10-22-2014

Estimation of global network statistics from incomplete data

Catherine A. Bliss
University of Vermont

Christopher M. Danforth
University of Vermont

Peter Sheridan Dodds
University of Vermont

Recommended Citation

Bliss CA, Danforth CM, Dodds PS. Estimation of global network statistics from incomplete data. PloS one. 2014 Oct 22;9(10):e108471.

OPEN ACCESS Freely available online

PLOS ONE

Estimation of Global Network Statistics from Incomplete Data

Catherine A. Bliss*, Christopher M. Danforth, Peter Sheridan Dodds

Department of Mathematics and Statistics, Vermont Complex Systems Center, The Computational Story Lab, and the Vermont Advanced Computing Core, University of Vermont, Burlington, Vermont, United States of America

Abstract

Complex networks underlie an enormous variety of social, biological, physical, and virtual systems. A profound complication for the science of complex networks is that in most cases, observing all nodes and all network interactions is impossible. Previous work addressing the impacts of partial network data is surprisingly limited, focuses primarily on missing nodes, and suggests that network statistics derived from subsampled data are not suitable estimators for the same network statistics describing the overall network topology. We generate scaling methods to predict true network statistics, including the degree distribution, from only partial knowledge of nodes, links, or weights. Our methods are transparent and do not assume a known generating process for the network, thus enabling prediction of network statistics for a wide variety of applications. We validate analytical results on four simulated network classes and empirical data sets of various sizes. We perform subsampling experiments by varying proportions of sampled data and demonstrate that our scaling methods can provide very good estimates of true network statistics while acknowledging limits. Lastly, we apply our techniques to a set of rich and evolving large-scale social networks, Twitter reply networks. Based on 100 million tweets, we use our scaling techniques to propose a statistical characterization of the Twitter Interactome from September 2008 to November 2008. Our treatment allows us to find support for Dunbar's hypothesis in detecting an upper threshold for the number of active social contacts that individuals maintain over the course of one week.

Citation: Bliss CA, Danforth CM, Dodds PS (2014) Estimation of Global Network Statistics from Incomplete Data. PLoS ONE 9(10): e108471. doi:10.1371/journal.pone.0108471

Editor: Tobias Preis, University of Warwick, United Kingdom

Received May 28, 2014; **Accepted** August 24, 2014; **Published** October 22, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. Relevant data have been deposited to Figshare: http://figshare.com/articles/Twitter_reply_networks/1152811.

Funding: The authors acknowledge the Vermont Advanced Computing Core and support by NASA (NNX-08AO96G) at the University of Vermont for Providing High Performance Computing resources that have contributed to the research results reported within this paper. CAB and PSD were funded by an NSF CAREER Award to PSD (\# 0846668). CMD and PSD were funded by a grant from the MITRE Corporation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: Catherine.Bliss@uvm.edu

Introduction

Data collected for complex networks is often incomplete due to covert interactions, measurement error, or constraints in sampling. Particular individuals may wish to remain hidden, such as members of organized crime, and individuals who are otherwise overt may have some interactions that they wish to remain hidden because those interactions are of a sensitive nature (e.g., romantic ties). In other instances, links may be erroneously inferred from spurious or noisy interactions. Furthermore, extremely large networks necessitate an understanding of how network statistics scale under various sampling regimes [1,2]. Explorations of empirically studied networks have largely ignored these biases and consequently, characterizations of the observable (sub)networks have been reported as if they represent the "true" network of interest.

When members of a population are drawn at random, each with equal selection probability, the sample statistic being studied is often a good estimate of the population statistic. Problematically, subsampling networks often induces bias: some individuals or interactions may be more likely to be selected [3]. Consider, for example, a network for which a random selection of links is observed. The collection of observed nodes in such a subnetwork is biased because large degree nodes are more likely to be included in the sample than nodes of small degree.

The development of techniques to correct sample estimates of population statistics is needed to enable more accurate portrayals of empirically studied large -scale networks and aid in efforts to model dynamics such as cascading failures and complex contagion [4–7].

A central confounding issue is that the errors introduced by biases in sampling may be exacerbated both by particular sampling strategies and by various underlying network topologies of the true network from which the subsamples are chosen [8–15]. Researchers have explored the effects of sampling by nodes [1,9,13,16–18]; sampling by edges or messages [1,2,18]; and graph exploration methods based on random walks, snowball sampling, and respondent driven sampling [1,19,20].

We organize our paper as follows. First, we outline some of the most common global network statistics. In the Methods and Materials section, we describe our data and sampling strategies. In the Analysis section, we describe scaling methods for global

Table 1. Summary of scaling techniques.

	Sampled nodes	Failed links	Sampled links	Sampled interactions
Predicted number of nodes ($\hat{N}$)	$\frac{n}{q}$	n	$\sum_{v_i \in V^*} \frac{1}{1-(1-q)^{d(v_i)}}$	$\sum_{v_i \in V^*} \frac{1}{1-(1-q)^{s(v_i)}}$
Predicted number of edges ($\hat{M}$)	$\frac{m}{q^2}$	$\frac{m}{q}$	$\frac{m}{q}$	$\sum_{e_i \in E^*} \frac{1}{1-(1-q)^{w(e_i)}}$
Predicted average degree $\left(\hat{k}_{\text{avg}}^{\text{obs}}\right)$	$\frac{k_{\text{avg}}^{\text{obs}}}{q}$	$\frac{k_{\text{avg}}^{\text{obs}}}{q}$	$\frac{2\hat{M}}{\hat{N}}$	$\frac{2\hat{M}}{\hat{N}}$
Predicted clustering ($\hat{C}$)	C	qC	$\frac{C}{q}$	–
Predicted max. degree $\left(\hat{k}_{\max}\right)$	$\frac{k_{\max}^{\text{obs}}}{q}$	$\frac{k_{\max}^{\text{obs}}}{q}$	$\frac{k_{\max}^{\text{obs}}}{q}$	$\frac{\hat{M}}{m} k_{\max}^{\text{obs}}$

doi:10.1371/journal.pone.0108471.t001

network statistics and apply our methods to four classes of simulated networks and six empirical datasets. We provide a summary of all our estimates in Table 1. In the subsequent section, we apply our methods to Twitter reply networks as both a case of scientific interest and demonstration of our methods. In the Discussion, we discuss the implications of our findings and suggest further areas of research.

Global network statistics

Real complex networks have come to be characterized by a range of functional network statistics. In this paper, we explore how descriptive measures such as the

- the number of nodes, N,
- the number of edges, M,
- degree distribution, P_k,
- the average degree, k_{avg},
- the max degree, $k_{\max}$,
- clustering coefficient, C, [21], and
- the proportion of nodes in the giant component, S,

scale with respect to missing network data. Based on our observations, we suggest predictor methods for inferring these network statistics from subsampled network data.

The most important structural feature of a network is the degree distribution, P_k, and this has been the focus of much previous work on subsampled networks. The classical Erdös-Rényi random graph model famously exhibits a Poisson degree distribution, $P_k = \frac{\lambda^k e^{-\lambda}}{k!}$ [22]. In contrast to Erdös-Rényi random networks, preferential attachment growth models describe a random process whereby new nodes attach with greater likelihood to nodes of large degree giving rise to a Power-law or Scale-free degree distribution, $Pr(k) \propto k^{-\gamma}$ [23–26]. Other distributions, such as lognormals and power-laws with exponential cutoffs may equally characterize the degree distributions of some empirical networks [27].

Previous work has explored how the degree distribution is distorted when the subnetwork is the induced subgraph on sampled nodes [9,10,13,17,18,28–30]. Han et al. [9] investigated the effect of sampling on four types of simulated networks: random graphs with (1) Poisson, (2) Exponential, (3) Power-law, and (4) Truncated normal distributions. They observed that degree distributions of sampled Erdös-Rényi random graphs appear to be linear on a log-log plot. Others have also suggested that subnetworks of Erdös-Rényi random graphs appear "power-law-like" and could be mistaken for a scale-free network [9,17]. Typically, scale-free networks have degree distributions which span several orders of magnitude and thus, subnetworks of Erdös-Rényi random graphs would not be classified as scale-free networks by most researchers. As warned in [27], further errors may be incurred when attempting to use linear regression to fit a power-law.

Stumpf and Wiuf [28] examined how degree distributions of Erdös-Rényi random graphs scale when subnetworks are obtained through uniform random sampling on nodes and "preferential sampling of nodes," whereby large degree nodes have a greater probability of being selected. They showed that Erdös-Rényi random graphs exhibit closure under subsampling by nodes (i.e., an Erdös-Rényi random graph sampled by nodes is again an Erdös-Rényi random graph). Erdös-Rényi random graphs did not exhibit closure under preferential sampling of nodes.

Stumpf et al. [13] suggested that the degree distribution of the subnetwork induced on randomly selecting nodes is independent of the proportion of nodes sampled and that the true degree distribution can only be determined by knowledge of the generating mechanism for the network. Unfortunately, this is often not known or fully understood.

Several researchers have explored techniques for estimating the true degree distribution from subnetwork data. We first examine the subnetwork degree distribution before examining attempts to solve for the true degree distribution in terms of the subnetwork degree distribution. We consider three cases. First, when links are sampled with probability q and the subnetwork is taken to be the network generated on sampled links, the probability that a node of degree i in the true network will become a node of degree k in the subnetwork $(k \leq i)$ is given by $\Pr(k|i) = \binom{i}{k} q^k (1-q)^{i-k}$. The subnetwork degree distribution can be determined by weighting these probabilities by P_i, the probability of node i appearing in the true network [31]. The subnetwork degree distribution is then given by

$$\tilde{P}_k = \begin{cases} \sum_{i=k}^{k_{\max}} \binom{i}{k} q^k (1-q)^{i-k} P_i, & \text{if } k > 0 \\ 0, & \text{if } k = 0. \end{cases} \quad (1)$$

Next, we consider subnetworks obtained by link failure. In these cases, all nodes are observed, only a proportion (q) of links are observed. This cases is nearly identical to Equation 1, except for the presence of nodes of degree zero.

$$\tilde{P}_k = \sum_{i=k}^{k_{\max}} \binom{i}{k} q^k (1-q)^{i-k} P_i, \text{ for } k \geq 0. \quad (2)$$

Lastly, we consider subnetworks obtained from the induced network on sampled nodes. In this case, the probability of observing a node is q. As such,

$$\mathbf{Pr}(v \text{ is observed and } \deg(v) \text{ is } k) = q \sum_{i=k}^{k_{\max}} \binom{i}{k} q^k (1-q)^{i-k} P_i.$$

We note that this is not the observed subnetwork degree distribution because when a subnetwork obtained from the induced network on sampled nodes is observed, the frequencies of nodes of degree k are computed relative to the number of observed nodes. This becomes

$$\mathbf{Pr}(\deg(v) \text{ is } k | v \text{ is observed}) = \frac{\mathbf{Pr}(v \text{ is observed and } \deg(v) \text{ is } k)}{q} = \sum_{i=k}^{k_{\max}} \binom{i}{k} q^k (1-q)^{i-k} P_i,$$

which is normalized. For added clarity, consider a network of N nodes and $M = 0$ edges. We observe that $\mathbf{Pr}(v \text{ is observed and } \deg(v) = 0) = q\binom{0}{0} q^0 (1-q)^0 P_0 = q$ whereas $\mathbf{Pr}(\deg(v) = k | v \text{ is observed}) = \frac{q\binom{0}{0} q^0 (1-q)^0 P_0 = q}{q} = 1$. The latter agrees with our observation, namely the (observed) network induced on sampled nodes will have all nodes of degree 0 and an observed probability distribution which is simply $P_0 = 1$.

Viewing Equation (1) as a system of k equations, we may derive an expression for the true degree distribution in terms of the observed subnetwork degree distribution. We refer the interested reader to Materials S1 for the derivation of this result:

Given a network with degree distribution P_j, with sampling fraction q, and the subnetwork degree distribution $\tilde{P}_i = \sum_{j=i}^{k_{\max}} \binom{j}{i} q^i (1-q)^{j-i} P_j$, we may solve for P_j in terms of the subnetwork degree distribution $\tilde{P}_i$. This yields

$$\hat{P}_k = \sum_{i=k}^{k_{\max}} \frac{(-1)^{i-k} \binom{i}{k} (1-q)^{i-k}}{q^i} \tilde{P}_i, \quad (3)$$

where $\hat{P}_k$ represents the predicted degree distribution and nodes of degree 0 are handled appropriately.

Verification of this result is also presented in Materials S1.

