Day-3 Session-1: On to General Theory of Graph Sampling

Li-Chun Zhang[1,2,3] **and** ***Melike Oguz-Alper***[2]

[1] ***University of Southampton (L.Zhang@soton.ac.uk)***
[2] ***Statistisk sentralbyraa, Norway***
[3] ***Universitetet i Oslo***

Graph, valued graph

$G = (U, A)$: order $|U| = N$, size $|A| = R$
Attaching values to U or A yields a *valued graph*
A graph is the *structure* of the corresponding valued graph

By default a graph is *directed*, or a *digraph*
A_{ij} = edges from i to j, $a_{ij} = |A_{ij}|$ and $A = \bigcup_{i,j \in U} A_{ij}$
simple graph if $|A_{ij}| \equiv 1$, *multigraph* otherwise
Out-edges $A_{i+} = \bigcup_{j \in U} A_{ij}$ and out-degree $a_{i+} = |A_{i+}|$
In-edges $A_{+i} = \bigcup_{j \in U} A_{ji}$ and in-degree $a_{+i} = |A_{+i}|$

Undirected graph if $A_{ij} \equiv A_{ji}$ & degree $d_i = a_{i+} = a_{+i}$
i.e. no distinction between out-edges and in-edges

Graph, valued graph

Edge $(ij) \in A$ is *incident* to nodes i and j, and vice versa
Two nodes $i, j \in U$ are *adjacent* if there exists at least one edge between them, or $a_{ij} + a_{ji} \geq 1$
NB. distinguish between $(ij) \in A$ and $(i, j) \in U$

$\alpha_i = \{j \in U : a_{ij} > 0\}$ are the *successors* of i
$\beta_i = \{j \in U : a_{ji} > 0\}$ are the *predecessors* of i
$\alpha_i \equiv \beta_i$ for undirected graphs
$a_{i+} = |\alpha_i|$ and $a_{+i} = |\beta_i|$ for simple graphs

An edge incident to the same node at both ends is a *loop* — whether or not loops are included in the definitions and notations above will be a matter of convention

Motif

$G(M) =$ subgraph *induced* by M, with node set $M \subset U$ and edge set $A \cap (M \times M) = \{A_{ij} : i, j \in M\}$

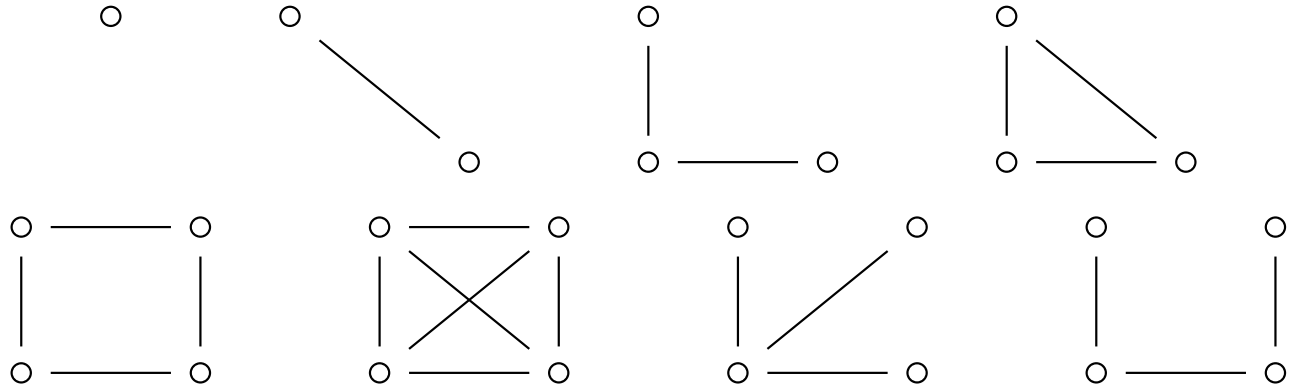
The specific characteristics of $G(M)$ is called *motif*
Denote motif by $[M]$, the *order* of $[M]$ is $|M|$

Induced motif $[M]$ if it can be determined based on the induced subgraph $G(M) = \big(M, A \cap (M \times M)\big)$ alone

- $[i, j : a_{ij}a_{ji} = 1]$ of $G(\{i, j\})$: node pair with 'mutual simple relationship', induced motif
- $[i : a_{i+} = 3]$ of $G(\{i\})$: node with out-degree 3, not induced as $A \cap (\{i\} \times U)$ needed beyond $A \cap (\{i\} \times \{i\})$

Motif

Some low-order induced motifs in undirected graphs:



Node ($\mathcal{K}_1$), 2-clique ($\mathcal{K}_2$), 2-star ($\mathcal{S}_2$), 3-clique (triangle, $\mathcal{K}_3$), 4-cycle ($\mathcal{C}_4$), 4-clique ($\mathcal{K}_4$), 3-star ($\mathcal{S}_3$) and 3-path ($\mathcal{P}_3$)

