ST5225: Statistical Analysis of Networks
Lecture 2: Network Sampling

WANG Wanjie
staww@nus.edu.sg

Department of Statistics and Applied Probability
National University of Singapore (NUS)

Sunday 21st January, 2018

Review

- Module information
- Introduction of Networks
  - Network Examples
  - Basic Concepts: Neighbor, Degree, Path, Distance, Subgraph, Connect, Component, Complete, Clique, Maximal clique, Adjacency matrix
  - Two algorithms: BFS and DFS

Graph Sampling

- Graph Sampling: Importance and Problems
- Induced-Subgraph Sampling
- Snowball Sampling
- Respondent-driven Sampling
- Trace-route Sampling
- Estimates
- Horvitz-Thompson Estimator

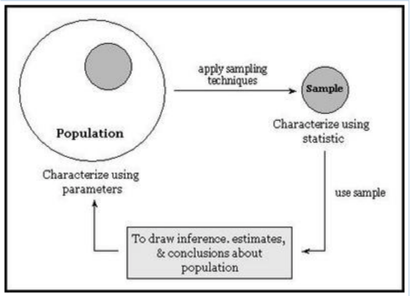
Relevant Chapter

Statistical Analysis of Network Data, Chapter 5.1–5.4

Graph Sampling

- Ideally, there is a census of the network. Every node is observed.
- However, usually it is impossible
  - Huge cost (time, money, human resource, etc.)
  - technical or social restrictions (not everyone want to enroll)
- Even if possible, it may be hard to analyze the whole network, due to the data size
- Therefore, sampling is important

Hopefully, we have

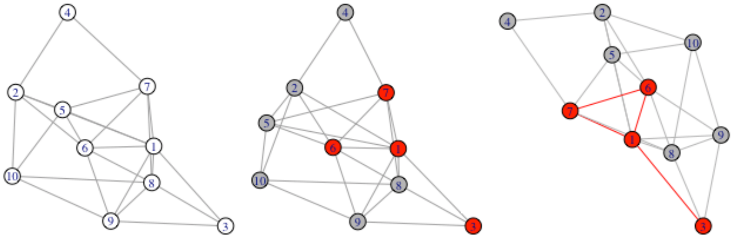


However, for graphs,

- How to get IID data? (the nodes and edges are obviously correlated)
- How do we infer the big network from the samples?

Induced-Subgraph Sampling

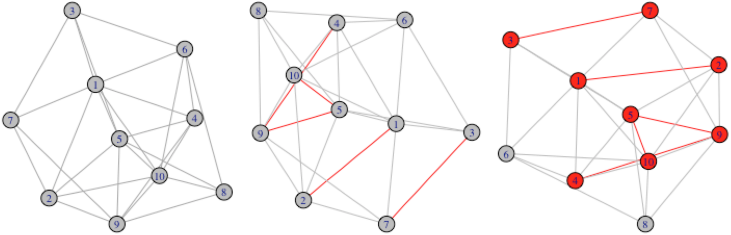
- A graph $G = (V, E)$
- We can sample on the set V, and then get the corresponding edges
- Procedure
  1. Uniformly sample a set $S = \{s_1, s_2, \cdots s_{n_s}\}$ of **nodes**
  2. Observe **edges** E_S between sampled nodes S
  3. The subgraph $G_S = (S, E_S)$ is the sampled graph
- Example: Facebook connection. Sampling a fraction of users from all the users, and observe their friendship connections.


4
2
7
5
6
1
10
8
9
3
4
2
7
5
6
1
10
8
9
3
4
2
10
5
6
7
9
1
8
3

Incident-Subgraph Sampling

- A graph $G = (V, E)$
- We can sample on the set of edges E (not the nodes!), and then get the corresponding nodes
- Procedure
  1. Uniformly sample a set $S = \{e_1, e_2, \cdots e_{n_s}\}$ of **edges**
  2. Observe **nodes** V_S incident to S
  3. The subgraph $G_S = (V_S, S)$ is the sampled graph
- Example: phone call connection – Sampling a fraction of all the phone calls (edges), and get the phone numbers (nodes). Bitcoin transactions.