Our derivation differs from Frank [29] by a factor of $\frac{1}{q}$,

$$\hat{P}_k = \sum_{i=k}^{k_{\max}} \frac{(-1)^{i-k} \binom{i}{k} (1-q)^{i-k}}{q^{i+1}} \tilde{P}_i. \quad (4)$$

Equation 4 solves $P_k' = q \sum_{i=k}^{k_{\max}} \binom{i}{k} q^k (1-q)^{i-k} P_i$, for P_i in terms of P_k', however P_k' is not the observed degree distribution. Neither of these derivations, however, are guaranteed to be non-negative [3] and their practicality of use is limited.

Model selection methods provide a different approach by employing maximum likelihood estimates to identify which type of degree distribution characterizes a true network, given only a subnetwork degree distribution [32]. Although these methods are able to discern that some network degree distributions may be better characterized by lognormal or exponential cutoff models instead of power-laws, only models selected *a priori* for testing form the candidate pool of possible distributions.

In contrast to the model selection technique proposed by Stumpf et al. [32], we explore a probabilistic approach which utilizes knowledge of the proportion of sampled network data (q) and the subnetwork degree distribution. In doing so, we desire an estimation that captures the qualitative nature of the degree distribution without making any assumptions about candidate models. We show that reasonably good estimates of P_k can be achieved with no knowledge of the generating mechanism. With a reasonable estimate of the degree distribution available, we are able to overcome a previously noted obstacle identified by Kolaczyk [3] who notes that predictors for network statistics (sampled by links) have proven more elusive because of the need for knowledge of the true degree distribution [3]. Our method can be used in conjunction with Hortiz-Thompson estimators to reasonably predict network statistics for cases where node selection is not uniform (i.e., subnetworks generated by sampled links or weights).

In the subsequent sections, we summarize this work and show how our method surmounts this obstacle. To our knowledge, scaling techniques for networks generated by sampled interactions (e.g., weighted networks) have not been addressed in the literature and given the interest in large, social networks derived from weighted, directed interactions, we find this analysis timely and relevant.

Materials and Methods

In this paper, we focus on four sampling regimes: (1) subnetworks induced on randomly selected nodes, (2) subnetworks obtained by random failure of links, (3) subneworks generated by randomly selected links, and (4) weighted subnetworks generated by randomly selecting interactions. Motivated by our work with Twitter reply networks [33] for which we have a very good approximation of the percent of messages which are obtained, we base our work on the assumption that the proportion of missing data is known. This is a critical assumption and one that we acknowledge may not always be satisfied in practice. Efforts to estimate the proportion of missing nodes or links are intriguing, but are beyond the scope of this paper.

Unweighted, undirected networks

Our data consist of simulated and empirical networks. We generate unweighted, undirected networks with $N = 2 \times 10^5$ nodes and average degree $k_{\text{avg}} = 10$ according to four known topologies: Erdös-Rényi random graphs with a Poisson degree distribution [22], Scale-Free random graphs with a power-law degree distribution [24,34], Small world networks [35], and Range dependent networks [36]. Erdös-Rényi, Scale-free, Small world, and Range dependent models were constructed with the CONTEST Toolbox for Matlab [37]. We note that the small world networks were set to have random rewiring probability

$p = 0.1$ and preferential attachment networks were set to have $d = 5$ new links when they enter the network. Range dependent networks were set to establish a link between nodes v_i and v_j with probability $\alpha\lambda^{|j-i|-1}$ where we set $\lambda = 0.9$ and $\alpha = 1$. As noted by [37], this choice of α ensures that nodes v_i and v_{i+1} are adjacent and $\lambda^{|j-i|-1}$ ensures that short range connections are more probable than long range connections. We also examine six well known empirical network datasets: *C. elegans* [35,38], Airlines [39], Karate Club [40], Dolphins [41], Condensed matter [42], and Powergrid [35].

We sample each of these simulated and empirical networks and examine the subnetwork induced on sampled nodes (Fig. 1), the subnetwork obtained by failing links (Fig. 2), and the subnetwork generated by sampled links (Fig. 3). For a given network, 100 simulated subnetworks are obtained for a given sampling strategy and subsampling percentage q, as q varies from 5% to 100% in increments of 5%.

Weighted, undirected networks

We examine the effects of uniformly increasing edge weight (Experiment 1, Cases I–V) as well as the distribution of edge weights (Experiment 2, Cases VI and VII) on the scaling of network statistics (Table 2).

Experiment 1: Uniform distribution of edge weights. In this set of experiments, we generate Erdös-Rényi networks with $N = 2000$ nodes and $k_{\text{avg}} = 6$. We assign each edge to have equal weight, w, where $w = 1$, 2, 3, 4, or 5 (corresponding to Cases I–V). We similarly generate Scale-free networks with $N = 2000$ nodes and $k_{\text{avg}} = 6$. We then sample each of the weighted, undirected networks by randomly selecting $q \sum_{e_i \in E(G)} w(e_i)$ interactions and examine the subnetwork generated by links with $w(e_j) > 0$ (Fig. 4). This procedure is repeated to generate one hundred simulated networks for each class and varying proportions of sampled interactions (q).

Experiment 2: Non-uniform distribution of edge weights. In this set of experiments, we explore how the distribution of weights on edges can impact scaling of global network statistics. As in the previous case, we first generate an Erdös-Rényi network with $N = 2000$ and $k_{\text{avg}} = 6$. We then add weights to edges in one of two ways. In Case VI, we assume "equal effort" in that all nodes will have an equal number of interactions distributed equally among their incident edges. This requirement ensures that all nodes have equal node strength and that effort is equally distributed to each neighbor. More specifically, for node $\deg(v_i) = k$, we set each of the k edges to have weight $\lceil \frac{30}{k} \rceil$. In Case VII, for each edge we select an integer weight between 1 and 9 from a uniform probability distribution. Certainly, other variants of the weight distribution exist and their analysis may provide additional insight in future studies.

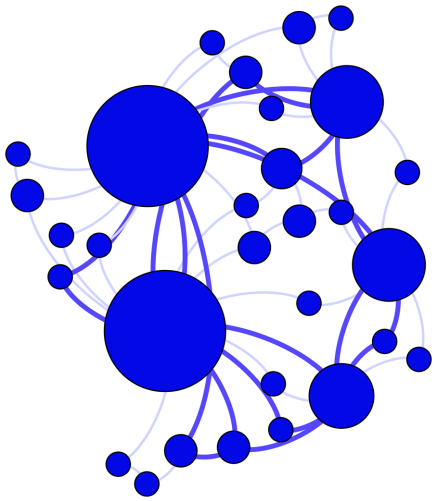

Figure 2. Failed link subnetwork. Hidden or missing links are depicted in grey. All nodes remain in the subnetwork and only visible or sampled links remain.
doi:10.1371/journal.pone.0108471.g002

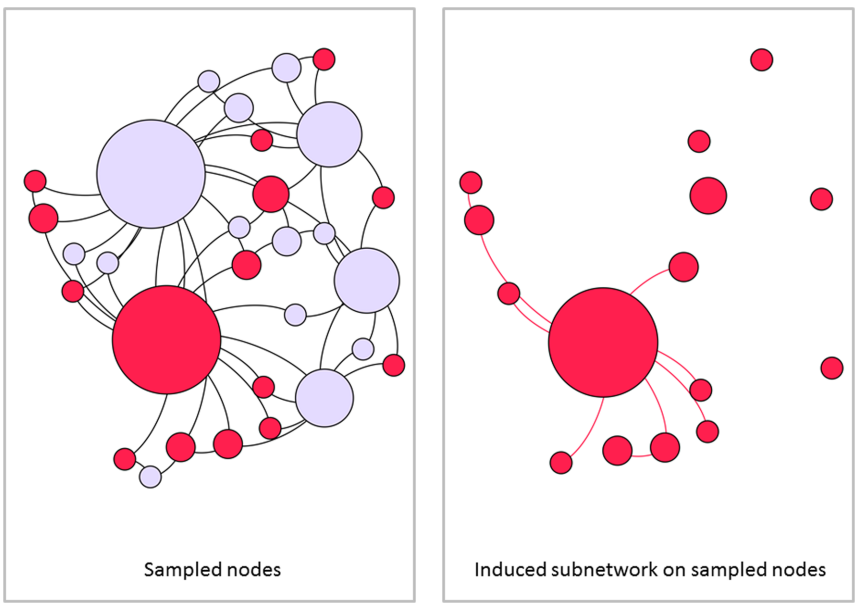

Figure 1. Node induced subnetwork on randomly sampled nodes. (Left) The true network is sampled by randomly selecting nodes (red). (Right) The node induced subnetwork consists of sampled nodes and edges whose endpoints both lie in the collection of sampled nodes.
doi:10.1371/journal.pone.0108471.g001

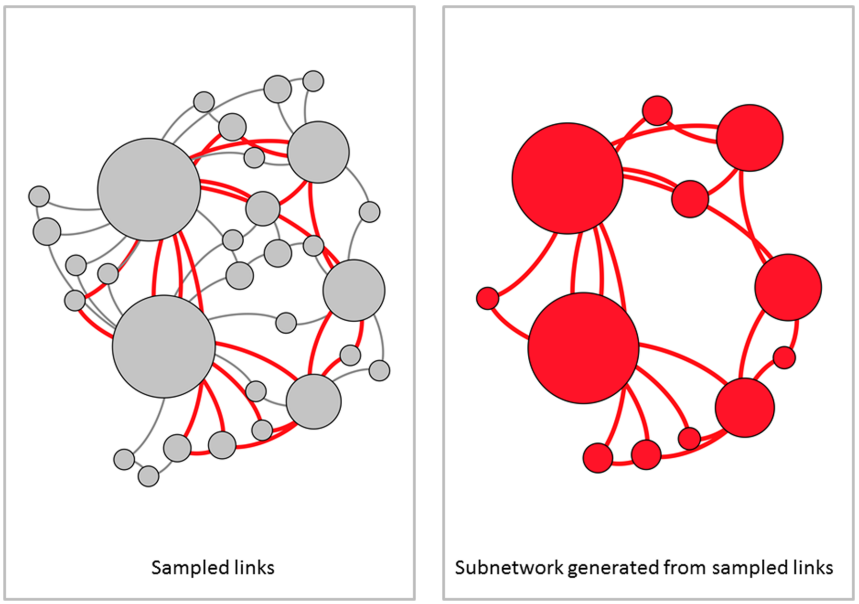

Figure 3. Subnetwork generated from sampled links. (Left) A network is sampled by randomly selecting links shown in red. (Right) The subnetwork consists of all sampled links and only nodes which are incident with the sampled links. In this type of sampling, no nodes of degree zero are included in the network. Large degree nodes are more likely to be included in the subnetwork.
doi:10.1371/journal.pone.0108471.g003

Weighted, directed networks–Twitter reply networks

Twitter reply networks [33] are weighted, directed networks constructed by establishing a directed edge between two individuals if we have a directed reply from a individual to another during the week under analysis. These networks are derived from over 100 million tweets obtained from the Twitter streaming API service during September 2008 to February 2009. We refer the interested reader to [33] for more information. The data for these networks is provided at http://www.uvm.edu/storylab/share/papers/bliss2014a/. During this time, we obtained between 25% to 55% of all tweets (Table S24 in Materials S1). Using the scaling methods developed in the Estimating global network statistics section, we predict global network statistics for the Twitter interactome during this period of time by viewing in- and out-network statistics separately (e.g., two distinct networks) to account for directionality.

Analysis

Sampling by nodes

Given a network, $G=(V, E)$, where V is the collection of nodes (or vertices) and E is the collection of links (or edges), we randomly select a portion of nodes q, where $0<q\leq 1$. The node induced subgraph on these randomly sampled nodes is given by $G^*=(V^*, E^*)$, where V^* represents the randomly selected nodes and E^* represents the edges in E for whom both endpoints lie in V^* (Fig. 1). This type of sampling occurs when a selected group, representative of the whole, is observed and all interactions between sampled individuals are known. This sampling strategy is well studied and we will only view key results here (see [3]).

Scaling of N, M, k_{avg}, C, k_{max}, and S

Given a subnetwork of size $n=qN$ known to be obtained by randomly selecting qN nodes, the number of nodes in the subsample clearly scales linearly with q (see Figs. S1a and S2a in

Table 2. Summary of weighted network experiments.