A *component* is an subgraph of connected nodes that are unconnected to the rest of graph – not induced motif

A *shortest path* over ordered M is not an induced motif, if it depends also on $A \setminus (M \times M)$ beyond $A \cap (M \times M)$ e.g. M of $\mathcal{S}_2$ vs. M of $\mathcal{K}_3$ above

Graph total, parameter

Let $y(M)$ be a function of valued subgraph $G(M)$
Let Ω contain all the relevant node sets M
Graph total of $y(M)$ over Ω: $\theta = \sum_{M \in \Omega} y(M)$
θ is of the *q-th order* if $|M| \equiv q$ for any $M \in \Omega$

Graph parameter μ: any function of $\{y(M) : M \in \Omega\}$
E.g. let θ_3 be the no. triangles in G and θ_2 that of 2-stars, then $\mu = \frac{\theta_3}{\theta_2 + \theta_3}$ is a 3rd-order graph parameter

Let $\mathfrak{m}$ be the motif (e.g. triangle) and each $\kappa \in \Omega$ an *occurrence* of $\mathfrak{m}$ in G. A corresponding graph total is

$$\theta = \sum_{\kappa \in \Omega} y_\kappa$$

Observation procedure (OP)

Given an *initial sample* $s_0 \subset U$ (of nodes), an OP is *induced* when A_{ij} is observed iff both $i \in s_0$ and $j \in s_0$, or *incident-reciprocal* when A_{ij} and A_{ji} are both observed if either $i \in s_0$ or $j \in s_0$. For digraphs, an OP is incident *forward* when A_{ij} is observed if $i \in s_0$, or *backward* when A_{ij} is observed if $j \in s_0$.

G : a $b \longrightarrow c \longrightarrow d$

Basic OP applied to $s_0 = \{b, d\}$

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & \boxed{0} \\ 0 & 0 & 0 & 1 \\ 0 & \boxed{0} & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 0 & 0 & 0 \\ \boxed{0} & \boxed{0} & \boxed{1} & \boxed{0} \\ 0 & 0 & 0 & 1 \\ \boxed{0} & \boxed{0} & \boxed{0} & \boxed{0} \end{bmatrix} \quad \begin{bmatrix} 0 & \boxed{0} & 0 & \boxed{0} \\ 0 & \boxed{0} & 1 & \boxed{0} \\ 0 & \boxed{0} & 0 & \boxed{1} \\ 0 & \boxed{0} & 0 & \boxed{0} \end{bmatrix} \quad \begin{bmatrix} 0 & \boxed{0} & 0 & \boxed{0} \\ \boxed{0} & \boxed{0} & \boxed{1} & \boxed{0} \\ 0 & \boxed{0} & 0 & \boxed{1} \\ \boxed{0} & \boxed{0} & \boxed{0} & \boxed{0} \end{bmatrix}$$

Induced Inc. forward Inc. backward Inc. reciprocal

As a convention, specify the observed edges as

$$A \cap s_{\text{ref}}$$

via a *reference set* s_{ref}, which explicates the parts of the *adjacency matrix* $[a_{ij}]$ which are observed given s_0 and the OP

Observation procedure (OP)

Multiwave: OP can be repeated *wave by wave*, such as in Network sampling; OP can be adaptive in addition, such as in ACS
Let s be the *seed sample*, $s_0 \subseteq s$, to which the OP is applied. E.g.

$$s_{\text{ref}} = s \times U \quad \text{if incident forward OP in digraphs}$$

Multistage: OP is multistage given different types of nodes
In a *multipartite* graph, node set U is partitioned into non-overlapping subsets, where edges only exist between nodes in different parts but not those in the same part, e.g. $\mathcal{B} = (F, \Omega; H)$

Multilayer: OP is multilayer given different types of edges
In a *multilayer* graph, edges are of different types (or dimensions), where an edge can exist between any pair of nodes in U

E.g. clinics and patients: two types nodes or two types of edges?
OP can be multilayer and does not involve different types of nodes
E.g. U = individuals, A = connections of kins or colleagues

Graph sampling (Zhang and Patone, 2017; Zhang, 2021)

Using $G = (U, A)$ instead of U, one may be interested in

- a finite population total (or mean), where the edges A provide more effectively access to the target population (part of U);
- or the structure of the edges, as in terms of motifs by Goodman (1961), Frank (1971, 1977, 1978, 1979, 1980, 1981, 2011)

Either way, graph sampling is a statistical approach to study real graphs. Three key elements:

I. Definition of sample graph, as a subgraph from given G

II. Basis of inference, can be different kinds of sampling probabilities associated with the given graph sampling method

III. Eligible sample motifs: a motif observed in the sample graph is ineligible for estimation, if the required probabilities for inference cannot be calculated. Identification of eligible sample motifs is the key to any feasible graph sampling strategy.

