Bayesian Inference of Global Statistics on Complex Networks using Random Walks (With Applications in Epidemiology)

Willow Kion-Crosby[1] and Alex Morozov[1]

[1]Department of Physics & Astronomy, Rutgers, The State University of New Jersey

Introduction

Figure 1: Social network example.

Many realistic systems are only accurately modeled by a network of nodes and edges. Examples are found in the study of

- Social networks (Facebook).
  1. Individuals are nodes and associations are edges.
  2. Given the network structure, how quickly does information flow through a population? (six-degrees of separation)
  3. Are there correlations between characteristics of individuals in realistic social networks?
- Epidemiology of contagious diseases and computer viruses.
  1. Individuals could be nodes in these model systems, but nodes could also be cities, major airports, or even abstract population states.
  2. The network may be fixed or dynamic. How does the network structure and dynamics affect the rate of contagion spreading?
  3. Is an outbreak even sustainable given the underlying network and node parameters?

In order to fully understand these physical processes, knowledge of the characteristics of the network's global structure is often necessary. These characteristics can typically be summarized through various global network statistics such as the degree distribution, $p(k)$, degree correlation measures, the clustering coefficient, etc.

Sampling via Random Walks (RW)

Each sample is taken after each step of the walk, and the probability to transition from a current node i to a neighboring node j is given through the edge weight w_{ji},

$$P(i \to j) = \frac{w_{ji}}{\sum_{j'} w_{j'i}}, \text{ where } w_i = \sum_j w_{ji} \tag{1}$$

Given that the average number of steps between subsequent visits to nodes with any property x is $\Delta\bar{\ell}_x$, and the ansatz that the distribution of return times is well-characterized by an exponential,

$$\Rightarrow P(\Delta\ell) = \frac{1}{\Delta\bar{\ell}_x} e^{-\Delta\ell/\Delta\bar{\ell}_x}, \tag{2}$$

the likelihood that nodes are visited with x a total of $\mathcal{K}_x$ times during a walk of total length ℓ is

$$P(\mathcal{K}_x|\Delta\bar{\ell}_x) = \sum_{\Delta\ell_0+\Delta\ell_1+\ldots+\Delta\ell_{\mathcal{K}_x}=\ell} \frac{1}{\Delta\bar{\ell}_x} e^{-\Delta\ell_0/\Delta\bar{\ell}_x} \times \frac{1}{\Delta\ell_x} e^{-\Delta\ell_1/\Delta\bar{\ell}_x} \times \ldots \times \frac{1}{\Delta\bar{\ell}_x} e^{-\Delta\ell_{\mathcal{K}_x}/\Delta\bar{\ell}_x}$$
$$= \binom{\ell}{\mathcal{K}_x} q_x^{\mathcal{K}_x} e^{-q_x \ell} \text{ where } q_x \equiv \frac{1}{\Delta\bar{\ell}_x}. \tag{3}$$

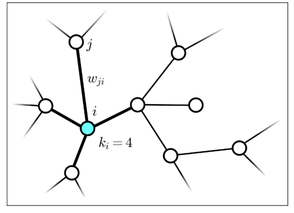

Figure 2: Infinite network with symmetric rates between nodes.

In general for networks with unbiased edge weights $q_x = p_x\langle w\rangle_x/\langle w\rangle$ is an exact equality where p_x is the fraction of nodes with property x, $\langle w\rangle_x$ is the average outward rate of nodes with property x, and $\langle w\rangle$ is the average outward rate over all nodes. The maximum likelihood value and standard error for the fraction p_x are then given by

$$\hat{p}_x = \frac{\sum_y \mathcal{K}_{x,w}/w}{\sum_{w'} \mathcal{K}_{w'}/w'} \text{ and } \sigma_{\hat{p}_x} = \frac{\hat{p}_x}{\sqrt{\mathcal{K}_x}}. \tag{4}$$

The presence of $1/w$ in this expression naturally accounts for the bias introduced in RW sampling due to the increased returns to hubs or nudes with a large connectivity.

Outward Rate Estimator

When x represents the outward rate of a node w_i, Eq. (4) yields the distribution estimator

$$\hat{p}_w = \frac{\mathcal{K}_w/w}{\sum_{w'} \mathcal{K}_{w'}/w'}, \tag{5}$$

which leads to an estimator for the average outward rate with standard error of

$$\langle\hat{w}\rangle = \frac{\ell}{\sum_w \mathcal{K}_w/w} \text{ and } \sigma_{\langle w\rangle} = \frac{\langle\hat{w}\rangle}{\sqrt{\ell}} \tag{6}$$

Estimating the Network Size

Before the sampling, label N_p nodes as ***pseudotargets*** and compute their average outward rate $\langle w\rangle_p$. Exploiting the direct dependence of the rate to find these targets on the network size, $q_p = N_p\langle w\rangle_p/N\langle w\rangle$, and estimator for the network size can be found to be

$$\hat{N} = \frac{\ell\langle w\rangle_p N_p}{\mathcal{K}_p\langle w\rangle} \text{ with } \sigma_N = \frac{\hat{N}}{\sqrt{\mathcal{K}_p}}. \tag{7}$$

As hubs are easy to find and designate as pseudotargets, the rate at which data points are collected can be controlled through both increasing N_p and $\langle w\rangle_p$ such that the error in the estimator rapidly vanishes.