Case	k_{avg}	w_{avg}	Distribution of weights
I	6	1.0	$w(e_j)=w_{avg}$ (uniform)
II	6	2.0	$w(e_j)=w_{avg}$ (uniform)
III	6	3.0	$w(e_j)=w_{avg}$ (uniform)
IV	6	4.0	$w(e_j)=w_{avg}$ (uniform)
V	6	5.0	$w(e_j)=w_{avg}$ (uniform)
VI	6	5.0	$s(v_i)=\lceil\frac{30}{k}\rceil$ (equal effort)
VII	6	5.0	$w(e_j)=randi\{1..9\}$ (randomized)

Note: $w(e_j)$ refers to the weight of edge e_j, $s(v_j)$ refers to the strength of node v_i) and $randi\{1..9\}$ refers to a randomly selected integers between 1 and 9 (inclusive).
doi:10.1371/journal.pone.0108471.t002

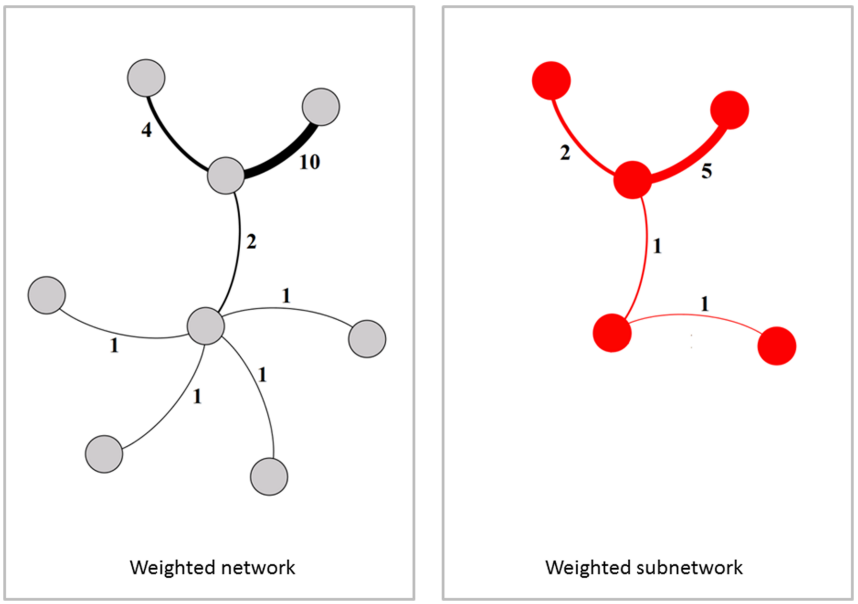

Figure 4. Weighted subnetwork generated from sampled interactions. (Left) An unsampled weighted network consists of nodes, links and weights representing the number of interactions represented by the link. (Right) Sampling by interacting produces a subsample whereby links are included in the subsample only if at least one interaction has been sampled. The subnetwork is the induced subgraph on these links with $w_i \geq 1$.
doi:10.1371/journal.pone.0108471.g004

Materials S1). The size of the true network is predicted by

$$\hat{N} = \frac{1}{q} n, \tag{5}$$

which shows good agreement with the true network statistic (Table S1 in Materials S1). Note that this result is independent of network type and is only dependent on q, the fraction of nodes subsampled, and n, the size of the subsample.

Given a network with N nodes and a subnetwork of n nodes, the probability of selecting edge e_{ij} is given by $\frac{n(n-1)}{N(N-1)}$. This is simply the probability that the two nodes, v_i and v_j, incident with the edge e_{ij}, are selected. The number of edges in the subnetwork is found by

$$m = \frac{n(n-1)}{N(N-1)} M, \tag{6}$$

where m represents the number of edges in the subnetwork and M represents the number of edges in the true network. For large networks, $m \approx q^2 M$. This agrees well with simulated results (Figs. S1b and S2b in Materials S1). The predicted number of edges is given by

$$\hat{M} = m \frac{N(N-1)}{n(n-1)}, \tag{7}$$

which scales as $\hat{M} \approx \frac{1}{q^2} m$ for large networks. This predictor shows good agreement with actual values (Table S2 in Materials S1).

The average degree, k_{avg}, is found by

$$k_{\text{avg}} = \frac{2M}{N}.$$

Given expressions for the expected number of edges (7) and the expected number of nodes (5), the expected average degree of a true network, $\hat{k}_{\text{avg}}$, based on an observed average degree of a subnetwork:

$$\hat{k}_{\text{avg}} = \frac{2\hat{M}}{\hat{N}} \tag{8}$$

$$= \frac{2m \frac{N(N-1)}{n(n-1)}}{\frac{n}{q}} \tag{9}$$

$$= \frac{2m}{n} \frac{N-1}{n-1} \tag{10}$$

$$= k_{\text{avg}}^{\text{obs}} \frac{N-1}{n-1} \tag{11}$$

$$\approx \frac{k_{\text{avg}}^{\text{obs}}}{q}, \tag{12}$$

where in line (10) we have assumed that $\hat{N} \approx N$, $N \gg 1$ and $n \gg 1$. Comparing this result to simulated subnetworks induced by subsampling nodes (Figs. S1c and S2c in Materials S1), we find very good agreement between the predicted average degree and true average degree (Table S3 in Materials S1), except for the small empirical networks (Karate club and Dolphins) sampled with low q. In these cases, we violate the assumption that $n \gg 1$ because subsamples of the Karate Club network degenerate to subnetworks

of 3 edges or less when $q \leq 0.20$. Similarly, subsamples of the Dolphin network degenerate to subnetworks of 3 edges or less when $q \leq 0.15$. When the observed number of edges in the subnetwork exceeds 3, our predicted $\hat{\boldsymbol{M}}$ has an error less than 5% (Table S3 in Materials S1).

The scaling of the max degree is highly dependent on network type, or more precisely, the relative frequency of high degree nodes. For networks with relatively few large hubs and many small nodes of small degree, $k_{\max}$ scales linearly with q and $\hat{k}_{\max} \approx \frac{k_{\max}}{q}$. For networks with many nodes of maximal degree $k_{\max}$ scales nonlinearly with q (Figs. S1d and S2d in Materials S1). An example of this would be a regular lattice. All nodes have the same (and hence maximal) degree. This pathological example is not often seen in practice. Simulated Small world networks begin as a regular lattice with random rewiring probability, p. Since our Small world networks have $p = 0.1$, our Small world networks exhibit this pathological behavior more so than several empirical Small world networks. We note that this is simply a matter of tuning p and not indicative of all Small world networks.

This distinction makes predicting the maximum degree more challenging since an accurate predictor ultimately relies on knowledge of the network type - knowledge one usually does not have in an empirical setting. Our proposed technique utilizes $\hat{k}_{\max} \approx \frac{k^{\text{obs}}_{\max}}{q}$, unless our algorithm detects a large number of nodes with degree similar to $k_{\max}$ and are assured that the subnetwork that has not degenerated to a small network ($n < 30$). More specifically, if our algorithm detects $n_{k_{\max}-1} k_{\max} - 1 > n_{k_{\max}} k_{\max}$, then we use the adjustment Equation 13, where $n_{k_{\max}-1}$ represents the number of nodes of degree $k_{\max} - 1$. In this case,

$$\hat{k}_{\max} \approx \frac{k^{\text{obs}}_{\max}}{1 - \frac{q}{\theta}}, \tag{13}$$

where $\theta =$ the number of nodes with degree greater than 75% of $k_{\max}$.

The rationale for this rough approximation is that the nodes which have high degree (>75% of the observed max. degree) may have been nearly equal contenders for losing a neighbor during subsampling. When all nodes have equal degree, the denominator of Equation 13 tends to $\hat{k}_{\max} \approx k^{\text{obs}}_{\max}$. Table S4 in Materials S1 presents the error for this predictor and demonstrates that our method performs reasonably well for most networks in our data set. To our knowledge, this is the first attempt to characterize how $k_{\max}$ scales with subsampling and we hope that future work improves upon our estimate.

We measure clustering using Newman's global clustering coefficient [21] $C_G = \frac{3 \times \tau_\Delta(G)}{\tau_3^+(G)}$, where $\tau_\Delta(G)$ denote the number of triangles on a graph and $\tau_3^+(G) = \tau_3(G) - 3\tau_\Delta(G)$, which is the number of vertex triples connected by exactly two edges (as in the notation used by [3]). Since the probability of selecting a node is q, both the number of triangles and connected vertex triples scale as q^3. Thus, $\hat{\tau}_\Delta(G) = \frac{1}{q^3}\tau_\Delta(G^*)$ and $\hat{\tau}_3^+(G) = \frac{1}{q^3}\tau_3^+(G^*)$ [43]. We then expect

$$\hat{C}_G \approx C^*_G. \tag{14}$$

This is supported by simulations (Figs. S1e and S2e in Materials S1) and small errors in $\hat{C}_G$ (Table S5 in Materials S1). We note that for small q, some subnetworks completely breakdown and no connected triples are present. In these situations, the clustering coefficient can not be computed nor can the true network's clustering coefficient be well predicted.

We next explore how the size of the giant component scales with the proportion of nodes sampled (Fig. S1f and S2f in Materials S1). For the Erdös-Rényi and Scale-free random graphs, the giant component emerges when the subnetwork has $k^{\text{sub}}_{\text{avg}} > 1$. This occurs when $qk_{\text{avg}} > 1$ and so for our simulated Erdös-Rényi and Scale-free networks, this occurs when $q = 0.10$ because the true networks have $k_{\text{avg}} = 10$. The thresholds for the emergence of the giant component in Small World and Range dependent networks are much higher. This may be due to the relatively large clustering coefficients of these networks. As suggested by Holme et al. [44], networks with a large clustering coefficient [35] are more vulnerable to random removal of nodes. We observe the same trend with Newman's global clustering coefficient.

In the case of the empirical networks, we find that the giant component emerges for q corresponding to $k^{\text{obs}}_{\text{avg}} > 1$. *C. elegans*, Airlines, and Condensed Matter networks are more resilient to random removal of nodes in that the giant component persists for small levels of q. This is most likely due to their relatively high average degrees, as compared to the other networks (heterogeneity of nodes' degrees in these networks). Heterogeneous networks demonstrate more resilience due to random removal of nodes at high levels of damage [45]. In general, it may be very difficult to predict the exact critical point at which the giant component emerges from subnetwork datasets.

Scaling of P_k. The complementary cumulative degree distribution (CCDF) becomes more distorted as smaller proportions of nodes are sampled, as shown in Figure S3 in Materials S1 and given by Equation 1. Subnetworks obtained by the induced graph on sampled nodes will often have $\tilde{P}_0 > 0$. This occurs when v_i is selected in sampling, but no neighbors of v_i are selected in the sample.

Our goal is to predict the degree distribution, given only knowledge of the proportion of nodes sampled (q) and the subnet degree distribution. We note that the probability that an observed node of degree k came from a node of degree $j \geq k$ in the true network is given by

$$\mathbf{Pr}(k|j) = \begin{cases} \binom{j}{k} q^k (1-q)^{j-k}, & \text{when } j \geq k \\ 0, & \text{when } j < k, \end{cases}$$

where q is the probability that a node's neighbor was included in the subsample and $1–q$ is the probability that a node's neighbor is not included in the subsample.

After normalizing, we find $\psi(j) = \frac{\mathbf{Pr}(k|j)}{c}$ describes the normalized probability that an observed node of degree k came from a node of degree j in the true network, where $c = \sum_{j=k}^{\infty} \mathbf{Pr}(k|j)$. Note that when $|1-q| < 1$ this series converges and we find $c = \sum_{j=k}^{\infty} \mathbf{Pr}(k|j) = \frac{1}{q}$. Thus,

$$\psi(j) = \begin{cases} q\binom{j}{k} q^k (1-q)^{j-k}, & \text{when } j \geq k \\ 0, & \text{when } j < k. \end{cases} \tag{15}$$

Let n_k represent the number of nodes of degree k. We compute

$$n_k\psi(k) = n_k\left(\frac{\binom{j}{k}q^k(1-q)^{j-k}}{c}\right) \tag{16}$$

$$= n_k\left(q\binom{j}{k}q^k(1-q)^{j-k}\right), \tag{17}$$

where we use Stirling's approximation to estimate the binomial coefficients for large j. We have taken care to include observed nodes of degree zero in this process (e.g., $k=0$ in Equation 16).

