Rapid Bayesian Inference of Global Network Statistics Using Random Walks

Willow B. Kion-Crosby and Alexandre V. Morozov
Department of Physics & Astronomy and Center for Quantitative Biology, Piscataway, New Jersey 08854, USA

(Received 3 December 2017; revised manuscript received 22 May 2018; published 20 July 2018)

We propose a novel Bayesian methodology which uses random walks for rapid inference of statistical properties of undirected networks with weighted or unweighted edges. Our formalism yields high-accuracy estimates of the probability distribution of any network node-based property, and of the network size, after only a small fraction of network nodes has been explored. The Bayesian nature of our approach provides rigorous estimates of all parameter uncertainties. We demonstrate our framework on several standard examples, including random, scale-free, and small-world networks, and apply it to study epidemic spreading on a scale-free network. We also infer properties of the large-scale network formed by hyperlinks between Wikipedia pages.

DOI: 10.1103/PhysRevLett.121.038301

Over the past few years, our lives have become increasingly dependent on large-scale networks. In addition to the original computer-based networks such as the World Wide Web and the Internet, many online social networks have emerged, notably Twitter and Facebook. Our professional and personal activities are influenced daily by knowledge-sharing online services such as Wikipedia and YouTube. More generally, complex networks describe a broad spectrum of systems in nature, science, technology, and society [1,2]. Many of these networks are large and evolving, making investigation of their statistical properties a challenging task. In particular, estimating the network size becomes nontrivial if the network is too large to visit every node. Consequently, predicting various network statistics, typically from random samples of limited size, has attracted considerable attention in the literature [3–11].

Here we develop a Bayesian approach to network sampling by random walks (RWs) [5,9]. Unlike previous results, our framework can be used to build posterior probability distributions for any network node-based quantity of interest. Our approach reproduces several previously known global network statistics estimators within a single formalism, automatically removes statistical biases caused by RW sampling [5,6], and yields standard results in the uniform sampling limit. Surprisingly, accurate estimates of various network properties, including its size, are obtained after examining only a small fraction of all network nodes.

Consider a RW on a network of N nodes with weighted edges $\{w_{ji}\}$, where w_{ji} is the rate of transition from node i to node j. At each step the walker will transition to a neighboring node with a probability $P(i \to j) = w_{ji}/\sum_{k\in\{nn\}_i} w_{ki}$, where the sum is over all nearest neighbors of node i. We subdivide all network nodes into sets S_x based on the value of some property x, such as the number of links connected to the current node, known as the node degree [1]; there are N_x nodes in each set and $\mathcal{N}_x$ distinct sets. We assume that the property in question is discrete; continuous properties can be discretized by binning. We focus on undirected networks with symmetric rates, $w_{ji} = w_{ij}$. In this case, the stationary probability for the RW to occupy node i, π_i, can be determined using the steady-state master equation [12,13]:

$$\sum_{j\in\{nn\}_i} [\pi_j P(j \to i) - \pi_i P(i \to j)] = 0. \tag{1}$$

Equation (1) is satisfied if $\pi_i \sim w_i = \sum_{k\in\{nn\}_i} w_{ki}$, where w_i is the total outward rate. For unweighted networks, the node's stationary probability is proportional to its degree k_i [14]. With normalization, the stationary probabilities become $\pi_i = w_i/\sum_{i=1}^{N} w_i$.

If the walker starts from a node with property x, the average number of steps between subsequent visits to any node within the set S_x, also known as the mean return time (RT), is given by [15]

$$\langle \ell \rangle_x = \frac{1}{\sum_{i\in S_x} \pi_i}. \tag{2}$$

In the case of undirected networks,

$$\langle \ell \rangle_x = \frac{\langle w \rangle}{p_x \langle w \rangle_x}, \tag{3}$$

where $p_x = N_x/N$ is the fraction of nodes with property x, $\langle w \rangle = N^{-1}\sum_{i=1}^{N} w_i$, and $\langle w \rangle_x = N_x^{-1}\sum_{i=1}^{N_x} w_i$.

