2015 48th Hawaii International Conference on System Sciences

Robustness of Network Centrality Metrics in the Context of Digital Communication Data

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

Social media data and other web-based network data are large and dynamic rendering the identification of structural changes in such systems a hard problem. Typically, online data is constantly streaming and results in data that is incomplete thus necessitating the need to understand the robustness of network metrics on partial or sampled network data. In this paper, we examine the effects of sampling on key network centrality metrics using two empirical communication datasets. Correlations between network metrics of original and sampled nodes offer a measure of sampling accuracy. The relationship between sampling and accuracy is convergent and amenable to non-linear analysis. Naturally, larger edge samples induce sampled graphs that are more representative of the original graph. However, this effect is attenuated when larger sets of nodes are recovered in the samples. Also, we find that the graph structure plays a prominent role in sampling accuracy. Centralized graphs, in which fewer nodes enjoy higher centrality scores, offer more representative samples.

1. Introduction

Robustness of network metrics is a significant issue due to the incomplete nature of large network datasets, especially those collected from online sources which provide a plethora of communication data. Often, we resort to dealing with a subset or "sample" of the original data. For example, the study of Twitter data entails trade-offs between data integrity, cost, and manageability [1]. One could theoretically employ the full data (i.e., the Twitter "firehose"), which incurs costs in obtaining the full data as well as requiring sufficient resources to manage and analyze the data. Alternatively, one may use the 1% sample, freely provided by Twitter which diminishes the integrity of results.

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.

Network algorithms that employ less than the whole data entail approximation using local structures from the ego-network [2] [3] and beyond [4]. They also include approximation for specific structures such as communities [5]) and approximation under desired errors bound constraints. [6].

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 sampling methodologies to address biases in the network samples [7]. In the area of cybersecurity, social network analysis of social media data offers improved assessments of threats by analyzing underground social media activities [8], dynamics between cyber-criminals [9], and topologies of dark networks [10]. However, these networks are less easily obtainable than many others reported in the literature.

Given that the use of sampled network data is ubiquitous, exactly how the metrics respond to sampled data continues to be a relevant issue. Some recent work includes [11] which explores several sampling methods; [12] which examines sampling in the context of information diffusion in Twitter data; [13] which also seeks to test sampling strategies for diffusion phenomena in networks; and [14] which explores is-

1530-1605/15 $31.00 © 2015 IEEE
DOI 10.1109/HICSS.2015.217

IEEE computer society

sues surrounding estimation of centrality statistics for whole and sampled networks of stylized topologies. This paper explores the effects of sampling on two empirical and temporal communication networks and also provides some insights into the specific effects of structural properties on the accuracy of the metrics in the samples.

2. Analysis

2.1. Data

The first dataset consists of 309,125 email exchanges between 3,188 members of a European university, spanning 83 days of activity [15]. We refer to this dataset as 'Email'. The second dataset comprises online messages exchanged between U.S. college students on a university-wide, Facebook-like system over a 193-day period [16]. We refer to this dataset as 'Message'. The network edges of both datasets are directed and weighted as there exist multiple communications between pairs of nodes. Both weighted and unweighted edges are considered in our analysis.

2.2. Descriptive Comparisons

Temporal graphs are typically aggregated into discrete time windows. For the Email data, we extract graphs for daily, weekly, bi-weekly, and tri-weekly windows. For the Message data, we extract daily graphs and windows of one week through seven weeks. In addition, we also consider the union graph for both datasets, i.e., all of the communications aggregated into one single graph.[1]

In Table 1, we compare basic network-level statistics from the weekly networks of each dataset. The 'm' column indicates the number of networks obtained from weekly windows; 'γ' indicates distributional skew. The correlations between the node counts and edge counts across the weeks are reported under r, ρ, and τ which respectively refer to Pearson's r correlation coefficient [17], Spearman's ρ (rho) rank correlation [18], and Kendall's τ rank correlation coefficient [19]. These correlations indicate the degree of association between two variables or vectors of interest (in this case weekly node counts and edge counts). The latter two correlations are rank-based measures, meaning they employ the rank of a quantity within its vector rather than the actual value. As such, they are more robust with data containing outliers than Pearson's r. We report the mean, standard deviation, and skewness of the node and edge counts for the weekly networks of each dataset. In addition, the correlation between node and edge counts across the weekly networks and the statistics for the density are reported. Multiple communications between pairs are tallied individually; that is, the edges may be considered weighted. Since density is ill-defined for graphs with multiple edges, we treat the edges as binary for that measure.

The 'Email' network is shorter in its data collection time span yet larger in average size per week. There is far more variance in the sizes of the 'Message' networks. In Figure 1, it is evident that activity for Message peaks after a short rise time then drops likely due to summer vacation for the students,[2] after which there is a slight resurgence in use. While the 'Message' weekly network node and edge count distributions consistently exhibit positive skew, the 'Email' weekly network is less consistent suggesting some sort of imbalance between node and edge counts. A positive skew denotes a distribution with a longer and fatter right tail and a mass that "leans" to the left; a negative skew denotes the reverse. The lower correlations for 'Email' compared to those of 'Message' seem to verify this inference. These observations may indicate a discrete period of time in which the network was inordinately more active. Since this paper explores the quality of samples, these datasets will be treated as the complete "ground truth". In fact, the Message data is complete as it represents all of the communication in the online system. The Email data however comprises only those communications between institution members and omits external ones.

