Graph sampling and estimation

- Population graph $G = (V, E)$
- Without loss of generality, write $V = [N_V] = \{1, \ldots, N_V\}$

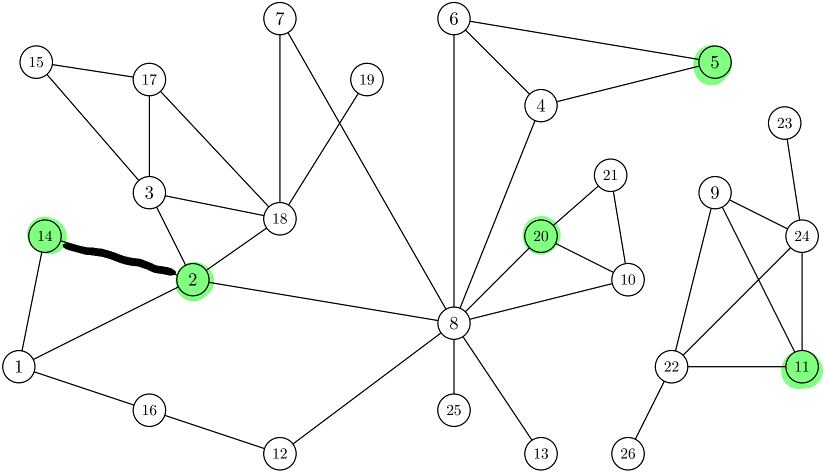
Either $(y_u)_{u \in \mathcal{U}}$ is unobserved or G is too big/complicated to compute $\tau = \sum_{u \in \mathcal{U}} y_u$:

- Randomly sample a subgraph $G^* = (V^*, E^*)$ from G **without replacement/duplicates**:
- That is, draw $V^* \subseteq V$ and $E^* \subseteq E$ according to some sampling scheme (see below) to get a random sample $S \subseteq \mathcal{U}$.
- Use Horvitz-Thompson approach

$$\hat{\tau} = \sum_{u \in S} \frac{y_u}{\pi_u}$$

for inclusion probabilities π_u, $u \in \mathcal{U}$.

INDUCED SUBGRAPH SAMPLING



$V^* = \{14, 2, 20, 5, 11\}$

$E^* = \{\{14, 2\}\}$

INDUCED SUBGRAPH SAMPLING

1. Sample n times **without replacement** from V to obtain V^*.
2. Add all edges joining vertices in V^*, i.e.,
   $E^* := \{\{u,v\} \in E : u, v \in V^*\}$.

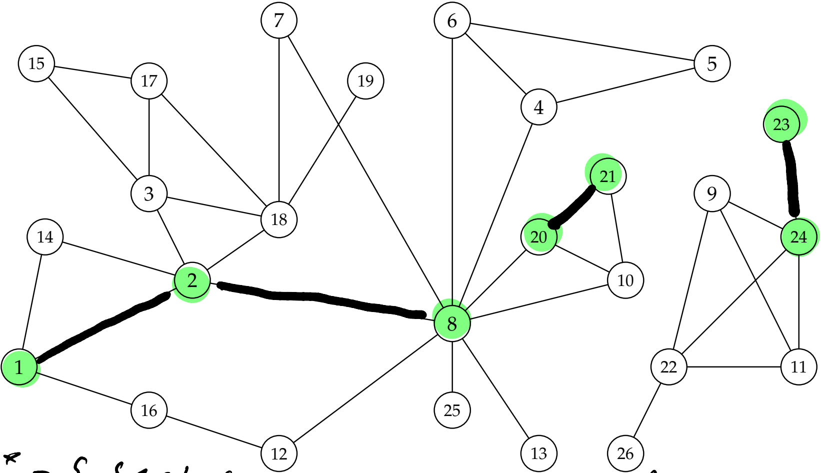
Vertex inclusion probabilities (as above):

- For $i \in V$, we have $\pi_i = \frac{n}{N_V}$.
- For $\{i,j\} \in V^{(2)}$, we have $\pi_{\{i,j\}} = \frac{n(n-1)}{N_V(N_V-1)}$.

in general :

$e = \{i,j\}$, $\pi_e = P(\text{edge } e \text{ is sampled}) \neq \pi_{\{i,j\}}$

INCIDENT SUBGRAPH SAMPLING



$E^* = \{ \{1,2\}, \{2,8\}, \{20,21\}, \{23,24\} \}$

$V^* = \{ 1, 2, 8, 20, 21, 23, 24 \}$

Incident subgraph sampling

1. Sample n times **without replacement** from E to obtain E^*.
2. Add all incident vertices
   $V^* := \{v \in V : \exists e \in E^* \text{ such that } v \in e\}$.