Incident-Subgraph Sampling



Snowball Sampling

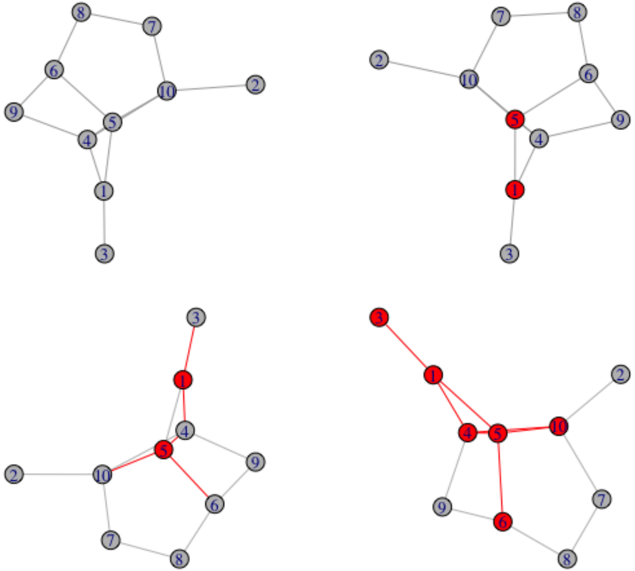
- We want to involve more info. about the data. An alternative way is to start with a set of targeted nodes, and enlarge the observed network

- Procedure
  1. Uniformly sample a set $S = S_1 = \{s_1, s_2, \cdots, s_{n_s}\}$ of nodes.
  2. Observe **all the edges** E_S incident to S.
  3. Let S_2 be the **set of neighbors** of nodes in S_1.
  4. Update the set to be $S = S_1 \cup S_2$. Update the set $E_S = \{(i,j) : i, j \in S\}$. Update the sampled graph as $G_S = (S, E_S)$.
  5. Repeat 1-4 until the number of sampled nodes is large enough (pre-selected size).

- The network becomes larger and larger, just like a *snowball*.

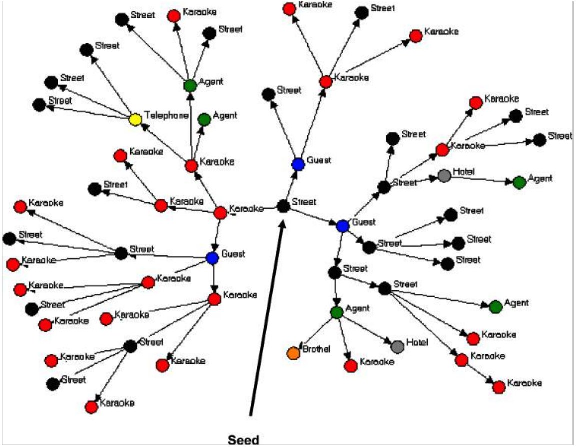
- To do a survey, start with some specific person, and then search over their friends.

Snowball Sampling



- A variant on snowball sampling
- Procedure
  1. Uniformly sample a set $S = S_1 = \{s_1, s_2, \cdots, s_{n_s}\}$ of nodes.
  2. For each node (respondent), a limited number of tokens (k) are given
  3. Each respondent give the token to his/her friend, and persuade them to enroll the study.
  4. According to the token, the link between the respondent and his/her friend will be record
  5. Repeat 1-4 until the number of sampled nodes is large enough (pre-selected size).
- Usually used for hard-to-find sub-populations – those may be illegal (drug users)
- Some incentive will be given for participation
- Note the degree will be restricted by k