In Figure 1, the relative weekly node and edge counts are charted with 'Message' having a longer time horizon of 28 weeks. The colors denote the measure category. Naturally, there is significant concordance

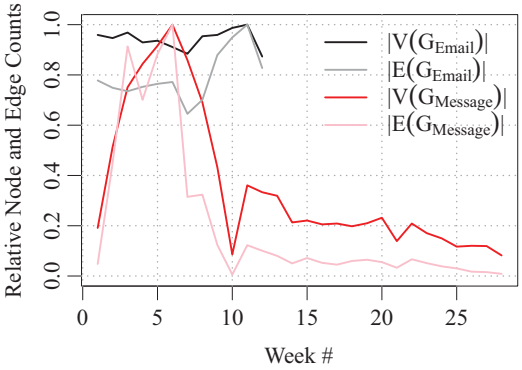


Figure 1. Weekly network sizes

between node and edge counts, $|V|$ and $|E|$.[3] However,

1. In this paper, "network" and "graph" are used interchangeably.

2. The first day of the dataset is April 19.

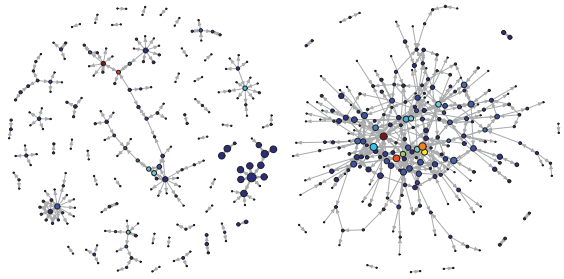
3. We also refer to $|V|$ and $|E|$ as n and e, respectively.

Table 1. Descriptive statistics of node and edge counts for weekly networks

		Node Count, $\|V\|$ Statistics			Edge Count, $\|E\|$ Statistics			Node/Edge Count Correlations			Density Statistics		
Data	m	μ	σ	γ	μ	σ	γ	r	ρ	τ	μ	σ	γ
Email	12	1798	73	-0.43	16891	2164	0.71	0.58	0.43	0.29	0.0019	0.0001	-0.40
Message	28	323	258	1.13	2137	3131	1.73	0.91	0.96	0.86	0.0080	0.0027	0.19

it appears that there is heightened activity near the end of the time horizon for 'Email' as implied by the lower correlations observed in Table 1. The first nadir in the weekly 'Message' counts coincides with the end of the spring semester. The resumed activity likely occurs during the summer sessions. Other variations in the temporal activity (in both datasets) are due to weekends and holidays.

In Figure 2, two of the smaller networks of each dataset are shown; these have comparable node counts ($|V| \approx 240$). These represent the daily communication from the 4th day of the Email data and the 24th day of the Message data. The node size and edge widths in the figure have been scaled by weighted degree centrality and edge weight, respectively. The node color denotes the weighted betweenness centrality; blue indicates low scores and red indicates high scores. These two network metrics are defined later in the paper. For



(a) Email (b) Message

Figure 2. Two example graphs

the first set of analysis, we include these two graphs as well as two larger example graphs approximately an order of magnitude larger than the small two graphs. The Email graph is sparse compared to the Message graph, which is not entirely surprising as the former graph spans a weekend day. It appears relatively more centralized than a random graph. In Table 2, basic graph-level statistics are reported. $|E|$ denotes the number of communications; $|E_{\text{w}}|$ is the count of weighted edges; μ and σ denote mean and standard deviation and the last column is unweighted density.

While the smaller graphs have comparable node sizes, their edge volumes are noticeably different as one might have guessed from the plots above. Interestingly, the counts of weighted edges differ by a larger ratio, yielding the higher magnitude of average weights in the email graph. Furthermore, the densities for both Message graphs are comparable while being different for the Email graphs due to heightened activity in the larger graph. Note that the size of the larger Message graph does not match that of the large Email graph as this is the largest obtainable from the data (i.e., the union graph of all the communication).[4]

Table 2. Network statistics on example networks

Graph	$\|V\|$	$\|E\|$	$\|E_{\text{w}}\|$	$\mu(E_{\text{w}})$	$\sigma(E_{\text{w}})$	$\frac{\|E\|}{\|V\|^2-\|V\|}$
Email (small)	243	650	224	2.90	8.55	0.0111
Message (small)	240	885	452	1.96	2.13	0.0154
Email (large)	2338	47991	12374	3.88	27.32	0.0088
Message (large)	1899	59835	20296	2.95	4.89	0.0166

2.3. Modeling Correlations of Centrality Metrics

Our random sampling method focuses on each individual communication or edge: email or message. For example, a pair of individuals may have exchanged m over the course of a week. For a graph with M total edges/communications, the probability that any one of the m emails or messages will be sampled is $\frac{m}{M}$. We assess the quality of the sampling by comparing the node-level network metric scores of the nodes of the graph induced by the edge-sampling and the scores of those same nodes in the context of their original network. The node-level network metrics we employ are degree and betweenness centralities.

Degree centrality is simply the sum of edges incident to a node or vertex and can indicate the activity level or popularity of a node [20]. More specifically, we employ total degree centrality for a directed network:

$$C_{\text{D}}(i) = \sum_{j=1}^{|V|} A_{i,j} + A_{j,i}$$

where $A_{i,j}$ (adjacency matrix) contains $\{0,1\}$ if there exists any edge from i to j; this is unweighted treatment of weighted edges. Betweenness centrality is

4. The larger Email graph is obtained from the communication of the 1st three weeks.

the sum of the proportions of shortest paths a node lies on between each pair (out of all shortest paths for each pair) [20], [21]. For betweenness centrality, we identify all the shortest paths in a graph, such that $\sigma_{s,t}$ is the number of shortest paths between nodes s and t. Betweenness centrality for node i then:[5]

$$C_{\mathrm{B}}(i) = \sum_{s \neq i \neq t} \frac{\sigma_{s,t}(i)}{\sigma_{s,t}}$$

The measure captures the extent to which a node acts as a bridge between distant sets of nodes or clusters of nodes. Both of these centrality metrics are described in Freeman's seminal work [20] and are two of the most widely studied network measures. One might assume that out of all network metrics, the centrality ones, particularly degree centrality, would be resilient to sampling effects, but we find that is not the case, especially for betweenness centrality which is dependent on more than just the local, ego-network.