The probability of making $\Delta\ell$ steps between subsequent visits to S_x, $P(\Delta\ell)$, is asymptotically exponential for arbitrary finite networks [16]:

$$P(\Delta\ell) \simeq q_x e^{-q_x \Delta\ell}, \tag{4}$$

where $q_x = \langle \ell \rangle_x^{-1} \ll 1$ is the hitting rate of the nodes within S_x. We find empirically that the exponential ansatz for $P(\Delta\ell)$ is sufficiently accurate for our purposes, although in principle our approach is not limited to it. The likelihood that during a single RW with $\ell \gg 1$ steps the walker has visited $\mathcal{K}_x$ nodes in S_x is then given by the Poisson distribution:

$$P(\mathcal{K}_x|q_x) = \frac{(q_x\ell)^{\mathcal{K}_x}}{\mathcal{K}_x!} e^{-q_x\ell}. \tag{5}$$

This likelihood function is maximized by $\hat{q}_x = \mathcal{K}_x/\ell$, which implies $\mathcal{K}_x \ll \ell$. Assuming a uniform prior for q_x in the [0, 1] range, the posterior probability density for q_x becomes

$$P(q_x|\mathcal{K}_x) = \frac{1}{B(\mathcal{K}_x, \ell)} q_x^{\mathcal{K}_x} e^{-q_x\ell}, \tag{6}$$

where $B(\mathcal{K}_x, \ell) = \int_0^1 dq_x q_x^{\mathcal{K}_x} e^{-q_x\ell} \simeq \mathcal{K}_x!/\ell^{\mathcal{K}_x+1}$ is a normalization constant. Thus Eq. (6) is closely approximated by a gamma distribution $\Gamma(q_x; \mathcal{K}_x + 1, \ell)$, which becomes Gaussian in the $\mathcal{K}_x \gg 1$ limit, with the mean $\bar{q}_x = \hat{q}_x$ and the standard deviation $\sigma_{q_x} = \hat{q}_x/\sqrt{\mathcal{K}_x}$.

This result in combination with Eq. (3) yields a maximum likelihood estimate (MLE) and a standard error for the probability p_x of the property x:

$$\hat{p}_x = \frac{\mathcal{K}_x/\langle w \rangle_x}{\sum_x^{\mathcal{N}_x} \mathcal{K}_x/\langle w \rangle_x} \quad \text{and} \quad \sigma_{p_x} = \frac{\hat{p}_x}{\sqrt{\mathcal{K}_x}}. \tag{7}$$

If the property of the node i is its total outward rate w_i discretized into $\mathcal{N}_w$ bins, Eq. (7) yields

$$\hat{p}_{w_i} = \frac{\mathcal{K}_{w_i}/w_i}{\sum_{j=1}^{\mathcal{N}_w} \mathcal{K}_{w_j}/w_j}, \tag{8}$$

where $\mathcal{K}_{w_j}$ is the number of visits to nodes with total outward rates in the bin j. For unweighted networks ($w_{ij} = 1, \ \forall \ i, j$), $\hat{p}_{w_i}$ reduces to $\hat{p}_{k_i}$, the network degree distribution [1].

For an arbitrary node property x, each set S_x can be additionally subdivided by the binned value of w, such that

$$\hat{p}_x = \sum_{j=1}^{\mathcal{N}_w} \hat{p}_{x,w_j} = \sum_{j=1}^{\mathcal{N}_w} \frac{\mathcal{K}_{x,w_j}}{w_j} \Big/ \sum_{j=1}^{\mathcal{N}_w} \frac{\mathcal{K}_{w_j}}{w_j}. \tag{9}$$

Here, $\mathcal{K}_{x,w_j}$ is the number of visits to nodes with both property x and the total outward rates in the bin j. Thus, the knowledge of $\mathcal{K}_{w_j}$, $\mathcal{K}_{x,w_j}$, and w_j is sufficient to compute the MLE of any property x and estimate its uncertainty [Eq. (7)]. Note that the division by w_j in Eq. (9) corrects for the bias introduced by RW sampling [5–7].

