graph sampling

SYSM 6302

CLASS 18

Graph Sampling

Given a population graph $G = (V, E)$

Obtain a sample graph $G^* = (V^*, E^*)$ ← often we think of G^* as a subgraph of G, but doesn't have to be (additional "incorrect" edges)

Consider a statistic of G: $\eta(G)$

Can we use G^* to get a reliable estimate?

$$\eta(G) \overset{?}{\approx} \hat{\eta} = \eta(G^*)$$

↑ the "plug-in" method

Suppose $\eta(G) = \frac{1}{N} \sum_{i \in V} k_i$ Average Degree $N = |V|$

① Sample n nodes without replacement

$V^* = \{i_1, i_2, \ldots, i_n\}$ with degrees $\{k_i^*\}_{i \in V}$

$$\Rightarrow \eta(G^*) = \frac{1}{n} \sum_{i \in V^*} k_i^* \quad \leftarrow \text{plug-in approach}$$

Consider two sampling methods:

(2.1) k_i^* is the actual degree of the node (i.e., observe <u>all</u> incident edges)

(2.2) k_i^* is the degree of the graph induced by V^* (i.e., only observe edges between sampled nodes)

→ (2.2) greatly underestimates $k_i \Rightarrow$ underestimates $\eta(G)$

In (2.2) $k_i^* \approx k_i \cdot \frac{n}{N}$

$\frac{n}{N}$ = fraction of edges expected to fall into the sample

k_i = actual degree

Adjusting an estimate for sample bias (Horvitz-Thompson Estimation)

Stepping back from networks...

Given a population U with values: $\{y_i\}_{i=1}^{N}$ (height, age, income, etc)

Total: $\tau = \sum_{i \in V} y_i$ **Mean:** $\mu = \frac{\tau}{N}$

Given a sample S with values: $\{y_i\}_{i=1}^{n}$ simple random sample w/ replacement

Sample mean: $\bar{y} = \frac{1}{n} \sum_{i \in S} y_i \rightarrow E[\bar{y}] = \mu$ and $E[N\bar{y}] = \tau$

expectation ⇒ Unbiased Estimate

If a different sampling technique is used could lead to biased estimates

$$\hat{\tau}_\pi = \sum_{i \in S} \frac{y_i}{\pi_i}$$

← unbiased the estimate through a weighted average (weighted by the probability π_i)

π_i: probability that i is in the sample ← inclusion probability ($\pi_i > 0$)

Define $z_i \in \{0, 1\}$, $z_i = \begin{cases} 1 & \text{if } i \text{ is in sample } S \\ 0 & \text{otherwise} \end{cases}$

$E[z_i] = P(z_i = 1) = \pi_i$

$$E\left[\hat{\tau}_\pi\right] = E\left[\sum_{i \in S} \frac{y_i}{\pi_i}\right] = E\left[\sum_{i \in U} \frac{y_i}{\pi_i} z_i\right] = \sum_{i \in U} \frac{y_i}{\pi_i} E[z_i] = \sum_{i \in U} y_i$$

Take-away: if we can quantify π_i, we can unbiased our estimates

↖ depends on sampling method

Induced Subgraph Sampling

n nodes sampled without replacement $\Rightarrow V^*$

$$E^* = \{(u,v) \mid u,v \in V^*, (u,v) \in E\}$$ ← Induced edges

Nodes selected uniformly : $\pi_i = \dfrac{\binom{N-1}{n-1}}{\binom{N}{n}} = \dfrac{\frac{(N-1)!}{(n-1)!\,(N-n)!}}{\frac{N!}{n!\,(N-n)!}} = \dfrac{n}{N}$