Respondent-driven Sampling

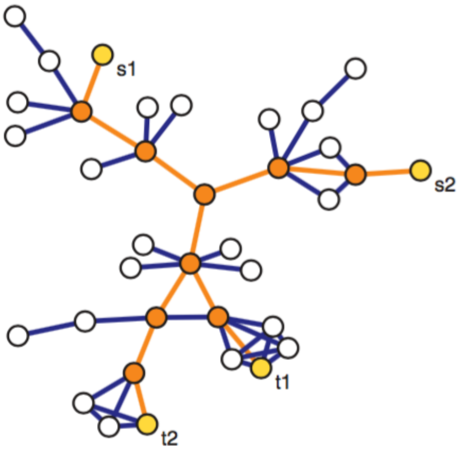


Johnston et al. (2006) Assessment of Respondent Driven Sampling for Recruiting Female Sex Workers in Two Vietnamese Cities: Reaching the Unseen Sex Worker

Trace-route Sampling

- Consider a set of nodes as "starting" points, and a set as "ending" nodes.
- Procedure
  1. Pick a set of source nodes, marked as S_1
  2. Pick a set of target nodes, marked as S_2
  3. For each (v_1, v_2), where $v_1 \in S_1$ and $v_2 \in S_2$, find a path from v_1 to v_2, and record all the nodes and edges traversed along the path.
  4. The corresponding subgraph is the sampling subgraph.
- Example: Six-degree separation. Choose one person A as the source node, another person B as the target node, and record everyone on the path that A gets in connection with B.

Trace-route Sampling



SAND, Figure 5.5

Summary

- 5 different sampling methods
  - Incident-subgraph sampling and induced-subgraph sampling select the set of nodes and edges first, respectively, and then get the corresponding subgraph
  - Snowball sampling, respondent-driven sampling, trace-route sampling start with a set of nodes, and expand to get the subgraph
- The choice of sampling method depend on the nature of the study
  - e.g. If we are trying to study the hidden sub-population, it is impossible to use incident-subgraph sampling or induced-subgraph sampling
- Are the estimates from different sampling methods the same?

- Recall that we have "plug-in" estimates, and, say, MLE, etc.
- Does "plug-in" estimate still work here?

Example:

- Induced subgraphs: for each node, the corresponding degree cannot be larger than the truth. Therefore, the corresponding estimation will be biased.
- Respondent-driven sampling: as there is a restriction on the degree, obviously there is some bia

For other estimates, there is also bias! Sometimes even unpredictable!

Estimates

- Recall that we have "plug-in" estimates, and, say, MLE, etc.
- Does "plug-in" estimate still work here?

Example:

- Induced subgraphs: for each node, the corresponding degree cannot be larger than the truth. Therefore, the corresponding estimation will be biased.
- Respondent-driven sampling: as there is a restriction on the degree, obviously there is some bia

For other estimates, there is also bias! Sometimes even unpredictable!

Solutions for the Bias??

- Just ignore it – Pretend that we get the whole graph
- (*) De-bias the estimates: Horvitz-Thompson estimator.
- Treat the unobserved part as "missing data", and model the data with the observed data and missing data

Horvitz-Thompson estimator

Assume that the population has size n $(X_1, X_2, \cdots, X_n)$, and the sample set is S. We are interested in the mean μ of the population, then the *Horvitz-Thompson estimator* is

$$\hat{\mu}_{HT} = \frac{1}{n} \sum_{i \in S} \frac{X_i}{\pi_i},$$

where π_i is the probability that X_i is included in the sample.

- An **unbiased** estimator. Can be viewed as a weighted mean
- When the inclusion probabilities are all equal (usual case), then $\pi_i = |S|/n$, and $\hat{\mu}_{HT} = \bar{X}$, the sample mean.
- When the inclusion probabilities are unequal (network sampling), the nodes/edges easier to be included has smaller weight, and those harder to be included has larger weight.

Proof of Unbiasness.

- Introduce Z_i, which are indicator variables indicating whether X_i is in the sample or not.

$$Z_i = \begin{cases} 1, & i \in S \\ 0, & i \notin S \end{cases}$$

According to the definition of π_i, $P(Z_i = 1) = \pi_i$, and $E[Z_i] = \pi_i$.

