D}})$ and density are perfectly correlated, it is not altogether surprising that the previous models fail to adequately account for the variance in the target measures $\mu(C_{\mathrm{B}})$. To address this concern, we employ an alternative polynomial model for WS graphs, selecting a different "fast" measure that offers more variation for this typology:

$$\mu(C_{\mathrm{B}}) = \beta_0 + \beta_1\sigma(C_{\mathrm{D}}) + \beta_1\sigma(C_{\mathrm{D}})^2 + \beta_3\sigma(C_{\mathrm{D}})^3 + \mathrm{N}(0,\sigma^2) \quad (4)$$

In this case, the 3rd order of the polynomial is appropriate for this kind of regression as the relationship between $\sigma(C_{\mathrm{D}})$ and $\mu(C_{\mathrm{B}})$ is a decay curve; the 2nd order polynomial exhibits a noticeably diminished fit. Again, we employ a least-squares multiple regression and assess the quality of the fit using the adjusted-R^2. The Bayesian Information Criterion produces comparable results. In the lower plot of Figure 2, we see that the model fit is decent for high neighborhood sizes ($k > 6$). Furthermore, the fits appear to stabilize with increasing network sizes.

An examination of larger simulated network sizes up to $n = 10{,}000$ (not shown in this paper) show the adj-R^2s have a lower bounded of 0.70. Also, each of the trends (with the exception of $k \in \{1, 2\}$) have a distinct nadir after which the fit improves. The point at which this occurs appears directly related to the network size. The nadir for $k = 3$ also displays mild recovery (with $n > 1000$, not shown). As for the lower k, their adj-R^2s exhibit more noticeable fluctuations with the higher n. Some of this fluctuation appears related to the simulation sample sizes, for which we used $m = 1000$, as lower sample sizes result in instability even for the higher values of k. Finally, there appears to be a distinct relationships among the neighborhood size k; whether or not the fits stabilize (with larger networks); and the network size at which we observe the nadir in adj-R^2.

For further exploration of "fast" parameter fit, we search through the space of all models produced by a set of seven parameters: $\mu(C_{\mathrm{D}})$, $\sigma(C_{\mathrm{D}})$, $\mu(C_{\Delta})$, $\sigma(C_{\Delta})$, n, d, and the additional parameters specific to the WS (r for rewiring) and SF (γ) networks. Here, we seek a combined model rather than separate models for each network size or other parameterization (such as k for WS graphs).

We offer model fits to each of $\mu(C_{\mathrm{B}})$, $\sigma(C_{\mathrm{B}})$, $\mu(C_{\Delta})$, and $\sigma(C_{\Delta})$ for each of the graph topologies. The additional sets of models reflect the employment of the WS generation parameter r, and the empirical γ measured from SF graphs. We select the top models primarily based on their adj-R^2s but also minimizing 1) the number of parameters and 2) use of the slower measure C_{Δ}. In the first section of Table 3, we report the selected models for each typology. The dependent network statistic variable, under column 'y', is predicted by other statistics indicated by the gray-filled table cells.

The models fit with varying degrees of success. Predicting betweenness centrality is complicated, requiring all six applicable parameters. Even still, the fit for $\sigma(C_{\mathrm{B}})$ is modest at best. Conversely, the mean clustering coefficient $\mu(C_{\Delta})$ appears completely predicted merely by density. However, this is not surprising as our supplementary analysis reveals that $\mu(C_{\mathrm{D}})$, $\mu(C_{\Delta})$, and d are highly correlated in Erdős-Rényi graphs, even at the lower densities. Finally, the s.d. of Δ, $\sigma(C_{\Delta})$, appears to defy prediction; our best model yields a relatively abysmal adj-$R^2 = 0.71$.

The somber outlook for prediction in WS networks established in Table 2 is not mitigated by these WS models. With the exception of $\mu(C_{\Delta})$, the WS models are poor to fair. While adding the generative rewiring parameter r assists the prediction, it remains uncertain whether a) one knows this ex-ante or b) this can be inferred from a graph exhibiting features of a WS type network.

