Sampling and design-based inference in finite networks

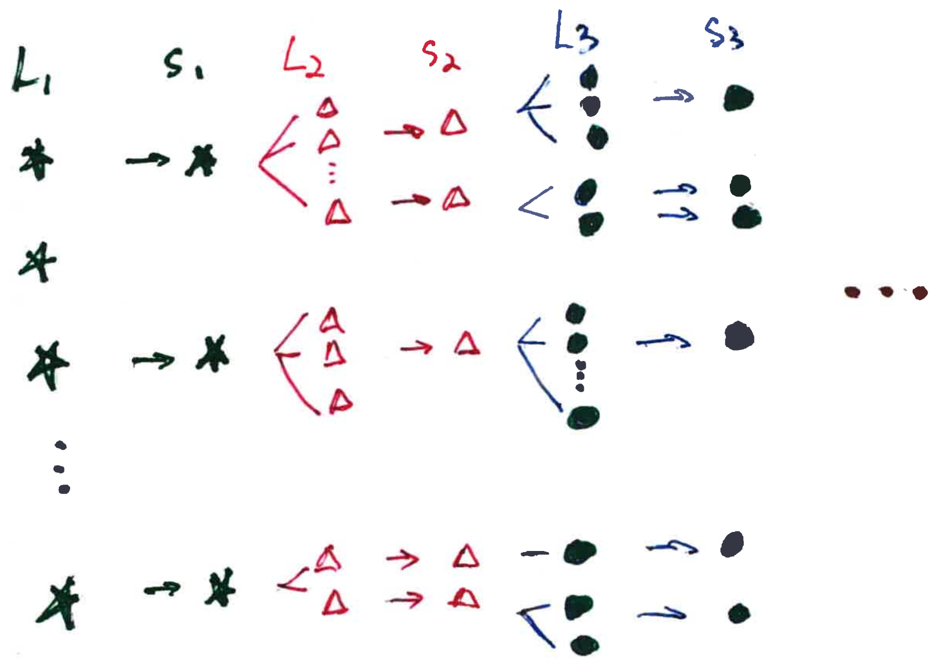
Li-Chun Zhang[1,2,3] and ***Melike Oguz-Alper***[2]

[1]*University of Southampton (L.Zhang@soton.ac.uk)*

[2]*Statistisk sentralbyraa, Norway*

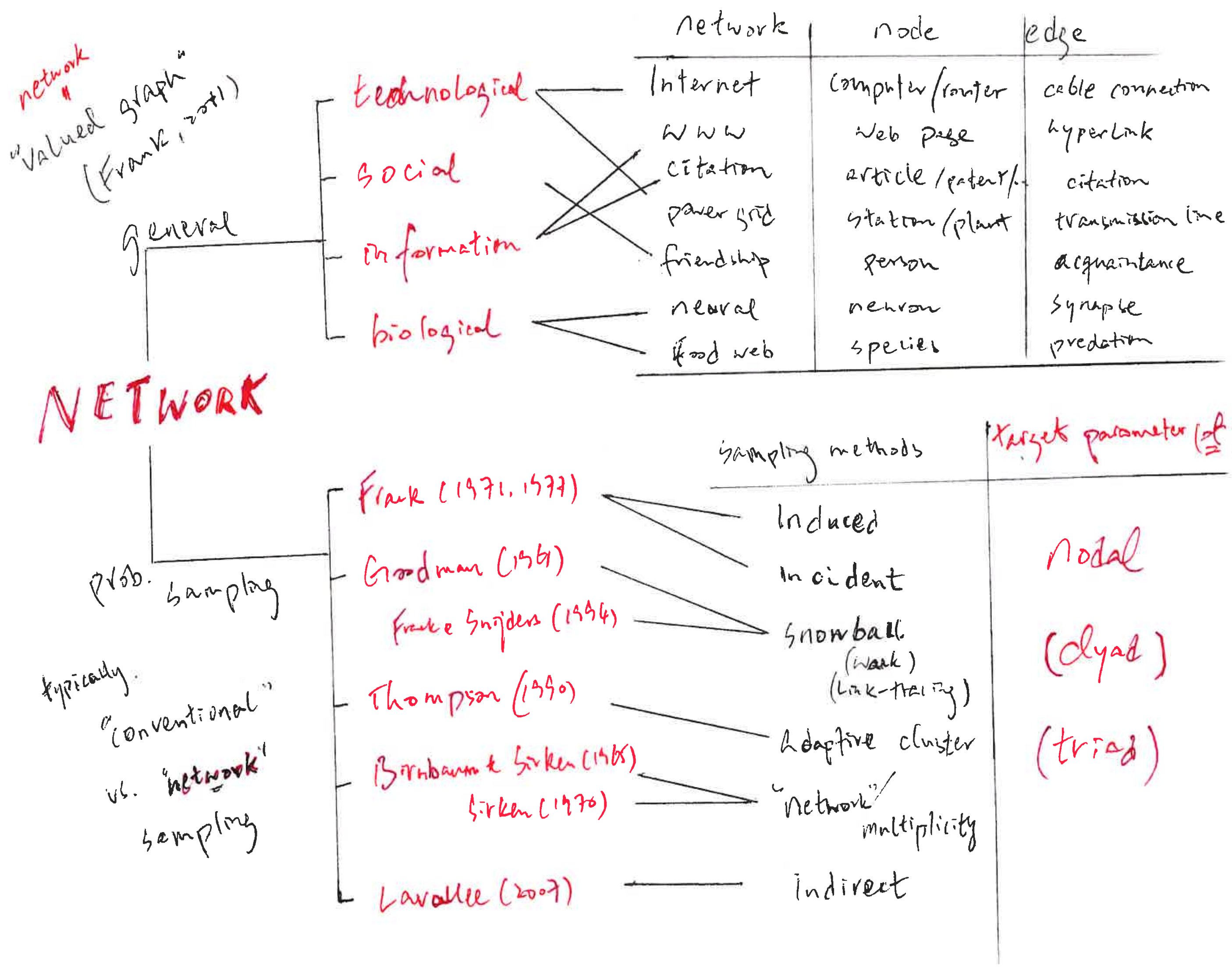
[3]*Universitetet i Oslo*

List-based multistage sampling:



NB. a special case of connections among units

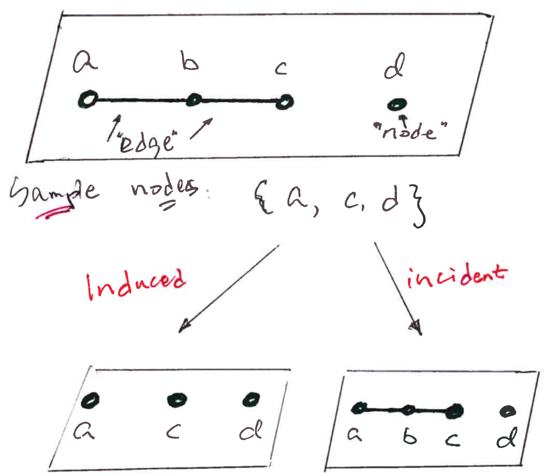
"Network" & unconventional sampling



A unified definition of Graph Sampling

Key features (Zhang and Patone, 2017)

- initial sample of nodes & observation procedure by edges



- **sample graph** defined in terms of edges included

NB. duality of incident relationship between edge and node

A unified definition of Graph Sampling

Graph: $G = (\mathcal{N}, A) =$ (Nodes, edges) [digraph by default]

Initial sample of nodes: $s_1 \subset \mathcal{N}$ [$p(s_1)$, π_i, π_{ij}, etc.]

Observation procedure: e.g.

- induced, incident (forward, backward, reciprocal), *ancestral*
- snowball propagation by same procedure or adaptive

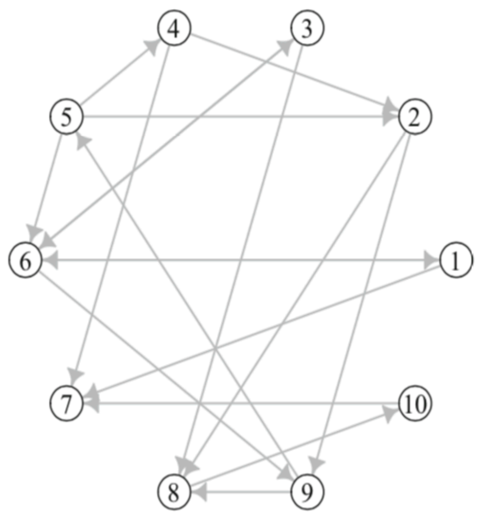
Included edges $A_s = A(s_2)$: *reference set* $s_2 \subseteq \mathcal{N} \times \mathcal{N}$

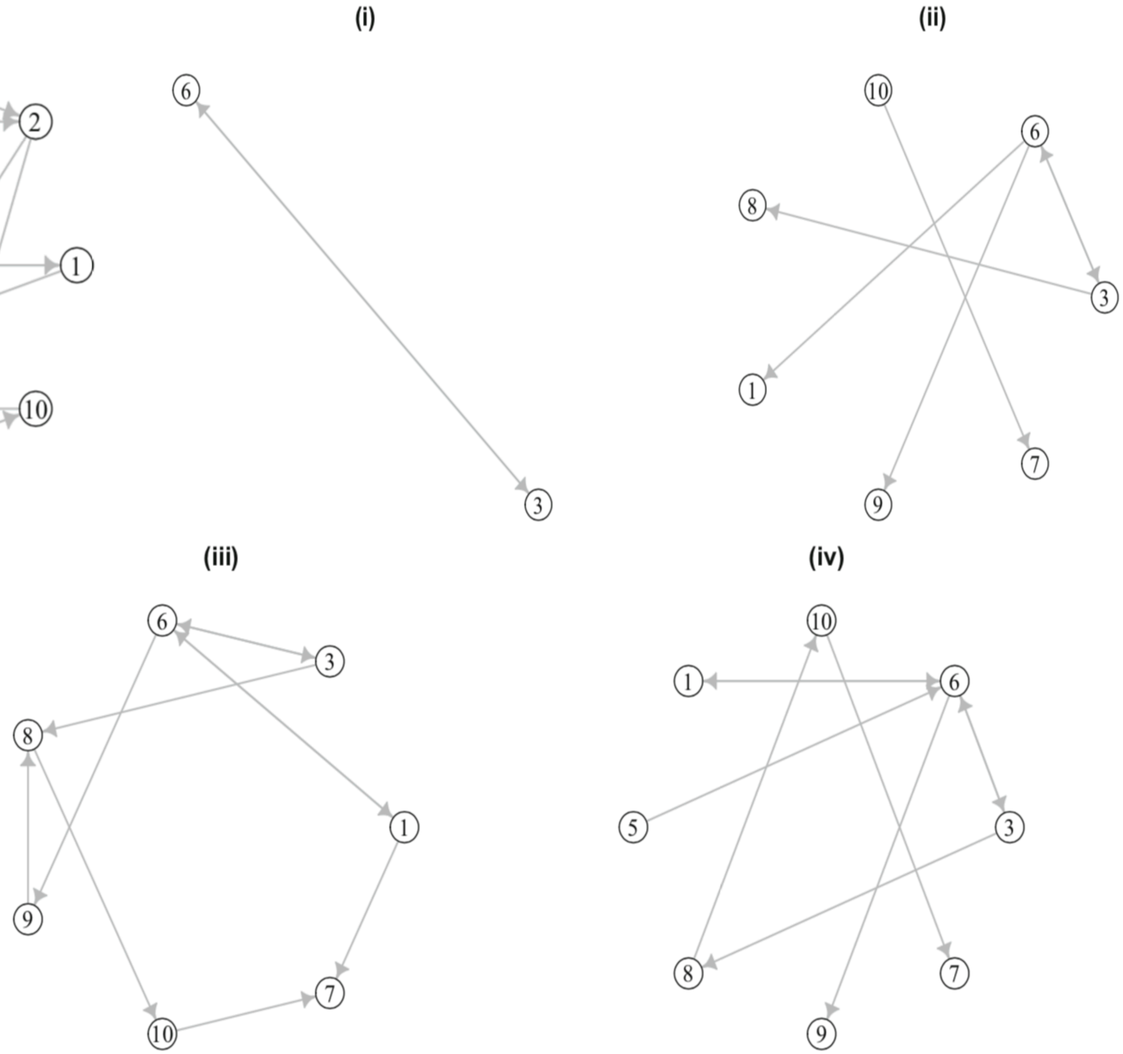
e.g. induced $s_2 = s_1 \times s_1$, inc. reciprocal $s_2 = s_1 \times \mathcal{N} \cup \mathcal{N} \times s_1$