$\binom{N-1}{n-1}$: # of ways that node i can be in the sample

$\binom{N}{n}$: # possible samples of size n

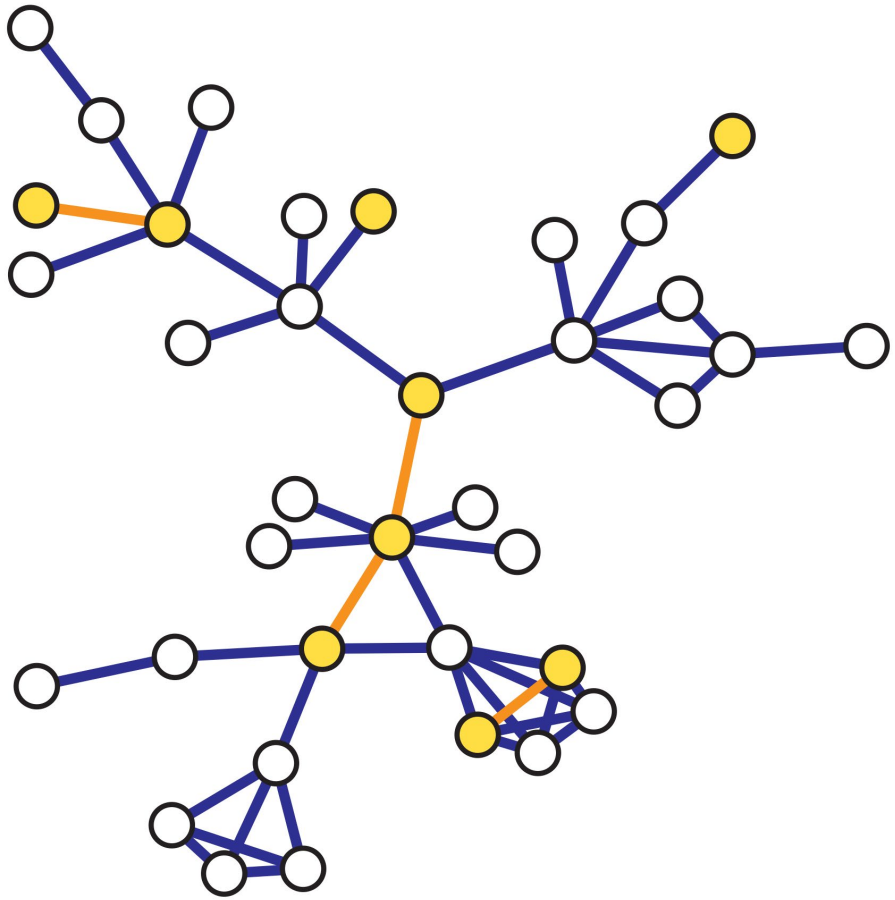
For node i and node j to both be sampled : $\pi_{ij} = \dfrac{n}{N} \cdot \dfrac{n-1}{N-1} = \dfrac{n(n-1)}{N(N-1)}$