The MLE of the average outward rate is given by

$$\widehat{\langle w \rangle} = \sum_{i=1}^{\mathcal{N}_w} w_i \hat{p}_{w_i} = \frac{\ell}{\sum_{j=1}^{\mathcal{N}_w} \mathcal{K}_{w_j}/w_j}, \tag{10}$$

where we used $\sum_{i=1}^{\mathcal{N}_w} \mathcal{K}_{w_i} = \ell$. The uncertainty of this estimate can be evaluated using $\sigma^2_{\langle w \rangle} = \sum_{i=1}^{\mathcal{N}_w} w_i^2 \sigma^2_{p_{w_i}}$ as well as Eqs. (7) and (8), to yield $\sigma_{\langle w \rangle} = \widehat{\langle w \rangle}/\sqrt{\ell}$, in accordance with the central limit theorem. Similarly, for an arbitrary property x

$$\widehat{\langle x \rangle} = \sum_x x \hat{p}_x \quad \text{and} \quad \sigma^2_{\langle x \rangle} = \sum_x x^2 \sigma^2_{p_x}. \tag{11}$$

Let us now suppose that the network nodes are divided into two sets: N_p randomly chosen nodes, which we shall refer to as *pseudotargets*, and all the rest. The pseudotarget nodes are drawn prior to exploring the network, so that their average outward rate $\langle w \rangle_p$ is known. Equations (3) and (5) can now be used to construct the posterior probability for the network size (assuming a uniform prior in the $[N_p, N_{\max}]$ range, where $N_{\max}$ denotes an upper limit on N):

$$P(N|\mathcal{K}_p) = \frac{N^{-\mathcal{K}_p} \exp\left(-\frac{N_p\langle w \rangle_p}{N\langle w \rangle}\ell\right)}{\sum_{\tilde{N}=N_p}^{N_{\max}} \tilde{N}^{-\mathcal{K}_p} \exp\left(-\frac{N_p\langle w \rangle_p}{\tilde{N}\langle w \rangle}\ell\right)}, \tag{12}$$

where $\mathcal{K}_p$ is the number of visits to pseudotargets. Note that using uniform priors in both Eqs. (6) and (12) does not affect the results as long as $\mathcal{K}_x$ and $\mathcal{K}_p$, respectively, are sufficiently large. Similar to Eq. (6), we find that this posterior probability rapidly becomes Gaussian as $\mathcal{K}_p$ increases, with

$$\hat{N} = \frac{\ell N_p \langle w \rangle_p}{\mathcal{K}_p \langle w \rangle} \quad \text{and} \quad \sigma_N = \frac{\hat{N}}{\sqrt{\mathcal{K}_p}}. \tag{13}$$

Using Eq. (10), we obtain

$$\hat{N} = \frac{N_p \langle w \rangle_p}{\mathcal{K}_p} \sum_{j=1}^{\mathcal{N}_w} \frac{\mathcal{K}_{w_j}}{w_j}. \tag{14}$$

Note that the error in $\hat{N}$ can be reduced either through increasing N_p or assigning highly connected nodes (network hubs) to be pseudotargets. In the $N_p = 1$ case, Eq. (13) recovers the network size estimator from Ref. [9].

The error estimate in Eq. (13), which is based on the exponential ansatz [Eq. (4)], may become too small if pseudotargets are placed too close to each other on the network, such that their RT statistics becomes correlated, i.e., dependent on the number of links between neighboring pseudotargets. However, we expect this effect to be

minimal in systems with $d_w < d_f$, where d_w is the RW dimension and d_f is the fractal dimension [15–17] and, more generally, in small-world networks, even if d_f is difficult to estimate accurately in such systems. Thus we expect our methodology to be applicable to real-life networks, which are predominantly small world and scale free [1,2]. By the same argument, choosing pseudotargets from a small random sample of the network nodes is preferable to clustered pseudotarget positioning, and, in fact, enables accurate predictions of the network size in highly disjoint and clustered systems.