Included nodes: $\mathcal{N}_s = s_1 \cup \mathrm{Inc}(A_s)$

Sample Graph: $G_s = (\mathcal{N}_s, A_s)$

Illustration: G and $s_1 = \{3, 6, 10\}$, $s_a = s_1 \cup \alpha(s_1)$





(iii) $s_1 = s_a \times s_a$

T-stage snowball sampling

Initial *seeds*: $s_{1,0} \subset \mathcal{N}$ with successors $\alpha(s_{1,0})$

- 1st-wave sample: $s_{1,1} = \alpha(s_{1,0}) \setminus s_{1,0}$ [seeds for 2nd-wave]
- 2nd-wave sample: $s_{1,2} = \alpha(s_{1,1}) \setminus (s_{1,0} \cup s_{1,1})$
- ... [if $s_{1,t} = \emptyset$, set $s_{1,t+1} = \cdots = s_{1,T} = \emptyset$]
- T-th stage sample: $s_{1,T} = \alpha(s_{1,T-1}) \setminus \big(\bigcup_{h=0}^{T-1} s_{1,h} \big)$

Sample of seeds: $s_1 = \bigcup_{t=0}^{T-1} s_{1,t}$

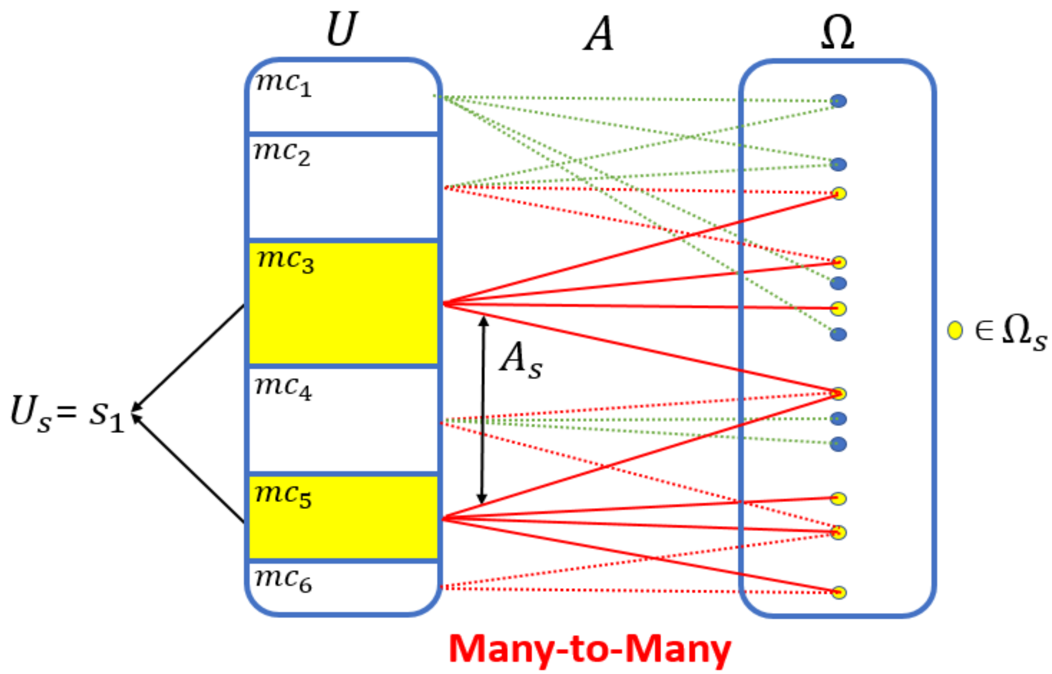
I. $s_2 = s_1 \times \mathcal{N} \mapsto A_s = \bigcup_{i \in s_1} \bigcup_{j \in \alpha_i} A_{ij}$

II. $s_2 = s_1 \times \mathcal{N} \cup \mathcal{N} \times s_1 \mapsto A_s = \bigcup_{i \in s_1} \bigcup_{j \in \alpha_i} (A_{ij} \cup A_{ji})$