Edge inclusion probabilities:

- For $e \in E \subseteq V^{(2)}$, $\pi_e = P(\text{edge } e \text{ is sampled}) = \frac{n}{N_E}$.

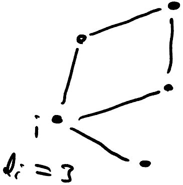
Note: If $\mathcal{U} = V^{(2)}$, we can only use this if

$$\tau = \sum_{\{i,j\} \in V^{(2)}} y_{\{i,j\}} = \sum_{e \in E} y_e$$

is an edge total! Otherwise, we would need to compute $\pi_{\{i,j\}} := P(\text{vertex pair } \{i,j\} \text{ is sampled})$, for all $\{i,j\} \in V^{(2)}$.

INCIDENT SUBGRAPH SAMPLING

1. Sample n times **without replacement** from E to obtain E^*.
2. Add all incident vertices
   $V^* := \{v \in V : \exists e \in E^* \text{ such that } v \in e\}$.

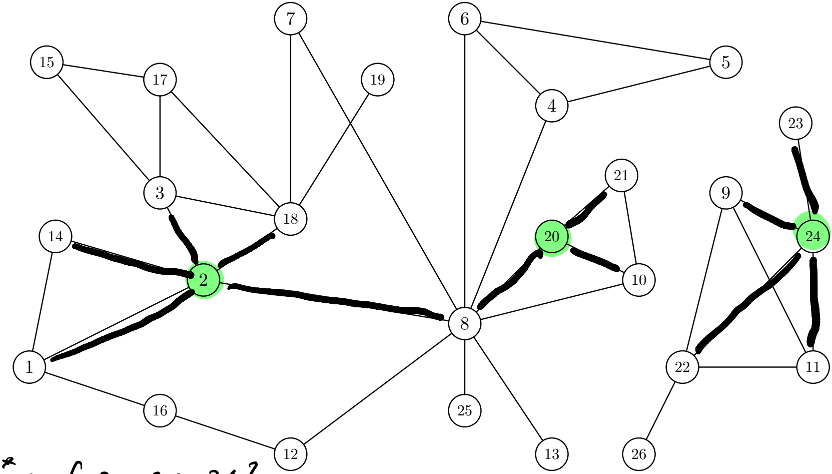


Vertex inclusion probabilities:

- For $i \in V$, $\pi_i = P(\text{vertex } i \text{ is sampled})$
- If $\boxed{d_i > N_E - n}$, then $\pi_i = 1$
- If $\boxed{d_i \leq N_E - n}$, then for $E_i := \{e \in E : i \in e\}$, $|E_i| = d_i$ and

$$\begin{aligned}\pi_i &= 1 - P(\text{no edge incident to } i \text{ is sampled}) \\ &= 1 - \frac{\#\text{ (unordered) samples of size } n \text{ drawn from } E \setminus E_i}{\#\text{ (unordered) samples of size } n \text{ drawn from } E} \\ &= 1 - \frac{\binom{N_E - d_i}{n} \cancel{n!}}{\binom{N_E}{n} \cancel{n!}}.\end{aligned}$$

UNLABELED STAR SAMPLING



$V^* = \{2, 20, 24\}$

$E^* = \{\{1,2\}, \{14,2\}, \dots\}$

Unlabeled star sampling

1. Sample n times **without replacement** from V to obtain V^*.
2. Add incident edges
   $E^* := \{e \in E : \exists v \in V^* \text{ such that } v \in e\}$.

Vertex inclusion probabilities:

- For $i \in V$, $\pi_i = \frac{n}{N_V}$.

Unlabeled star sampling

1. Sample n times **without replacement** from V to obtain V^*.
2. Add incident edges
   $E^* := \{e \in E : \exists v \in V^* \text{ such that } v \in e\}$.