The SF networks fare better than either topology for most predictions. This should not be surprising considering the SF statistics overlap those of ER but in a far limited density range. However, the overlap diminishes with increasing network size and some of the centrality σ statistics. In some instances, particularly the prediction of betweenness statistics (i.e., $\mu(C_{\mathrm{B}})$ and $\sigma(C_{\mathrm{B}})$), measuring the power law exponent γ improves the fit. For the clustering coefficient statistics, γ offers very little improvement. Also, we notice that the betweenness predictions for SF do not seem to require clustering coefficient predictors unlike ER and WS graphs. This is not to say that the clustering coefficient predictors do not improve the fit, but the improvement is superficial. Considering how WS graphs respond very differently to the predictors in this table, one might entertain the possibility that the manner in which these network are predicted by "fast" measures offers another means of classifying these topologies.

2.7. Non-Linear Models

If we wanted to avoid the calculation of even the "medium" measures such as $\mu(C_{\Delta})$ in estimation, we could further explore the relationship between a "slow" measures such as $\mu(C_{\mathrm{B}})$ and a "fast" one such as density.

For expository purposes, we focus solely on $\mu(C_{\mathrm{B}})$ of Erdős-Rényi (ER) networks. In Figure 3, we plot network density d against $\mu(C_{\mathrm{B}})$ for three sets of $m = 5000$ samples of Erdős-Rényi networks of sizes $n \in \{100, 200, 300\}$. In the left plot, the colors denote the Erdős-Rényi network sizes of $n \in \{50, 100, 200\}$. In the right plot, the colors distinguish the two fitting procedures.

The pattern displays a notable spike-decay pattern at the lower densities and transforms into a line at some density level. Other similar patterns are evident with

Table 3. Model search results

	Erdős-Rényi							Watts-Strogatz								Scale-Free							
y	R^2	$\overline{D}$	$\hat{D}$	$\overline{\Delta}$	$\hat{\Delta}$	n	d	R^2	$\overline{D}$	$\hat{D}$	$\overline{\Delta}$	$\hat{\Delta}$	n	d	r	R^2	$\overline{D}$	$\hat{D}$	$\overline{\Delta}$	$\hat{\Delta}$	n	d	γ
$\mu(C_{\mathrm{B}})$	0.96	■	■	■	■	■	□	0.27	■	■	■	■	■	■	□	0.97	■	■	□	□	■	■	□
$\sigma(C_{\mathrm{B}})$	0.82	■	■	■	■	■	□	0.68	■	■	■	■	■	■	□	0.98	■	■	□	□	■	■	□
$\mu(C_{\Delta})$	1.00	□	□	□	□	□	□	0.88	■	■	□	□	■	■	□	0.94	■	■	□	□	□	■	□
$\sigma(C_{\Delta})$	0.71	■	■	□	□	■	□	0.51	■	■	□	□	■	■	□	0.88	□	■	□	□	■	■	□
$\mu(C_{\mathrm{B}})$		□	□	□	□	□	□	0.47	■	■	■	■	■	■	■	0.97	■	■	□	□	■	□	■
$\sigma(C_{\mathrm{B}})$		□	□	□	□	□	□	0.81	■	■	■	■	■	■	■	0.98	■	■	□	□	■	■	■
$\mu(C_{\Delta})$		□	□	□	□	□	□	0.92	□	□	□	□	■	■	■	0.95	■	■	□	□	■	■	□
$\sigma(C_{\Delta})$		□	□	□	□	□	□	0.77	■	■	□	□	■	■	■	0.90	□	■	□	□	■	■	□

Note: R^2 = Adj-R^2, $\overline{X} = \mu(X)$, $\hat{X} = \sigma(X)$, $n = |V|$, s = sampling prob., r = WS rewiring prob., γ = SF parameter

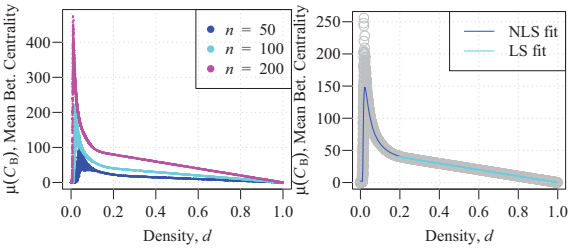


Figure 3. Density $d \times \mu(C_{\mathrm{B}})$

statistics of other centrality measures. In fact, $\sigma(C_{\mathrm{B}})$ exhibits an almost identical pattern. On the other hand, $\sigma(C_{\mathrm{D}})$ displays a distinct inverted-U pattern which is far more easily tractable than the pattern for $\mu(C_{\mathrm{B}})$. Some centrality statistics display more complicated patterns; for instance, the relationship between $\sigma(C_{\mathrm{C}})$ and d contains three optima (not shown).