In the case of a complete unweighted network in which each node is connected to all N nodes (including itself), RW sampling reduces to uniform sampling with replacement. In this limit, Eq. (7) yields $\hat{p}_x = \mathcal{K}_x/\ell$ and $\sigma_{p_x} = \sqrt{\mathcal{K}_x}/\ell$, consistent with the standard results based on binomial sampling. Moreover, $\hat{N} = \ell N_p/\mathcal{K}_p$ in this case, reproducing the classic Lincoln-Petersen estimator of biological population sizes by the mark and recapture method [18] (the differences between uniform sampling with and without replacement can be neglected in the $\mathcal{K}_p \ll N_p$ limit). These results remain valid for any network in which the total outward rate w is the same for every node. Note that the key difference between RW sampling and uniform sampling is that the former preferentially visits the nodes with larger w values, so that, given ℓ, σ_{p_x} is smaller for RW if $\langle w \rangle_x > \langle w \rangle$, and vice versa.

We have implemented the above theoretical framework as follows: for each network, N_p pseudotargets are randomly drawn and their $\langle w \rangle_p$ is computed. Commencing the RW from one of these pseudotargets, we record ℓ, $\mathcal{K}_p$, $\{\mathcal{K}_w\}$, and $\{\mathcal{K}_{x,w}\}$ for a desired set of node properties x. At each step in the RW, Eqs. (7)–(14) can then be used to infer various network properties.

To verify the validity of our algorithm on standard model systems, we have studied three unweighted, undirected networks: an Erdos-Renyi (ER) random graph [19], a scale-free (SF) random graph [1], and a small-world (SW) network [20]. Each network has $N = 10^6$ nodes. The ER network was constructed by randomly assigning $\lceil N \log(N)/2 \rceil$ edges between nodes, the SF network by the preferential attachment method [1] with $m = 2$ edges attached to new nodes, and the SW network as described in Ref. [21], with the shortcut probability $p = 1/2$.

For each network, $N_p = 10^3$ pseudotargets were randomly drawn and the network was subsequently explored with a RW for $\ell = 10^5$ steps, visiting at most 10% of all nodes. Besides network size and the node degree distribution, we have focused on posterior probabilities of the average degree of nearest-neighbor nodes, which is a measure of network degree assortativity,

$$\langle k_{nn} \rangle_i \equiv k_i^{-1} \sum_{j \in \{nn\}_i} k_j, \tag{15}$$

the clustering coefficient [3],

$$C_i \equiv \frac{2y}{k_i(k_i - 1)}, \tag{16}$$

where y is the total number of links shared by the nearest neighbors of node i, and a measure of the degree inhomogeneity [6]

$$\rho_i \equiv \sum_{j \in \{nn\}_i} (k_i^{-1/2} - k_j^{-1/2})^2. \tag{17}$$

A comprehensive summary of the inferred network statistics can be found in the Supplemental Material [22]. Although network topologies of these three systems are quite different, all statistics we have considered are predicted accurately. As an extreme example of network size inference in a highly disjoint system, we have considered two clusters connected by a single link [22]. Accurate prediction of the total network size is still possible in such a system if (i) pseudotargets are chosen as a random subset of all network nodes to minimize correlation effects and (ii) $\langle k \rangle$ is similar in each cluster. The latter requirement can be relaxed if pseudotargets are chosen, e.g., among network hubs within a narrow range of k; this extension will be addressed in future work.

Next, we have constructed a generalized ER network with $N = 10^6$ nodes and weighted edges. After placing all the edges as in the unweighted ER network, a loop was added to each node with probability $p = 1/2$. All loops and edges were then assigned a symmetric weight $w_{ij} = w_{ji}$ drawn from an exponential distribution with unit mean. For this system, we have collected statistics on each node's total outward rate, w_i, loop weight, $w_i^{\text{loop}} = w_{ii}$ (note that $w_{ii} = 0$ for nodes without loops), outward rate averaged over all nearest neighbors of node i, $\langle w_{nn} \rangle_i$, and average nearest-neighbor loop weight, $\langle w_{nn}^{\text{loop}} \rangle_i$. We have again employed a RW with $\ell = 10^5$ steps and $N_p = 10^3$ randomly drawn pseudotargets. Although the RT distribution for this system deviates from purely exponential due to loops, all the network statistics we have considered are again predicted accurately [22]. Thus our methodology is equally applicable to studies of weighted networks with loops.