Tag and Recapture

In the case of a complete network, $\langle w\rangle = \langle w\rangle_p = N - 1$, RW sampling is identical to uniform sampling of a population of size N. Provided that $N \gg 1$, the pseudotarget drawing followed by the RW sampling processes result in two sets, A and B, virtually consisting of only unique elements of respective sizes $N_A = N_p$ and $N_B = \ell$. Further recognizing that $N_{AB} = \mathcal{K}_p$ is the number of intersecting elements in these sets, Eq. (7) recovers the classic formula for estimating the size of a population given two independent samplings, $\hat{N} = N_A N_B/N_{AB}$ [1]. Our formalism additionally predicts a standard error for this classic result, $\sigma_N = N_A N_B/N_{AB}^{3/2}$, as well as the full posterior distribution for the population size.

Tracking Traffic Driven Epidemics

Hidden Metric

For the following epidemics example, the network structure has been generated using a hidden metric consisting of a 1-dimensional circle [2]. This method of generation results in a network with not only the scale-free and small-world properties, but additionally develops local cluster structures. Generation procedure:

- Assign each node a uniformly drawn location on this hidden metric, $\theta \in [0, 2\pi)$,
- and an expected degree, κ, drawn from a power-law distribution,

$$p(\kappa) \sim \kappa^{-\gamma}.$$

- Each pair of nodes are then linked with a probability based on these two parameters,

$$p \sim (1 + d(\theta, \theta')/\eta\kappa\kappa')^{-\alpha},$$

where $\eta \equiv (\alpha - 1)/2\langle k\rangle$, $d(\theta, \theta')$ is the geodesic distance between the two nodes on the hidden metric, and α is a tunable parameter. For our network construction we set $N = 10^5$, $\gamma = 2.6$, and $\alpha = 2$. Even with the local structures present in this network, the RT distribution is very well approximated by the exponential ansatz for our set of $N_p = 1000$ pseudotargets.

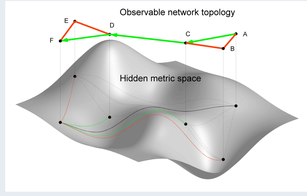

Figure 3: Hidden metric space.

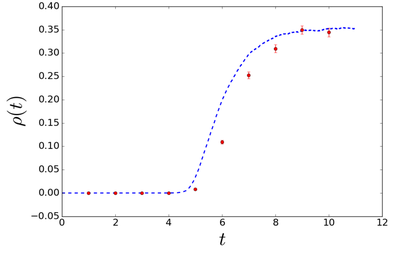

Figure 4: Fraction of infected nodes.

At regular intervals, RW sampling was performed for $\ell = 10^4$ steps during which the following statistics were tracked:

- The number of packets currently occupying a node in the network, ω_i, expected to be $\sim (2N)k_i/N\langle k\rangle$ in steady-state.
- The total fraction of infected nodes in the network, $\rho(t)$.
- Several other statistics relevant to epidemiology:
  1. The network size, N.
  2. The average node degree, $\langle k\rangle$.
  3. The average degree of neighboring nodes, $\langle\langle k_{nn}\rangle\rangle$.
  4. The network clustering coefficient, $\langle C\rangle$.

The results of estimating these statistics from a single walk are shown in the table below. The estimation of $\rho(t)$ at the end of each of the intervals alongside the true simulation values are shown in Fig. 4. Note that during the interval $t = 4..8$ when the epidemic spread is the most rapid, the estimation is below the true value as the statistic is based on a range of values of $\rho(t)$.

$\hat{N}$	N	$\langle\hat{k}\rangle$	$\langle k\rangle$	$\langle\widehat{\langle k_{nn}\rangle}\rangle$	$\langle\langle k_{nn}\rangle\rangle$	$\langle\hat{C}\rangle$	$\langle C\rangle$	$\langle\hat{\omega}\rangle$	$\langle\omega\rangle$
$\pm\sigma_N$		$\pm 2\sigma_{\langle k\rangle}$		$\pm 2\sigma_{\langle\langle k_{nn}\rangle\rangle}$		$\pm 2\sigma_{\langle C\rangle}$		$\pm 2\sigma_{\langle\omega\rangle}$	
1.01×10^5	10^5	8.02	8.14	64.6	67.1	0.251	0.255	2.00	2.00
$\pm\ 0.08 \times 10^5$		± 0.16		± 4.2		± 0.011		± 0.045	

Contagion Packet Dynamics

On a network generated using a hidden metric, an epidemic is simulated through the exchange of contagion packets between nodes:

- Packets move from node i to node j on the network,
- j becomes infected with probability β if i is infected in this time step.
- Each infected node recovers with a rate $\mu = 1$