The initial points of the curve in Figure 3 lie at $\mu(C_{\mathrm{B}}) = 0.0$, and the greatest variance of the statistic is observed at the same density the maximum of the statistic is observed (i.e., the top of the spike). We hypothesize that, at this density, multiple components exist and have the potential to become easily connected through the introduction of very few edges, thereby drastically increasing the betweenness centralities of the nodes participating in those edges. However, if those key edges do not exist then the average betweenness centrality is reduced.

The critical density at which the non-linear curve becomes strictly linear is easily predicted by a polynomial function of n. Due to space concerns, we regretfully omit this function from this paper. The cyan colored line shows the fitted line. The spike however does not seem readily amenable to a straightforward fitting function. With some effort, we find the following non-linear function to offer a surprisingly good fit (the blue curve) considering only density d was employed:[4]

$$\mu(C_{\mathrm{B}}) = M(1-\exp\{-(\boldsymbol{d}/\tau_1)^{\alpha}\})+N\exp\{-(\boldsymbol{d}/\tau_2)^{\beta}\}+o \quad (5)$$

The function is a double-exponential sum; the first addend term fits the rise of the curve, and the second fits the decay; a single exponential term does not suffice. Simpler non-linear formulas displayed far inferior fits. Non-linear least squares (NLS) regression was employed to fit the equation to the data. Table 4 shows the model coefficients and its log-likelihood. We employ $n = 5000$ sampled Erdős-Rényi networks for this estimation.

Table 4. Model estimation for $n = 100$

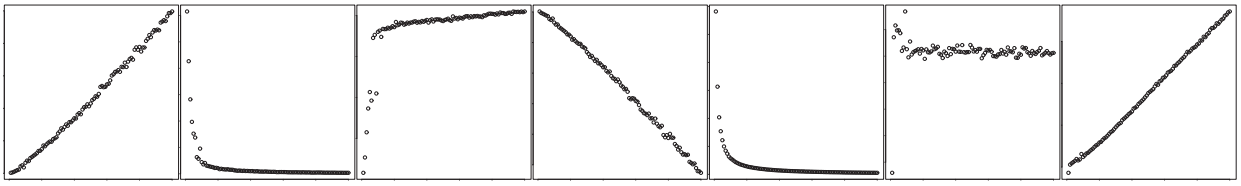
Parameter	Estimate	Parameter	Estimate
M	169.000*** (16.435)	N	193.196*** (14.981)
τ_1	0.033*** (0.003)	τ_2	0.014*** (0.000)
α	−1.352*** (0.069)	β	−5.046*** (0.282)
o	−166.746*** (16.283)		
m	5000	$\mathcal{L}$	9.05

All six parameters of the model reach statistical significance at the $p < 0.001$ level. In Figure 3, we superimpose the model fit onto the plot, and it yields a superior adj-R^2 of 0.938 ($r = 0.968$, RMSE = 302). A comparable linear polynomial model using the predictors from Table 3 at best produces an adj-$R^2$of 0.838 ($r = 0.921$, RMSE = 977). However, the NLS and polynomial LS model attempt different fittings. The NLS model having only one predictor finds a smooth curve while the LS model offers a more precise fit to all the data points. For fairer comparison, we predict the NLS curve with the same best performing polynomial LS model. Of course, we find an improved adj-R^2 of 0.9 ($r = 0.95$, RMSE = 800). The model still remains shy of the performance of NLS. The NLS model performs better and is more parsimonious as it employs fewer parameters and data.[5] Still, the presence of heteroskedasticity (which is obviously visible in the spike-decay) warrants search for a more sophisticated model, such as weighted NLS.