For measuring accuracy of centrality metrics in a sampled network, we employ both Pearson's r correlation and Kendall's τ correlation. In Figure 3, the correlations for edge sampling probabilities, $s \in [0, 1]$, are plotted for the small and large example graphs. For this plot, we sample from fixed sampling probabilities (50 distinct values). For each edge sampling probability, we draw 5 sets of edges of counts corresponding to the sampling probability. Centrality metrics are then computed from the subgraph induced by each edge set.

The larger graphs (i.e., those having more nodes) appear to produce better correlations particularly at lower sampling levels. In order to better understand the relationships between graph characteristics and the quality of the correlation, we explore both linear and non-linear models.

A simple linear (least-squares) model is first tested on the pairs of small and large graphs from each dataset:

$$r(C_y, C_y^{\mathrm{s}}) = \beta_0 + \beta_1 n + \beta_2 s + \beta_3 (n \times s) \qquad (1)$$

where C_y represent the original degree or betweenness centrality scores (i.e., $y \in \{\mathrm{D}, \mathrm{B}\}$) of the nodes found in the graph induced by the edge samples and C_y^{s} refers to the centrality metrics of those nodes in the induced graph sample. The model consists of two main effects for network size n and the edge sampling probability s and their interaction.[6] Table 3 reports the model coefficients/parameters. In lieu of the parameter name (e.g., β_1) we show the covariate which the parameter scales. The standard errors (or s.e.) for the coefficients appear underneath the model coefficients/parameters within parentheses; they represent the uncertainty inherent in the coefficient. For presentation purposes, the coefficients and standard errors for n and $(n \times s)$ are scaled by a factor of 10^4 and 10^3, respectively. Naturally, the sampling probability

Table 3. Predicting correlation (r)

	Degree Centrality		Betweenness Centrality	
Predictor	Email	Message	Email	Message
β_0	0.491***	0.669***	0.180***	0.554***
	(0.015)	(0.015)	(0.020)	(0.016)
n	1.381***	1.095***	2.800***	1.683***
	(0.091)	(0.109)	(0.115)	(0.117)
s	0.618***	0.416***	0.854***	0.537***
	(0.026)	(0.026)	(0.033)	(0.027)
$n \times s$	−0.156***	−0.128***	−0.278***	−0.193***
	(0.015)	(0.019)	(0.019)	(0.020)
AIC	−828	−870	−642	−803
BIC	−807	−849	−621	−782
Adj-R^2	0.642	0.461	0.730	0.554
n	500	500	481	497

*** $= p < 0.001$

s has a prominently positive (and significant) effect on improving the correlation across both centralities and both datasets. Surprisingly, so does n, the network node size controlling for identical levels of s. This is due to the fact that the actual number of sampled edges varies depending on the network size. The standardized effect size of s (not shown) is actually double that of n. The Adj-R^2 roughly indicates proportion of variance explained by the models and demonstrate that they modestly fit the data. Later on, we will exert more precise controls in the model by incorporating exactly the number of edges sampled.

The apparent convergent relationship between sampling level and correlation induces us to consider a non-linear relationship. So, we employ the following model fit using non-linear least squares (NLS):

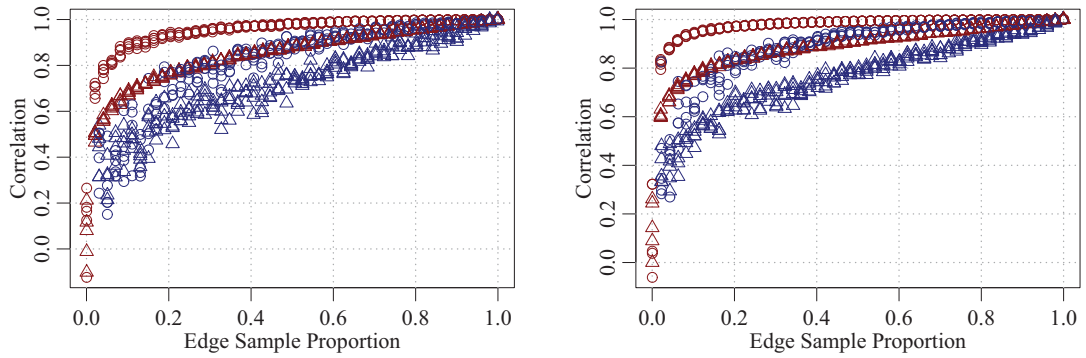
$$r(C_y, C_y^{\mathrm{s}}) = \beta_0 + \beta_1 n + \beta_2 \left(1 - s^{-\alpha}\right) + \beta_3 n \times \left(1 - s^{-\alpha}\right) \qquad (2)$$

The $1 - s^{-\alpha}$ term produces a curve similar to that delineated by the empirical relationship between s and r. We also maintain the interaction term. The results are shown in Table 4. The parameters for the n and interaction terms are respectively scaled by 10^5 and 10^4 for presentation purposes.