- For $e = \{i, j\} \in E \subseteq V^{(2)}$, **(edge (!)** inclusion probabilities)

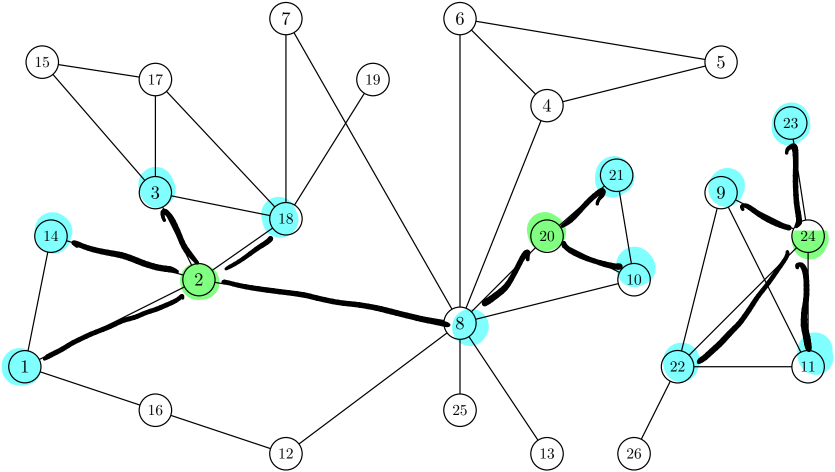
$$\begin{aligned} \pi_e &= P(\text{edge } e = \{i, j\} \text{ is sampled}) \\ &= 1 - P(\text{neither vertex } i \text{ nor vertex } j \text{ is sampled}) \\ &= 1 - \frac{\binom{N_V - 2}{n}}{\binom{N_V}{n}}. \end{aligned}$$

- **However**, for **vertex pairs** $\{i, j\} \in V^{(2)}$, we have

$$\pi_{\{i,j\}} = P(\text{vertex pair } \{i, j\} \text{ is sampled}) = \frac{n(n-1)}{N_V(N_V - 1)},$$

as in induced subgraph sampling.

Labeled star sampling



Labeled star sampling

1. Sample n times **without replacement** from V to obtain V_0^*.
2. Add incident edges
   $E^* := \{e \in E : \exists v \in V_0^* \text{ such that } v \in e\}$.
3. Add the vertices incident to E^*
   $V^* := \{v \in V : \exists e \in E^* \text{ such that } v \in e\}$.

Edge inclusion probabilities (as in the unlabeled case):

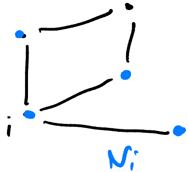
- For $e \in E$,

$$\pi_e = 1 - \frac{\binom{N_V - 2}{n}}{\binom{N_V}{n}}.$$

LABELED STAR SAMPLING

1. Sample n times **without replacement** from V to obtain V_0^*.
2. Add incident edges
   $E^* := \{e \in E : \exists v \in V_0^* \text{ such that } v \in e\}$.
3. Add the vertices incident to E^*
   $V^* := \{v \in V : \exists e \in E^* \text{ such that } v \in e\}$.

$d_i = 3$



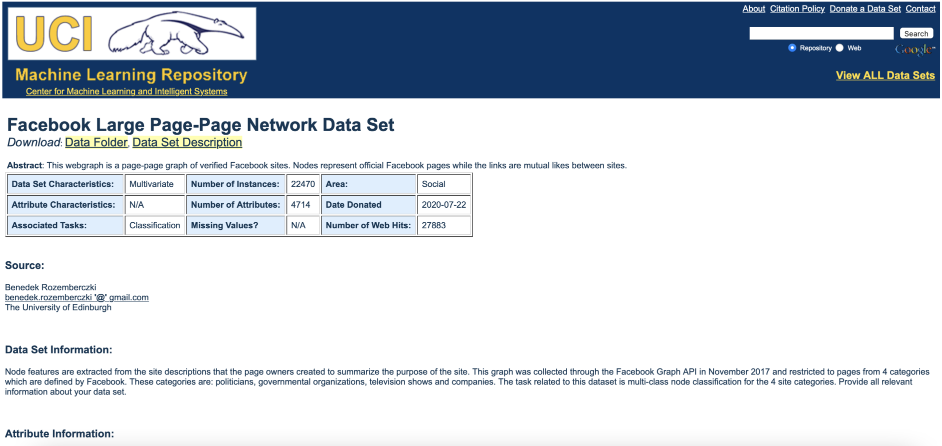
Vertex inclusion probabilities:

- For $i \in V$, define $\underline{N_i :=} \{j \in V : \text{dist}(i,j) \leq 1\}$. Thus, $|N_i| = d_i + 1$ and

$$\begin{aligned}
\pi_i &= P(\text{vertex } i \text{ is sampled}) \\
&= 1 - P(\text{no vertex from } N_i \text{ is contained in } V_0^*) \\
&= 1 - \frac{\binom{N_V - (d_i+1)}{n}}{\binom{N_V}{n}}.
\end{aligned}$$

Example: How political are TV-shows?

Example:
How political are TV-shows?



Facebook Large Page-Page Network Data Set

Download: Data Folder, Data Set Description

Abstract: This webgraph is a page-page graph of verified Facebook sites. Nodes represent official Facebook pages while the links are mutual likes between sites.

Data Set Characteristics:	Multivariate	**Number of Instances:**	22470	**Area:**	Social
Attribute Characteristics:	N/A	**Number of Attributes:**	4714	**Date Donated**	2020-07-22
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	27883

Source:

Benedek Rozemberczki
benedek.rozemberczki '@' gmail.com
The University of Edinburgh

Data Set Information:

Node features are extracted from the site descriptions that the page owners created to summarize the purpose of the site. This graph was collected through the Facebook Graph API in November 2017 and restricted to pages from 4 categories which are defined by Facebook. These categories are: politicians, governmental organizations, television shows and companies. The task related to this dataset is multi-class node classification for the 4 site categories. Provide all relevant information about your data set.

Attribute Information:

`https://archive.ics.uci.edu/ml/datasets/`

`Facebook+Large+Page-Page+Network`

EXAMPLE: POLITIZATION OF TV-SHOWS

- Population graph: $22\,470$ Facebook pages of politicians, government agencies, TV-shows and companies.
- Links/edges are mutual likes (sparse).
- Year: 2019
- Question: How strongly connected are TV-shows and political agents (politicians or government agencies)?
- Formally: For an edge $e \in E$, set $y_e = 1$ if e connects a TV-show with a political agent, and set $y_e = 0$ otherwise. Compute

$$\tau := \sum_{e \in E} y_e, \quad \mu := \frac{\tau}{N_E}.$$

- $\mathcal{U} = E$
- Assume that we don't have access to the edge set E directly, but only through querying vertex neighbors.
- Consider both known and unknown N_E ($N_E = 171\,002$).

1.) Induced subgraph sampling

- Draw V^* of size n uniformly from V without replacement.
- Choose $E^* = \{\{u, v\} \in E : u \in V^* \wedge v \in V^*\}$.
- Here, the vertex pair inclusion probability $\pi_{\{i,j\}}$ equals the edge inclusion probability $\pi_e = \pi = \frac{n(n-1)}{N_V(N_V-1)}$.
- Horvitz-Thompson estimate

$$\hat{\tau} = \sum_{e \in E^*} \frac{y_e}{\pi_e} = \frac{1}{\pi} \sum_{e \in E^*} y_e$$
$$= \frac{N_V(N_V-1)}{n(n-1)} \cdot \# \{e \in E^* : e \text{ is a TV/politics pair}\} .$$

For moderately large n we often find $\hat{\tau} = 0$.

Estimating μ and N_E

- If N_E is known,

  $$\hat{\mu} = \frac{\hat{\tau}}{N_E}$$

  is an unbiased estimator for μ.