In Figure 4, we display the non-linear function parameters for the various spike-decay regions of $\mu(C_{\mathrm{B}})$ for additional networks of sizes $n \in \{50, 100, \ldots, 4950, 5000\}$. The x-axes contain the network sizes n and the y-axes contain the fitted parameters and their scales vary across parameters. We omit axes in order to present all of the parameters side-by-side. The unlabeled x-axes consistently span the range of network sizes and the unlabeled y-axes contain

4. We also explored rational functions as a separate fitting alternative. Such functions on d would appear in the form of $f(d) = \frac{g(d)}{h(d)}$ where $g(d)$ and $h(d)$ are polynomial functions. However, we were unable to find one that parsimoniously fit the Erdős-Rényi $\mu(C_{\mathrm{B}})$ statistic.

5. Recall that our linear models employ a full interaction model. A listed model with 3 predictors (colored cells) can have up to $1+\sum_{i=1}^{m}\binom{m}{i}$ or $1+3+3+1$ or 7 actual coefficients/parameters; some of the predictor combinations may be collinear. The best LS model we use for comparison could potentially have 64 parameters.



(a) $n \times M$ (b) $n \times \tau_1$ (c) $n \times \alpha$ (d) $n \times N$ (e) $n \times \tau_2$ (f) $n \times \beta$ (g) $n \times o$

Figure 4. Relationship between network node count n and hyper-parameters

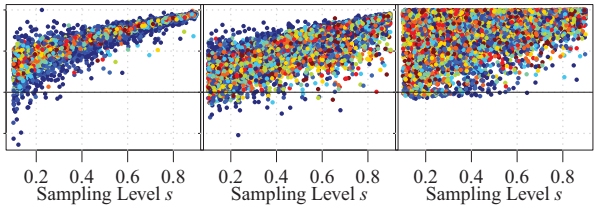
different scales across the parameter plots. We observe that the relationships between the parameters are either strictly linear or convergent, and thus at least amenable to a simple function. In other words, we can use the function in Eq. 5 to model $\mu(C_{\mathrm{B}})$ of much larger Erdős-Rényi networks.

2.8. Network Sampling

The enormity of online and large social networks often necessitates sampling sub-graphs from the networks of interests. In fact, much of social media freely available to researchers happens to be a sub-sample (of a distinct point in or range of time), e.g., Twitter's streaming API provides only 1% of the tweets. The characteristics of the networks can dictate the sampling strategy. For example, a diffusive, snowball sampling of a network will yield a very different sampling than say sampling a random set of nodes and their edges. Some of the consequences of sampling (and similar sources of bias such as the effects of node or edge removal) have been investigated in recent network literature, e.g., [36]. Borgatti et al [37] examine the effects of "unintentional sampling" (i.e., missing network data) among other errors. Frantz [38] in particular demonstrates how error (which includes edge removal) varies with different network topologies. However, some of his topologies, namely, cellular, and core-periphery, differ from the ones examined in this paper. Other work delves into the properties of these subgraphs [39]. We endeavor to add to this line of research by examining the predictability of the complete networks' centrality measures and statistics from the analogous statistics for the sub-samples, all in the context of the three topologies.

Our sampling strategy here is straightforward. For each randomly generated network, we merely sample without replacement some proportion of nodes, $s \in [0.1, 0.9]$, and the edges between those nodes. Thus, we mimic those situations in which a sample of ego-centric networks are obtained from a population. Consequently, we obtain "sampled" analogs of the centrality measures of interest. In Figure 5, we examine the correspondence between the original betweenness centrality (C_{B}) and the sampled version (C_{B}^s) for each of the topologies. The x-axis denotes sub-sampling levels, $s \in [0.1, 0.9]$, across networks of different sizes, $n \in \{50, 100, \ldots, 1000\}$.

In the leftmost plot (ER), we see that there exists



(a) ER (b) WS (c) SF

Note: dark red indicates the smallest network ($n = 50$), dark blue indicates the large networks ($n = 1000$). The y-axis denotes the correlation level, with the horizontal line indicating correlation $r = 0.0$ and the top of the graph indicating $r = +1.0$.

Figure 5. Correlation of C_{B} at varying sub-sampling levels

some correspondence between the sampled betweenness centrality scores of ER graphs and the original scores. The correlation is unsurprisingly highest at the higher sampling levels. We also observe the variance (estimated by the thickness of the color band) is wider for smaller networks, which is encouraging since most on-line network data consists of much larger networks. The findings are even less precise for WS networks. While some linear pattern exists, the variance remains high even for large networks. And, for SF networks, there appears to be no distinct pattern.

