eecs589

ADVANCED COMPUTER NETWORKS

Rasti, A.H., et al., "Respondent-driven Sampling for Characterizing Unstructured Overlays," *Proc. of IEEE Infocom '09*, Apr. 2009

Querying a Network

What is the degree/connectivity distribution of nodes in a P2P network?

How many friends in an online social network like X?

What is the average temperature reading in a sensor network?

What is the average speed of cars on a vehicular network?

If We Have a Central List of Nodes

For an accurate estimate, we must sample uniformly at random

Generate a set of uniformly distributed random numbers

Use the random numbers as indices into the list

Access the nodes at those locations to obtain its readings

With or Without Replacement?

Do we allow for multiple samplings of the same node? Or do we only sample unique node each time?

With dynamic graphs, where nodes come and go, sampling without replacement can lead to bias towards short-lived nodes

Example [S+09]:

- we want to know the average number of files per node
- half of the nodes is long-lived and hold a lot of files
- the other half is short-lived and hold only a few files
- without replacement, we will sample more of the short-lived nodes and conclude erroneously that most nodes hold only a small number of files

Without a Central List

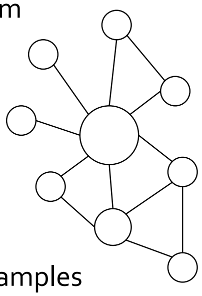
How do you estimate various characteristics of a network, $G = (V, E)$, with no central list of nodes?
V: vertices, E: edges

Random walk: start at any node,

- choose one of its neighbors uniformly at random
  Transition function from x to y:

$$P(x,y)=\begin{cases}\frac{1}{\text{degree}(x)}, & y \text{ is a neighbor of } x\\ 0, & \text{otherwise}\end{cases}$$

- step to chosen neighbor
- repeat
- some portion of visited nodes are selected as samples



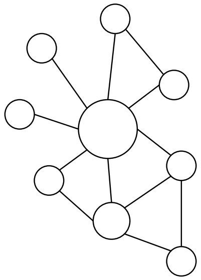
Problem with Random Walk

Higher-degree nodes visited more often

Stationary distribution, $\pi(x)$, at any particular node x is the probability of being at node x (how often x is visited)

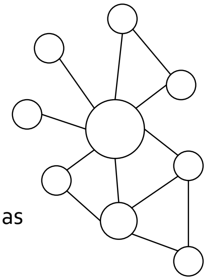
Stationary distribution of the walk at any particular node x is proportional to the degree of x, $\pi(x) \propto \text{degree}(x)$

How to correct for this inherent bias?



Metropolis-Hastings Algorithm

Correct for the bias by modifying the random walk, by altering the transition function to any desired stationary distribution, e.g., uniform, $\pi(x) = 1/|V|$

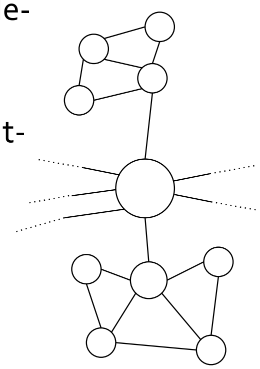


$$Q(x,y)=\begin{cases}\overbrace{P(x,y)}^{\text{tentatively select a neighbor uniformly at random}}\overbrace{\min\left(\frac{\text{degree}(x)}{\text{degree}(y)},1\right)}^{\text{probability of accepting transition, correcting the bias}}, & \text{if } x \neq y\\ \underbrace{1-\sum_{z\neq x} Q(x,z)}_{\text{probability of remaining at current node (taking a self-edge)}}, & \text{if } x = y\end{cases}$$

Problem with Metropolis-Hastings

Random walk can get "stuck in a cul-de-sac" for network with highly-skewed node degrees and highly skewed local clustering coefficients (tiered or transit-stub networks)

Local clustering coefficient of a vertex: how well connected are the vertex's immediate neighbors to each other