Node sample: $\mathcal{N}_s = s_1 \cup \alpha(s_1)$

Birnbaum & Sirken (1965): Multiplicity sampling

Example: s_1 of medical centres (U), access to patients (Ω)



BIG: bipartite incidence graph $G = (U, \Omega; A)$

- bipartition (U, Ω) of $\mathcal{N}$, edges only between U and Ω
- e.g. $(U, \Omega) =$ (parents, children) in Lavalleè (2007)

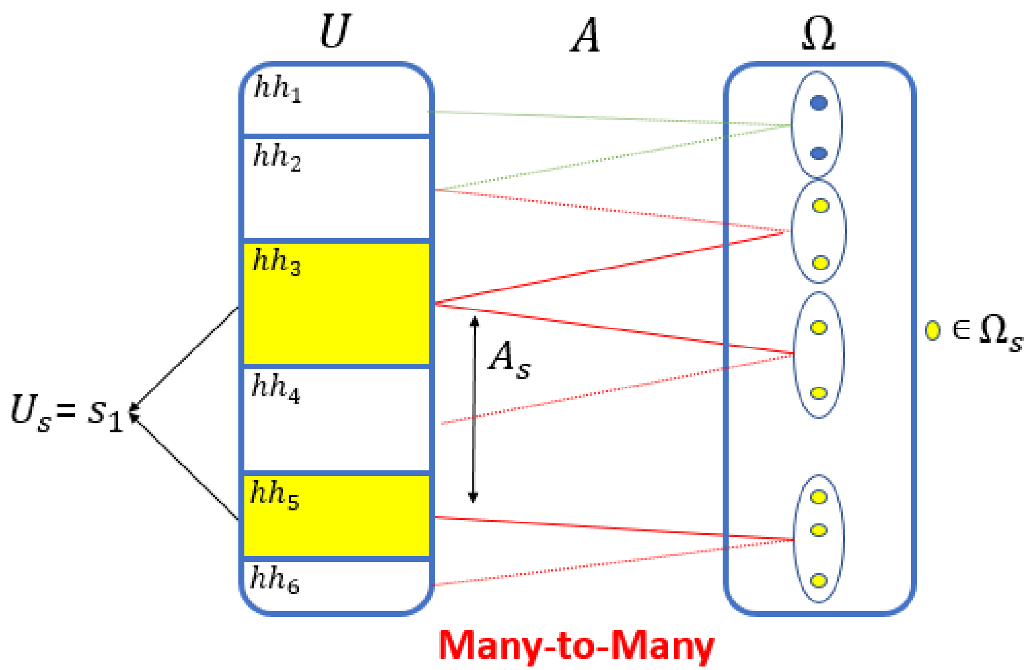
Sirken (2005): Network sampling

Example: s_1 of household (U), access to siblings (Ω)

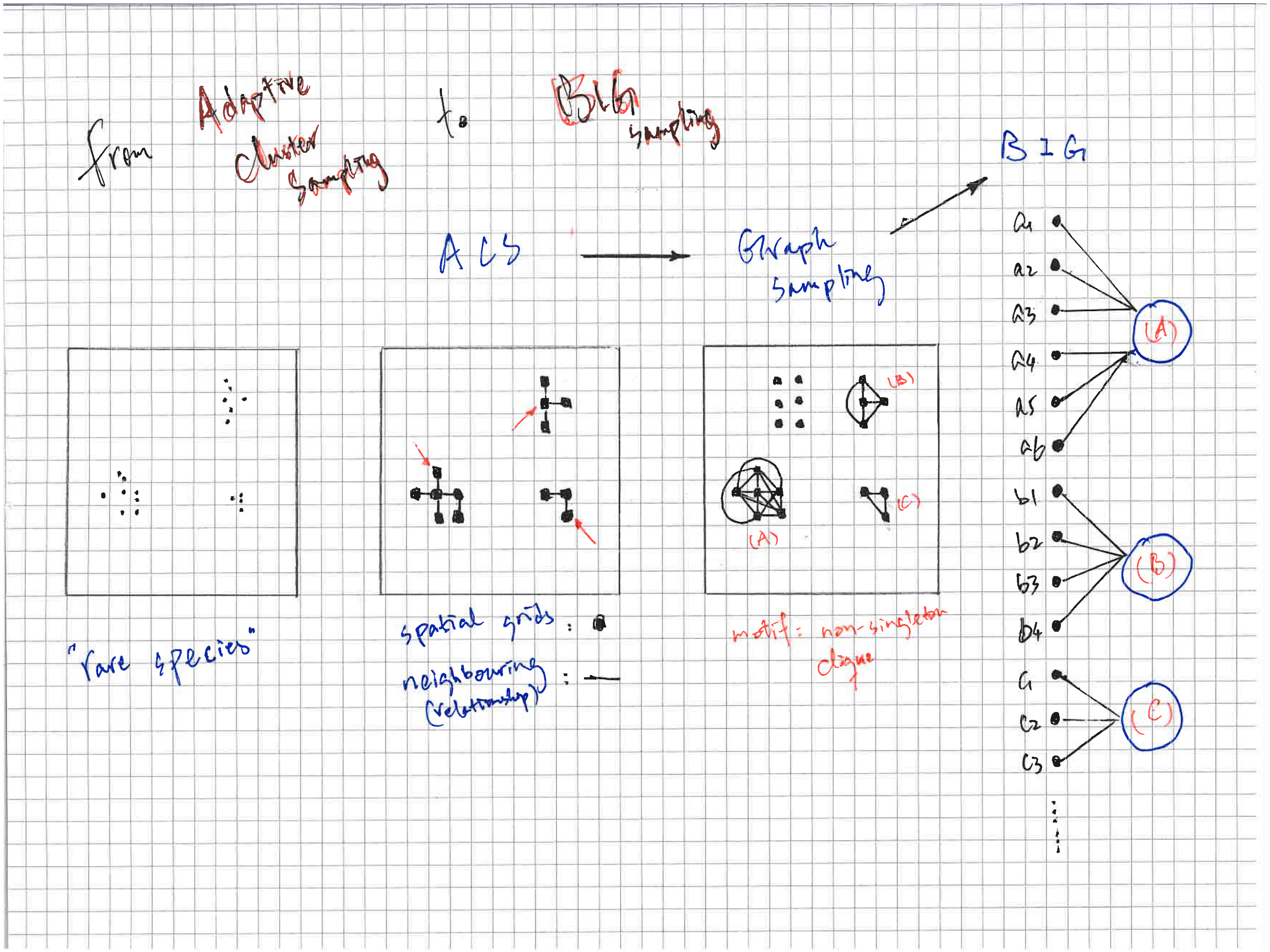
E.g. sampling in *projection-relation graph*:

- projection edges from U to P (persons): $\mathcal{N} = U \cup P$
- relation edges $a_{ij} = a_{ji}$ for $i, j \in P$ if i and j are siblings

Can use BIG with $\mathcal{N} = U \cup \Omega$ [*hypernode* $k \in \Omega$]



Thompson (1990): Adaptive cluster sampling (ACS)



BIG sampling

Any representation of *sampling* in finite graph/network

- e.g. multiplicity/indirect sampling, "network" sampling, ACS
- e.g. induced, incident, snowball sampling (Frank 1971, ..., 2011)

BIG representation $G = (U, \Omega; A)$ for *estimation*

- sampling units U, measurement ***motifs*** Ω, incidence edges A
- ***ancestral*** observation for $\boxed{\text{design-based inference}}$: need to know all the nodes in U that could lead to the observed motifs in Ω_s

  NB. generalise the notion "multiplicity" (Birnbaum & Sirken, 1965)
- solution: use $s_2^* = s_1 \times s_1$ under T-stage snowball sampling

$\mathcal{C}_q$ = the set of all M of order q, $M \subset \mathcal{N}$ and $|M| = q$

Zhang & Patone (2017) define q-th order ***graph total***

$$\theta = \sum_{M \in \mathcal{C}_q} y(M)$$

Graph parameter = a function of graph totals

[Similarly for network totals and network parameters]

Motif: a node set M of specific characteristics, $M \subseteq \mathcal{N}$

NB. a motif $[M]$ may or may not have a fixed order, giving rise to graph totals with or without a given order

e.g. graph order $|\mathcal{N}|$: 1st-order, graph size $|A|$: 2nd-order

e.g. $[M]$ = connected components, without fixed order

Example: Triads, i.e. $|M| = 3$

The no. triads of size 3, 2, 1, respectively, in undirected simple graph:

$$\theta_{3,3} = \sum_{M \in \mathcal{C}_3} a_{ij} a_{jh} a_{ih} \qquad [M = \{i, j, h\}]$$

$$\theta_{3,2} = \sum_{M \in \mathcal{C}_3} a_{ij} a_{ih}(1 - a_{jh}) + a_{ij} a_{jh}(1 - a_{ih}) + a_{ih} a_{jh}(1 - a_{ij})$$

$$\theta_{3,1} = \sum_{M \in \mathcal{C}_3} a_{ij}(1 - a_{jh})(1 - a_{ih}) + a_{ih}(1 - a_{ij})(1 - a_{jh}) + a_{jh}(1 - a_{ij})(1 - a_{ih})$$

Relationship to the mean and variance of degrees (Frank, 1981):

$$\mu = \sum_{d=1}^{N} \frac{N_d}{N} d = \frac{2R}{N} \qquad Q = \sum_{d=1}^{N} d^2 N_d \qquad \sigma^2 = \frac{Q}{N} - \mu^2$$

$$R = \frac{1}{N-2}\big(\theta_{3,1} + 2\theta_{3,2} + 3\theta_{3,3}\big)$$

$$Q = \frac{2}{N-1}\big(\theta_{3,1} + N\theta_{3,2} + 3(N-1)\theta_{3,3}\big)$$

Two network HT-estimators

BIG sampling: Ω = population set of $[M]$, Ω_s = sample set of $[M]$

For convenience: enumerate the motifs as $k = 1, 2, ...$ in Ω and Ω_s

Yhat: HT-estimator of graph total $\theta = \sum_{k\in\Omega} y_k$

$$\hat{\theta}_y = \sum_{k\in\Omega} \delta_k y_k / \pi_{(k)}$$

δ_k = inclusion indicator and $\pi_{(k)}$ = inclusion probability of motif

NB. $\pi_{(k)}$ for distinction to inclusion probability π_j of unit $j \in U$

NB. Under T-stage snowball sampling, a motif $[M]$ is observed

if $M \subseteq s_1$, where $M = \{i_1, ..., i_q\}$

or if $M_{(h)} \subseteq s_1$, where $M_{(h)} = M \setminus \{i_h\}$ and $1 \leq h \leq q$

(Zhang and Patone, 2017)

Two network HT-estimators

Zhang and Patone (2017) show that

$$\pi_{(k)} = \sum_{h=1}^{q} \Pr\Big(M_{(h)} \subseteq s_1\Big) - (k-1)\Pr\Big(M \subseteq s_1\Big)$$

where e.g. $\Pr\big(M \subseteq s_1\big) = \pi_{(i_1)(i_2)\cdots(i_q)}$ is joint inclusion probability

