STATISTICAL INFORMATION FROM RANDOM WALKS

BRUNO RIBEIRO AND DON TOWSLEY

1. Introduction

Let $G = (V, E)$ be an undirected, unweighted graph with N vertices. We define E such that if v_i and v_k are connected by an edge, then both $(v_i, v_k) \in E$ and $(v_k, v_i) \in E$. Function $d(v_i)$ gives the degree of vertex v_i.

We seek to find an unbiased estimator for functions of the form

$$\sum_{\forall e \in E} f(e). \tag{1.1}$$

Functions of the form $\sum_{\forall e \in E} f(e)$ are very useful to compute graph characteristics. For instance, the fraction of vertices in G that have degree h can be written as

$$\sum_{\forall (v_i, v_k) \in E} g_h(v_i, v_k), \text{ where } g_h(v_i, v_k) = \begin{cases} \frac{1}{d(v_i)N} & \text{if } d(v_i) = h \\ 0 & \text{otherwise} \end{cases} .$$

Note that a more accurate estimator would also use the degree information in v_k. Graph assortativity is another example of such function.

In what follows we present three unbiased estimators of equation (1.1).

2. Randomly chosen vertices

TODO: Show estimator.

TODO: present FI for degree dist.

3. Randomly chosen edges

Let $\mathcal{U}$ be a set with m edges chosen uniformly at random, with replacement, from G. The next lemma shows that $\sum_{\forall U \in \mathcal{U}} f(U)/(m/|E|)$ is an unbiased estimator of eq. (1.1).

Lemma 1. $E[\sum_{\forall U \in \mathcal{U}} f(U)]/(m/|E|) = \sum_{\forall e \in E} f(e)$.

Proof. The proof is quite straightforward:

$$\frac{E[\sum_{\forall U \in \mathcal{U}} f(U)]}{m/|E|} = \frac{\sum_{\forall U \in \mathcal{U}} E[f(U)]}{m/|E|} = \frac{\sum_{\forall U \in \mathcal{U}} \left(\sum_{\forall e \in E} f(e)/|E|\right)}{m/|E|} = \sum_{\forall e \in E} f(e)$$

□

TODO: Present FI for degree dist.

4. Edges chosen in a Random Walk

In this section we estimate eq.(1.1) using the edges sampled in a random walk over G. Let $\Gamma = \{\epsilon_i | i = 1, \ldots, n\}$ be an $(n+1)$-step random walk over G starting in steady state. A random walk starts in steady state if either one of the following conditions is true: (1) ϵ_1 is chosen uniformly at random from E; or (2) $\epsilon_1 = (v, u)$ and v is chosen with probability $d(v)/|E|$ from V. We say an edge (v, u) is chosen at step i of the random walk if vertex v is chosen at step i and vertex u is chosen at step $i + 1$.

Lemma 2. *The probability that an edge $e \in E$ is chosen at the i-th step of a random walk (starting in steady state) is* $1/|E|$.

Proof. We just need to prove that the probability of choosing an edge (v, u) in the graph at the i-th step of the random walk is $1/|E|$. In steady state, vertex v is chosen at the i-th step with probability $d(v)/|E|$. Thus, edge (v, u) is chosen with probability $p = (d(v)/|E|) \cdot 1/d(v) = 1/|E|$. □

Each edge in the random walk is chosen with probability $1/|E|$ but two edges in the same random walk are clearly not chosen independently. However, the next lemma shows that because expectation is a linear operator, all functions of the form presented in equation (1.1) can be estimated from random walks.

Lemma 3. $\sum_{\forall \epsilon \in \Gamma} f(\epsilon)/(n/|E|)$ *is an unbiased estimator of* $\sum_{\forall e \in E} f(e)$.

Proof. Estimator $\sum_{\forall \epsilon \in \Gamma} f(\epsilon)/(n/|E|)$ is an unbiased estimator of $\sum_{\forall e \in E} f(e)$ if

$$E\left[\sum_{\forall \epsilon \in \Gamma} f(\epsilon)/(n/|E|)\right] = \sum_{\forall e \in E} f(e).$$

As expectation is a linear operator,

$$E\left[\sum_{\forall \epsilon \in \Gamma} f(\epsilon)/(n/|E|)\right] = \frac{E\left[\sum_{\forall \epsilon \in \Gamma} f(\epsilon)\right]}{(n/|E|)} = \frac{\sum_{\forall \epsilon \in \Gamma} E\left[f(\epsilon)\right]}{(n/|E|)}. \tag{4.1}$$

Edges in a random walk are chosen with probability $2/|E|$, then

$$E\left[f(\epsilon)\right] = \sum_{\forall e \in E} f(e) \frac{1}{|E|}. \tag{4.2}$$

Replacing eq. (4.2) into eq. (4.1) we have

$$E\left[\sum_{\forall \epsilon \in \Gamma} f(\epsilon)/(n/|E|)\right] = \frac{\sum_{\forall \epsilon \in \Gamma} \sum_{\forall e \in E} f(e) \frac{1}{|E|}}{(n/|E|)} = \sum_{\forall e \in E} f(e).$$

□

Note that edge samples obtained in a random walk are dependent. While this dependency this does not affect the unbiasedness of the estimates, we will see in Section 5 that (in most graphs) this dependency significantly reduces the statistical information obtained about the true value of equation (1.1). Section 5 exemplifies this decrease in estimation accuracy over a real social network. We see that estimates of the graph degree distribution using 1,000 randomly sampled edges are much more accurate than the same estimates using 1,000 edges sampled in a random walk. We then look at K independent random walks with $1,000/K$ steps each (K

is chosen such that $1,000/K$ is an integer). We show that the amount of statistical information about the original value of eq. (1.1) decreases with K. We then use the Orkut social network to show that such decrease in estimation accuracy can be significant.

5. Statistical information from sampling

Remark: Our examples are based on the snowball samples of the Orkut social network collected by Mislove et al. . The impact of their snowball sampling is probably to significantly decrease the mixing time when compared to the "full" Orkut network.

6. Random graphs and statistical information

Samples from random graphs have greater statistical information than samples from general graphs. Show using the data processing inequality?!?