-

$$\begin{aligned} E(\hat{\mu}_{HT}) &= E[\frac{1}{n}\sum_{i \in S} X_i/\pi_i] = E[\frac{1}{n}\sum_{i=1}^{n} X_i Z_i/\pi_i] \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i E[Z_i]/\pi_i \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i \pi_i/\pi_i \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i = \mu \end{aligned}$$

Induced Subgraph Sampling

- Vertex inclusion probabilities (assuming sampling n vertices):

$$\pi_i = \frac{n}{|V|}$$

- Edge inclusion probabilities:

$$\pi_{i,j} = \frac{n(n-1)}{|V|(|V|-1)}$$

Incident Subgraph Sampling

- Edge inclusion probabilities (assuming sampling n edges):

$$\pi_{i,j} = \frac{n}{|E|}$$

- Vertex inclusion probabilities :

$$\pi_i = \begin{cases} 1 - \frac{\binom{|E|-d_i}{n}}{\binom{|E|}{n}}, & \text{if } n \leq |E| - d_i \\ 1, & \text{if } n > |E| - d_i \end{cases}$$

Other Sampling Methods

- For snowball sampling and respondent-driven sampling, it is impossible to calculate the inclusion probabilities, hence impossible to get the ”de-biased estimates”
- For trace Route sampling method, the inclusion probabilities can be approximated (see Page 137 of SAND), yet the betweenness centrality for each edge is required, which is quite difficult.
- HT estimator is not perfect! Network analysis is more complicated than we thought.

Descriptive Statistics

- Degree Distribution
- Centrality
  - Closeness
  - Betweenness
  - Eigenvector
- Cohesion
  - Cliques, k-cores
  - Connectivity
  - Local Density
- Graph Partition

Relevant Chapter

Statistical Analysis of Network Data, Chapter 4.1– 4.3

Degree

Recall:

- Degree of a node i:

$$d_i = \text{the number of edges incident on the node } i$$

- For directed graphs,

  $d_i^{in} = \#\text{edges pointing in towards } i, \quad d_i^{out} = \#\text{edges pointing out from } i.$

  We use $d_i^{tot} = d_i^{in} + d_i^{out}$ to denote the number of all the edges incident to i for directed graphs.

- Degree Sequence/Degree Vector: A vector containing the degrees of each node

If we look into the degree statistic...

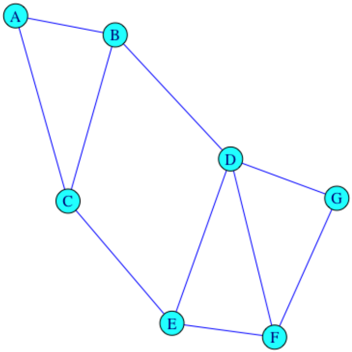
- Degree of each node is a good summary statistic, which is also called *degree centrality*
- For each node, it shows how important this node is.
- For the whole network, we are more interested in the *degree distribution*

Degree Distribution

Given a network graph $G = (V, E)$, define f_d to be the fraction of vertices $v \in V$ with degree $d_v = d$. The collection $\{f_d\}_{d \geq 0}$ is called the *degree distribution* of G, which is simply the distribution from the degree vector.

- Just as other dist., we can learn its mean, median, standard deviation, quantiles, etc.
- The shape of the distribution also give some information
- If we have the population network, the degree dist. is the empirical distribution from the network. If we only have a sampling network, the degree dist. needs to be estimated.

