Statistics for Data Science, WS2023

Chapter 5:

Statistical Network Analysis

Networks are everywhere

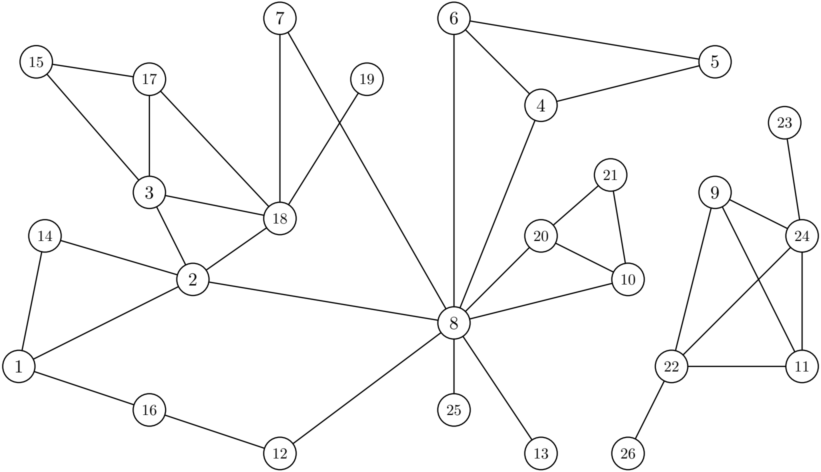
- social networks
- computer networks (WWW)
- electricity grid
- street maps
- gene regulatory networks
- etc.

Often the full network is unknown or too big to compute or extract the characteristics of interest

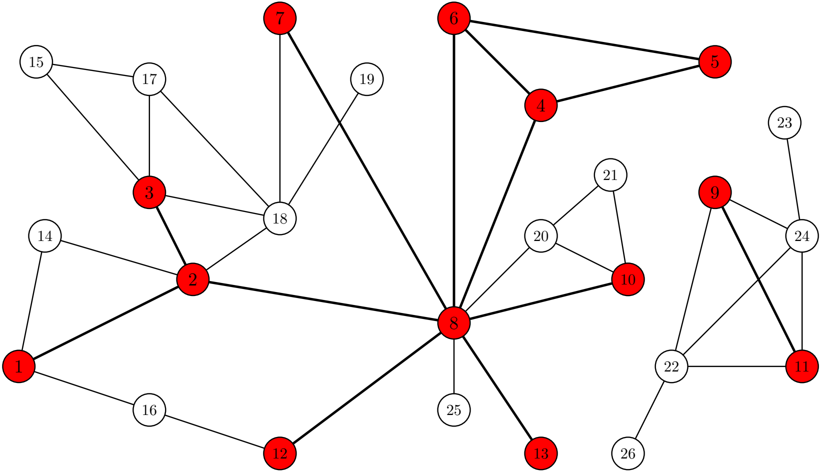
$$\Downarrow$$

Random sampling + statistical inference

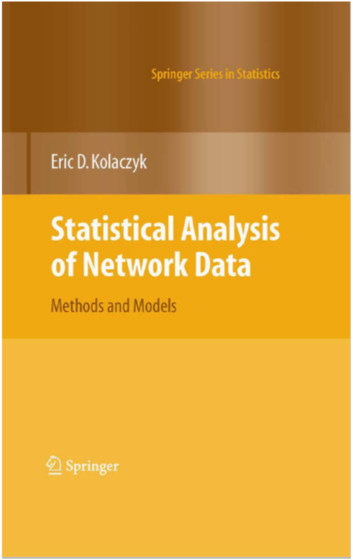
Sampling from a hidden network



Sampling from a hidden network



Following Kolaczyk (2009)



Graph notation

Definition 5.1

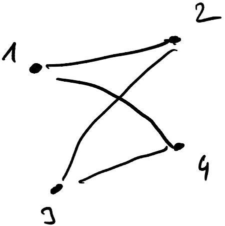
An undirected graph is given by a pair $G = (V, E)$ where
$V \subseteq \mathbb{N}$ is the set of *vertices* (or *nodes*) and
$E \subseteq V^{(2)} := \{\{v, u\} : v, u \in V, v \neq u\}$ is the set of (undirected)
edges or *links*. We write $N_V := |V|$, $N_E := |E|$ and
$V = \{v_1, \ldots, v_{N_V}\}$.