In terms of inclusion prob. in initial seed sample $s_{1,0}$, we have

$$\pi_{(i_1)(i_2)\cdots(i_q)} = \sum_{L \subseteq M} (-1)^{|L|} \bar{\pi}(L),$$

where $\bar{\pi}(L)$ is the (exclusion) probability of $L \cap s_1 = \emptyset$:

$$\bar{\pi}(L) = \Pr(R_L \cap s_{1,0} = \emptyset) = \bar{\pi}_{R_L} = \sum_{D \subseteq R_L} (-1)^{|D|} \pi_D$$

where $R_L = \bigcup_{i \in L} R_i$ and R_i is the ancestors of i up to the $T-1$ steps, and π_D is joint inclusion probability of the nodes (in D) in $s_{1,0}$

Two network HT-estimators

Birnbaum and Sirken (1965): provided $\sum_{i \in U} P_{ik} = 1$, $\forall k \in \Omega$,

$$\theta = \sum_{k \in \Omega} y_k = \sum_{k \in \Omega} \big(\sum_{i \in U} P_{ik} \big) y_k = \sum_{i \in U} \big(\sum_{k \in \Omega} P_{ik} y_k \big) = \sum_{i \in U} z_i$$

Zhat based on $\boxed{z_i = \sum_{k \in \Omega} P_{ik} y_k}$ with P_{ik}'s constant of s_1:

$$\hat{\theta}_z = \sum_{i \in s_1} z_i / \pi_i = \sum_{i \in U} z_i \delta_i / \pi_i$$

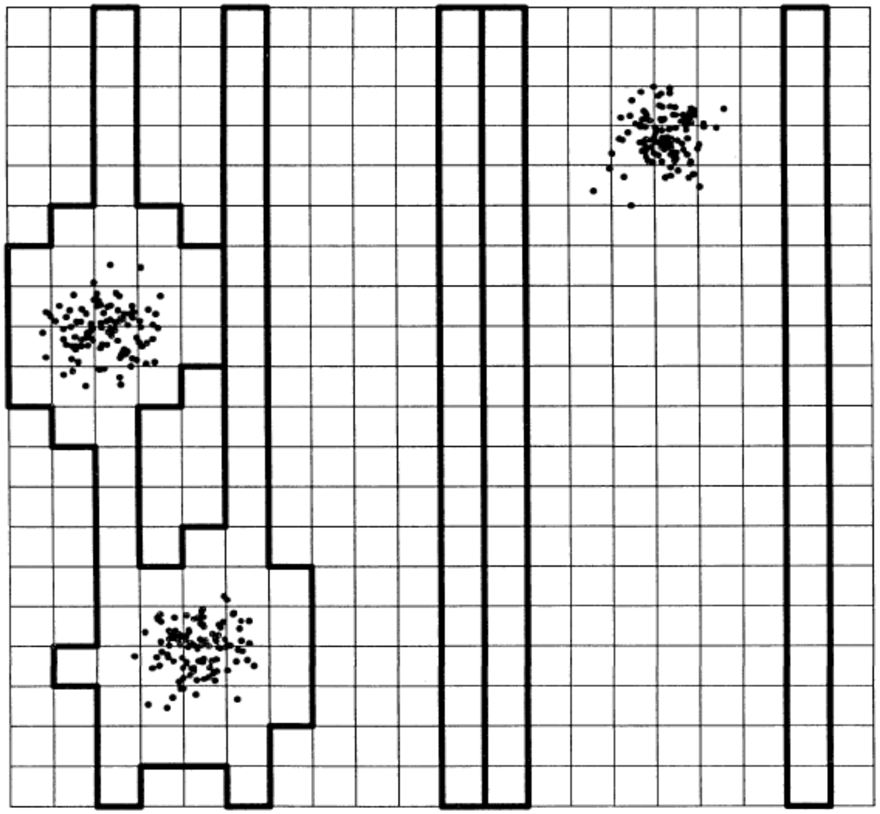
NB. *Equal-share weight*, given multiplicity $m_k = |A_{+k}|$ in BIG:

$$P_{ik} = m_k^{-1} \quad \text{if } |A_{ik}| > 0, \qquad P_{ik} = 0 \quad \text{otherwise}$$

NB. *pps-share weight*: $P_{ik} \propto \pi_i$ if $|A_{ik}| > 0$, $P_{ik} = 0$ otherwise