A toy example

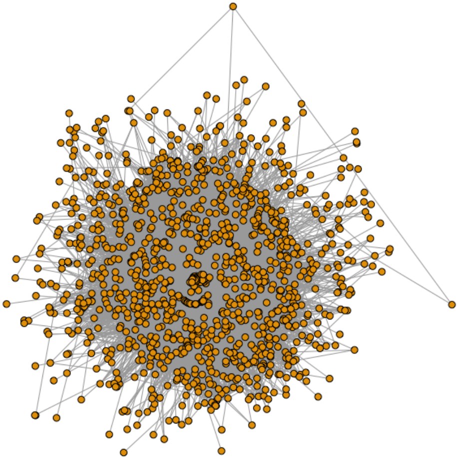


For this graph, the degree sequence is $(2, 3, 3, 4, 3, 3, 2)$. Therefore, the degree distribution is

$$f_2 = 2/7, \quad f_3 = 4/7, \quad f_4 = 1/7.$$

Degree Distribution (Examples)

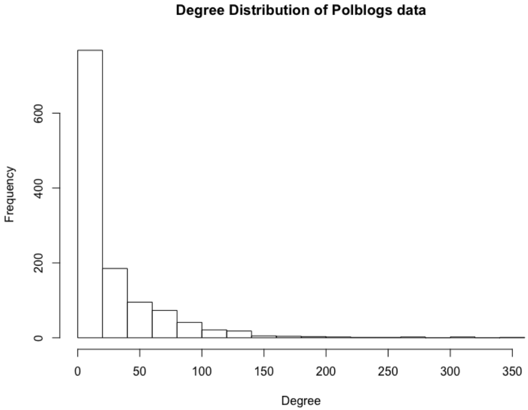
Politics Blogs Data: Blogs are labeled according to the political stand: liberal (0) or conservative (1). Links between blogs were automatically extracted from a crawl of the front page of the blog.



Adamic and Glance (2005), "The political blogosphere and the 2004 US Election"

Degree Distribution (Examples)

Degree Distribution for Politics Blogs Data:



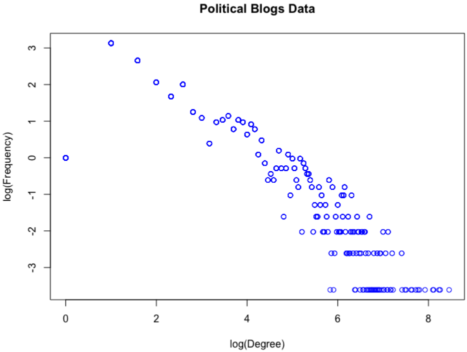
- It is right-skewed and heavy-tailed.
- It's quite normal for real data sets (but not required!!!). If not so, please double check your data.
- We need a better scale.

Degree Distribution (Examples)

Log-log scale of Degree Distribution:

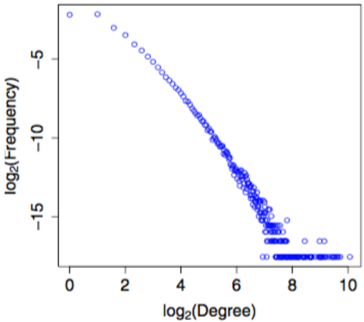
- y-axis: $\log_2$(Frequency)
- x-axis: $\log_2$(Degree)

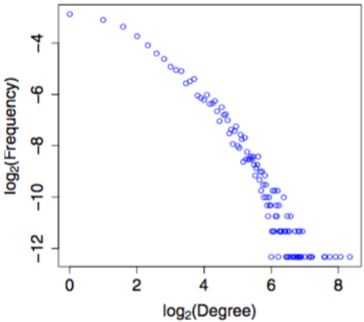
Political Blogs Data:



Remark. Very close to a *linear* relationship with *negative* coefficient!

Two more examples from Textbook (Page 82):





Remark.

- Still quite *linear* !
- Say that $\log_2(f_d) \approx -\alpha \log_2(d) + C = \log_2(2^C d^{-\alpha})$, so

$$f_d \propto d^{-\alpha} \text{ Power-law degree distributions}$$

An approximation of degree distribution:

$$f_d \propto d^{-\alpha}$$

- The decreasing speed is $\log f_d \approx -\alpha \log(d)$. Compare to other dist:
  - Gaussian dist. $d \sim N(0, \sigma^2)$:

$$f_d = C \exp^{-d^2/2\sigma^2} \Longrightarrow \log f_d \approx -Cd^2 < -\alpha \log(d)$$

  - Exponential dist. $d \sim Exp(\lambda)$:

$$f_d = \lambda e^{-\lambda d} \Longrightarrow \log f_d \approx -Cd < -\alpha \log(d)$$

  - In all, f_d decreases slower than the exponential tail and Gaussian tail, which means heavy-tail
- The parameter α is an important quantity to evaluate the network.
  - Larger α means faster decreasing, which means a network with fewer "Hot nodes"

How to figure out α?