π_{ij}: probability to have edge (i,j) in sample

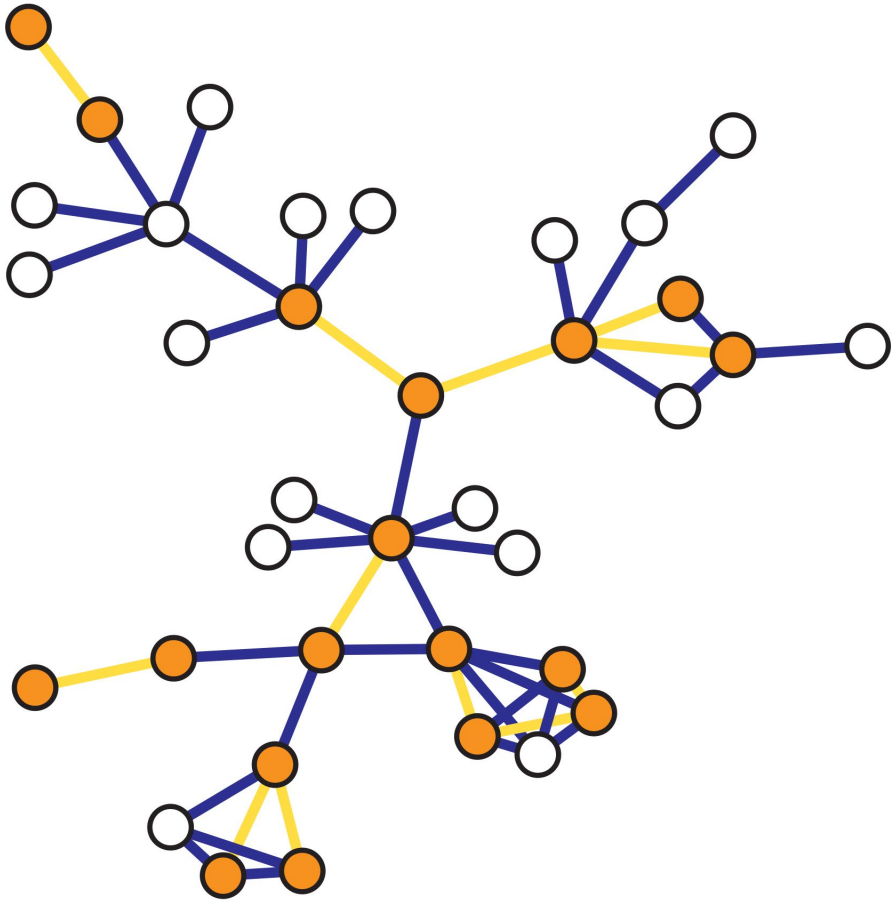
$\frac{n}{N}$: prob of one node

$\frac{n-1}{N-1}$: prob. of next node

These do require that we know N



Induced



Incident

Incident Subgraph Sampling

n edges are sampled without replacement $\longrightarrow E^*$

All nodes incident to sampled edges are observed $\longrightarrow V^*$

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

(n: # edges sampled; $|E|$: out of so many total edges)

Requires knowledge of M

$$\pi_i = P(\text{vertex } i \text{ is sampled}) = 1 - P(\text{no edge incident to node } i \text{ is sampled})$$

$$= \begin{cases} 1 - \dfrac{\binom{M-k_i}{n}}{\binom{M}{n}} & \text{if } n \leq M-k_i \\ 1 & \text{if } n > M-k_i \end{cases}$$

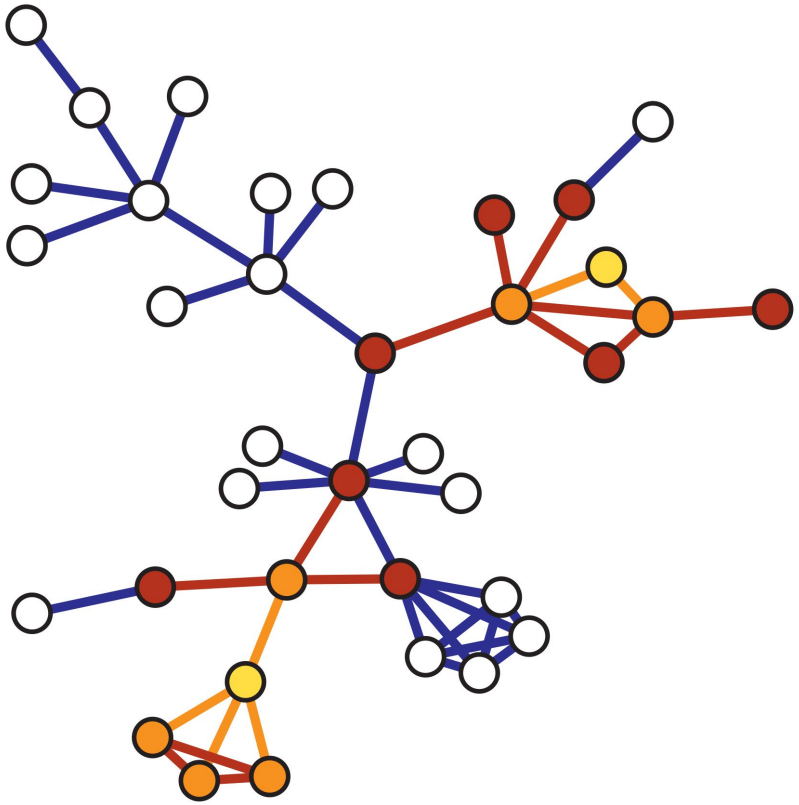
Vertex inclusion probabilities are non-uniform, depending on its degree

requires knowledge of k_i

Snowball Sampling

1. An initial set of nodes V_0 is selected
2. Set $V^* = V_0$
3. Follow edges incident to V^* and add other endpoints to V^*
4. Repeat 3 until some stopping condition is met

This is the process used by web-crawlers



Link Tracing

→ After the selection of the initial sample a subset of edges incident to the sample are followed to additional nodes. (according to a predesignated rule)

→ Snow ball sampling is a special case in which all edges are followed