For networks with nodes of large degree (e.g., hubs), one can further speed up the computation and reduce floating point arithmetic errors by mapping back observed nodes of degree k to the expected value of the distribution obtained in Equation 15:

$$E(j) = \frac{1}{c}\sum_{j=k}^{\infty} j\binom{j}{k}q^k(1-q)^{j-k} \tag{18}$$

$$= q\frac{1-q+k}{q^2} \tag{19}$$

$$\approx \frac{k}{q}, \text{for } k \gg 1, \tag{20}$$

where $c \approx \frac{1}{q}$. In making use of $E(j) \approx \frac{k}{q}$, we perform a separate calculation for nodes of degree zero: $\left\{n_0\frac{(1-q)^j}{\sum_j(1-q)^j}\right\}_{j=1}^{4k_{\max}^{\text{obs}}}$. In all cases, we assume a finite network. We limit our calculations to $4k_{\max}^{\text{obs}}$ as a rough estimate on the upper bound needed for the sum in Equation 15.

Figure S4 in Materials S1 reveals the predicted degree distribution for subnets induced on varying levels of randomly selected nodes. To test the goodness of fit for the estimated degree distribution and the true P_k, we apply the two sample Kolmogorov-Smirnov test. Figure S16 in Materials S1 shows the D test statistics for the predicted degree distributions for both estimation methods (Equations 16 and 18), as well as the D_{crit} computed from $c(\alpha)\sqrt{\frac{n_1+n_2}{n_1n_2}}$, where $c(0.05)=1.36, n_1=k_{\max}$ and $n_2=\hat{k}_{\max}$. For most networks, $D \leq D_{\text{crit}}$ for $q \geq 0.3$, suggesting that when at least 30% of network nodes are sampled, our methods provide an estimated degree distribution which is statistically indistinguishable from the true degree distribution. Although we reject the null hypothesis for the preferential attachment case, for all $q \neq 1$, we wish to point out the potential for bias in the Kolmogorov-Smirnov test with large n [46]. As shown, D_{crit} values are quite low and the bias in this test is due to large n_1 and n_2. The statistical power in this test leads to the detection of statistically significant differences, even when the absolute difference is negligible. Thus, we caution the interpretation of this statistical test and place more interest in the value $D = \max|F_{i,\text{true}} - F_{i,\text{predicted}}|$, where F_{true} and $F_{\text{prediction}}$ represent the true and predicted CDFs.

Link failure

We now turn our attention to link failure. As in the previous cases, we denote the true, unsampled network as $G=(V,E)$. Some proportion, q of links remain "on" (or present in the sample) and $1-q$ are hidden or undetected by sampling. $E^* \subseteq E$ consists of precisely the links that remain "on" and $V^*=V$ (Fig. 2). Figures S5–S6 demonstrate how network statistics scale in this sampling regime.

In this case we may use the estimator to predict the number of nodes, $\hat{N}=n$ and we may predict the number of edges by $\hat{M}=\frac{m}{q}$. The average degree is found by

$$\hat{k}_{\text{avg}} = \frac{2\hat{M}}{\hat{N}} \tag{21}$$

$$= \frac{2m}{qn} \tag{22}$$

$$= \frac{k_{\text{avg}}^{\text{obs}}}{q}. \tag{23}$$

Using Newman's global clustering coefficient $C_G = \frac{3\times\tau_\Delta(G)}{\tau_3^+(G)}$ [21], we note that $q^3\tau_\Delta(G)=\tau_\Delta(G^*)$ and $q^2\tau_3^+(G)=\tau_3^+(G^*)$ because each edge is selected with probability q. Thus,

$$\begin{aligned} C_G^* &= \frac{3\times\tau_\Delta(G^*)}{\tau_3^+(G^*)} \\ &= \frac{3q^3\times\tau_\Delta(G)}{q^2\tau_3^+(G)} \\ &= qC_G. \end{aligned}$$

Thus,

$$\hat{C}_G = \frac{1}{q}C_G^*. \tag{24}$$

We compute the maximum degree with the same method as described in in the subsection on sampling by nodes because the number of neighbors of a node scales the same in both cases. Using these estimates, we find relatively low error in the predicted the network measures for $N, M, k_{\text{avg}}, k_{\max}$, and C_G (Tables S6–S10 in Materials S1).

Several networks' giant components exhibit similar patterns of resilience when sampling by nodes or failing links. Comparing the resilience of the proportion of nodes in the giant component under sampling by nodes vs. failing links, we see that Erdös-Rényi random graphs, random graphs with preferential attachment, Airlines, Condensed matter, *C. elegans*, and Powergrid networks all perform relatively similarly under the two sampling regimes. A noticeable difference is seen in Small world, Range dependent, Karate club, and Dolphin networks. In the case of Small world and Range dependent networks, the regularity of the underlying lattice in these networks means that each time a node is not observed, this also means that k_{avg} edges are also missing. Given that the majority of nodes have roughly the same degree for these

networks, subsampling fractures the giant component quickly (i.e., for q around 0.7 and 0.8 respectively). In the case of the small Karate club and Dolphins networks sampled by nodes, the proportion of nodes in the giant component increases with decreasing q. In these cases, the network consists of relatively few nodes, which are connected. In contrast, when examining the failing links case, we have all nodes present, but these nodes are missing almost all links and the network is highly disconnected.

Figure S7 in Materials S1 reveals the distortion of the CCDF when links fail in a network and all nodes remain known to the observer. Clearly, there are nodes of degree zero that are observed in this sampling regime. The predicted degree distribution is obtained by the methods described under sampling by nodes (including the treatment of observed nodes of degree zero) and presented in Fig. S8. The results of the two sample Kolmogorov-Smirnov test reveal that the estimated degree distribution and the true degree distribution are statistically indistinguishable for $q \geq 0.3$ for most networks (Fig. S17 in Materials S1). As previously noted, the large number of observations in degree distribution for the random graph grown with preferential attachment leads to high statistical power and a low D_{crit}.

Sampling by links

The problem of missing links may also manifest itself in another manner. In contrast to the case when all nodes are known and some links are hidden, we now consider subnetworks generated by sampled links and the nodes incident to those links (Fig. 3). This type of sampling occurs in many social network settings, such as networks constructed from sampled email exchanges or message board posts. In this case, we have data pertaining to messages (links). Nodes (individuals) are only discovered when a link (email) which connects to them is detected.

In this case, edges are sampled uniformly at random and we may use our previous estimator, $\hat{M} = \frac{m}{q}$. Node inclusion is biased, however, in that nodes of high degree will be detected with greater probability than nodes of low degree precisely because they are more likely to have an incident edge sampled.

To motivate an appropriate predictor, we must first consider how the number of nodes in a subnetwork obtained by the subnetwork generated by sampled links scales with q (Figs. S9a and S10a in Materials S1). To do this, let us consider the probability that a node is included in such a subsample. If the number of edges not sampled (M–m) is less than the degree $k(v_i)$ of node v_i, then we can be certain that our node of interest will be detected in sampling. On the other hand, if $M-m \geq k(v_i)$, then the probability of v_i being in the subnetwork scales nonlinearly with q. Using the framework set forth by Kolaczyk [3], observe that there are $\binom{M-k}{m}$ ways of choosing m edges from the $M-k$ edges not incident with node v_i and there are $\binom{M}{m}$ total ways of choosing m edges from all M. Thus, we have

$$\begin{aligned}\mathbf{Pr}(v_i \text{ is sampled}) &= 1 - \mathbf{Pr}(\text{no edge incident to } v_i \text{ is sampled}) \\ &= \begin{cases} 1 - \dfrac{\binom{M-k(v_i)}{m}}{\binom{M}{m}}, & \text{if } m \leq M-k(v_i) \\ 1, & \text{if } m > M-k(v_i). \end{cases}\end{aligned}$$

The Horvitz-Thompson estimator given by

$$\hat{N} = \sum_{v_i \in V^*} \frac{1}{\pi_i}, \tag{25}$$

where $\pi_i = \mathbf{Pr}(v_i \text{ is sampled})$.

Kolaczyk [3] warns that this may not be a useful result, due to the fact that the true degree of a given node is likely to be unknown. We overcome this limitation by using our predicted degree distributions obtained by the techniques previously mentioned. Observe that when sampling by links, no nodes of degree zero will be observed. We also note that in the case when $k \ll M$ and m, we may make the following approximation which is less computationally burdensome:

$$\begin{aligned}\frac{\binom{M-k}{m}}{\binom{M}{m}} &= \frac{(M-k)!M-m)!}{M!(M-m-k)!} \\ &= \frac{(M-m)(M-m-1)(M-m-2)...(M-m-(k-1))}{M(M-1)(M-2)...(M-(k-1))} \\ &= \left(\frac{M-m}{M}\right)\left(\frac{M-1-m}{M-1}\right)...\left(\frac{M-(k-1)-m}{M-(k-1)}\right) \\ &= \left(1-\frac{m}{M}\right)\left(1-\frac{m}{M-1}\right)...\left(1-\frac{m}{M-(k-1)}\right) \\ &\approx (1-q)^{k(v_i)} \text{for } k(v_i) \text{ relatively small compared to } m \text{ and } M.\end{aligned}$$

This is simply the probability that a node of degree $k(v_i)$ loses all edges during subsampling $q^0(1-q)^k$ and thus $\mathbf{Pr}(\text{not detecting } v_i) \approx (1-q)^{k(v_i)}$. Thus,

$$\hat{N} = \sum_{v_i \in V^*} \frac{1}{\pi_i} \tag{26}$$

$$\sum_{v_i \in V^*} \frac{1}{1-\mathbf{Pr}(\text{not detecting } v_i)} \tag{27}$$

$$= \sum_{v_i \in V^*} \frac{1}{1-(1-q)^{k(v_i)}} \tag{28}$$

We apply these methods to our simulated and empirical networks.

Once $\hat{N}$ and $\hat{M}$ have been computed, the average degree is simply $\hat{k}_{\text{avg}} = \frac{2\hat{M}}{\hat{N}}$. The max degree scales roughly linearly for preferential attachment models and many of the empirical networks, however scales sublinearly in networks with a high proportion of nodes of similar degree (e.g. the regular lattice structure seen in Small world and Range dependent networks). Clustering scales approximately as $\hat{C} = \frac{c}{q}$ and the giant component shows a critical threshold which varies according to network type

and average degree. The relative errors of our predictors are summarized in Tables S11–S15 in Materials S1. The scaling of P_k and the predicted degree distribution are presented in Figs. S11 and S12.

To test the goodness of fit for the estimated degree distribution and the true P_k, we again compute $D=\max|F_{i,\text{true}}-F_{i,\text{predicted}}|$, two sample Kolmogorov-Smirnov test statistic (Fig. S18 in Materials S1). This figure shows that reasonable results are achieved when $q>0.50$, a noticeable increase in the percent of network knowledge needed, as compared to other sampling strategies (sampling by nodes and failing links).

Sampling by interactions

Lastly, we consider the case of sampling by interactions in the special case of a weighted network (Fig. 4). In this case, we begin with $G=(V,E)$, where E is a set of edges, e_j, with weight $w(e_j)$. The weight on an edge represents the number of interactions between two vertices. An alternative representation is simply a network with multiple edge between two such vertices, one for each interaction. A subnetwork generated by $q\sum_{e_j\in E}w(e_j)$ sampled interactions is simply a sampled collection of multi-edges and the nodes incident to these edges (e.g., the subnetwork generated by links with nonzero weight and nodes incident to those edges).

To consider how the number of nodes scales, we consider a similar formulation as discussed in the previous section for the probability that a given node is selected when sampling by links, however instead of the degree of a node, $k(v_i)$, we are now interested in the strength of a node. The strength of a node is given by $s(v_i)=\sum_{e_j\in\mathcal{N}(v_i)}w(e_j)$, where $\mathcal{N}(v_i)$ denotes the neighborhood of vertex v_i [47]. Let $L=\sum_{e_j\in E}w(e_j)$ represent network load and $\ell=qL$, the number of sampled interactions. If the number of interactions which are not sampled $(L-\ell)$ is less than the strength of a node $s(v_i)$, then we can be certain that node v_i will be detected in sampling.