$\{v, u\} = \{u, v\}$

GRAPH NOTATION

For (an undirected) graph $G = (V, E)$, its *adjacency matrix* $A \in \mathbb{R}^{N_V \times N_V}$ is given by

$$A_{ij} = \begin{cases} 1, & \text{if } \{v_i, v_j\} \in E, \\ 0, & \text{else.} \end{cases}$$



$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

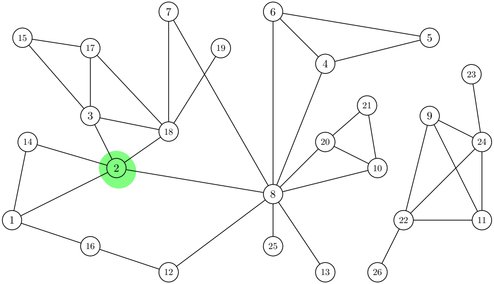
Degree

For $G = (V, E)$ and $v \in V$, the *degree* d_v of v is given by the number of vertices adjacent to v, i.e.,

$$d_v := \sum_{u \in V} \mathbb{1}_E(\{u, v\}) = \sum_{j=1}^{N_V} A_{ij},$$

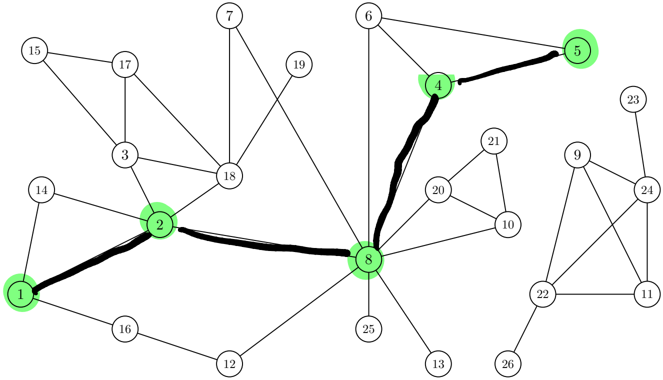
where $i \in [N_V]$ is such that $v_i = v$.

$d_2 = 5$



PATH

For $G = (V, E)$ a *path* is a sequence of adjacent vertices in which no vertex occurs twice, i.e., $(v_1, v_2, \ldots, v_l) \in V^l$ is a path of length $l - 1$ if $\{v_i, v_{i+1}\} \in E$ for $i = 1, \ldots, l - 1$ and $v_i \neq v_j$ for all $i \neq j$.

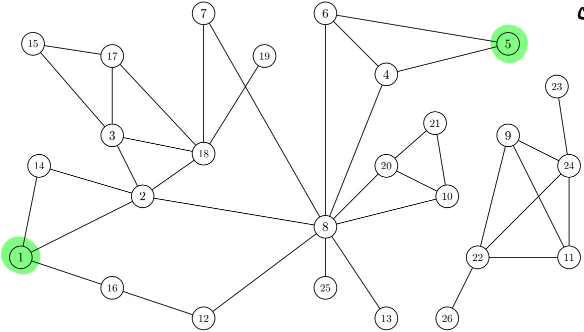


GEODESIC DISTANCE

For $G = (V, E)$ and $v, u \in V$, the *geodesic distance* $\mathrm{dist}(v, u)$ between v and u is the length of the (or a) shortest path starting in v and ending in u, and $\mathrm{dist}(v, v) := 0$.

If there is no path from v to u, we set $\mathrm{dist}(v, u) = \infty$.

$d(1,5) = 4$



Centrality

For $G = (V, E)$ and $v \in V$, the *closeness centrality* $c_{Cl}(v)$ of v is defined as

$$c_{Cl}(v) := \frac{1}{\sum_{u \in V} \text{dist}(v, u)}.$$

For a vertex $v \in V$, the *betweenness centrality* $c_B(v)$ of v is defined as

$$c_B(v) := \sum_{\substack{\{s,t\} \in V^{(2)} \\ v \notin \{s,t\}}} \frac{\sigma(s, t|v)}{\sigma(s, t)},$$

where $\sigma(s, t)$ is the number of shortest paths between s and t, and $\sigma(s, t|v)$ is the number of all those paths that also pass through v. By convention we set $\frac{0}{0} = 0$.

Sampling from a finite population

Course evaluation!