In Table 5, the correlations for five centrality measures for a sub-sampling rate of $s = 0.5$ are reported. Both the mean and s.d. are averaged across both

Table 5. Correlation of measures between original and sub-sample

	$\mu(r(C_\cdot, C_\cdot^s))$				
topo	C_{D}	C_{B}	C_{C}	C_{E}	C_{Δ}
ER	0.679 (0.190)	0.667 (0.198)	0.674 (0.193)	0.673 (0.195)	0.489 (0.250)
WS	0.594 (0.225)	0.577 (0.239)	0.548 (0.237)	0.489 (0.291)	0.477 (0.258)
SF	0.810 (0.218)	0.822 (0.222)	0.543 (0.284)	0.677 (0.347)	0.619 (0.299)

network sizes and samples. It is quite clear that the high σ's reflect the semi-scattered display in the above figure. Interestingly, the mean degree and betweenness correlations remain high for SF graphs.

Naturally, we investigate how well sub-sampled centrality statistics can estimate centrality statistics of the original network sample. We entertain two scenarios. In the first, prior knowledge of the original network is unavailable, meaning the sub-sampling s level and full network size n are unknown. In the second scenario, that information is considered available for use in

estimation.

For each scenario and dependent centrality statistic, we again perform exhaustive searches selecting models based on their fit (i.e., adj-R^2), their model sizes (i.e., number of parameters), while penalizing models for their reliance on the slower measures, i.e., betweenness centrality and clustering coefficient statistics. The "fast" measures would then be the C_{D} statistics, n^s, and d^s (the size and density of the sub-sampled network). For models having relatively low adj-R^2, the best one is reported, with slight consideration of model size. The results for each topology are presented in Table 6.

Naturally, the models that are handicapped by the absence of the size of the original network and the sampling level for the sub-sampled network underperform those models that include this prior knowledge. In the full model, the sub-sample network size n^s is comparable to an interaction term of n and s (i.e., $n{\times}s$). However, we find again that the mean clustering coefficient $\mu(C_{\Delta})$ retains its relationship to solely density even across the divide between sub-sampled and the original network sample while the standard deviation requires the full complement of predictors. Also, more often than not, the slower measures are required to achieve even a modest fit. Interestingly, some of these models do not require the dependent variable's analog to be present.

Despite the inclusion of the prior parameters (n and s), some of the models still require more than a few additional predictors. One might surmise the model rank (i.e., number of predictors or solid colored cells) is an indicator of the complexity of the network statistic and subsequently the measure itself.

For the WS models lacking the prior network information, the fits resemble those for ER graphs. One notable exception is the difference in fits for $\mu(C_{\mathrm{B}})$ and to a lesser extent $\sigma(C_{\mathrm{D}})$. Prediction of mean betweenness appears particularly difficult in WS networks while its s.d. prediction remains comparably effective. Moreover, the addition of the prior network parameters n and s does not appear to improve the prediction whereas ER network statistics are better predicted by these two pieces of information. Finally, we observe that WS offers an improved fit for $\sigma(C_{\mathrm{D}})$ over that of ER which further highlights the topological differences. Also, the efficiency of the minimal prior-informed model for $\mu(C_{\Delta})$ no longer holds for WS graphs which requires 5 predictors for an adequate fit.

For the SF models, the overall fits are only fair at best, so we report the maximally fitting models disregarding the case of over-fitting to emphasize their mediocrity. This is somewhat surprising since the distributions of centrality statistics from SF networks overlap heavily with those of ER networks. This discrepancy certainly bears further investigation, and we might expect to find that either the regions of non-overlap reveal crucial differences or a more subtle element is responsible.

3. Conclusion and Discussion

This paper outlines the first steps towards nodal centrality score and statistic estimation using correlations and both LS and NLS models. Specifically, we estimate distributions of "slow" centrality metrics from "fast" metrics for networks with known prior size and known or unknown density under an approach different than other estimation approaches discussed in the introduction. However, the quality of estimation has been shown to vary between topologies. Several correlations are effective for ER and SF, but less so for WS networks. Likewise, correlations between only a subset of the measures appear successful; generally, the clustering coefficient fails in its association with the centrality.