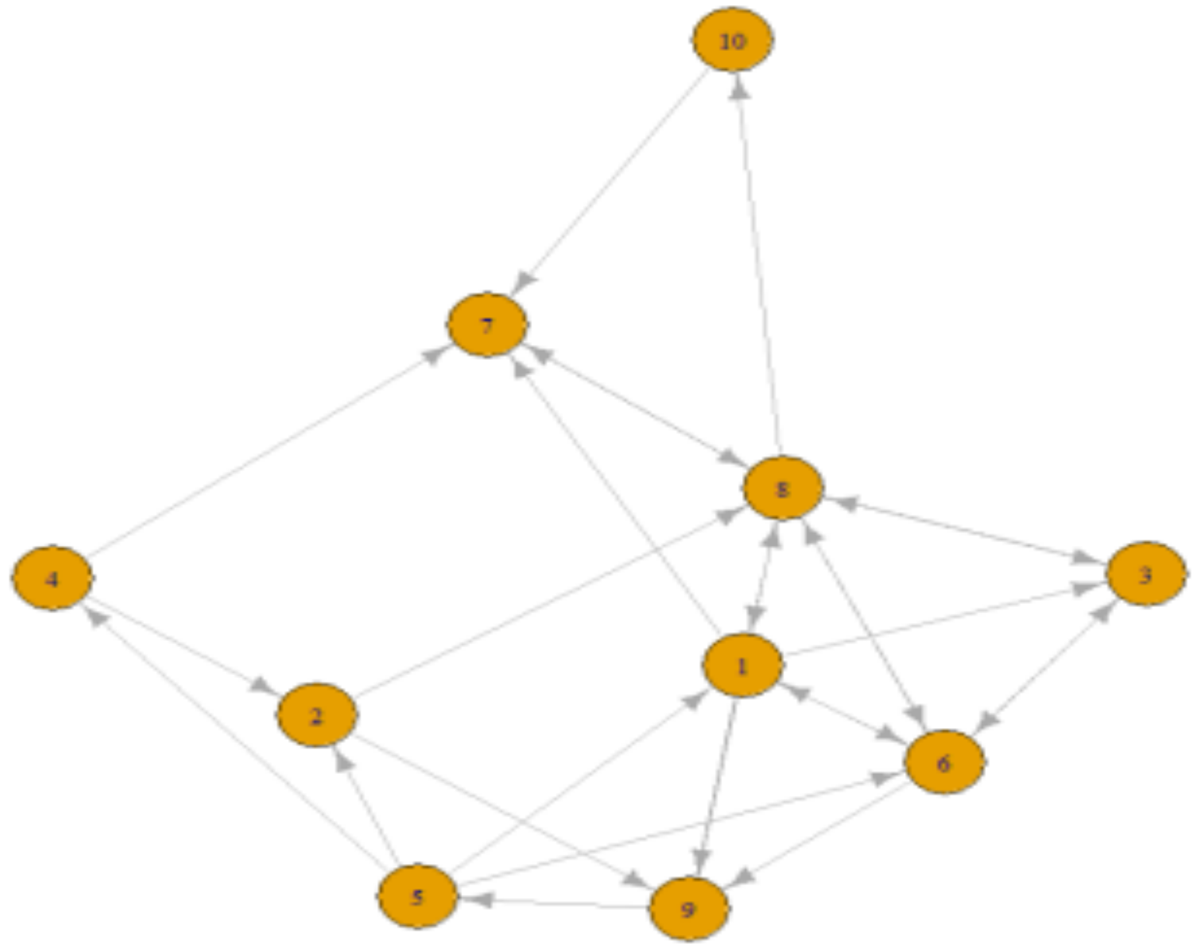
NB. $\hat{\theta}_z$ much easier to calculate than $\hat{\theta}_y$ provided m_k

Example (Thompson, 1991): Two-stage ACS



	RRMSE (%)		
$\|s_1\|$	$\hat{\theta}_{SCS}$	$\hat{\theta}_z^{eq}$	$\hat{\theta}_y$
1	143.9	*112.1*	*112.1*
2	96.8	75.4	72.5
4	64.4	50.1	43.6
6	49.1	38.3	29.1
10	32.2	25.1	12.3

An example of graph sampling: SRS of s_1, $|s_1| = 3$



An example of graph sampling: SRS of s_1, $|s_1| = 3$

```
# --------------------------------------------------------------
# Triad types in a directed graph (Davis & Leinhardt, 1972)
# --------------------------------------------------------------
g1     003 A,B,C                    empty graph
g2     012 A--+B,C                  graph with a single directed age
g3     102 A+--+B,C                 graph with a mutual connection between two vertices
g4    021D A+--B--+C                out-star
g5    021U A--+B+--C                in-star
g6    021C A--+B--+C                triple, directed line
g7    111D A+--+B+--C               triple
g8    111U A+--+B--+C               triple
g9    030T A--+B+--C, A--+C         triple and transitive
g10   030C A+--B+--C, A--+C         triple
g11    201 A+--+B+--+C              triple
g12   120D A+--B--+C, A+--+C        triple and transitive
g13   120U A--+B+--C, A+--+C        triple and transitive
g14   120C A--+B--+C, A+--+C        triple and transitive
g15    210 A--+B+--+C, A+--+C       triple and transitive
g16    300 A+--+B+--+C, A+--+C triple, complete and transitive graph
# --------------------------------------------------------------
```

An example of graph sampling: SRS of s_1, $|s_1| = 3$

$s_2^* = s_1 \times s_1$, $s_2 = s_1 \times U \cup U \times s_1$

		RRMSE (%)		
	Parameter	$\hat{\theta}_y(s_2^*)$	$\hat{\theta}_y(s_2)$	$\hat{\theta}_z^{eq}(s_2)$
1st-order	Indegree	331.261	*26.022*	
2nd-order	Density	0.041	0.003	0.004
	Reciprocity	0.118	0.013	0.016
3rd-order	g6	333.053	73.600	81.478
	g7	375.735	96.397	104.520
	g8	540.774	108.593	116.406
	g9	771.335	149.723	160.095
	g10	540.774	136.630	142.923
	g11	771.335	172.970	190.091
	g12	1 095.445	211.943	230.090
	g13	1 095.445	211.943	230.090
	g14	540.774	122.138	131.251
	g15	771.335	172.970	190.091
	g16	1 095.445	211.943	230.090
	Transitivity	0.084	*0.028*	*0.028*