Sampling from finite population

- Population or universe $\mathcal{U} = \{1, \ldots, N\}$, with N known
- Characteristics of interest $y_i \in \mathbb{R}, i \in \mathcal{U}$
- Population total and average $\tau := \sum_{i \in \mathcal{U}} y_i, \mu := \frac{\tau}{N}$
- Draw a **random** sample $S = (i_1, \ldots, i_n) \in \mathcal{U}^n$
- We **observe** $y_{i_1}, \ldots, y_{i_n}$ (duplicates possible!) $(1, 2) \neq$

Goal: Estimate τ and/or μ. $(2, 1)$

Sampling from finite population

$$y_i \in \mathbb{R}, i \in \mathcal{U} = \{1, \ldots, N\}, \quad \tau := \sum_{i \in \mathcal{U}} y_i, \quad \mu := \frac{\tau}{N}$$

random sample $\quad S = (i_1, \ldots, i_n) \in \mathcal{U}^n$

natural choice: (why?)

$$\tilde{\mu} = \frac{1}{n} \sum_{j=1}^{n} y_{i_j}, \quad \tilde{\tau} = N\tilde{\mu}$$

How to compute $\mathbb{E}[\tilde{\tau}]$ and $\mathrm{Var}[\tilde{\tau}]$?

SAMPLING FROM FINITE POPULATION

$$y_i \in \mathbb{R}, i \in \mathcal{U} = \{1, \dots, N\}, \quad \tau := \sum_{i \in \mathcal{U}} y_i, \quad \mu := \frac{\tau}{N}$$

random sample $\quad S = (i_1, \dots, i_n) \in \mathcal{U}^n$

For $i \in \mathcal{U}$, define

$$\boxed{Z_i := \sum_{j=1}^{n} \mathbb{1}_{\{i_j\}}(i)}$$

number of times individual i is sampled

and $\pi_i := \mathbb{E}[Z_i]$, $\pi_{ij} := \mathbb{E}[Z_i Z_j]$.

$\pi_{ij} \neq \pi_{i_j}$

If $\pi_i > 0$, define

$$\hat{\tau} := \sum_{j=1}^{n} \frac{y_{i_j}}{\pi_{i_j}}.$$

Horvitz-Thompson estimate

SAMPLING FROM FINITE POPULATION

$$y_i \in \mathbb{R}, i \in \mathcal{U} = \{1, \dots, N\}, \quad \tau := \sum_{i \in \mathcal{U}} y_i, \quad \mu := \frac{\tau}{N}$$

random sample $\quad S = (i_1, \dots, i_n) \in \mathcal{U}^n$

For $i \in \mathcal{U}$, define

$$Z_i := \sum_{j=1}^{n} \mathbb{1}_{\{i_j\}}(i) \qquad \text{number of times individual } i \text{ is sampled}$$

and $\pi_i := \mathbb{E}[Z_i]$, $\pi_{ij} := \mathbb{E}[Z_i Z_j]$.

Because of

$$\boxed{\sum_{i \in \mathcal{U}} \frac{y_i}{\pi_i} Z_i = \sum_{j=1}^{n} \sum_{i \in \mathcal{U}} \frac{y_i}{\pi_i} \mathbb{1}_{\{i_j\}}(i) = \sum_{j=1}^{n} \frac{y_{i_j}}{\pi_{i_j}} = \hat{\tau},}$$

we have

$$\mathbb{E}[\hat{\tau}] = \mathbb{E}\left[\sum_{i \in \mathcal{U}} \frac{y_i}{\pi_i} Z_i\right] = \sum_{i \in \mathcal{U}} \frac{y_i}{\pi_i} \overset{\substack{\pi_i \\ \shortparallel}}{\mathbb{E}[Z_i]} = \sum_{i \in \mathcal{U}} y_i = \tau.$$

Sampling from finite population

$$\mathrm{Var}[\hat{\tau}] = \mathrm{Var}\Big(\sum_{i\in\mathcal{U}} \frac{y_i}{\pi_i} Z_i\Big) = \sum_{i,j\in\mathcal{U}} \mathrm{Cov}\Big(\frac{y_i}{\pi_i} Z_i, \frac{y_j}{\pi_j} Z_j\Big)$$

$$= \sum_{i,j\in\mathcal{U}} \frac{y_i}{\pi_i}\frac{y_j}{\pi_j} \underbrace{\mathrm{Cov}(Z_i, Z_j)}$$

$$= \underbrace{E(Z_i Z_j)}_{= \pi_{ij}} - \underbrace{E(Z_i)E(Z_j)}_{\pi_i \cdot \pi_j}$$

$$= \sum_{i,j\in\mathcal{U}} y_i y_j \Big(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\Big)$$

HW: Find an unbiased estimator for $\mathrm{Var}[\hat{\tau}]$.