Since correlations themselves offer only ranking (distributions) of "slow" measures using "fast" measures, we continued our investigation with predictive models for "slow" centrality statistics: mean and standard deviation. In conjunction with a correlation, these statistics will allow us to sample target centrality scores from "slow" measures using Eq. 1. For prediction of these statistics, we explored both modeling of polynomial and non-polynomial functions using LS and NLS, respectively. The LS models appear adequate for ER and SF, but not for WS. While predictive power is enhanced with the inclusion of generating parameters such as rewiring r, these would not be necessarily available for centrality prediction in empirical networks. Still, it remains possible to infer r assuming we are certain our test network qualifies as a WS topology. In addition, we noticed that the predetermined neighborhood size k allows us to distinguish which sizes induce better prediction providing us with a prior estimation on how effective our approach will be for empirical WS networks. The LS models could be enhanced if we employed higher (polynomial) degrees for the main effect terms as we did in Eqs. 2 and 4.

The NLS model was presented for only one statistic, $\mu(C_{\mathrm{B}})$, under one topology, ER. Its effectiveness encourages us to examine how similar functions might fit other centrality statistics. Preliminary findings reveal that WS graphs appear less tractable under this approach. We also examined an alternative model fitting criterion, the Bayesian Information Criterion (BIC), and found its results consistent with the adj-R^2.

Another criterion we ought to investigate is how accurately we can estimate top scoring nodes (as characterized by one or more measures). Top scoring node lists often appear in network findings as a way of highlighting important actors or entities.

Clearly, the next steps include testing our estimation strategies on empirical networks and larger networks. Furthermore, sampling issues on empirical networks have been receiving some attention, e.g. [40]. Studies

Table 6. Sampled network model search results for each topology

	Erdős-Rényi											Watts-Strogatz											Scale-Free										
y	R^2	$\overline{D}$	$\hat{D}$	$\overline{B}$	$\hat{B}$	$\overline{\Delta}$	$\hat{\Delta}$	n^s	d^s	n	s	R^2	$\overline{D}$	$\hat{D}$	$\overline{B}$	$\hat{B}$	$\overline{\Delta}$	$\hat{\Delta}$	n^s	d^s	n	s	R^2	$\overline{D}$	$\hat{D}$	$\overline{B}$	$\hat{B}$	$\overline{\Delta}$	$\hat{\Delta}$	n^s	d^s	n	s
$\mu(C_{\mathrm{D}})$	0.76											0.75											0.72										
$\sigma(C_{\mathrm{D}})$	0.75											0.83											0.72										
$\mu(C_{\mathrm{B}})$	0.75											0.37											0.73										
$\sigma(C_{\mathrm{B}})$	0.61											0.59											0.77										
$\mu(C_{\Delta})$	1.00											0.95											0.65										
$\sigma(C_{\Delta})$	0.66											0.72											0.64										
$\mu(C_{\mathrm{D}})$	1.00											0.98											0.74										
$\sigma(C_{\mathrm{D}})$	0.99											0.94											0.74										
$\mu(C_{\mathrm{B}})$	0.95											0.39											0.83										
$\sigma(C_{\mathrm{B}})$	0.91											0.63											0.84										
$\mu(C_{\Delta})$	1.00											0.95											0.69										
$\sigma(C_{\Delta})$	0.92											0.82											0.65										

Note: R^2 = Adj-R^2, $\overline{X} = \mu(X), \hat{X} = \sigma(X), n^s = n$ of $G^s, d^s =$ density of $G^s, n = n$ of $G, s =$ node-sampling prob.

show that real-world networ ks rarely fall cleanly into one of the topological categories so this complication must be considered in the analysis [41]. We might apply models from each of the topologies to a single empirical network, aggregating the results. Alternatively, we could expand the topological definitions to include hybrids. Such an approach would then require a means to assess the position of a tested, empirical network in this space of topologies. The space between WS and ER should reflect noisier WS networks as would the space between SF and ER for SF networks. For example, we observe that our sample set of smaller ER graphs contain more networks that passed the test for the power law degree distribution than larger ones (about 7-to-1 and vice versa for WS networks). That is, larger WS graphs appear to converge into or overlap with the SF topology while larger ER graphs diverge from them.

We demonstrated that our estimation approaches scale with network size, but for a limited range of sizes. Despite the duration required to produce simulated data of much larger networks, we plan on devoting some effort towards assessing whether or not the estimation remains consistent for larger networks.