- If $N_E = \sum_{e \in E} 1$ is unknown, we can estimate it as a population total using the Horvitz-Thompson approach

  $$\hat{N}_E := \sum_{e \in E^*} \frac{1}{\pi_e} = \frac{|E^*|}{\pi}, \qquad y_e = 1$$

  with

  $$\mathbb{E}[\hat{N}_E] = \sum_{e \in E} y_e = N_E,$$

  and set

  $$\hat{\mu} = \frac{\hat{\tau}}{\hat{N}_E} = \frac{\#\{e \in E^* : e \text{ is a TV/politics pair}\}}{|E^*|}.$$

2.) Labeled star sampling

- Draw V_0^* of size n uniformly from V without replacement.
- Choose $E^* = \{\{u, v\} \in E : u \in V_0^* \vee v \in V_0^*\}$.
- Take $V^* := \{v \in V : \exists e \in E^* \text{ such that } v \in e\}$.
- Edge inclusion probability

$$\pi_e = \pi = 1 - \frac{\binom{N_V - 2}{n}}{\binom{N_V}{n}} = 1 - \frac{(N_V - n)(N_V - n - 1)}{N_V(N_V - 1)}.$$

- Horvitz-Thompson estimate

$$\hat{\tau} = \sum_{e \in E^*} \frac{y_e}{\pi_e} = \frac{1}{\pi} \sum_{e \in E^*} y_e$$
$$= \frac{\#\{e \in E^* : e \text{ is a TV/politics pair}\}}{\pi}.$$

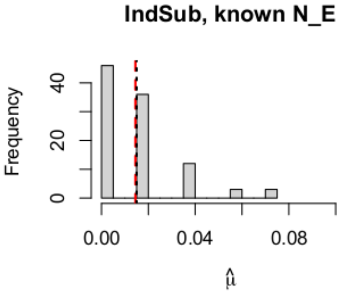
For the same n as in 1.) this searches through a lot more edges and thus has much higher chance of discovering any TV/politics edges.

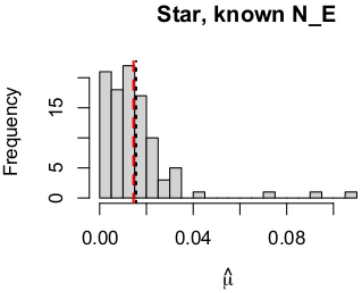
MC simulations

- Investigate the sampling distribution of $\hat{\mu}$ in our different scenarios.
- For the same $n = |V_0^*|$, unlabeled star sampling is computationally more expensive than induced subgraph sampling.
- We simulate such that both sampling designs have similar runtime and compare their statistical performance.

MC SIMULATIONS

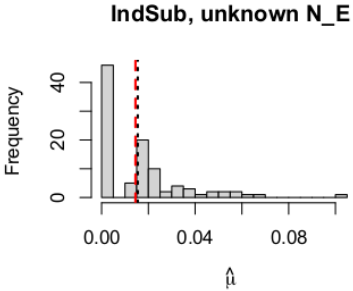
$MC = 100$	induced subgraph	labeled star
$n = \lvert V_\bullet^* \rvert$	400	50
avg. $\lvert E^* \rvert$	53.17	768.49
avg. runtime	0.49	0.38

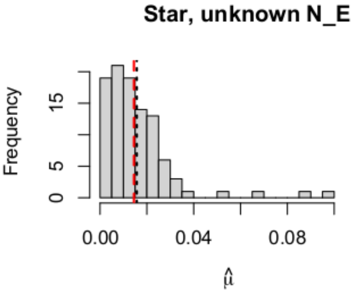




MC SIMULATIONS

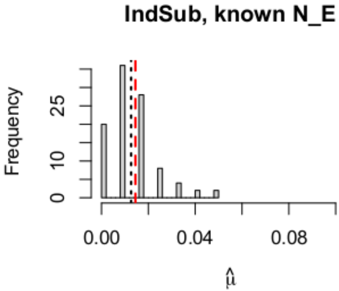
$MC = 100$	induced subgraph	labeled star
$n = \|V^*\|$	400	50
avg. $\|E^*\|$	53.17	768.49
avg. runtime	0.49	0.38