SAMPLING **WITH** REPLACEMENT

$$y_i \in \mathbb{R}, i \in \mathcal{U} = \{1, \ldots, N\}, \quad \tau := \sum_{i \in \mathcal{U}} y_i$$

draw n times uniformly from $\mathcal{U}$ **with replacement** to obtain $S = (i_1, \ldots, i_n) \in \mathcal{U}^n$

For $i \in \mathcal{U}$, define

$$Z_i := \sum_{j=1}^{n} \mathbb{1}_{\{i_j\}}(i) \qquad \text{number of times individual } i \text{ is sampled.}$$

We have

$$(Z_1, \ldots, Z_N) \sim \text{Multinomial}\left(n; \overbrace{\frac{1}{N}, \ldots, \frac{1}{N}}^{N \text{ times}}\right).$$

In particular, $\pi_i = \mathbb{E}[Z_i] = \frac{n}{N}$, $\pi_{ii} = \mathbb{E}[Z_i^2] = \frac{n(N+n-1)}{N^2}$, and $\pi_{ij} = \mathbb{E}[Z_i Z_j] = \frac{n(n-1)}{N^2}$ for $i \neq j$.

Multinomial Distribution

p.m.f : $P(Z_1 = z_1, Z_2 = z_2, \dots, Z_N = z_N) = \circledast$

$$\sum_{i=1}^{N} z_i = n, \qquad z_0 \in \mathbb{N}_0$$

draw n-times with replacement

$$U = \{1, 2, \dots, N\} \qquad p_i = \frac{1}{N}$$

$$p_1 \quad p_2 \quad \dots \quad p_N$$

Example: $S = (1, 3, 5, 3, 3, 1)$

$n = 6$

	Z_1	Z_2	Z_3	Z_4	Z_5	$\dots$	Z_N
$z =$	2	0	3	0	1		0

probability of drawing $S = p_1^2 \cdot p_3^3 \cdot p_5^1$

probability of observing z? Different samples can produce the same z!

How many?

$$p_1^2 \cdot p_3^3 \cdot p_5^1 \cdot \frac{6!}{2!\,3!\,1!}$$

$$= n! \prod_{i=1}^{N} \frac{p_i^{z_i}}{z_i!} = (*)$$

$$p_i = \frac{1}{N} \Rightarrow (*) = n! \prod_{i=1}^{n} \frac{\left(\frac{1}{N}\right)^{z_i}}{z_i!}$$

$$= n! \frac{\left(\frac{1}{N}\right)^{\sum_{i=1}^{N} z_i}}{z_1! \cdots z_N!} \qquad \sum z_i = n$$

$$= \frac{n!}{N^n} \frac{1}{z_1! \cdots z_N!}$$

SAMPLING **WITHOUT** REPLACEMENT

$$y_i \in \mathbb{R}, i \in \mathcal{U} = \{1, \ldots, N\}, \quad \tau := \sum_{i \in \mathcal{U}} y_i$$

draw n times uniformly from $\mathcal{U}$ **without replacement** to obtain $S = (i_1, \ldots, i_n) \in \mathcal{U}^n$

For $i \in \mathcal{U}$, define $Z_i := \sum_{j=1}^n \mathbb{1}_{\{i_j\}}(i) \in \{0, 1\}$. We have

$$P(Z_i = 1) = 1 - \frac{\#\text{ samples of size } n \text{ not containing individual } i}{\#\text{ samples of size } n}$$

$$= 1 - \frac{\cancel{(N-1)}\cancel{(N-2)} \cdots \cancel{\cdots}(N-n)}{N\cancel{(N-1)} \cancel{\cdots} \cancel{(N-n+1)}} = \frac{n}{N}$$

$$\Rightarrow \quad \pi_i = \mathbb{E}[Z_i] = P(Z_i = 1) = \frac{n}{N} \text{ and } \pi_{ii} = \mathbb{E}[Z_i^2] = \mathbb{E}[Z_i] = \pi_i.$$