Still, prediction through linear modeling assumes errors to be Gaussian distributed which is not always the case with some of these topologies. Further exploration will require more advanced modeling, perhaps Bayesian regression or WNLS. We might also be able to eliminate some of the model variance using normalized centrality measures thereby reducing the need to employ network size n as a parameter. However, normalization incurs additional complications for networks containing weighted edges. It would also behoove us to confirm our findings with statistics from much larger networks. Additionally, we could infer LS models within each simulated network sample and then collate models hoping to find regularities in the coefficients similar to those observed in the NLS hyper-parameters. Preliminary analysis shows this may be possible for smaller (more parsimonius) LS models.

For sub-samples, the models perform fairly, achieving adj-R^2s generally between 0.6 and 0.8. The fits are visibly improved with the inclusion of prior network knowledge specifically the complete network size n and consequently the sampling level. Still, this information may not always be available. We observe that the sub-sampled models appear to outperform the direct prediction of network statistics. For example, the prediction of $\mu(C_{\mathrm{B}})$ for WS graphs from sub-samples has a higher fit than $\mu(C_{\mathrm{B}})$ prediction from non-C_{B} terms. We surmise that the sub-sampled C_{B} statistics (in conjunction with those of sub-sampled C_{D} and C_{Δ}) offer more information than the C_{D} and C_{Δ} statistics of the whole network. Still, one might expect the error to exhibit heteroskedasticity in the sense that predictions from smaller sub-samples will be more inaccurate. This bears further investigation.

Finally, we recognize that there is an explicit trade-off between time and cost saved through estimation and the fidelity of estimates. Exactly how much error is incurred for a unit of time saved? Future work will explore measuring this trade-off.

Acknowledgments

This material is based upon work supported by the Army Research Office under Award No. W911NF1-310155 and the Minerva Initiative (N000141310835). We thank the HICSS anonymous reviewers for their valuable comments.

References

[1] M. Kas, M. Wachs, K. M. Carley, and L. R. Carley, "Incremental algorithm for updating betweenness centrality in dynamically growing networks," in *The 2013 IEEE/ACM ASONAM*, 2013.

[2] U. Brandes and C. Pich, "Centrality estimation in large networks," *International Journal of Bifurcation and Chaos*, vol. 17, no. 7, pp. 2303–2318, 2007.

[3] M. G. Everett and S. P.Borgatti, "Ego-network betweenness," *Social Networks*, vol. 27, no. 1, pp. 310–38, 2005.

[4] P. Pantazopoulos, M. Karaliopoulos, and I. Stavrakakis, "On the local approximations of node centrality in internet router-level topologies," in *The 7th IWSOS '13 (IFIPP)*, 2013.

[5] J. Pfeffer and K. M. Carley, "k-centralities: Local approximations of global measures based on shortest paths," in *Proc. of the WWW Conf. 2012 (LSNA 2012)*, 2012, pp. 1043–1050.

[6] S. Y. Chan, I. Leung, and P. Li., "Fast centrality approximation in modular networks," in *Proc. of ACM/CNIKM*, 2009, pp. 31–38.

[7] D. Eppstein and J. Wang, "Fast approximation of centrality," *Journal of Graph Algorithms and Applications*, vol. 8, no. 1, pp. 39–45, 2004.

[8] T. Zheng, M. J. Salganik, , and A. Gelman, "How many people do you know in prison?: Using overdispersion in count data to estimate social structure in networks," *Journ. of the Amer. Stat. Assoc.*, vol. 101, no. 474, 2006.

[9] M. Motoyama, D. McCoy, K. Levchenko, S. Savage, and G. M. Voelker, "An analysis of underground forums," in *Proc. of ACM/Internet Measurement Conf., 2011*, 2011.

[10] M. Yip, N. Shadbolt, and C. Webber, "Structural analysis of online criminal social networks," in *Proc. of IEEE/Intelligence and Security Informatics*, 2012, pp. 60–65.

[11] J. Xu and H. Chen, "The topology of dark networks," *Comm. of the ACM*, vol. 51, no. 10, pp. 58–65, 2008.

[12] J. Pfeffer, "Modeling and calibrating real world interpersonal networks," in *IEEE NSW*, 2011.