On the other hand, if $L-\ell\geq s(v_i)$, then there are at most $\binom{L-s(v_i)}{\ell}$ ways of choosing ℓ interactions from the $L-s(v_i)$ interactions not involving node v_i. As an upper bound, we assume that the $L-s(v_i)$ interactions are distributed over $L-s(v_i)$ edges (weight of 1 on each edge) which maximizes the number of ways these could be chosen. There are at most $\binom{L}{\ell}$ total ways of choosing ℓ (distinct, labeled) interactions from all L. Letting $\mu(i)$ represent the probability that v_i is sampled, we have

$$\mu_i=1-\mathbf{Pr}(\text{no interaction incident to } v_i \text{ is sampled})=\begin{cases}1-\dfrac{\binom{L-s(v_i)}{\ell}}{\binom{L}{\ell}}, & \text{if } \ell\leq L-s(v_i)\\ 1, & \text{if } \ell>L-s(v_i).\end{cases}$$

Thus, our Horvitz-Thompson estimator is,

$$\hat{N}=\sum_{v_i\in V^*}\frac{1}{\mu_i}, \tag{29}$$

where $\mu_i=\mathbf{Pr}(v_i \text{ is sampled})$. This can be well approximated by

$$\mu_i=1-(1-q)^{s(v_i)}. \tag{30}$$

It should be noted that the strength of a node is merely predicted. Thus, effort must be made to predict the node strength distribution in the same spirit as was previously done for the degree distribution. To predict the node strength distribution, we modify Equation 17 and predict an observed node of strength s to be of strength $\frac{s}{q}$ in the true network. Applying this corrector to our subsampled weighted networks, we find low relative error in the predicted number of nodes for most networks (Tables S16 and S17 in Materials S1). An exception to this is Case I (Erdös-Rényi) for $q<0.55$. We predict the node strength to be $\frac{s}{q}\geq 2$ and yet in this case, the true network is unweighted (e.g., $w(e_j)=1,\forall e_j\in E$). If there is knowledge that the network is unweighted, this example shows that the techniques from sampling by edges subsection will yield much better results.

We now consider how the number of edges in the subnetwork scales with the proportion of sampled interactions. The probability of selecting an edge $e_j\in E$ is equal to 1-$\mathbf{Pr}$(not selecting edge e_j). Notice that when the $\ell>L-w(e_j)$, the edge e_j is certain to be included in the subsample. When $\ell\leq L-w(e_j)$, the probability of not selecting edge e_j is simply the number of ways of selecting the $L-w(e_j)$ interactions ℓ at a time, which are not on edge e_j divided by the number of ways of selecting ℓ weights from L.

$$\mathbf{Pr}(e_j \text{ is sampled})=1-\mathbf{Pr}(\text{no interaction along } e_j \text{ is sampled})=\begin{cases}1-\dfrac{\binom{L-w(e_j)}{\ell}}{\binom{L}{\ell}}, & \text{if } \ell\leq L-w(e_j)\\ 1, & \text{if } \ell>L-w(e_j).\end{cases}$$

Thus, our Horvitz-Thompson estimator is,

$$\hat{M}=\sum_{e_j\in E^*}\frac{1}{\lambda_j}, \tag{31}$$

where $\lambda_j=\mathbf{Pr}(e_j \text{ is observed})$, which is well approximated by

$$\lambda_j=1-(1-q)^{w(e_j)}. \tag{32}$$

Again, we must have knowledge of the edge weights, or be able to predict them with reasonable accuracy. To do this, we predict an edge of weight $w(e_j)$ in the subnetwork to be of edge weight $\frac{w(e_j)}{q}$ in the true network.

As the weights on edges tends to 1 (the unweighted network case), we retrieve our result for how edges scale when links are sampled (synonymous with weights in the case where $w_i=1$):

$$\begin{aligned}\lim_{w(e_j)\to 1}\mathbf{Pr}(e_j \text{ is observed}) &= \lim_{w(e_j)\to 1} 1-\mathbf{Pr}(w(e_j))\\ &= \lim_{w(e_j)\to 1} 1-\frac{\binom{L-w(e_i)}{\ell}}{\binom{L}{\ell}}\\ &= 1-\frac{\binom{M-1}{m}}{\binom{M}{m}}\\ &= 1-\frac{M-m}{M}\\ &= \frac{m}{M}\\ &= q,\end{aligned}$$

where q is the proportion of sampled links. Thus, when the weights on edges tends to 1, our Horvitz-Thompson estimator is

$$\begin{aligned}\hat{M} &= \sum_{e_j\in E^*}\frac{1}{\lambda_j},\\ &= \frac{m}{q},\end{aligned}$$

which recovers our previous result for scaling of edges when sampling by links. The scaling of network statistics is demonstrated in Fig. S13. The results of applying our estimation techniques to the node strength and degree distribution are shown in Figs. S14 and S15. The relative error incurred for the predicted number of edges is presented in Tables S18 and S19 in Materials S1.

Having found suitable predictors for N and M, the average degree may be predicted by,

$$\hat{k}_{\text{avg}} = \frac{2\hat{M}}{N}.$$

Applying these scaling techniques, we obtain reasonably low error for both networks in both experiments 1 and 2 (Tables S20 and S21 in Materials S1).

To estimate $k_{\max}$, we recognize that the observed max degree will need to be scaled by roughly the proportion of missing edges. Using $\frac{\hat{M}}{m}$ as our scaling factor, we find relatively high error for both networks (Tables S22 and S23). This is due to errors in the $\hat{M}$ hindering accuracy in $\hat{k}_{\max}$.

Estimating the size of the Twitter interactome

We now consider the weighted, directed network of replies whereby a link from node v_i to node v_j represents the existence of at least one reply directed from v_i to v_j and the weight on this edge represents the number of messages sent in the time period under consideration. We apply our methods to reply networks constructed from tweets gathered during the ten week period from September 9, 2008 to November 17, 2008, a period for which we have a substantially higher percentage of all authored messages.

For each of these weeks, we receive between 20–55% of all messages posted on Twitter and similarly believe that we receive approximately 20–55% of all replies posted in this period (Table S24 in Materials S1). We apply our previously developed methods to estimate the number of nodes, edges, strengths on these edges, average degree, max degree, and distribution of node strength. To help validate our predictions, we also predict the number of nodes, edges, average degree, and max degree by performing 100 sampling experiments in which a proportion q of the observed messages used for subnetwork construction. These sampling experiments essentially "hide" some of the messages from our view and thus allow us to consider how further subsampling impacts the inferred networks statistics. Curve fitting over this region of q allows us to extrapolate the network statistic to a predicted value over increased percentages of observed messages. We use this to validate with our estimated statistic using the methods from the previous section.

Number of nodes. Since our reply networks are directed, we consider both the number of nodes which make a reply (N_{repliers}) and the number of nodes which receive a reply (N_{receiver}). As expected from our previous discussion, the number of nodes scales nonlinearly with the proportion of observed messages (Fig. 5). We fit models of the form $\mathrm{N} = ax^b$ to observed data and in doing so find an excellent fit ($R^2 \approx 0.99$) for all weeks over the subsampled region (Fig. 5). Extrapolating these fitted models to $q = 1$, we find excellent agreement with our predicted number of nodes obtained from Equations 29 and 30. The predicted number of nodes from both methods agree to within $\pm 5\%$. Figure 6 reveals that the predicted number of nodes is nearly double the number of observed nodes.

Strength of nodes. Figure 7 depicts a log-log plot of the predicted node strength distribution. This plot reveals that there are fewer nodes in the high strength region than would be expected in a scale-free distribution. Figure 8 reveals that low degree nodes dominate the dataset and that many of these low degree nodes often have low average edge weight ($w_{\text{avg}} \approx 1.5$). We find a peak in the average weight per edge as a function of degree around $k \approx 10^2$. This peak is more pronounced for out-going edges. Beyond this value, a limiting factor may prevent increases in the weight per edge, a result also noted by Gonçalves et al. [48].

Number of edges. The number of edges can be predicted using Equations 31 and 32. We present our results in Figure 9. In all cases, the number of edges increases throughout the period of the study. Figure 10 depicts the predicted edge weight and degree distributions. The edge weight distribution shows that very few ($<.001\%$) edges have weight greater than 10^2. The degree distribution of the observed subnetwork can be rescaled by reassigning nodes of degree k, to nodes of degree $\frac{\hat{M}}{m}k$. Figure 10 demonstrates a slightly heavier tail in the in-degree distribution as compared to the out-degree distribution. The degree distribution reveals that fewer than .01% of the nodes have more than 10^2 distinct neighbors. This value is approximately Dunbar's number, a value suggested to be the upper limit on the number of active social contacts for humans [49].

Average degree. Once the number of nodes and edges have been predicted for the network, we may simply compute the average degree as $\hat{k}_{\text{avg,in}} = \frac{\hat{M}}{\hat{N}_{\text{receivers}}}$ and $\hat{k}_{\text{avg,out}} = \frac{\hat{M}}{\hat{N}_{\text{repliers}}}$. Upon doing so, we find that the average degree for Twitter reply

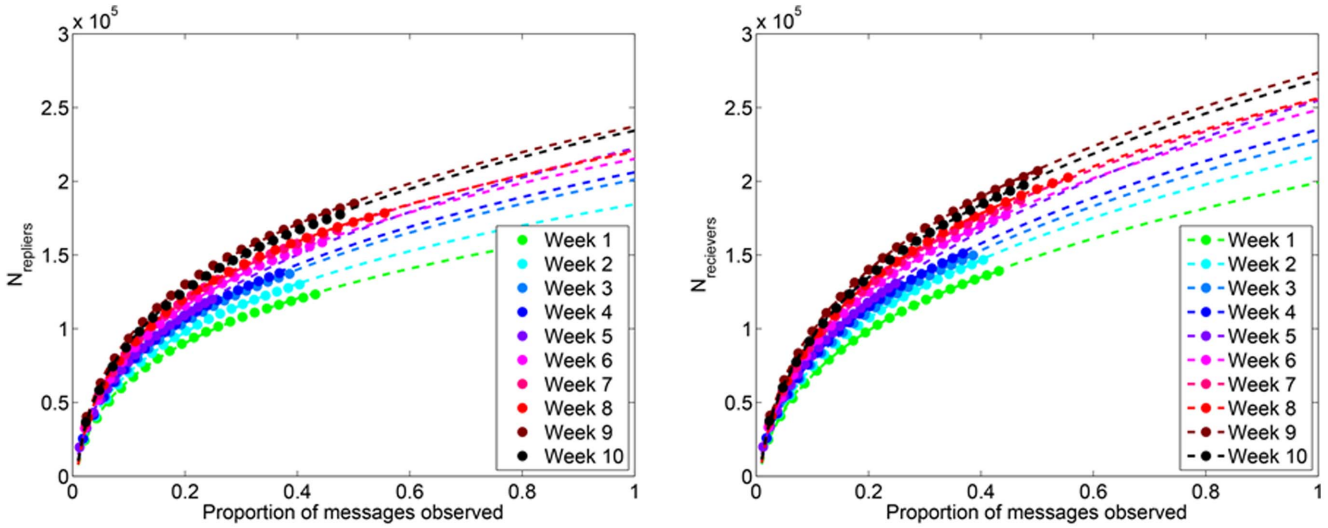

Figure 5. Number of nodes in Twitter reply subnetworks. (Left) The quantity N_{repliers} is shown for Weeks 1 to 10, where each data point (dot) represents the average over 100 simulated subsampling experiments. The dashed line represents the best fitting model of the form $N_{\text{repliers}} = ax^b$ to the observed data. We extrapolate this model to predict N_{repliers}. (Right) The same as panel, except for $N_{\text{receivers}}$.
doi:10.1371/journal.pone.0108471.g005

networks is between 4 and 5 (Fig. 11). We find that the average in-degree is less than the average out-degree (Fig. 12).

Maximum degree. The maximum degree simply scales in proportion to the probability of edge inclusion. Since the probability of edge inclusion is no longer q, as in the case of sampling by links, we may approximate the probability of edge inclusion by $\frac{m}{M}$ and thus $\hat{k}_{\max} = \frac{M}{m} k_{\max}^{\text{obs}}$. The predicted maximum degree for Twitter reply networks is shown in Figures 13 and 14.

Discussion

Network measures derived from empirical observations will often be poor estimators of the true underlying network structure of the system. We have explored four sampling regimes: (1) subnetworks induced on randomly sampled nodes, (2) subnetworks obtained when all nodes are known and some links fail or are hidden, (3) subnetworks generated from randomly sampled links, and (4) weighted subnetworks generated by randomly sampled interactions. We have described how network statistics scale under these regimes via sampling experiments on simulated and empirical networks. Our paper advances an understanding of how network statistics scale, and more importantly how to correct for missing data when the proportion of missing nodes, links or interactions is known.