As NLS does not report Adj-R^2, we estimate it by replacing the α with its parametric value and employing an LS model. These models containing the non-linear term offers substantially superior fits to those in Table 3 as indicated by the AIC, BIC, and Adj-R^2. The one exception is the fit to the correlation of betweenness centrality. The dispersion of the corre-

5. Again, we describe the unweighted version of betweenness.

6. Other graph-level measures such as density or edge count cannot be employed in tandem with n as there are only two networks in the data being modeled at this point.



(a) Email (b) Message

Note: ○ and △ are Pearson's r and Kendall's τ correlations, respectively; network size colors are small and large.

Figure 3. Degree centrality correlations for sampled graphs

Table 4. Non-linear fit

	Degree Centrality		Betweenness Centrality	
Predictor	Email	Message	Email	Message
β_0	1.005***	1.005***	0.904***	0.990***
	(0.005)	(0.004)	(0.015)	(0.005)
n	0.696*	0.296	5.046***	0.941**
	(0.282)	(0.233)	(0.807)	(0.331)
$1 - s^{-\alpha}$	1.322***	0.315***	$4.7e3$	0.441***
	(0.167)	(0.024)	$1.4e6$	(0.040)
α	0.145***	0.316***	0.000	0.301***
	(0.015)	(0.015)	(0.021)	(0.018)
$n \times (1 - s^{-\alpha})$	−3.876***	−1.260***	$-1.5e4$	−1.809***
	(0.429)	(0.077)	$4.5e6$	(0.139)
AIC	−1575	−1909	−637	−1561
BIC	−1549	−1883	−612	−1536
Adj-R^2	0.920	0.933	0.728	0.903
n	500	500	481	497

*** $= p < 0.001$, ** $= p < 0.01$, * $= p < 0.05$

lation exhibited in the Email dataset renders the non-linear model less effective. In fact, the linear model is slightly superior.

We observe that the parameter (α) that describes the shape of convergence is consistent for Message. However, the effect of node count n varies considerably across the models and datasets highlighting the sensitivity of both these models and the centrality metrics to n.

2.4. Considering Further Data

We now consider samples from all of the available data comprising the varying window sizes. To reiterate, for 'Email', the window sizes are daily, 1–3 weeks, and a union graph comprising all of the available data days. For 'Message', the windows sizes are daily, 1–7 weeks, and union of all days. In Figures 4 and 5, the edge sampling levels (i.e., actual sampled edge counts) are plotted against both r and τ for up to 5,000 sampled edges. For these samples, we draw 1000 sets of edges, each set having anywhere between 20 and full edge count number of edges. Centrality metrics are then computed from the subgraph induced by each edge set.

The colors denote the rank of the network size within each dataset with dark blue indicating networks of relatively low node count and dark red indicating high node count. Pearson's r appear to converge (or reach higher correlations) more quickly than Kendall's τ; this difference highlights the skewness of the degree centrality distribution. The finding is modestly optimistic: at 5,000 edge samples of the aggregated, union graph or roughly 2.5% of all communications, the mean r lies around 0.62, whereas the mean τ is slightly lower, $\approx$0.60. The contrast in the convergence between 'Email' and 'Message' is due to the smaller network sizes of 'Message'. The correlation patterns for betweenness centrality are similar to those for degree centrality, but these converge at a lower correlation point and exhibit higher variance.

2.5. Modeling Correlations with Edge Sampling Levels

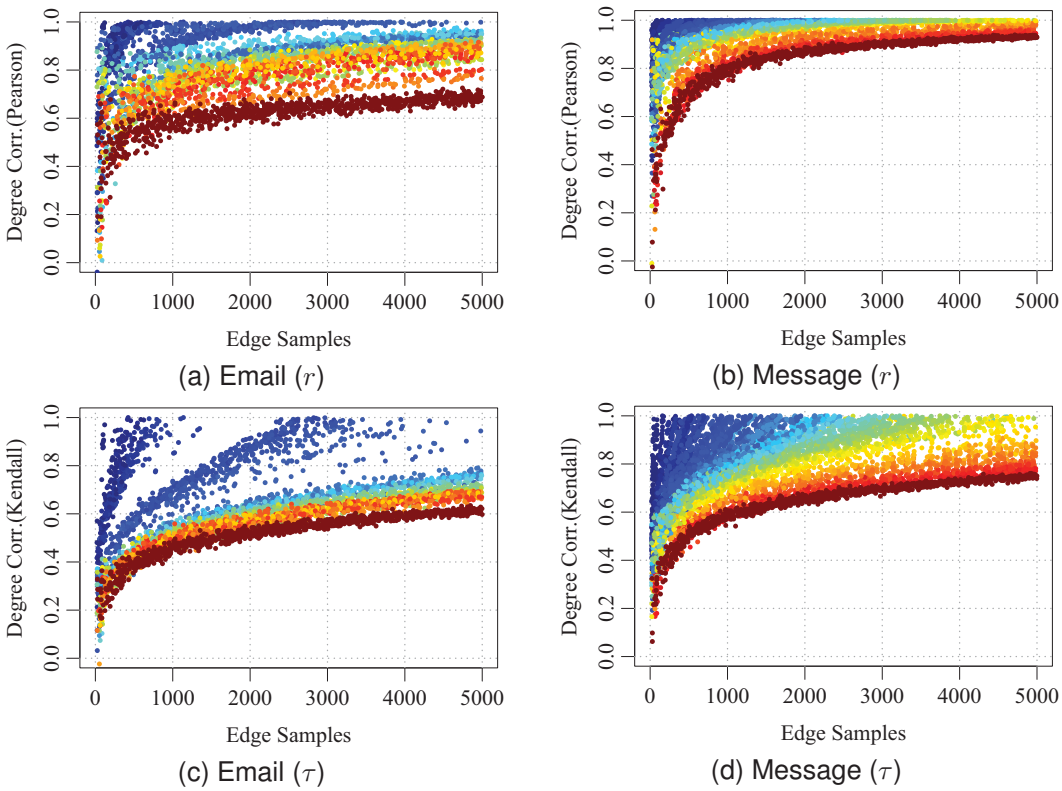
The convergence patterns are again amenable to a non-linear fit (NLS). The following models for fitting Pearson's r appear to offer a parsimonious fit.