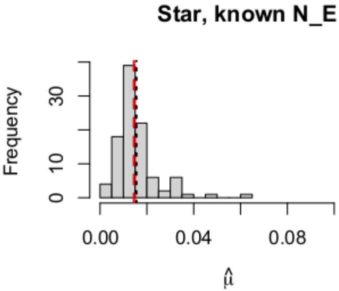




MC SIMULATIONS

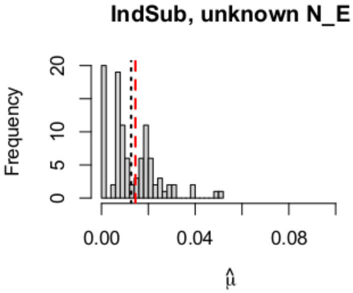
$MC = 100$	induced subgraph	labeled star
$n = \lvert V^* \rvert$	600	100
avg. $\lvert E^* \rvert$	124.11	1550.64
avg. runtime	1.27	1.74

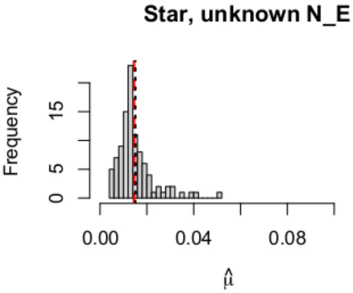




MC SIMULATIONS

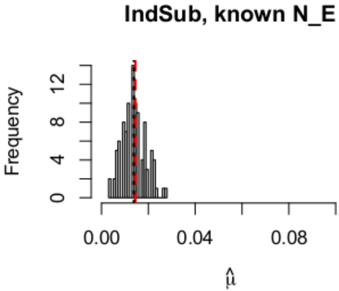
$MC = 100$	induced subgraph	labeled star
$n = \|V^*\|$	600	100
avg. $\|E^*\|$	124.11	1550.64
avg. runtime	1.27	1.74

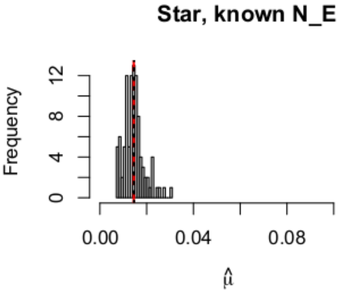




MC SIMULATIONS

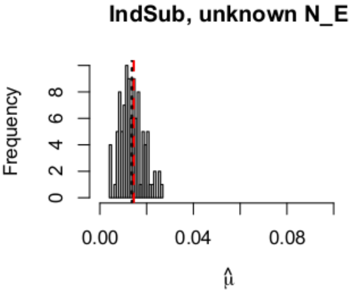
$MC = 100$	induced subgraph	labeled star
$n = \|V^*\|$	1500	300
avg. $\|E^*\|$	767.68	4505.11
avg. runtime	17.51	16.51

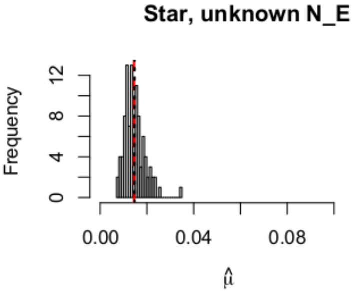




MC SIMULATIONS

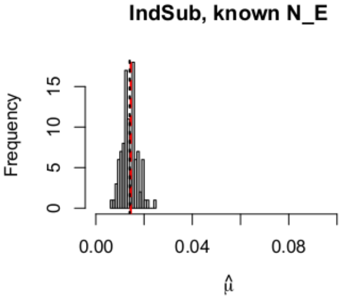
$MC = 100$	induced subgraph	unlabeled star
$n = \|V^*\|$	1500	300
avg. $\|E^*\|$	767.68	4505.11
avg. runtime	17.51	16.51

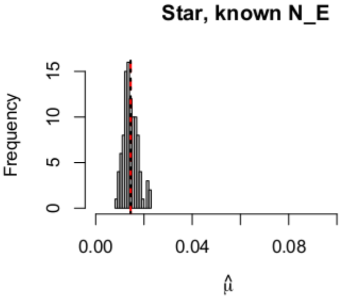




MC SIMULATIONS

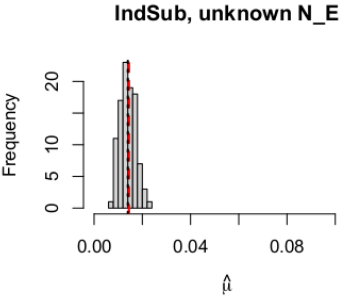
$MC = 100$	induced subgraph	labeled star
$n = \|V^*\|$	2800	800
avg. $\|E^*\|$	2663.96	11839.38
avg. runtime	87.41	91.21