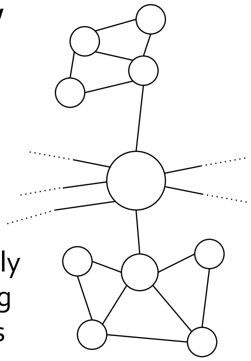


Respondent-driven Sampling

Instead of modifying the random walk, reweight the sampled values

An importance sampling estimator weighted by the (actual) stationary distribution π

Importance sampling: sample more frequently those values that have more impact, resulting in biased sampling, but reweight the samples to correct the bias



Respondent-driven Sampling

Given property of interest X, partition all possible values of X into m groups: $\{R_1, \ldots, R_m\}$

V is accordingly also partitioned into m groups: $\{V_1, \ldots, V_m\}$, where $V_i = \{v \in V: X(v) \in R_i\}$

Example: X is positive integer value and we group by value: $V_i = \{v \in V: X(v) = i\}$

For stationary distribution $\pi(x)$, $0 \le \pi(x) \le 1$, $\sum_{v \in V} \pi(v) = 1$

$E\left(\frac{1}{\pi(v)} X(v)\right) = \sum_{v \in V} X(v)$, the population total

Respondent-driven Sampling

Consider an n-step random walk that visits the set of nodes $T = \{t_1, \ldots, t_n\}$ starting from a node randomly selected according to the stationary distribution, where individual node may be visited more than once (i.e., with replacement), and let

$T_i = T \cap V_i$

The Hanson-Hurwitz Estimator

For any node property X, the Hanson-Hurwitz estimator is:

$$\hat{S}(X) := \frac{1}{n} \sum_{v \in T} \frac{X(v)}{\pi(v)}$$

Since $E\left(\frac{X(v)}{\pi(v)}\right) = S(X)$, where $S(X) = \sum_{v \in V} X(v)$ is the population total, thus $E\left(\hat{S}(X)\right) = S(X)$ and $\hat{S}(X)$ is an unbiased and consistent estimator of the population total:

$$S(X) := \sum_{v \in V} X(v)$$

[https://onlinecourses.science.psu.edu/stat506/node/15]

Respondent-driven Sampling

From the group memberships and node degrees observed during a random walk, we can estimate $p_i = |V_i|/|V|$, the proportion of nodes in group i of node property X

When $X = I_{V_i}$ is an indicator whether a node is in group i, i.e.,

$$I_{V_i}(v) = \begin{cases} 1, & \text{if } v \in V_i, \\ 0, & \text{otherwise,} \end{cases}$$

then $\hat{S}(I_{V_i})$ estimates $|V_i|$

When $X = 1$, then $\hat{S}(1)$ estimates $|V|$

Respondent-driven Sampling

Given $\pi(x) \propto \text{degree}(x)$, we can estimate p_i as :

$$\hat{p}_i = \frac{|V_i|}{|V|} = \frac{\hat{S}(I_{V_i})}{\hat{S}(1)} = \frac{\sum_{v\in T_i} \frac{1}{\pi(v)}}{\sum_{u\in T} \frac{1}{\pi(u)}} = \frac{\sum_{v\in T_i} \frac{1}{\text{degree}(v)}}{\sum_{u\in T} \frac{1}{\text{degree}(u)}}$$

$\hat{p}_i$ is consistent (converges to the true value of p_i) as n grows

Sampling Techniques Evaluation

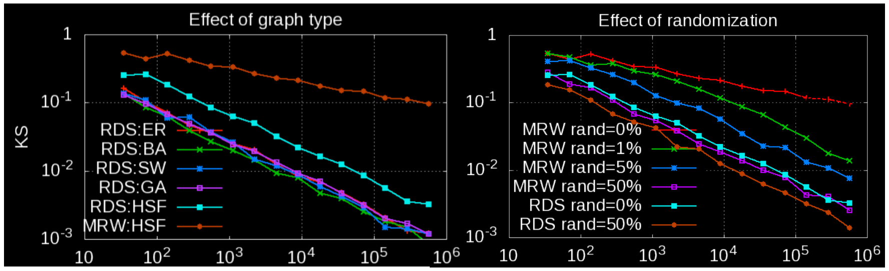
Use synthetic graphs and Gnutella overlay snapshots

Synthetic graphs allow for:

1. accurate evaluation since the distribution of the sampled property (ground truth) is known
2. identify separate effects of graph properties and graph dynamics on the accuracy and efficiency of the techniques

Performance metric: Kolmogorov-Smirnov (KS) statistic, the maximum vertical distance between the plots of two functions with values in $[0, 1]$, such as the CDFs of estimated and actual distributions of X

Degree Distribution

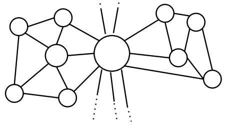


Better estimates as number of samples increases

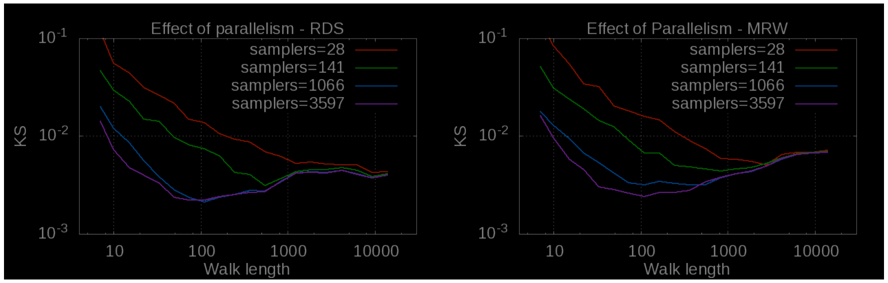
Except for HSF graph, MH follows a similar trend as RDS, but with 2e-3 lower accuracy

HSF graph has “cul de sac” causing MH to get stuck

Rewiring removes “cul de sac”



Number of Walkers vs. Walk Length

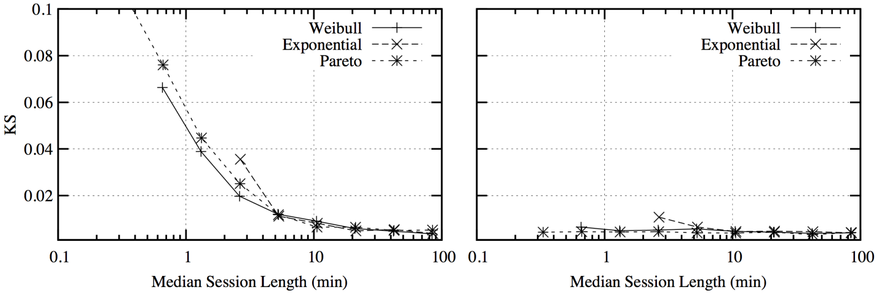


Too many walkers bias towards nodes near starting point, resulting in low accuracy at short walk length

Longer walks see more churn

Alternative: start sampling only after r steps, for subsequent n steps [GMS06]

Lifetime-related Properties



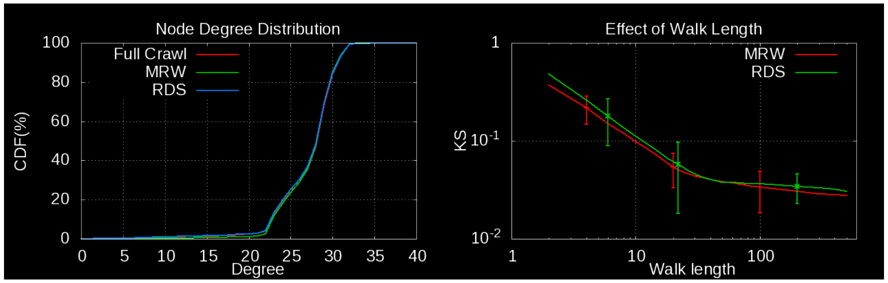
(a) Tech:RDS Prop:Deg (b) Tech:RDS Prop:RTT

Short-lived nodes usually have lower node degrees, leading to biased sampling even with RDS if churn rate is too high (session length below 10 mins)

Properties less correlated with session length, such as RTT, do not show similar bias

Gnutella Snapshots

Snapshots of top-level Gnutella overlay collected back-to-back once every 7 minutes



MH and RDS estimates are roughly the same, Gnutella network is not HSF

Longer walks see more churn