[13] L. C. Freeman, "Centrality in social networks: Conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215–239, 1979.

[14] G. Sabidussi, "The centrality index of a graph," *Psychometrika*, vol. 69, pp. 581–603, 1966.

[15] T. Opsahl, F. Agneessens, and J. Skvoretz, "Node centrality in weighted networks: Generalizing degree and shortest paths," *Social Networks*, vol. 32, no. 3, pp. 245–251, 2010.

[16] P. Boldi and S. Vigna, "Axioms for centrality," *CoRR*, vol. 1308.2140, 2013.

[17] T.-H. Wei, "The algebraic foundations of ranking theory," Ph.D. dissertation, University of Cambridge, 1952.

[18] C. Berge, *Théorie des graphes et ses applications*. Paris, France: Dunod, 1958.

[19] P. Bonacich, "Factoring and weighting approaches to status scores and clique identification," *Journal of Mathematical Sociology*, vol. 2, no. 5, pp. 113–120, 1972.

[20] ——, "Power and centrality: A family of measures," *Amer. Journ. of Sociology*, vol. 92, pp. 1170–1182, 1987.

[21] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[22] P. Erdős and A. Rényi, "On random graphs," *Publ. Math. (Debrecen)*, vol. 6, pp. 290–297, 1959.

[23] ——, "On the evolution of random graphs," *Publ. Math. Inst. Hung. Acad. Sci.*, vol. 5, no. 17, pp. 17–61, 1960.

[24] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Rev. of Modern Physics*, vol. 74, no. 1, pp. 47–97, 2002.

[25] M. Newman, *The Structure and Dynamics of Networks*. Princeton, NJ: Princeton University Press, 2006.

[26] M. Barthélémy, , and L. A. N. Amaral, "Small-world networks: Evidence for a crossover picture," *Phys. Rev. Lett.*, vol. 82, no. 15, pp. 31–80, 1999.

[27] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley, "Classes of small-world networks," *PNAS*, vol. 97, no. 21, pp. 11 149–11 152, 2000.

[28] Q. Telesford, K. Joyce, S. Hayasaka, J. Burdette, and P. Laurienti, "The ubiquity of small-world networks," *Brain Connect*, vol. 1, no. 5, pp. 367–75, 2011.

[29] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, "Graph structure in the web," *Computer Networks*, vol. 33, no. 1-6, pp. 309–320, 2000.

[30] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[31] A. Clauset, C. Shalizi, and M. Newman, "Power-law distributions in empirical data," *SIAM Review*, vol. 51, no. 4, pp. 661–703, 2009.

[32] U. Brandes, "A faster algorithm for betweenness centrality," *Journal of Mathematical Sociology*, vol. 25, no. 2, pp. 163–177, 2001.

[33] T. W. Valente, K. Coronges, C. Lakon, and E. Costenbader, "How correlated are network centrality measures?" *Connections*, vol. 28, no. 1, pp. 16–26, 2008.

[34] A. H. Dekker, "Centrality in social networks: Theoretical and simulation approaches," in *Proc. of SimTecT*, 2008, pp. 33–38.

[35] K. Okamoto, W. Chen, and X.-Y. Li, "Ranking of closeness centrality for large-scale social networks," in *Proc. of FAW '08*, 2008, pp. 186–195.

[36] 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.

[37] S. P. Borgatti, K. M. Carley, and D. Krackhardt, "On the robustness of centrality measures under conditions of imperfect data," *Social Networks*, vol. 28, pp. 124–136, 2006.

[38] T. L. Frantz and K. M. Carley, "Relating network topology to the robustness of centrality measures," Carnegie Mellon University, Tech. Rep. CMU-ISRI-05-117, 2005.

[39] P. Wang, J. C. S. Lui, J. Zhao, B. F. Ribeiro, D. Towsley, and X. Guan, "Efficiently estimating subgraph statistics of large networks in the dark," *CoRR*, vol. 1306.5288, 2013.

[40] J.-S. Lee and J. Pfeffer, "Robustness of network centrality metrics in the context of digital communication data," in *Proc. of HICSS 2015*, 2015.

[41] H. Akkermans and R. Bakhshi, "Toward a next generation of network models for the web," in *Proc. of WebSci 2013*, 2013, pp. 1–10.