After validating our approach on model systems, we have demonstrated its effectiveness in a more realistic setting, by tracking an epidemic spreading on a scale-free network in the traffic-driven epidemiological (TDE) model [23]. Following Ref. [23], we have generated the underlying network using a hidden-metric approach, which employs a tunable parameter α to control the degree of local node clustering [24,25]. The number of links in each node is drawn from a power-law distribution, $p_{k_i} \sim k_i^{-\gamma}$.

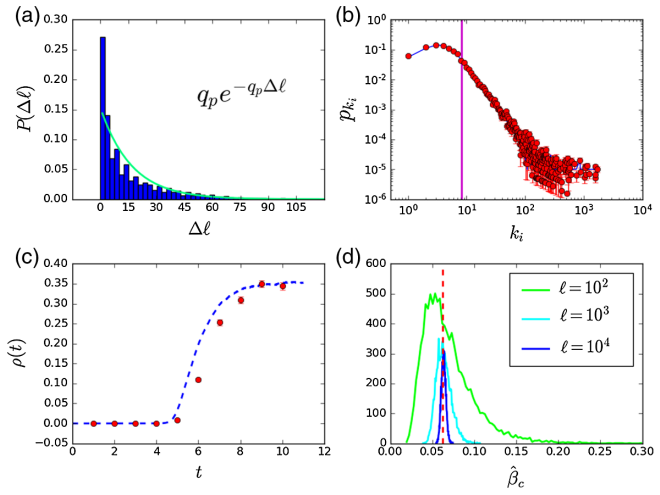

FIG. 1. Epidemic spreading statistics. (a) Pseudotarget RT distribution. Equation (4) parametrized by exact q_p is shown as a cyan solid line. (b) MLE $\pm 2\sigma$ (red circles with error bars) for the node degree distribution; exact distribution is shown as a blue solid line and its average is shown as a vertical line. (c) MLE $\pm 2\sigma$ (red circles with error bars) for the fraction of infected nodes $\rho(t)$ computed at unit time intervals, with the exact value shown as a dashed blue line. (d) Histograms of β_c MLEs obtained using 10^4 independent runs with $\ell = 10^2$, 10^3, 10^4 steps. Exact value is shown as a vertical dashed line.

For our network, we have chosen $N = 10^5$, $\gamma = 2.6$, and $\alpha = 2$ (which leads to significant clustering).

Epidemic propagation was simulated through the exchange of W contagion packets between nodes (see Ref. [23] for details). Briefly, each node can be in either susceptible or infected state; the simulation starts with a single infected node. When a packet moves from node i to node j on the network, node j becomes infected with the spreading probability β if node i was infected; infected nodes can also recover with rate μ, set to 1 without loss of generality. We have focused on the case in which contagion packets perform RWs between randomly assigned initial and destination nodes. Once a packet reaches its destination, it is removed and a new packet is added to keep W constant. On average, each packet moves once per unit simulation time. Under this choice of packet dynamics, there is a critical value of $\beta_c = (\langle k\rangle^2/\langle k^2\rangle)N/W$ above which a sustained epidemic outbreak is observed [23]. We have set $W = 2N$ and $\beta = 7 \times 10^{-1} \gg \beta_c = 6.24 \times 10^{-2}$ in the simulation.

We have used a single RW with $\ell = 10^4$ steps and $N_p = 10^3$ pseudotargets to verify the validity of our exponential ansatz [Fig. 1(a)] and predict the node degree distribution [Fig. 1(b)]; several other statistics relevant to the study of epidemics on networks [26] are listed in Table I. In addition, we have tracked time-dependent evolution of the fraction of infected nodes $\rho(t)$ [Fig. 1(c)]. We have assumed that nodes can be queried much faster than the time scales on which the epidemic spreads, and thus matched ℓ steps of our RW sampling to the unit time interval in the TDE model [Fig. 1(c), Table I]. Finally, we have predicted β_c using the evolving system's snapshot, again under the assumption that RW sampling is fast compared to the time scales of the epidemics [Fig. 1(d)].