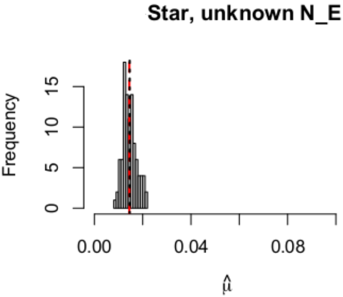




MC SIMULATIONS

$MC = 100$	induced subgraph	labeled star
$n = \|V^*\|$	2800	800
avg. $\|E^*\|$	2663.96	11839.38
avg. runtime	87.41	91.21





Towards uncertainty quantification for graph sampling

TOWARDS UNCERTAINTY QUANTIFICATION

- We continue with the Facebook example.
- For sufficiently large n, the sampling distributions of our estimators look symmetric and bell shaped with no serious outliers.
- This motivates a normal approximation.
- A rigorous mathematical motivation is beyond the scope of this course.
- Recall the approximate Gaussian CI

$$\boxed{CI_\alpha = \hat{\mu} \pm q^{(N)}_{1-\frac{\alpha}{2}} \hat{se}.}$$

- We need an estimate $\hat{se} = \hat{se}(\hat{\mu})$ of the standard error of our estimator $\hat{\mu}$.

TOWARDS UNCERTAINTY QUANTIFICATION

- Recall homework: For $\boxed{\mathcal{U} = [N]}$ and $S = (i_1, \ldots, i_n) \in \mathcal{U}^n$,

$$\hat{se}^2 = \sum_{k=1}^{n} \sum_{l=1}^{n} y_{i_k} y_{i_l} \left(\frac{1}{\pi_{i_k} \pi_{i_l}} - \frac{1}{\pi_{i_k i_l}} \right)$$

  is unbiased for $\mathrm{Var}[\hat{\tau}]$, where $\hat{\tau} = \sum_{j=1}^{n} \frac{y_{i_j}}{\pi_{i_j}}$.

- Therefore,

$$\hat{se}^2 / N^2 = \frac{1}{N^2} \sum_{k=1}^{n} \sum_{l=1}^{n} y_{i_k} y_{i_l} \left(\frac{1}{\pi_{i_k} \pi_{i_l}} - \frac{1}{\pi_{i_k i_l}} \right)$$

  is unbiased for $\mathrm{Var}[\hat{\mu}] = \mathrm{Var}[\hat{\tau}/N] = \mathrm{Var}[\hat{\tau}]/N^2$.

Towards uncertainty quantification

- Now: $\boxed{\mathcal{U} = E}$, $N = N_E$, $S = E^*$ and

$$\widehat{se}^2(\hat{\mu}) \quad \cancel{\widehat{se}^2} := \frac{1}{N_E^2} \sum_{e \in E^*} \sum_{f \in E^*} y_e y_f \left(\frac{1}{\pi_e \pi_f} - \frac{1}{\pi_{ef}} \right)$$

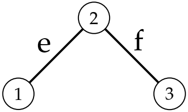
  is unbiased for $\mathrm{Var}[\hat{\mu}]$, where $\hat{\mu} := \frac{1}{N_E} \sum_{e \in E^*} \frac{y_e}{\pi_e}$.
- If N_E is unknown, we estimate it as before by $\hat{N}_E = |E^*|/\pi$.
- We have already computed edge inclusion probabilities π_e.
- We need also edge-pair inclusion probabilities π_{ef} for $e, f \in E$.