- Fitting directly with, say, least squares regression.
  Problem. The noise at the high degrees will cause much trouble
- Fitting α with cumulative densities $F(d) = P(\text{degree} \leq d)$. Given the definition, the probability for nodes with high degrees won't be affected much by the noise.
  The tail probabilities have the form
  $$1 - F(d) \sim d^{-(\alpha-1)}.$$
  Consider a linear regression to estimate α.
- Instead of linear regression, use the estimates in other forms.

Distance

Recall:

- *Distance between two nodes i, j*:

$$d(i,j) = \text{the length of the shortest path between } i \text{ and } j.$$

- If i and j are unconnected, define $d(i,j) = \infty$

New:

- Average distance between nodes (in the same component)

$$\bar{d} = \frac{1}{n(n-1)} \sum_{i,j} d(i,j).$$

- Diameter of a graph.

Diameter

Given a graph $G = (V, E)$, the *diameter* of G is defined as

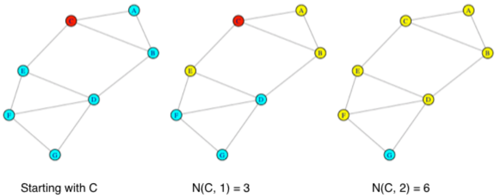
$$diam(G) = \max_{i,j} \min d(i,j),$$

which is the *maximum* geodesic distance between two nodes.

Diameter

Typically, the diameter of a graph is low. The intuition is as follows:

- Given an arbitrary starting node i, let $N(i, j)$ be the number of nodes which are reachable in a j-step path.



- Let $\bar{d}$ be the average degree $\Rightarrow$ $\bar{d}$ is the average of number of neighbors of a node.
- $N(i, 1) \approx \bar{d}$, $N(i, 2) \approx \bar{d}(\bar{d} - 1)$, $\cdots$, $N(i, r) \approx \bar{d}(\bar{d} - 1)^{r-1} \approx \bar{d}^r$.
- Assume at step r, all the n nodes are covered, where $n \approx \bar{d}^r$. Then $r \approx Diam(G)$.
- Therefore, $Diam(G) \approx \log(n)/\log(\bar{d})$, which is at the rate of $\log(n)$.
- $\log(n)$ is quite small compared to n.

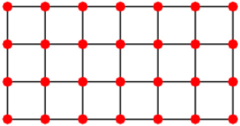
Diameter

- For most real data networks, the diameter is small.
  - A famous example is *Six-degree separation (small-world network)*. Assuming the average degree (connection one person has) is 40, then the diameter is approximately

$$\frac{\log(\text{whole world population})}{\log(\bar{d})} = \frac{\log(3,490,333,715)}{\log(40)} = 5.96.$$

- However, there are counter-examples
  - Recall the two-dimensional lattice network:

This network has 28 nodes, with average degree approximately 3. However, the diameter is 9, which is much larger than $\log(28)/\log(3) = 3$.



  - Usually, a p-dim lattices have diameters at $O(n^{1/p})$
- The real-world network can be thought of as in a *high-dimensional* space

Summary

When you have a real data network,

- Check whether this network is simple, connected, directed/undirected
- Get a summary of the number of nodes, edges, and the corresponding properties
- Examine the degree distribution. Have a plot of the degree distribution, and get the corresponding stats
- Take a look at the diameter of the graph.

All of these show the properties of the whole graph. However, sometimes we are interested in the role of *a node* in the graph.