Next, we have examined the network formed by hyperlinks between English articles on Wikipedia. Links connecting an article to itself were disregarded, multiple links between articles were counted as one, and automatic redirects were disallowed, resulting in an unweighted, undirected, loopless network consisting of all English articles, redirect pages, and disambiguation pages [27]. To assign pseudotargets, the first 5000 pages were drawn from Wikipedia's static HTML dumps. A single randomly chosen link was then taken from each of these pages and the node it pointed to was designated as a pseudotarget, resulting in $N_p = 4769$. This procedure increases the likelihood that the pseudotargets are hubs with a large number of links, facilitating collection of the network statistics since $\mathcal{K}_p$ grows more rapidly [4,9,14].

We have focused on several statistics that facilitate comparison with known properties of Wikipedia: the size of each page in bytes, ν (as provided by Wikipedia), and two variables $\chi_r, \chi_d \in \{0, 1\}$ representing whether a page is a redirect or a disambiguation page, respectively. The quantities $\langle\chi_r\rangle$, $\langle\chi_d\rangle$, and $\langle\nu_a\rangle \equiv \langle(1-\chi_r)\nu\rangle$ then give the fraction of redirect pages, disambiguation pages, and the average storage space in bytes of English articles (Wikipedia excludes redirect pages from its estimates of the number of articles [27]). The RW was run for $\ell = 5 \times 10^4$ steps, with the resulting predictions shown in Table II and Fig. 2.

TABLE I. TDE model statistics summary. Shown are MLE and 95% confidence interval ($\pm 2\sigma$) for each quantity, followed by exact values for the TDE model system. All predictions are based on a single representative RW with $\ell = 10^4$ steps corresponding to the unit time interval in the TDE model.

$\hat{N}$		$\widehat{\langle k\rangle}$		$\widehat{\langle\langle k_{nn}\rangle\rangle}$		$\widehat{\langle C\rangle}$		$\widehat{W/N}$	
$\pm 2\sigma_N$	N	$\pm 2\sigma_{\langle k\rangle}$	$\langle k\rangle$	$\pm 2\sigma_{\langle\langle k_{nn}\rangle\rangle}$	$\langle\langle k_{nn}\rangle\rangle$	$\pm 2\sigma_{\langle C\rangle}$	$\langle C\rangle$	$\pm 2\sigma_{W/N}$	W/N
1.01×10^5	10^5	8.02	8.14	64.6	67.1	0.251	0.255	2.000	2.000
$\pm 0.08 \times 10^5$		± 0.16		± 4.2		± 0.011		± 0.045	

TABLE II. Wikipedia statistics summary. Shown are MLE $\pm 2\sigma$ for each quantity. All predictions are based on a single trial with $\ell = 5 \times 10^4$ steps. N_a and N_r are the total number of English articles and redirect pages in Wikipedia, as of Dec. 2017 [28].

$\hat{N}$ $\pm 2\sigma_N$	$\widehat{\langle k \rangle}$ $\pm 2\sigma_{\langle k \rangle}$	$\widehat{\langle \chi_r \rangle}$ $\pm 2\sigma_{\langle \chi_r \rangle}$	$\widehat{\langle \chi_d \rangle}$ $\pm 2\sigma_{\langle \chi_d \rangle}$	$\widehat{\langle \nu_a \rangle}$ $\pm 2\sigma_{\langle \nu_a \rangle}$	$\widehat{\langle \nu \rangle}$ $\pm 2\sigma_{\langle \nu \rangle}$	$\hat{N}(1-\widehat{\langle \chi_r \rangle})$ $\pm 2\sigma_{N(1-\langle \chi_r \rangle)}$	N_a	$\hat{N}\,\widehat{\langle \chi_r \rangle}$ $\pm 2\sigma_{N\langle \chi_r \rangle}$	N_r	$\hat{N}\,\widehat{\langle \nu_a \rangle}$ $\pm 2\sigma_{N\langle \nu_a \rangle}$
13.4×10^6 $\pm 1.2 \times 10^6$	47.7 ± 0.4	0.6009 ± 0.0197	0.0399 ± 0.0047	2670 ± 40 bytes	2720 ± 40 bytes	5.35×10^6 $\pm 0.56 \times 10^6$	5.3×10^6	8.05×10^6 $\pm 0.79 \times 10^6$	8.0×10^6	35.8 ± 3.3 GB