Example: Sector labour flows 2015Q1-2017Q1

$|\mathcal{N}| = 263$

$|A| = 31120$, $a_{ij} \in A$ if labour flow from i to j

Density = 0.45, Reciprocity = 0.73

$s_2^* = s_1 \times s_1$, $s_2 = s_1 \times U \cup U \times s_1$

	RRMSE (%)					
	$\|s_1\| = 3$			$\|s_1\| = 6$		
Parameter	$\hat{\theta}_y(s_2^*)$	$\hat{\theta}_y(s_2)$	$\hat{\theta}_z^{eq}(s_2)$	$\hat{\theta}_y(s_2^*)$	$\hat{\theta}_y(s_2)$	$\hat{\theta}_z^{eq}(s_2)$
Indegree	75.01	31.76		47.84	22.12	
Mutual Edges	91.20	37.27	37.42	57.42	26.01	26.27
Density	75.01	31.76	31.89	47.84	22.12	22.34
Reciprocity	62.20	14.00	14.03	31.35	8.49	8.57

BIG sampling **with replacement (WR)**

- $p_i = \Pr(\delta_i = 1)$ for $i \in U$
- $y_{\alpha_i} = y_k$ for $k = \alpha_i$ and $p_{(k)} = \sum_{i \in \beta_k} p_i = p_{\beta_k}$
- Hansen-Hurwitz (HH) estimators

$$\tilde{\theta}_z = \frac{1}{n} \sum_{i=1}^{n} \frac{z_i}{p_i} \qquad \text{and} \qquad \tilde{\theta}_y = \frac{1}{n} \sum_{i=1}^{n} \frac{y_{\alpha_i}}{p_{\beta_k}} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_k}{p_{(k)}}$$

Result: $V(\tilde{\theta}_z) \geq V(\tilde{\theta}_y)$, where the equality holds if $P_{ik} = p_{(k)}^{-1} p_i$ for $i \in \beta_k$ and 0 otherwise. $\square$

NB. equal-probability $s_1 \mapsto \tilde{\theta}_z$ with equal-share weights

BIG sampling **without replacement (WOR)**

- $\pi_i = \Pr(\delta_i = 1)$ and $\pi_{ij} = \Pr(\delta_i \delta_j = 1)$ for $i, j \in U$
- $\pi_{(k)} = \Pr(\delta_k = 1)$ and $\pi_{(k)(l)} = \Pr(\delta_k \delta_l = 1)$ for $k, l \in \Omega$

Result: For HT-estimators $\hat{\theta}_y$ and $\hat{\theta}_z$ with $P_{ik} \propto \pi_i$,

$$V(\hat{\theta}_z) - V(\hat{\theta}_y) =$$

$$\sum\sum_{k \neq l \in \Omega} y_k y_l \Big(\sum_{i \in \beta_k} \sum_{j \in \beta_l} \frac{\pi_{ij}}{\pi_i \pi_j} P_{ik} P_{jl} - \frac{\pi_{(k)(l)}}{\pi_{(k)} \pi_{(l)}} \Big)$$

NB. cluster sampling as special case $V(\hat{\theta}_z) = V(\hat{\theta}_y)$

To explore: scope of finite network sampling theory

More observation procedures, greater scope of application

Function of network totals of definite orders: **yes**

e.g. density, reciprocity, transitivity, etc.

e.g. "structural equivalence" ["similarity", Pearson corr.]

Parameters based on geodesic: **feasible?**

e.g. "closeness" centrality: inverse of mean of invserse geodesics

Measures based on fixed-point-equation: **impossible?**

e.g. Katz centrality: $\mathbf{x}_{N\times 1} = \alpha A\mathbf{x} + \boldsymbol{\beta}_{N\times 1}$

e.g. "regular equivalence" btw $i, j \in \mathcal{N}$: $\boldsymbol{\sigma}_{N\times N} = \alpha A\boldsymbol{\sigma} + \boldsymbol{I}_{N\times N}$

[1] Birnbaum, Z.W. and Sirken, M.G. (1965). *Design of Sample Surveys to Estimate the Prevalence of IRareDiseases: Three Unbiased Estimates*. Vital and Health Statistics, Ser. 2, No.11. Washington:Government Printing Office.

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

[3] Frank, O. (1977a). Estimation of graph totals. *Scandinavian Journal of Statistics*, 4:81–89.

[4] Frank, O. (1977b). A note on Bernoulli sampling in graphs and Horvitz-Thompson estimation. *Scandinavian Journal of Statistics*, 4:178–180.

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

[6] 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.

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

[8] Frank, O. (1980a). Estimation of the number of vertices of different degrees in a graph. *Journal of Statistical Planning and Inference*, 4(1):45–50, 1980.

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

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

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

[12] Frank O. and Snijders T. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, 10:53–53.

[13] Goldenberg, A., Zheng, A.X., Fienberg, S.E. and Airoldi, E.M. (2010). A Survey of Statistical Network Models. *Foundations and Trends in Machine Learning*, 2:129–233.

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

[15] Klovdahl, A. S. (1989). Urban social networks: Some methodological problems and possibilities. In M. Kochen (ed.) *The Small World*. Norwood, NJ: Ablex Publishing, pp. 176–210.

[16] Lavalleè, P. (2007). *Indirect Sampling*. Springer.

[17] Newman, M.E.J. (2010). *Networks: An Introduction*. Oxford University Press.

[18] Sirken, M.G. (2005). *Network Sampling*. In *Encyclopedia of Biostatistics*, John Wiley & Sons, Ltd. DOI: 10.1002/0470011815.b2a16043

[19] Snijders, T. A. B. (1992). Estimation on the basis of snowball samples: How to weight. *Bulletin de Methodologie Sociologique*, 36:59–70.

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

[21] Thompson, S. K. (1991). Adaptive cluster sampling: Designs with primary and secondary units. *Biometrics*, 47:1103–1115.