$$r(C_y, C_y^s) = \beta_0 + \beta_1 n + \beta_2 e + [\beta_3 n_s^{\alpha_1}] + \beta_4 (1 - e_s^{\alpha_2}) \quad (3)$$

The bracketed section of the model [...] is omitted for betweenness centrality models due to the lack of predictability (i.e., collinearity with other predictors). The covariates are as follows:

Covariate	Description
n	original network size, $\lvert V \rvert$
e	original network edge count, $\lvert E \rvert$
n_s	sampled node count
e_s	sampled edge/communication count

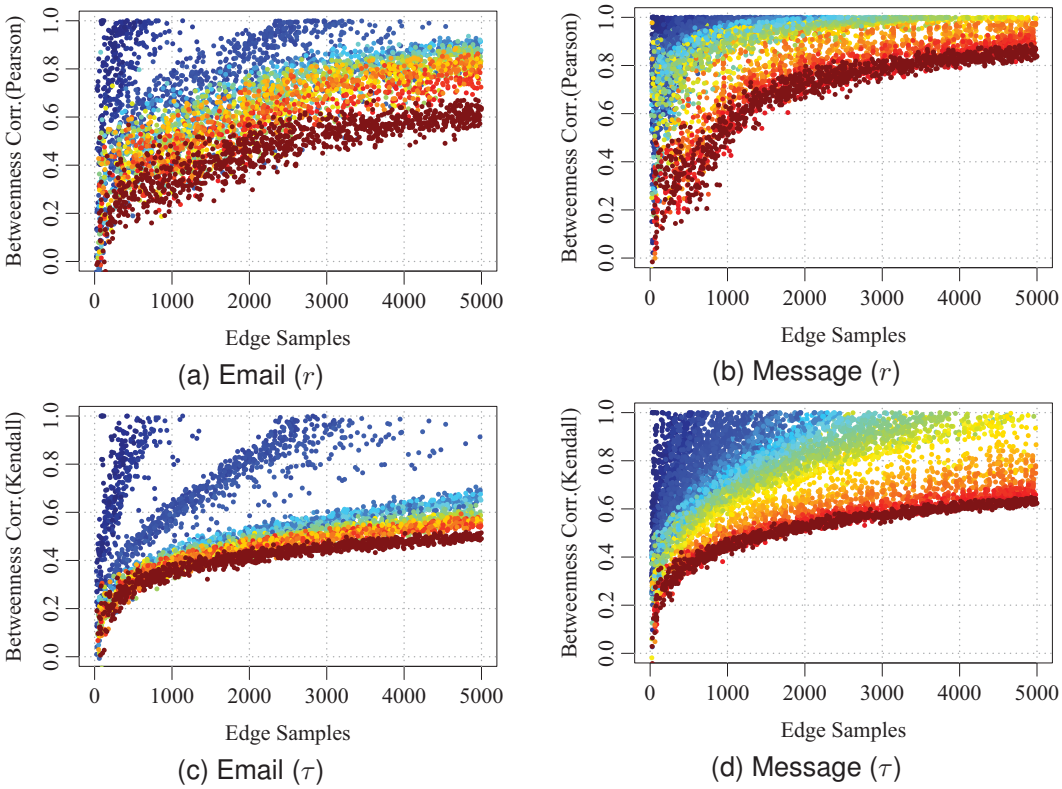
The original network graph-level measures n and e are included in the model as above findings show that



(a) Email (r) (b) Message (r)

(c) Email (τ) (d) Message (τ)

Figure 4. Degree centrality correlations for windowed networks



(a) Email (r) (b) Message (r)

(c) Email (τ) (d) Message (τ)

Figure 5. Betweenness centrality correlations for windowed networks

the overall network and edge sizes impact the quality of the sampling. The non-linear coefficients are α_1 and α_2. The n_s term is omitted from the betweenness centrality correlation model as it correlates too highly with the e_s non-linearly treated term.

In Table 5, we present the model fits to the correlations of degree and betweenness centralities for each of the datasets. For presentation purposes, the β coefficients (and standard errors) for n, e, and $n_s^{\alpha_1}$ are scaled by factors of 10^3, 10^7, and 10^1, respectively.