Sample graph

Sampling from $G = (U, A)$ has generally two parts:

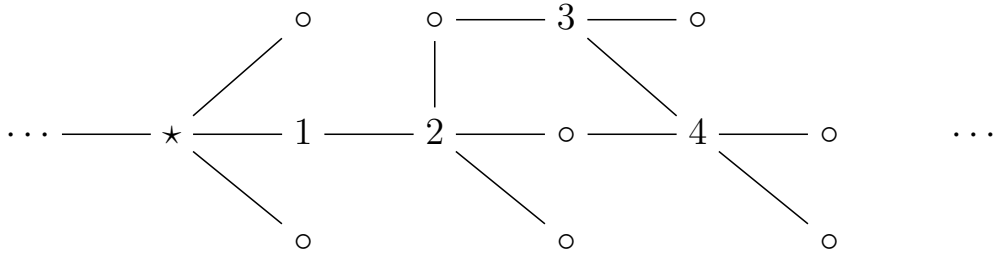
- apply a given design to select an initial sample of nodes, $s_0 \subset U$;
- apply a specified OP that makes use of the edges in A.

Sample graph $G_s = (U_s, A_s)$ with seed sample s, where

$$A_s = A \cap s_{\text{ref}} \quad \text{and} \quad U_s = s \cup \text{Inc}(A_s)$$

NB. Graph sampling can depend on values associated with G, and values associated with G_s are also observed

NB. Zhang and Patone (2017) consider OP based on adjacent nodes; definition here allows for non-adjacent nodes, e.g. walk 1, 2, 3, 4:



Basis of inference

HTE of θ based on *sample graph* inclusion probability:

$$\hat{\theta} = \sum_{M \in \Omega} y(M)\delta_{[M]}/\pi_{(M)}$$

where $\delta_{[M]} = 1$ if the motif $[M]$ is observed in the sample graph G_s and 0 otherwise, and $\pi_{(M)} = \Pr(\delta_{[M]} = 1)$
In particular, any induced motif $[M]$ is observed in G_s iff

$$M \times M \subseteq s_{\text{ref}}$$

Basis of inference other than $\pi_{(M)}$, e.g.

- modified $\pi^*_{(i)} = \pi_i$ for terminal node i under ACS (Thompson, 1990)
- stationary probabilities under random walk (e.g. Thompson, 2006a)
- selection probability of the ordered sample under Adaptive Web Sampling (Thompson, 2006b)

Graph sampling strategy

Given sample graph G_s and basis of inference, an observed motif in G_s is *eligible* for estimation or inference iff the required sampling probabilities can be calculated.
Examples of ineligible motifs and feasible strategies:

- indirect sampling with F = clinics and Ω = patients:
  if unknown out-of-sample clinics of sample patients
  Strategy: cannot use modified HTE or restricted β^*_κ...
  however, can use sample-dependent β'_κ
- ACS with $F = \Omega = U$: terminal nodes $\bigcirc$ or $\circ$
  Strategy: HTE*, restricted β^*_κ, sample-dependent $\beta^\dagger_\kappa$
- LIS with F' consisting of detectability partitions
  Strategy: sampling probabilities w.r.t. $\mathcal{B}' = (F', \Omega; H')$
  approximated by those w.r.t. $\mathcal{B}^* = (F^*, \Omega_s; H^*)$

[1] Frank, O. (1971). *Statistical inference in graphs*. Stockholm: Försvarets forskningsanstalt.

[2] Frank, O. (1977). Survey sampling in graphs. *Journal of Statistical Planning and Inference*, 1(3):235–264.

[3] Frank, O. (1978). Estimation of the number of connected components in a graph by using a sampled subgraph. *Scandinavian Journal of Statistics*, 5:177–188.

[4] Frank, O. (1979). Sampling and estimation in large social networks. *Social networks*, 1(1):91–101.

[5] Frank, O. (1980). Sampling and inference in a population graph. *International Statistical Review/Revue Internationale de Statistique*, 48:33–41.

[6] Frank, O. (1981). A survey of statistical methods for graph analysis. *Sociological methodology*, 12:110–155.

[7] Frank, O. (2011). Survey sampling in networks. *The SAGE Handbook of Social Network Analysis*, pages 389–403.

[8] Goodman, L.A. (1961). Snowball sampling. *Annals of Mathematical Statistics*, 32:148–170.

[9] Thompson, S.K. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association*, 85:1050–1059.

[10] Thompson, S.K. (2006a). Targeted random walk designs. *Survey Methodology*, 32, 11–24.

[11] Thompson, S.K. (2006b). Adaptive Web Sampling. *Biometrics*, 62, 1224–1234.

[12] Zhang, L.-C. (2021). Graph sampling: An introduction. *The Survey Statistician*, 83:27-37.

[13] Zhang, L.-C. and Patone, M. (2017). Graph sampling. *Metron*, **75**, 277-299. `DOI:10.1007/s40300-017-0126-y`