A major obstacle to generating scaling techniques for subnetworks generated by sampled links or interactions has previously been the lack of a practical method for estimating the true degree distribution or node strength distribution. Problematically, the random selection of links creates a biased sample of nodes whereby hubs are more likely to be detected, and nodes of small degree are

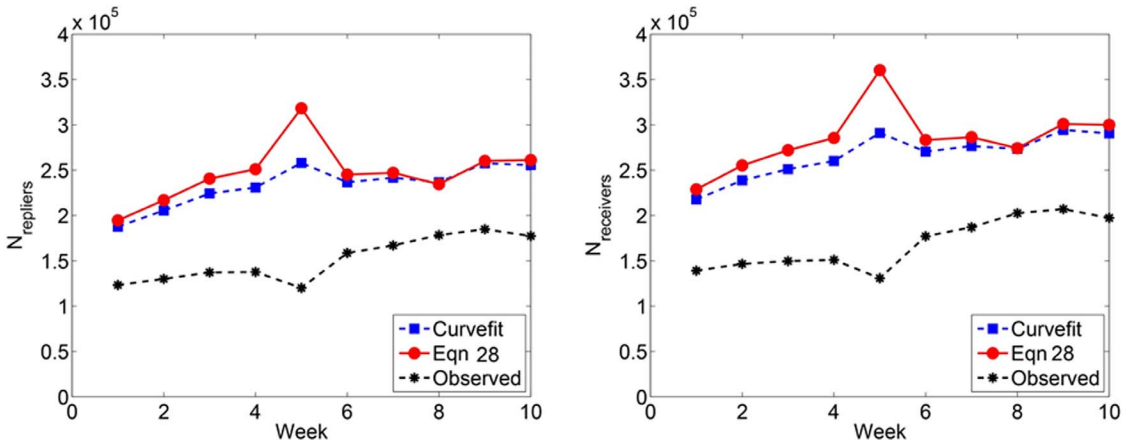

Figure 6. Predicted number of nodes in Twitter reply networks. The number of nodes observed for each week is depicted, along with the predicted number of nodes obtained from curve fitting (Fig. 5) and Equation 28. The predicted number of nodes is nearly double the number of observed nodes. The relatively low proportion of messages received for Week 5 (<25%) may be creating greater inaccuracies in the predictors for that week.
doi:10.1371/journal.pone.0108471.g006

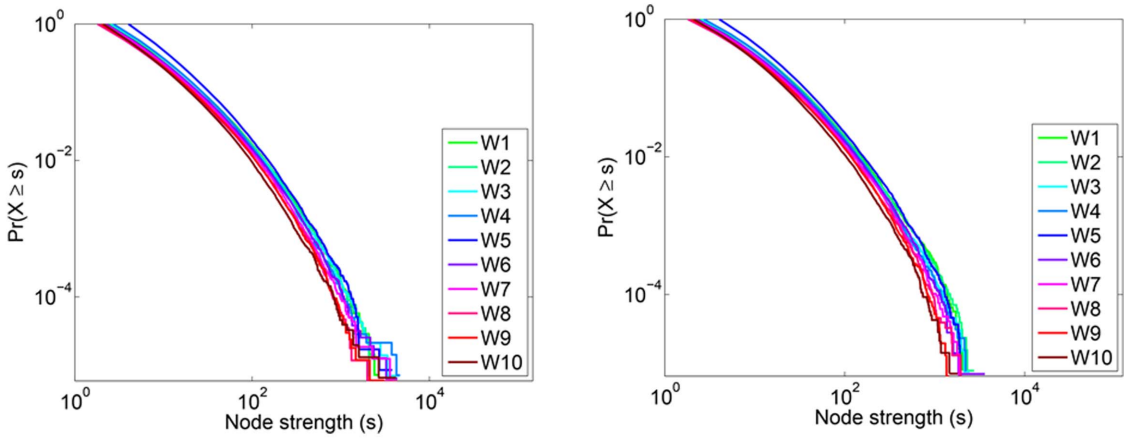

Figure 7. Predicted P_s for Twitter reply networks. (Left) The node strength distribution for in-coming interactions. (Right) The node strength distribution out-going interactions. In both cases, the distribution is heavy tailed, but falls off faster than would be expected in a scale-free distribution.
doi:10.1371/journal.pone.0108471.g007

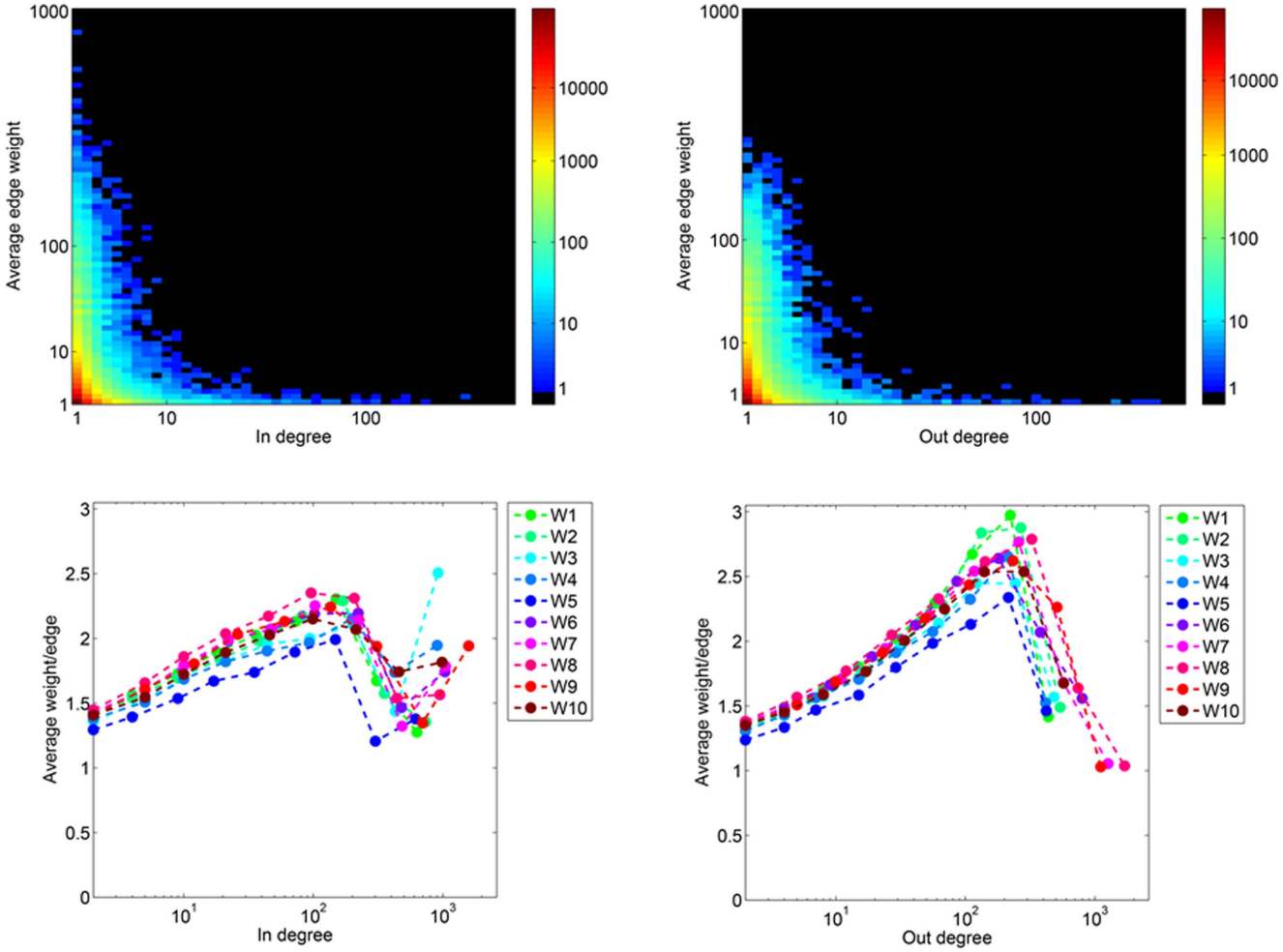

Figure 8. In, Out-degree vs. Average edge weight for Twitter reply networks. (Top, left) The average in-coming edge weight for each node of degree k is depicted in a logarithmically binned heatmap. (Top, right) The same as (a), except for out-going edges. (c.) The average weight per edge for in-coming edges as a function of k_{in} shows a gradual increase to $k_{in} \approx 10^2$ with a peak of approximately 2.2 interactions per edge. (d.) The average weight per edge for out-going edges as a function of k_{out} shows a gradual increase to $k_{out} \approx 10^2$ with a peak of between 2.5 and 3 interactions per edge.
doi:10.1371/journal.pone.0108471.g008

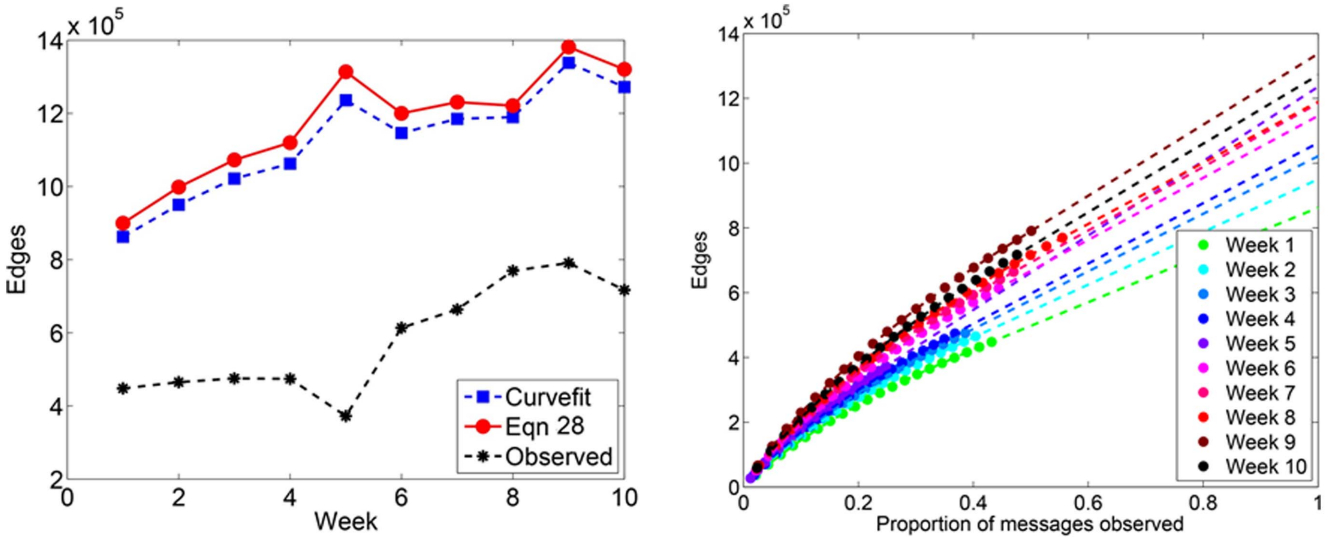

Figure 9. Predicted number of edges in Twitter reply networks. (Left) A small proportion of observed messages for Week 5 (<25%) may explain the spike in the estimated number of edges for that week. (Right) Each data point represents the number of directed edges observed, averaged over 100 simulated subsampling experiments. The dashed line extrapolates the predicted number of edges for greater proportions of sampled data.
doi:10.1371/journal.pone.0108471.g009