Table 5. Edge-sampling NLS model

	Degree Centrality		Betweenness Centrality	
Predictor	Email	Message	Email	Message
Intercept	-1.305^{***}	-0.354^{**}	-0.653^{***}	0.230^{***}
	(0.141)	(0.137)	(0.036)	(0.018)
n	-0.146^{***}	-0.323^{***}	-0.226^{***}	-0.300^{***}
	(0.003)	(0.008)	(0.004)	(0.006)
e	-0.504^{*}	33.133^{***}	1.613^{***}	30.414^{***}
	(0.250)	(1.329)	(0.321)	(1.530)
$n_s^{\alpha_1}$	0.000	0.136^{***}		
	(0.000)	(0.038)		
α_1	2.296^{***}	0.490^{***}		
	(0.186)	(0.029)		
$1-e_s^{-\alpha_2}$	2.587^{***}	1.249^{***}	3.184^{***}	2.556^{***}
	(0.119)	(0.120)	(0.034)	(0.172)
α_2	0.359^{***}	0.560^{***}	0.107^{***}	0.056^{***}
	(0.020)	(0.046)	(0.005)	(0.006)
AIC	-16970	-29993	-11766	-19219
BIC	-16918	-29937	-11727	-19177
Adj-R^2	0.783	0.736	0.794	0.631
N	5000	8000	5000	8000

$^{***} = p < 0.001$, $^{**} = p < 0.01$, $^{*} = p < 0.05$

The Adj-R^2s are again obtained by fitting a linear regression employing the α's found through the non-linear regression; these demonstrate reasonable model fits. Also, these models now include sampled edge count, instead of sampling probabilities as these data covariates offer superior model fits. They also include the node count of the sample (n_s); tests reveal superior fit when this covariate is included and also attenuated by the exponent α_1. The interaction models are omitted as they do not change the direction of coefficients' predictions and leave the models more readily interpretable.

As we examine the directions/signs of the model parameters/predictors, we find that now network node size n is no longer positively associated with improving correlations, while controlling for the other predictors. This is expected: increased node size of the original graph while holding the other sampled graph characteristics constant reduces the extent to which the sampled graph is representative of the original graph.

However, increased edge volume e, which indicates more dense networks, seems to negatively impact the correlation. Denser networks might appear to improve the correlation, but as we control for the edge sample count, again, having more edges in the original graph reduces the extent to which the sampled graph mirrors the original. This is only true for degree centrality in the Email network. The opposite effect is found for the Message data and the betweenness metrics for both data sets suggesting that impact of initial edge volume depends on the topology of the network. Also, the magnitude of e's effect varies considerably between the datasets, being much higher for 'Message'. This difference might be due to 'Message' networks all having greater densities than the 'Email' networks.

As expected, when the sampled graph contains more edges, the correlations improve. The edge sample attenuation parameter α_2 is naturally positive or else it would define a monotonically decreasing trend. Subsequently, the positive coefficient on the number of sampled edges, e_s, denotes a positive contribution towards correlation. As the correlations focus on node-level network measures, an increase in node count penalizes them, as denoted by the negative coefficient on n. However, the node count on the sampled networks apparently assists the degree centrality correlations for just the 'Message' networks.

2.6. Model Fitting with Graph-Level Centralization

It stands to reason that sampling correlations might be a function of the network topology and might be more accurate for graphs with heterogeneous structural properties (e.g., a centralized graph). So, we extend the models of Equation 2 to include degree and betweenness centralization to incorporate graph-level properties that indicate certain aspects of graph topology. Specifically, normalized versions of these centralization metrics lie between 0.0 and 1.0 and define the spectrum of heterogeneity of a centrality metric [20]. A degree centralization of 0.0 means that all nodes have equal degree centrality scores while a centralization of 1.0 means that a single node has a link to all other nodes which have no links to one another. Similarly, a betweenness centralization of 1.0 means a single node has the highest possible betweenness centrality of all graphs of having the same node count. The calculation of centralization for a centrality metric y and a given graph $G \in \{V, E\}$ is:

$$C_y^* = \frac{\sum_{v \in V}(C_y(v^*) - C_y(v))}{\max(\sum_{v \in V} C'_y(v^*) - C'_y(v))}$$

where y is the network centrality metric, i.e., D(egree) or B(etweenness); v^* is the vertex with the highest cen-

trality (C_y) in the respective graph; and the normalizing denominator $\max(\ldots)$ is the maximum theoretical possible centralization score within the distribution of graphs having the same node count $|V|$ but any number of edges.[7]

The NLS model incorporating centralization is then:

$$r(C_y, C_y^{\mathrm{s}}) = \beta_0 + \beta_1 n + \beta_2 e + [\beta_3 n_{\mathrm{s}}^{\alpha_1}] + \beta_4 \left(1 - e_{\mathrm{s}}^{\alpha_2}\right) + \beta_5 C_y^* + \{\beta_6(\langle x \rangle \times C_y^*)\} \quad (4)$$

where C_{D}^* and C_{B}^* are the degree and betweenness centralizations of the original network being sampled and $\langle x \rangle$ can be either the original network node count n or edge count e. The bracketed section $[\ldots]$ is not present in all of the models due to their high collinearity with other covariates. The curly parentheses section $\{\ldots\}$ appears in augmented models. The models for correlations of degree centralities for the Email network appear in Table 6. For presentation purposes, the coefficients and standard errors for n, e, and $\langle x \rangle \times C_{\mathrm{D}}^*$ are scaled by 10^3, 10^6, and 10^3 for these models. Also, the coefficient and s.e. for $n_{\mathrm{s}}^{\alpha_1}$ are scaled by 10^1.