We find that Wikipedia contains 13.4 million pages, each of which is connected on average to 48 other pages. The majority of Wikipedia pages, 60%, are redirect pages, and 4% are disambiguation pages. We estimate the total number of English articles (including disambiguation pages) to be 5.35×10^6, and the total number of redirect pages to be 8.05×10^6, within the confidence intervals of the values reported by Wikipedia [28] (Table II). We find the total size of English articles in Wikipedia to be 35.8 GB, in reasonable agreement with the Wikipedia statement that text alone accounts for 27.6 GB of the storage space of English articles [29].

Figure 2(a) demonstrates that the assumption of the exponential RT distribution is reasonable for Wikipedia, with some enrichment for short RTs due to the choice of network hubs as pseudotargets. Figure 2(b) shows how the estimate of the total number of Wikipedia pages evolves as $\mathcal{K}_p$ increases. As in many other Internet-based networks [30], the degree distribution of Wikipedia pages is scale free [Fig. 2(c)]. In contrast, the distribution of page sizes is not scale free, and the size of an average Wikipedia page is only 2.7 kB [Fig. 2(d), Table II].

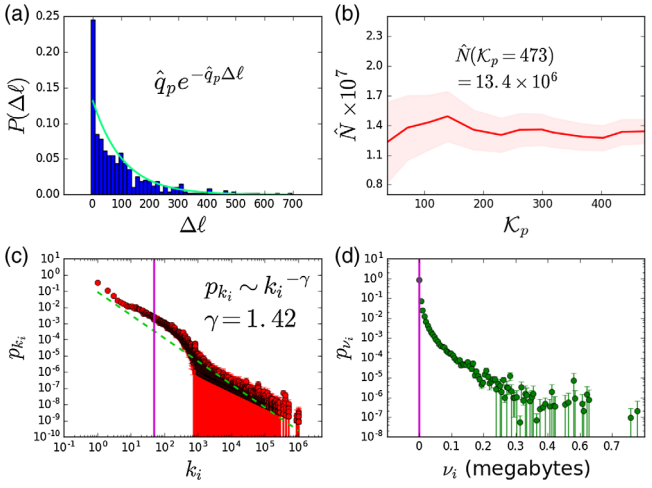

FIG. 2. Wikipedia network statistics. (a) Pseudotarget RT distribution. Equation (4) parametrized by $\hat{q}_p$ is shown as a cyan solid line. (b) MLE $\pm 2\sigma$ for N as a function of $\mathcal{K}_p$. (c) MLE $\pm 2\sigma$ for the degree distribution of Wikipedia pages of all types. Power-law fit is shown as a green dashed line. Average degree is shown as a vertical line. (d) MLE $\pm 2\sigma$ for the distribution of Wikipedia page sizes. Average size is shown as a vertical line.

In conclusion, we have presented a general Bayesian approach to inferring various network properties, including its size, by using RWs that visit only a small fraction of all network nodes. Our approach works for both weighted and unweighted undirected networks, and remains accurate in the presence of loops and node clustering. Our main assumption, that of the exponentiality of the RT distribution, appears to hold in all the cases we have examined explicitly, and can be relaxed if necessary. Our future work will focus on extending this methodology to directed and time-dependent networks.

[1] R. Albert and A. L. Barabási, Statistical mechanics of complex networks, Rev. Mod. Phys. **74**, 47 (2002).

[2] M. E. J. Newman, *Networks: An Introduction* (Oxford University Press, New York, 2010).

[3] M. E. J. Newman, Mixing patterns in networks, Phys. Rev. E **67**, 026126 (2003).

[4] S. H. Lee, P.-J. Kim, and J. Hawoong, Statistical properties of sampled networks, Phys. Rev. E **73**, 016102 (2006).

[5] S. Yoon, S. Lee, S.-H. Yook, and Y. Kim, Statistical properties of sampled networks by random walks, Phys. Rev. E **75**, 046114 (2007).