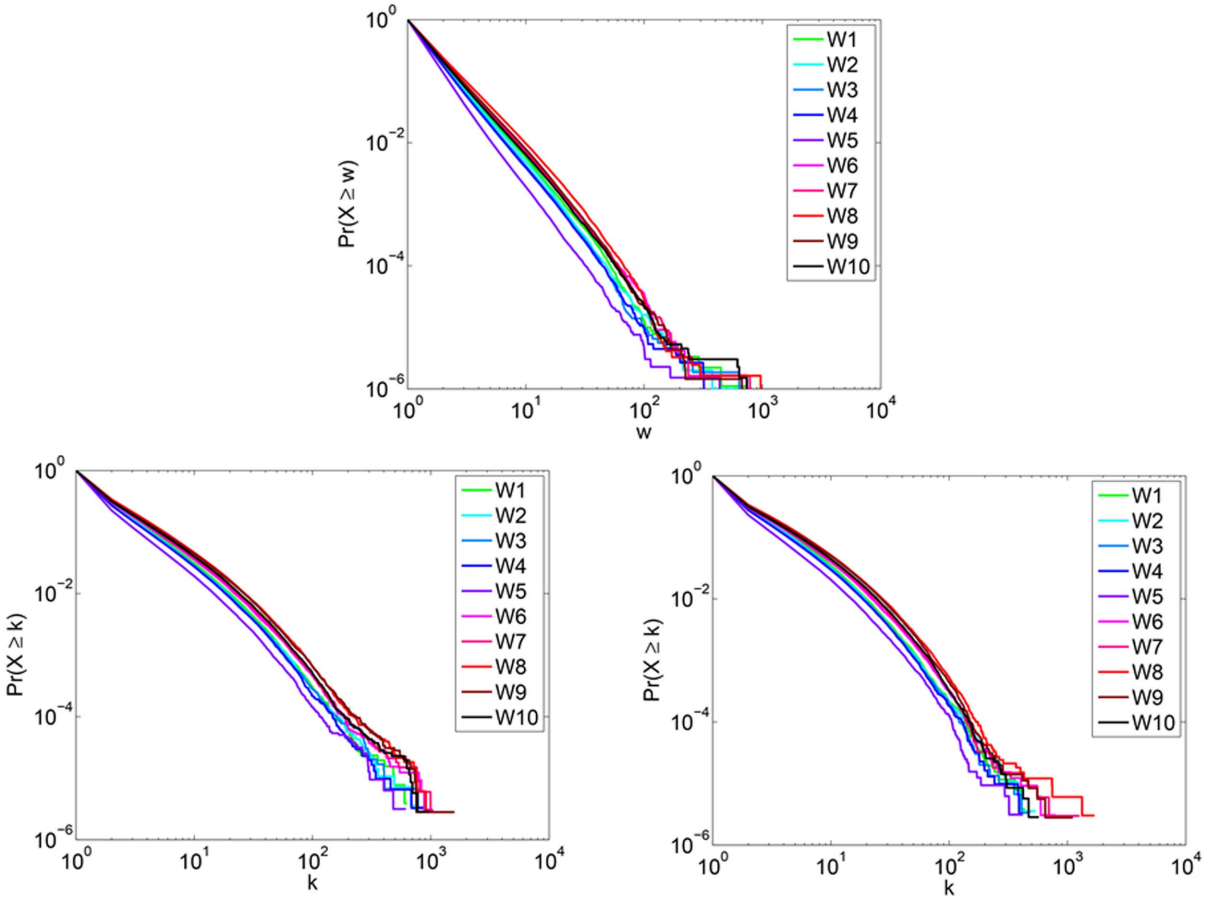

Figure 10. Predicted edge weight and degree distributions for Twitter reply networks. (Top) The predicted edge weight distribution. (Bottom, left) Predicted $Pr(k_{in})$ and (Bottom, right) $Pr(k_{out})$ for Twitter reply networks.
doi:10.1371/journal.pone.0108471.g010

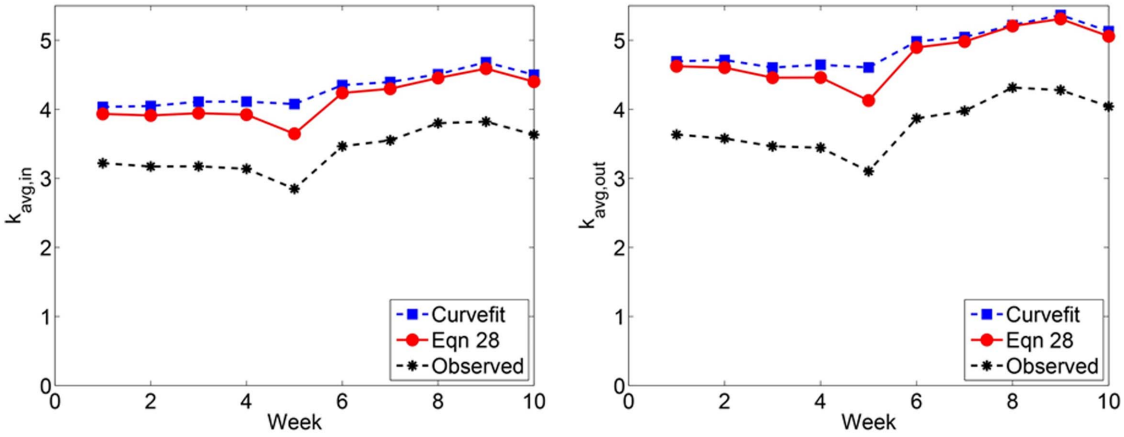

Figure 11. (Left) Predicted $k_{avg,in}$ and (Right) $k_{avg,out}$ in Twitter reply networks.
doi:10.1371/journal.pone.0108471.g011

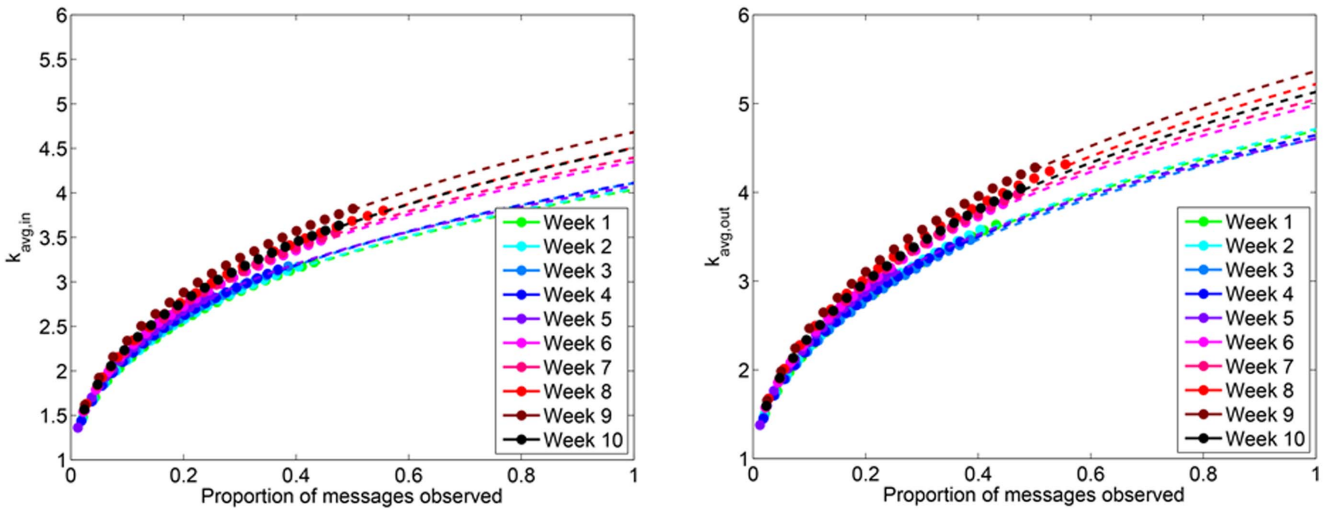

Figure 12. (Left) $k_{avg,in}$ and (Right) $k_{avg,in}$ for Twitter reply networks. Each data point represents the observed average in- and out-degree, averaged over 100 simulated subsampling experiments. The dashed line extrapolates the predicted number of edges for greater proportions of sampled data.
doi:10.1371/journal.pone.0108471.g012

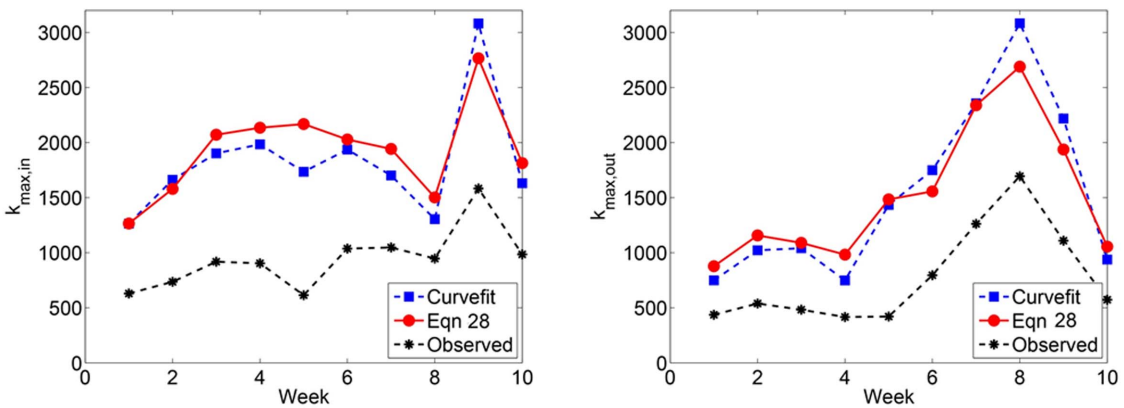

Figure 13. (Left) Predicted $k_{max,in}$ and (Right) $k_{max,out}$ in Twitter reply networks.
doi:10.1371/journal.pone.0108471.g013

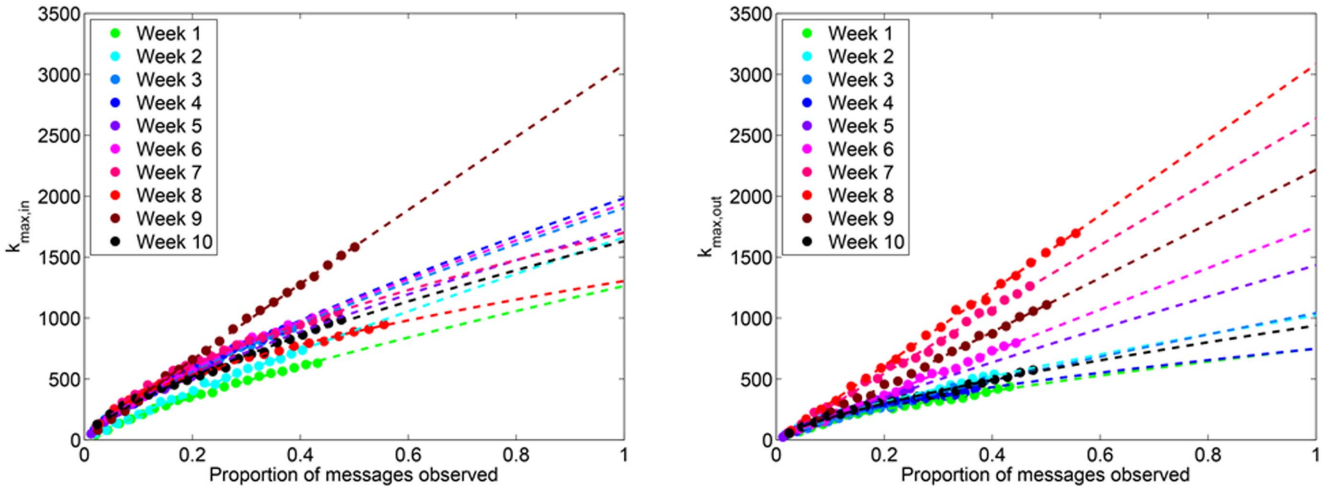

Figure 14. (Left) $k_{max,in}$ and (Right) $k_{max,in}$ for Twitter reply networks. Each data point represents the observed maximum in- and out-degree, averaged over 100 simulated subsampling experiments. The dashed line extrapolates the predicted number of edges for greater proportions of sampled data.
doi:10.1371/journal.pone.0108471.g014

more likely to go undetected. Although scaling methods have been suggested, they are based on knowledge of (or a reasonable estimate of) the degree or node strength distribution [3]. In this paper, we have overcome this obstacle by our proposed scaling techniques for the degree distribution and apply this to several simulated and empirically derived networks with reasonably good results.

Very few studies have addressed the missing data problem in empirically studied networks, such as those constructed from tweets. An exception is work by Morstatter et al. [2] who compared network statistics for the current Twitter's Spritzer ($\approx$1% of all tweets) to the full Firehose (100% of all tweets), however no methods for scaling from data collected via the API were suggested.

We concluded our work by applying our derived scaling methods to Twitter reply networks. Our work supports Dunbar's hypothesis which suggests that individuals maintain an upper limit of roughly 100–150 contacts each week [49]. Further evidence for this hypothesis comes from previous work in link prediction. Bliss et al. [50] detect the Resource Allocation index to often evolve a large, positive weight–thus contributing heavily (and positively) in the prediction of new links. This index considers the amount of time and attention one individual has as a "social resource" to spend in the social network and assumes that each node will distribute its resource equally among all neighbors. Although the presence of hubs is suggestive of preferential attachment, it is clear that the constraints of time and attention limit truly scale-free behavior in weekly Twitter reply networks. We find that the number of individuals who make replies is less than the number of individuals who receive replies.