Table 6. Predicting degree centrality correlation with centralization (Email)

Predictor	1	$\langle x \rangle = n$	3	$\langle x \rangle = n$	$\langle x \rangle = e$
Intercept	0.479***	−0.478***	−0.347***	−0.335***	−0.383***
	(0.049)	(0.046)	(0.029)	(0.026)	(0.029)
n	−0.171***	−0.121***	−0.174***	−0.150***	−0.192***
	(0.003)	(0.004)	(0.002)	(0.003)	(0.003)
e	0.172***	0.723***	0.293***	0.782***	1.561***
	(0.020)	(0.029)	(0.018)	(0.029)	(0.138)
$n_{\mathrm{s}}^{\alpha_1}$	−114.623***	−0.097			
	(26.614)	(0.067)			
α_1	−0.965***	0.468***			
	(0.074)	(0.079)			
$1 - e_{\mathrm{s}}^{-\alpha_2}$	2.110***	2.671***	1.975***	1.944***	2.001***
	(0.229)	(0.071)	(0.022)	(0.019)	(0.023)
α_2	0.051***	0.143***	0.186***	0.175***	0.192***
	(0.011)	(0.011)	(0.004)	(0.004)	(0.004)
C_{D}^*	0.209***	1.640***	0.207***	1.485***	0.388***
	(0.014)	(0.063)	(0.014)	(0.062)	(0.024)
$\langle x \rangle \times C_{\mathrm{D}}^*$		−0.762***		−0.691***	−0.007***
		(0.034)		(0.033)	(0.001)
AIC	−17162	−17383	−16840	−17272	−16923
BIC	−17104	−17318	−16795	−17219	−16871
Adj-R^2	0.792	0.801	0.778	0.796	0.781
N	5000	5000	5000	5000	5000

*** $= p < 0.001$

In the Email data, looking at the main effect of the centralization β coefficient, we find that degree centralization appears to bolster the sampling accuracy. However, the effect is attenuated in larger (node-count-wise) and more voluminous (communication-wise) networks; the interaction coefficient is negative. Furthermore, the non-centralization coefficients remain consistent in all of the models.

Table 7. Centralization coefficients from various models

Dataset	Predictor	1	$\langle x \rangle = n$	3	$\langle x \rangle = n$	$\langle x \rangle = e$
Email	C_{D}^*	0.209***	1.640***	0.207***	1.485***	0.388***
		(0.014)	(0.063)	(0.014)	(0.062)	(0.024)
Message	C_{D}^*	−0.035*	−0.029	−0.115***	−0.111***	−0.073***
		(0.014)	(0.028)	(0.015)	(0.029)	(0.018)
Email	C_{B}^*	0.223***	−1.514***	0.184***	−1.657***	1.250***
		(0.039)	(0.152)	(0.038)	(0.147)	(0.071)
Message	C_{B}^*	0.173***	−0.093***	0.111***	−0.273***	−0.048***
		(0.010)	(0.027)	(0.009)	(0.025)	(0.014)

*** $= p < 0.001$

The coefficients of analogous models for Email (betweenness) and Message (degree and betweenness) appear in Table 7. Only the centralization coefficients are reported for presentation purposes. While we find that degree centralization penalizes sampling accuracy for the Message data, the effects are at least consistent and likely due to structural differences between the two datasets. Furthermore, the story of betweenness centralization is more complicated. While it has a positive effect in the non-interaction models, its effects are inconsistent in the presence of an interaction term. These findings beg further, more subtle analyses.

2.7. Additional Centralization

We also examine the inclusion of the centralizations of two other prominent centrality metrics: closeness and eigenvector centrality. Closeness centrality measures how close a node is to all other nodes based on their shortest-paths (or geodesics) [22] [20]. In this paper, the harmonic (or inverse) variant of closeness is employed [23] [24]:

$$C_{\mathrm{C}}(v) = \sum_{u \in V \setminus v} \frac{1}{d_{u,v}}$$

where $d_{u,v}$ is the shortest path length between the vertices u and v. Eigenvector centrality of a node measures its relative popularity [25]. A node has high eigenvector centrality if it is connected to other nodes having high eigenvector centrality. Eigenvector centrality is typically the absolute value of the 1st eigenvector of the system [26] [27]:

$$Ax = \lambda x$$

where A is the graph in an adjacency matrix format. As eigenvector centrality is ill-defined for weighted, directed graphs, we transform our edges to be symmetric (using the minimum) and unweighted. While the models in Table 8 are analogous to those found in Table 6, only centralization coefficients are reported.

7. Typically, the "star" network produces the maximum centralization for C_{D} and C_{B} of undirected and unweighted graphs.

Table 8. Models with four centralizations

	Degree Centrality		Betweenness Centrality	
Predictor	Email	Message	Email	Message
C^*_{D}	0.088^{***}	0.050^{**}	-0.079^{**}	-0.070^{*}
	(0.016)	(0.017)	(0.026)	(0.032)
C^*_{B}	-0.333^{***}	-0.014^{*}	0.057	0.177^{***}
	(0.022)	(0.006)	(0.036)	(0.012)
C^*_{C}	1.689^{***}	$-0.065^{\hat{}}$	-0.914^{***}	$0.122^{\hat{}}$
	(0.151)	(0.033)	(0.244)	(0.065)
C^*_{E}	-0.326^{***}	-0.006	-1.372^{***}	-0.033
	(0.055)	(0.030)	(0.090)	(0.059)
AIC	-17017	-30480	-12190	-19969
BIC	-16952	-30410	-12124	-19899
Adj-R^2	0.785	0.759	0.811	0.667
n	5000	7934	4989	7912

$^{***} = p < 0.001$, $^{**} = p < 0.01$, $^{*} = p < 0.05$, $^{\hat{}} = p < 0.10$

The coefficients highlight some interesting patterns. For example, in the prediction of degree centrality correlations r, degree centralization coefficients are positive while betweenness centralization coefficients are negative. Conversely, the prediction of betweenness centrality r correlation is supported by betweenness centralization but hampered by degree centralization. Eigenvector centralization in general appears to detract from sampling accuracy. While closeness centralization has significant effects, it seems to have an unpredictable pattern. That is, sampling accuracies of network metrics are supported by the inclusion of their centralization analogue, but they are potentially penalized by structures reflecting an increase in other centralization metrics.