[6] E. Estrada, Quantifying network heterogeneity, Phys. Rev. E **82**, 066102 (2010).

[7] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, Walking in Facebook: A case study of unbiased sampling of OSNs, *Proceedings of the 29th Conference on Information Communication, INFOCOM'10, Piscataway, NJ, USA* (IEEE Press, Bellingham, 2010), p. 2498.

[8] V. Zlatic, A. Gabrielli, and G. Caldarelli, Topologically biased random walk and community finding in networks, Phys. Rev. E **82**, 066109 (2010).

[9] C. Cooper, T. Radzik, and Y. Siantos, Estimating network parameters using random walks, *Proceedings of the 2012 Fourth International Conference on Computational Aspects of Social Networks (CASoN), 2012* (IEEE, New York, 2012), pp. 33–40, https://ieeexplore.ieee.org/document/6412374/.

[10] C. A. Bliss, C. M. Danforth, and P. S. Dodds, Estimation of global network statistics from incomplete data, PLoS One **9**, e108471 (2014).

[11] Y. Zhang, E. D. Kolaczyk, and B. D. Spencer, Estimating network degree distributions under sampling: An inverse problem, with applications to monitoring social media networks, Ann. Appl. Stat. **9**, 166 (2015).

[12] N. G. van Kampen, *Stochastic Processes in Physics and Chemistry* (Elsevier, Amsterdam, 2007).

[13] P. L. Krapivsky, S. Redner, and E. Ben-Naim, *A Kinetic View of Statistical Physics* (Cambridge University Press, Cambridge, England, 2010).
[14] J. D. Noh and H. Rieger, Random walks on complex networks, Phys. Rev. Lett. **92**, 118701 (2004).
[15] S. Condamin, O. Benichou, and M. Moreau, Random walks and Brownian motion: A method of computation for first-passage times and related quantities in confined geometries, Phys. Rev. E **75**, 021111 (2007).
[16] E. M. Bollt and D. ben-Avraham, What is special about diffusion on scale-free nets?, New J. Phys. **7**, 26 (2005).
[17] S. Condamin, O. Benichou, V. Tejedor, R. Voituriez, and J. Klafter, First-passage times in complex scale-invariant media, Nature (London) **450**, 77 (2007).
[18] A. J. Webster and R. Kemp, Estimating omissions from searches, Am. Stat. **67**, 82 (2013).
[19] P. Erdos and A. Renyi, On the evolution of random graphs. II, Bull. Inst. Int. Stat. **38**, 343 (1961).
[20] D. J. Watts and S. H. Strogatz, Collective dynamics of 'small-world' networks, Nature (London) **393**, 440 (1998).
[21] M. E. J. Newman, C. Moore, and D. Watts, Mean-field solution of the small-world network model, Phys. Rev. Lett. **84**, 3201 (2000).
[22] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.121.038301 for additional network inference results.
[23] S. Meloni, A. Arenas, and Y. Moreno, Traffic-driven epidemic spreading in finite-size scale-free networks, Proc. Natl. Acad. Sci. U.S.A. **106**, 16897 (2009).
[24] M. A. Serrano, D. Krioukov, and M. Boguna, Self-similarity of complex networks and hidden metric spaces, Phys. Rev. Lett. **100**, 078701 (2008).
[25] M. Boguna, D. Krioukov, and K. C. Claffy, Navigability of complex networks, Nat. Phys. **5**, 74 (2009).
[26] L. Pellis, F. Ball, S. Bansal, K. Eames, T. House, V. Isham, and P. Trapman, Eight challenges for network epidemic models, Epidemics **10**, 58 (2015).
[27] https://en.wikipedia.org/wiki/Wikipedia:What_is_an_article?.
[28] https://stats.wikimedia.org/EN/TablesWikipediaEN.htm.
[29] https://en.wikipedia.org/wiki/Wikipedia:Size_in_volumes.
[30] M. Faloutsos, P. Faloutsos, and C. Faloutsos, On power-law relationships of the internet topology, SIGCOMM Comput. Commun. Rev. **29**, 251 (1999).