We have examined the case in which these packets are performing RWs from randomly assigned initial and destination nodes with $2N$ packets traversing the network at any given time. The packets traverse a link with rate 1 such that on average $2N$ packets have moved once per simulation time unit. Under this choice of packet dynamics there is a critical value for β, below which there can be no sustained epidemic outbreak, and that this critical value is based on the network connectivity through $\beta_c \sim \langle k\rangle^2/\langle k^2\rangle$. In our reconstruction of this simulation, we have chosen $\beta = .7 \gg \beta_c$, and set and maintained the initial fraction of infected nodes to $1/N$ in order to guarantee an eventual outbreak.[3]

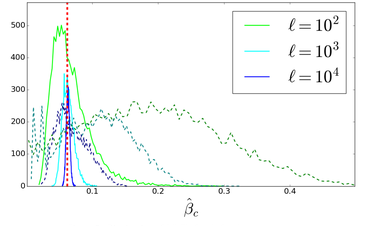

Figure 5: Distribution of MLE's for β_c for three walk lengths

10^4 independent RWs were run each for $\ell = 10^2$, 10^3, and 10^4 steps, and the quantity $\hat{\beta}_c \equiv \langle\hat{k}\rangle^2/(\langle\hat{k}^2\rangle\langle\hat{\omega}\rangle)$ was computed for each to obtain the distribution of MLE values using this methodology. The histograms of the resulting values are shown in Fig. 5 for all three walk lengths, the longest of which allows for a maximum of 10% of the network to be sampled from. As a constrast, three uniform samplings were also performed with $n = \ell$ samples to match all three walks, and β_c was estimated using this method.

Generalized Erdős-Rényi

GER Network Construction

We have constructed a generalized ER network with $N = 10^6$ nodes and weighted edges. After placing all the edges as in an 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.

The RT distribution for this system deviates from purely exponential since many returns occur after a single step due to loops (Fig. 6). Nonetheless, all the network statistics we have considered are predicted accurately excepting the tail of the distribution since those rare events were not observed. Thus our methodology is equally applicable to studies of weighted networks with loops.

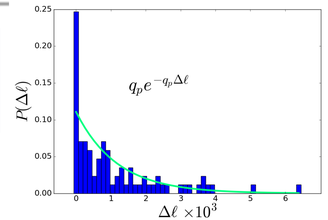

Figure 6: RT distribution for generalized ER network.

Wikipedia

Wikipedia as an Undirected Network

Finally, 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.

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. We have focused on several statistics that facilitate comparison with known properties of Wikipedia: the size of each page in bytes, ν, and two variables $\chi_r, \chi_d \in \{0, 1\}$ representing whether a page is a redirect or a disambiguation page, respectively. The quantities $\langle\chi_r\rangle$, $\langle\chi_d\rangle$, $\langle\chi_r\chi_d\rangle$, and $\langle\nu_a\rangle \equiv \langle(1 - \chi_r)\nu\rangle$ then give the fraction of redirect pages, disambiguation pages, both redirect and disambiguation pages, and the average storage space in bytes of English articles (Wikipedia excludes redirect pages from its estimates of the number of articles). The RW was run for $\ell = 5 \times 10^4$ steps, with the resulting predictions shown in Fig. 7. We find that Wikipedia contains 13.4 million pages, each of which is connected to 48 other pages on average. 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 million, and the total number of redirect pages to be 8.05 million, within the confidence intervals of the values reported by Wikipedia: 5.5 and 8.0 million, respectively. We find the total size of English articles in Wikipedia to be 35.8 gigabytes (GB), in reasonable agreement with the Wikipedia statement that text alone accounts for 27.6 GB of the storage space of English articles.

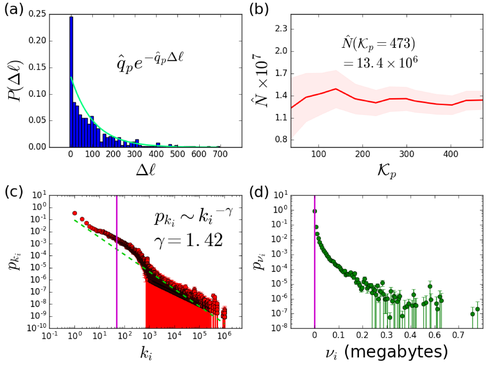

Figure 7: Wikipedia network results.

Conclusion

In conclusion, we have presented a general Bayesian approach to collecting various network statistics, including the size of the network, 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. 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] Anthony J Webster and Richard Kemp.
Estimating omissions from searches.
2013.

[2] M. Angeles Serrano, Dmitri Krioukov, and Marián Boguñá.
Self-similarity of complex networks and hidden metric spaces.
Phys. Rev. Lett., 100:078701, Feb 2008.

[3] Sandro Meloni, Alex Arenas, and Yamir Moreno.
Traffic-driven epidemic spreading in finite-size scale-free networks.
Proceedings of the National Academy of Sciences, 106(40):16897–16902, 2009.