3. Conclusion

In this paper, we seek to understand how the sampling of edges affects the quality of network centrality metrics of sampled online communication networks compared to those of the complete data. This issue is especially relevant for online networks for which a complete enumeration of ties is often difficult to obtain or non-existent. In order to understand sampling effects, we draw random samples of edges of varying levels from two distinct empirical datasets consisting of online communication data. This sampling strategy mimics those cases in which the data is streamed (or temporal) and incomplete in an unstructured manner.

Our initial findings confirm the obvious: larger samples improve the accuracy in the node-level centrality metrics while conversely higher graph densities actually lead to degradation in the correlation coefficients. The correlations' non-linear relationships with the sample size (either edge sample size or sample proportion) render the data amenable to NLS models. Through these models, we learn that while the coefficients (or effect sizes) of the edge sampling curvature parameter and node size of the initial graph varies modestly across centrality metrics and datasets, the effect of edge count (in the original graph) is inconsistent for the Email data. This inconsistency indicates that the volume of edges can have considerable influence on the accuracy of centrality metrics. In some cases (as with degree centrality in the Email data), greater edge volume is detrimental.

We also learn that aspects of graph topology can significantly impact the correlation accuracy. However, the effects are mixed. More centralized networks which entail greater heterogeneity in the degree centrality scores than decentralized networks yield more accurate centrality metrics but only for the Email data while betweenness centralization enhances accuracy for both datasets. Furthermore, some ancillary results have shown that the effects of centralization vary depending on the type of accuracy measure used (Pearson vs. Kendall). The varied yet significant impact of centralization on sample accuracy warrants further investigation.

While the inclusion of multiple temporal windows in tandem in the models might warrant some caution as the networks across window sizes will undoubtedly be correlated, their inclusion in each model offers some general expectations for how various temporal partition sizes will affect sampling accuracy. The very different levels of accuracy show that network metrics can be very sensitive to data sampling. Therefore, the metrics are not necessarily robust for sampled data. This finding is sobering for the study of social media; in this context, researchers are almost always limited to incomplete, sampled data.

In certain cases, e.g., Twitter 1%, the process of sampling is unknown and may contain systematic biases. In the context of Twitter, the data available for most researchers via Twitter's Streaming API is always a sample. However, Twitter does not share their sampling strategy with the public, and recent research [28] shows evidence that their samples are not representative of the overall data available on Twitter (i.e., the "firehose").

Finally, most network data is incomplete or sampled from a larger whole in one way or another; this applies to social media data as well as other kinds of real-life data. Even supposedly discrete networks (e.g., students of a classroom or employees of an organization) will have consequential linkages outside the assumed boundaries. There is almost always "leakage" in such real-world systems. Consequently, future research must study these effects and constraints in more detail and strive to develop more robust metrics.

Acknowledgments

This material is based upon work supported by the United States Army Research Office under Award No. W911NF1310155. We thank the HICSS anonymous reviewers for their valuable comments.

References

[1] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley, "Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's Firehose," in *Proceeedings of ICWSM*, 2013.

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

[3] P. Pantazopoulos, M. Karaliopoulos, and I. Stavrakakis, "On the local approximations of node centrality in internet router-level topologies," in *The 7th International Workshop on Self-Organizing Systems (IFIP IWSOS'13)*, 2013.

[4] J. Pfeffer and K. M. Carley, "k-centralities: Local approximations of global measures based on shortest paths," in *Proceedings of the WWW Conference 2012, 1st International Workshop on Large Scale Network Analysis (LSNA 2012)*, 2012, pp. 1043–1050.

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

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

[7] 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," *Journal of the American Statistical Association*, vol. 101, no. 474, 2006.

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

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

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

[11] J. Leskovec and C. Faloutsos, "Sampling from large graphs," in *KDD 2006*, 2006.

[12] M. D. Choudhury, Y.-R. Lin, H. Sundaram, K. S. Candand, L. Xie, and A. Kelliher, "How does the data sampling strategy impact the discovery of information diffusion in social media?" in *ICWSM 2010*, 2010.

[13] M. E. Mehdiabadi, H. R. Rabiee, and M. Salehi, "Sampling from diffusion networks," in *Proceedings of the 2012 International Conference on Social Informatics*, 2012, pp. 106–112.

[14] J.-S. Lee and J. Pfeffer, "Estimating centrality statistics for complete and sampled networks: Some approaches and complications," in *Proceedings of the 48th Hawaii International Conference on System Sciences 2015*, 2015.

[15] J.-P. Eckmann, E. Moses, , and D. Sergi, "Entropy of dialogues creates coherent structures in e-mail traffic," *Proceedings of the National Academy of Science*, vol. 101, no. 40, pp. 14 333–14 337, 2004.

[16] T. Opsahl and P. Panzarasa, "Clustering in weighted networks," *Social Networks*, vol. 31, no. 2, pp. 155–163, 2009.

[17] K. Pearson, "Notes on regression and inheritance in the case of two parents," *Proceedings of the Royal Society of London*, vol. 58, pp. 240–242, 1895.

[18] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904.

[19] M. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1-2, pp. 81–89, 1938.

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

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

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

[23] A. Bavelas, "A mathematical model for group structure," *Human Organization*, vol. 7, pp. 16–30, 1948.

[24] 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.

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

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

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

[28] F. Morstatter, J. Pfeffer, and H. Liu, "When is it biased? assessing the representativeness of Twitter's Streaming API," in *Proceeedings of WWW WebScience*, 2014.