For $i \neq j$, we have

$$
\begin{aligned}
\pi_{ij} &= \mathbb{E}[Z_i Z_j] \overset{!}{=} \mathbb{E}[Z_i + Z_j - 1 + \overbrace{(1 - Z_i)(1 - Z_j)}^{1 - z_i - z_j + z_i z_j}] \\
&= P(Z_i = 1) + P(Z_j = 1) - 1 + P(Z_i = 0 = Z_j) \\
&= 2\frac{n}{N} - 1 + \frac{(N-2)\cdots(N-n-1)}{N \cdot (N-1)\cdots(N-n+1)} \\
&= \frac{2n(N-1) - N(N-1) + (N-n)(N-1-n)}{N(N-1)} \\
&= \frac{2n(N-1) - N(N-1) + N(N-1) - n(N-1) - Nn + n^2}{N(N-1)} \\
&= \frac{n(N-1) - Nn + n^2}{N(N-1)} = \frac{n(n-1)}{N(N-1)}.
\end{aligned}
$$

Sampling from finite population

$$y_i \in \mathbb{R}, i \in \mathcal{U} = \{1, \ldots, N\}, \quad \tau := \sum_{i \in \mathcal{U}} y_i, \quad \mu := \frac{\tau}{N}$$

random sample $\quad S = (i_1, \ldots, i_n) \in \mathcal{U}^n$

For $i \in \mathcal{U}$, define

$$Z_i := \sum_{j=1}^{n} \mathbb{1}_{\{i_j\}}(i) \qquad \text{number of times individual } i \text{ is sampled}$$

and $\pi_i := \mathbb{E}[Z_i]$, $\pi_{ij} := \mathbb{E}[Z_i Z_j]$.

If $\pi_i > 0$, define

$$\hat{\tau} := \sum_{j=1}^{n} \frac{y_{i_j}}{\pi_{i_j}}.$$

Horvitz-Thompson estimate

Graph sampling designs

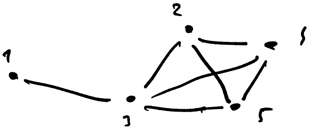
Graph characteristics as totals

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

We want to estimate a population total, for example:

- $\mathcal{U} := V$, $(y_u)_{u \in \mathcal{U}}$, $\tau = \sum_{u \in \mathcal{U}} y_u$ $\mu = \frac{\tau}{N_V}$.
    - vertex characteristics, e.g., y_i is gender, age, etc.
    - degree $y_i = d_i \Rightarrow \tau = 2N_E$
- $\mathcal{U} := V^{(2)}$, $(y_u)_{u \in \mathcal{U}}$, $\tau = \sum_{u \in \mathcal{U}} y_u$.
    - $y_u = y_{\{i,j\}}$ is the proportion of shortest paths between i and j passing through a given vertex $k \in V$ and $y_{\{i,j\}} = 0$ if $k \in \{i, j\} \Rightarrow \tau = c_B(k)$
    - edge characteristics/weights, e.g., number of phone calls between two phone numbers $\Rightarrow \tau$ is the total number of phone calls
    - $y_{\{i,j\}} = \mathbb{1}_E(\{i,j\}) \Rightarrow \tau = \sum_{e \in E} 1 = N_E$ ($\mathcal{U} = E$)
    - $y_{\{i,j\}} = \mathbb{1}_E(\{i,j\}) \mathbb{1}_{y_i = y_j}$ (e.g., y_i gender) $\Rightarrow \tau =$ number of same sex friendships ($\mathcal{U} = E$)

Graph characteristics as totals



We want to estimate a population total, for example:

- Number of connected triangles in the graph:

$$\mathcal{U} = V^{(3)} := \{ \{i, j, l\} : i \neq j,\ j \neq l,\ i \neq l \}$$

$$y_u = \mathbb{1}_E(\{i,j\}) \cdot \mathbb{1}_E(\{j,l\}) \cdot \mathbb{1}_E(\{i,l\})$$

$$= \begin{cases} 1 & \text{if triangle} \\ 0 & \text{if not} \end{cases}$$

$$\{2, 3, 5\} = \{3, 2, 5\} = \{5, 3, 2\} = \ldots$$

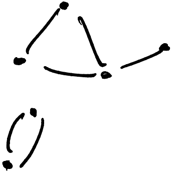
Graph sampling and estimation

- Population graph $\quad G = (V, E)$
- 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, $u \in \mathcal{U}$.