One limitation of our work is that our scaling methods are based upon the assumption that the sampling fraction, q is known, while in practice this need not be the case. In cases where one may establish an upper and lower bound for q, our methods could be used to help establish bounds for the predicted network measures. In some cases, particularly when sampling by links or interactions, small changes in q may have relatively little impact on the predicted statistics, especially for large q. Future work that seeks to classify subnetworks by network class based on signature subsampling properties may also prove to be fruitful. With some knowledge of network class or generative model, methods for estimating q may be possible. Additionally, efforts to predict structural holes in networks from localized information may also greatly advance the field [51].

To our knowledge, this is the first attempt provide scaling methods for $k_{\max}$. While our scaling techniques for predicting $k_{\max}$ perform well for several networks, they did not perform as well on simulated networks with a regularized structure. Our rewiring probability for the simulated Small world networks was quite low, with $p = 0.1$. Our methods perform well on other networks which are known to exhibit to Small world structure, such as our empirical networks Powergrid and *C. elegans*. Future work which detects and accounts for motif distributions may improve upon our efforts here.

With an increased interest in large, networked datasets, we hope that continued efforts will aid in the understanding of how subsampled network data can be used to infer properties of the true underlying system. Our methods advance the field in this direction, not only adding to the body of literature surrounding sampling issues and Twitter's API [2], but also to the growing body of literature on incomplete network data.

Supporting Information

Materials S1 Supporting figures and tables. Derivation of Equation (2). Figure S1: Scaling of statistics for simulated subnetworks induced on sampled nodes. Figure S2: Scaling of statistics for empirical subnetworks induced on sampled nodes. Figure S3: CCDF distortion for subnetworks induced on sampled nodes. Figure S4: Predicted CCDF from subnetworks induced on sampled nodes. Figure S5: Scaling of subnetwork statistics for simulated networks obtained by failing links. Figure S6: Scaling of subnetwork statistics for empirical networks obtained by failing. Figure S7: CCDF distortion for subnetworks obtained by failing links. Figure S8: Predicted CCDF from subnetworks obtained by failing links. Figure S9: Scaling of subnetwork statistics for simulated networks induced on sampled links. Figure S10: Scaling of subnetwork statistics for empirical networks induced on sampled links. Figure S11: CCDF distortion for subnetworks induced on sampled links. Figure S12: Predicted CCDF from subnetworks induced on sampled links. Figure S13: Scaling of subnetwork statistics for simulated networks induced on sampled interactions.

Figure S14: Predicted node strength distribution for weighted, simulated networks. Figure S15: Predicted degree distribution for weighted, simulated networks. Figure S16: Kolmogorov-Smirnov two sample test for true CDF and predicted CDF from subnetworks induced on sampled nodes. Figure S17: Kolmogorov-Smirnov two sample test for true CDF and predicted CDF from subnetworks obtained by failing links. Figure S18: Kolmogorov-Smirnov two sample test for true CDF and predicted CDF from subnetworks generated by sampled links. Table S1: Error in $\hat{N}$ when sampling by nodes. Table S2: Error in $\hat{M}$ when sampling by nodes. Table S3: Error in $\hat{k}_{\text{avg}}$ when sampling by nodes. Table S4: Error in $\hat{k}_{\text{max}}$ when sampling by nodes. Table S5: Error in $\hat{C}$ when sampling by nodes. Table S6: Error in $\hat{N}$ when failing links. Table S7: Error in $\hat{M}$ when failing links. Table S8: Error in $\hat{k}_{\text{avg}}$ when failing links. Table S9: Error in $\hat{C}$ when failing links. Table S10: Error in $\hat{k}_{\text{max}}$ when failing links. Table S11: Error in $\hat{N}$ when sampling by links. Table S12: Error in $\hat{M}$ when sampling by links. Table S13: Error in $\hat{k}_{\text{avg}}$ when sampling by links. Table S14: Error in $\hat{C}$ when sampling by links. Table S15: Error in $\hat{k}_{\text{max}}$ when sampling by links. Table S16: Error in $\hat{N}$ when sampling by interactions in an Erdös-Rényi random graph. Table S17: Error in $\hat{N}$ when sampling interactions in a Scale-free weighted network. Table S18: Error in $\hat{M}$ when sampling by interactions in an Erdös-Rényi random graph. Table S19: Error in $\hat{M}$ when sampling interactions in a Scale-free weighted network. Table S20: Error in $\hat{k}_{\text{avg}}$ when sampling by interactions in an Erdös-Rényi random graph. Table S21: Error in $\hat{k}_{\text{avg}}$ when sampling interactions in a Scale-free weighted network. Table S22: Error in $\hat{k}_{\text{max}}$ when sampling by interactions in an Erdös-Rényi random graph. Table S23: Error in $\hat{k}_{\text{max}}$ when sampling interactions in a Scale-free weighted network. Table S24: Number of messages from September 2008-November 2009.
(PDF)

Author Contributions

Conceived and designed the experiments: CAB CMD PSD. Performed the experiments: CAB. Analyzed the data: CAB CMD PSD. Contributed reagents/materials/analysis tools: CAB CMD PSD. Wrote the paper: CAB. Edited the manuscript: CAB CMD PSD.

References

1. Leskovec J, Faloutsos C (2006) Sampling from large graphs. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, KDD '06, pp. 631–636. doi:http://doi.acm.org/10.1145/1150402.1150479. URL http://doi.acm.org/10.1145/1150402.1150479.
2. Morstatter F, Pfeffer J, Liu H, Carley KM (2013) Is the sample good enough? Comparing data from Twitters streaming API with Twitters firehose. Proceedings of ICWSM.
3. Kolaczyk ED (2009) Statistical Analysis of Network Data: Methods and Models. New York, NY: Springer Publishing Company, Inc., 1st edition.
4. Weng L, Menczer F, Ahn YY (2013) Virality prediction and community structure in social networks. Scientific Reports 3.
5. Hines P, Balasubramaniam K, Sanchez EC (2009) Cascading failures in power grids. Potentials, IEEE 28: 24–30.
6. Pahwa S, Scoglio C, Scala A (2014) Abruptness of cascade failures in power grids. Scientific reports 4.
7. Cotilla-Sanchez E, Hines PD, Danforth CM (2012) Predicting critical transitions from time series synchrophasor data. Smart Grid, IEEE Transactions on 3: 1832–1840.
8. Costenbader E, Valente TW (2003) The stability of centrality measures when networks are sampled. Social Networks 25: 283–307.
9. Han JDJ, Dupuy D, Bertin N, Cusick ME, Vidal M (2005) Effect of sampling on topology predictions of protein-protein interaction networks. Nature Biotechnology 23: 839–944.
10. Stumpf MPH, Wiuf C, May RM (2005) Subnets of scale-free networks are not scale-free: Sampling properties of networks. Proceedings of the National Academy of Sciences of the United States of America 102: 4221–4224.
11. Kossinets G (2006) Effects of missing data in social networks. Social Networks 28: 247–268.
12. Wiuf C, Stumpf MPH (2006) Binomial subsampling. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science 462: 1181–1195.
13. Stumpf MPH, Thorne T, de Silva E, Stewart R, An HJ, et al. (2008) Estimating the size of the human interactome. Proceedings of the National Academy of Sciences 105: 6959–6964.
14. Frantz T, Cataldo M, Carley K (2009) Robustness of centrality measures under uncertainty: Examining the role of network topology. Computational and Mathematical Organization Theory 15: 303–328.
15. Martin S, Carr RD, Faulon JL (2006) Random removal of edges from scale free graphs. Physica A: Statistical Mechanics and its Applications 371: 870–876.
16. de Silva E, Thorne T, Ingram P, Agrafioti I, Swire J, et al. (2006) The effects of incomplete protein interaction data on structural and evolutionary inferences. BMC Biology 4: 39.
17. Lakhina A, Byers J, Crovella M, Xie P (2003) Sampling biases in IP topology measurements. In: Proceedings of IEEE Infocom. URL http://www.cs.bu.edu/faculty/crovella/paper-archive/infocom03-graph-bias.pdf.
18. Lee SH, Kim PJ, Jeong H (2006) Statistical properties of sampled networks. Physical Review E 73: 016102.
19. Frank O, Snijders T (1994) Estimating the size of hidden populations using snowball sampling. Journal of Official Statistics 10: 53–53.
20. Biernacki P, Waldorf D (1981) Snowball sampling: Problems and techniques of chain referral sampling. Sociological Methods and Research 10: 141–163.
21. Newman MEJ (2003) Mixing patterns in networks. Physical Review E 67: 026126.
22. Erdös P, Rényi A (1960) On the evolution of random graphs. Magyar Tud Akad Mat Kutató Int Közl 5: 17–61.
23. de Solla Price DJ (1965) Networks of scientific papers. Science 149: 510–515.
24. Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286: 509–512.
25. Simon HA (1955) On a class of skew distribution functions. Biometrika 42: 425–440.
26. Yule GU (1925) A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS. Philosophical Transactions of the Royal Society of London Series B, Containing Papers of a Biological Character 213: 21–87.
27. Clauset A, Shalizi C, Newman M (2009) Power-law distributions in empirical data. SIAM Review 51: 661–703.
28. Stumpf MPH, Wiuf C (2005) Sampling properties of random graphs: the degree distribution. Physical Review E 72: 036118.
29. Frank O (1980) Estimation of the number of vertices of different degrees in a graph. Journal of Statistical Planning and Inference 4: 45–50.
30. Platig J, Girvan M, Ott E (2013) Robustness of network measures to link errors. Bulletin of the American Physical Society 58.
31. Cohen R, Erez K, Ben-Avraham D, Havlin S (2000) Resilience of the internet to random breakdowns. Physical Review Letters 85: 4626.
32. Stumpf M, Ingram P, Nouvel I, Wiuf C (2005) Statistical model selection methods applied to biological networks. Transactions on Computational Systems Biology III: 65–77.
33. Bliss CA, Kloumann IM, Harris KD, Danforth CM, Dodds PS (2012) Twitter reciprocal reply networks exhibit assortativity with respect to happiness. Journal of Computational Science 3: 388–397.
34. Price DDS (1976) A general theory of bibliometric and other cumulative advantage processes. Journal of the American Society for Information Science 27: 292–306.
35. Watts DJ, Strogatz SH (1998) Collective dynamics of small-world networks. Nature 393: 440–442.
36. Grindrod P (2002) Range-dependent random graphs and their application to modeling large small-world Proteome datasets. Physical Review E 66: 066702.
37. Taylor A, Higham DJ (2009) CONTEST: A controllable test matrix toolbox for MATLAB. ACM Transactions on Mathematical Software 35: 26:1–26:17.
38. White J, Southgate E, Thompson J, Brenner S (1986) The structure of the nervous system of the nematode *C. Elegans*. Philosophical Transactions of the Royal Society of London 314: 1–340.
39. Woolley-Meza O, Grady D, Thiemann C, Bagrow JP, Brockmann D (2013) Eyjafjallajökull and 9/11: The impact of large-scale disasters on worldwide mobility. PloS one 8: e69829.
40. Zachary WW (1977) An information flow model for conflict and fission in small groups. Journal of Anthropological Research: 452–473.
41. Lusseau D, Schneider K, Boisseau O, Haases P, Slooten E, et al. (2003) The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. Behavioral Ecology and Sociobiology 54: 396–405.
42. Newman MEJ (2001) The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences 98: 404–409.
43. Frank O (1978) Sampling and estimation in large social networks. Social Networks 1: 91–101.

44. Holme P, Kim BJ, Yoon CN, Han SK (2002) Attack vulnerability of complex networks. Physical Review E 65: 056109.
45. Barrat A, Barthlemy M, Vespignani A (2008) Dynamical processes on complex networks. Cambridge University Press.
46. Goldstein ML, Morris SA, Yen GG (2004) Problems with fitting to the power-law distribution. The European Physical Journal B-Condensed Matter and Complex Systems 41: 255–258.
47. Barrat A, Barthelemy M, Pastor-Satorras R, Vespignani A (2004) The architecture of complex weighted networks. Proceedings of the National Academy of Sciences of the United States of America 101: 3747–3752.
48. Gonçalves B, Perra N, Vespignani A (2011) Modeling users' activity on Twitter networks: Validation of Dunbar's Number. PLoS one 6.
49. Dunbar RIM (1995) Neocortex size and group size in primates: A test of the hypothesis. Journal of Human Evolution 28: 287–296.
50. Bliss CA, Frank MR, Danforth CM, Dodds PS (2014) An evolutionary algorithm approach to link prediction in dynamic social networks. Journal of Computational Science.
51. Bagrow JP, Desu S, Frank MR, Manukyan N, Mitchell L, et al. (2013) Shadow networks: Discovering hidden nodes with models of information flow. arXiv preprint, arXiv:13126122.