Edge pair inclusion probabilities

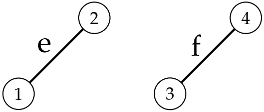
$$\pi_{ef} = P(\text{ e and f are sampled })$$
$$= P(\, e, f \in E^*)$$

- There are two different kinds of edge pairs (e, f):

A



B



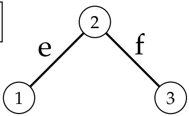
EDGE PAIR INCLUSION PROBABILITIES

Induced subgraph sampling:

$n = |V^*|$

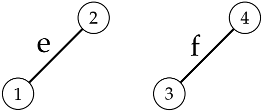
$$\pi_{ef} = P(e \in E^* \text{ and } f \in E^*) =$$

A



$$= P(1,2,3 \in V^*) = \frac{\binom{N_V-3}{n-3}}{\binom{N_V}{n}} = \frac{n(n-1)(n-2)}{N_V(N_V-1)(N_V-2)}.$$

B



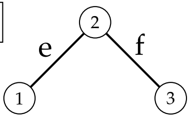
$$= P(1,2,3,4 \in V^*) = \frac{\binom{N_V-4}{n-4}}{\binom{N_V}{n}} = \frac{n(n-1)(n-2)(n-3)}{N_V(N_V-1)(N_V-2)(N_V-3)}.$$

Edge pair inclusion probabilities

Labeled star sampling:

$$\pi_{ef} = P(e \in E^* \text{ and } f \in E^*) =$$

$\boxed{\text{A}}$



$$= P(\underline{\{1,2,3 \in V^*\}} \cup \underline{\{1,2 \in V^*, 3 \notin V^*\}} \cup \underline{\{1,3 \in V^*, 2 \notin V^*\}}$$
$$\cup \underline{\{2,3 \in V^*, 1 \notin V^*\}} \cup \underline{\{2 \in V^*, 1,3 \notin V^*\}})$$

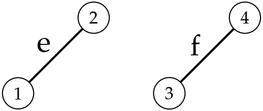
$$= \frac{\binom{N_v-3}{n-3} + 3\cdot\binom{N_v-3}{n-2} + \binom{N_v-3}{n-1}}{\binom{N_v}{n}}$$

Edge pair inclusion probabilities

Labeled star sampling:

$$\pi_{ef} = P(e \in E^* \text{ and } f \in E^*) =$$

$\boxed{\text{B}}$

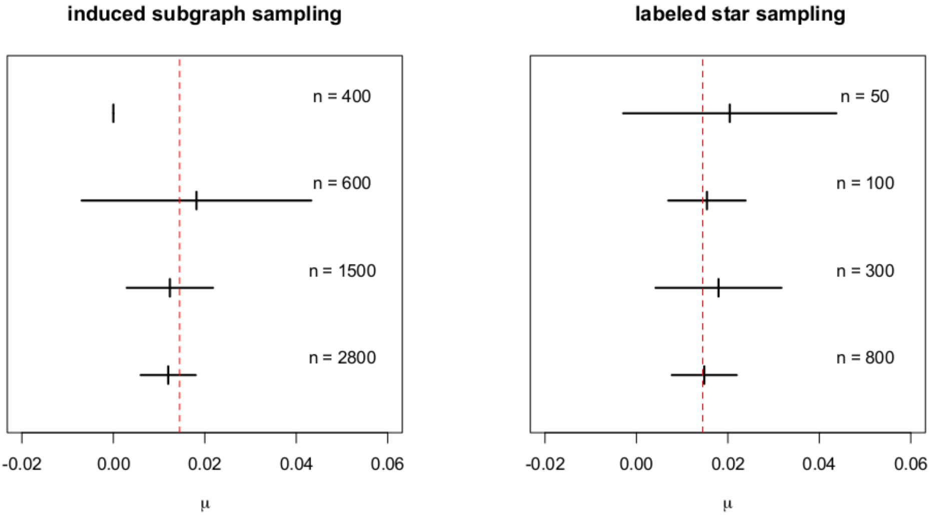


$$= P(\underline{\{1,2,3,4 \in V^*\}}) + 4 \cdot P(\underline{\{1,2,3 \in V^*, 4 \notin V^*\}})$$
$$+ 4 \cdot P(\underline{\{1,3 \in V^*, 2,4 \notin V^*\}\}})$$

$$= \frac{\binom{N_v - 4}{n-4} + 4 \binom{N_v - 4}{n-3} + 4 \cdot \binom{N_v - 4}{n-2}}{\binom{N_v}{n}}$$

95% CI FOR μ

$\mu =$ proportion of edges that connect a TV show with a politician or government institution.


induced subgraph sampling
n = 400
n = 600
n = 1500
n = 2800
μ
labeled star sampling
n = 50
n = 100
n = 300
n = 800
μ