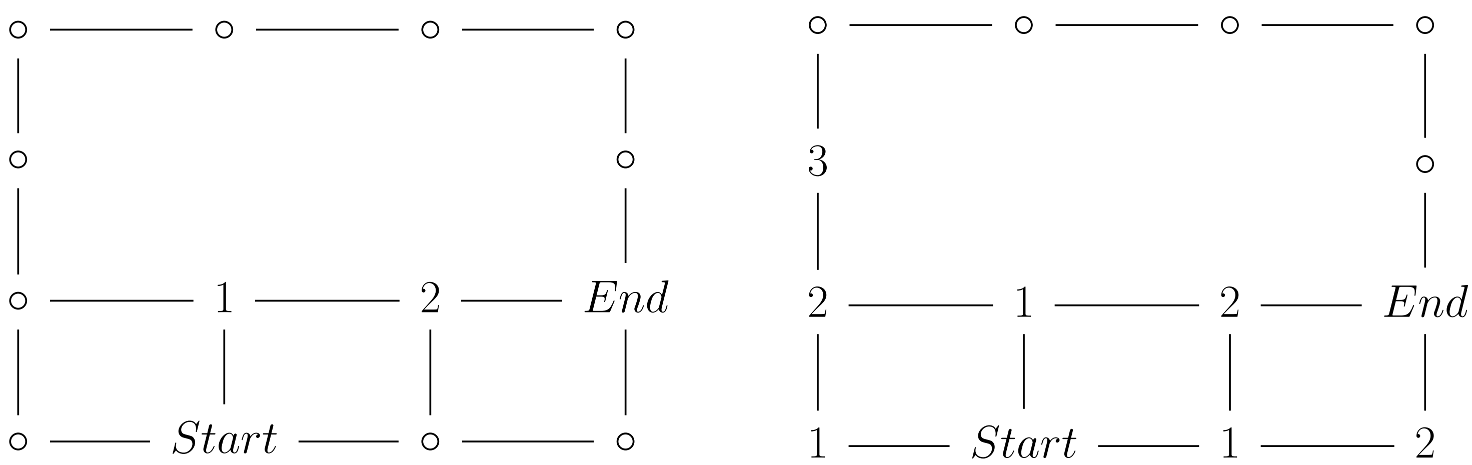
Day-3 Session-2: Snowball Sampling and Targeted Random Walk Sampling

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*

Breadth-first search (BFS) or depth-first search (DFS)



Depth-first (left): if possible go up, right, down *or* left
Breadth-first (right): if possible, go up, right, down *and* left

BFS or DFS: basic *graph search* algorithms
Can exhaust a finite connected graph in the end
Probabilistic sampling methods if non-exhaustive:

- T-wave snowball sampling as analogy to BFS
- T-step targeted random walk as analogy to DFS

T-wave snowball sampling (TSBS)

T*-wave* incident OP starting from s_0: for $t = 1, ..., T$, let

$$s_t = \alpha(s_{t-1}) \setminus \bigcup_{r=0}^{t-1} s_r \qquad \text{and} \qquad s = \bigcup_{t=0}^{T-1} s_t$$

be the t-th wave sample and the seed sample, respectively

- Directed graphs: $s_{\text{ref}} = s \times U$ and $A_s = \bigcup_{i \in s} \bigcup_{j \in \alpha_i} A_{ij}$
- Undirected: $s_{\text{ref}} = s \times U \cup U \times s$ and $A_s = \bigcup_{i \in s} \bigcup_{j \in \alpha_i} (A_{ij} \cup A_{ji})$

$$\begin{tikzcd}
\cdots \arrow[r, dash] & h_1 \arrow[r, dash] & j_3 \arrow[r, dash] \arrow[d, dash] \arrow[dr, dash] & i_3 \arrow[d, dash] \arrow[dl, dash] \arrow[dr, dash] & h_2 \arrow[r, dash] \arrow[d, dash] & \cdots \\
\cdots \arrow[r, dash] & j_1 \arrow[r, dash] & i_1 \arrow[r, dash] & i_2 \arrow[r, dash] & j_2 \arrow[r, dash] & \cdots
\end{tikzcd}$$

Triangle κ of $M = \{i_1, i_2, i_3\}$ under TSBS...

Basis of inference: Sample inclusion probabilities

Goodman (1961): mutual best friendship (out-degree $\equiv 1$)
Frank and Snijders (1994): 1SBS in digraphs generally
Zhang and Patone (2017): sample inclusion probabilities of kth order induced motifs under TSBS

G :

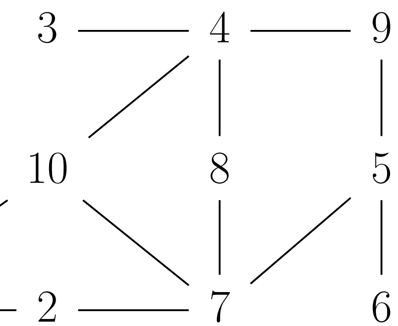


Illustration of $\pi_{(3)}$ under 2SBS:

- $^{[t]}$th wave ancestors: $\nu_3^{[0]} = \{3\}, \nu_3^{[1]} = \{4\}, \nu_3^{[2]} = \{8, 9, 10\}$, ...
- ancestors of node 3 under 2SBS: $R_3 = \{3, 4, 8, 9, 10\}$
- SRS, $|s_0| = 2$: $\quad \pi_{(3)} = 1 - \binom{5}{2} / \binom{10}{2} = \frac{7}{9}$ $\quad$ whereas $\pi_3 = \frac{2}{10}$

Distances to a motif

Let φ_{ij} be the geodesic distance from node i to node j in G

$$\begin{tikzcd}
\cdots & 1 & 2 \arrow[d, no head] & \cdots & 4 \arrow[r, no head] \arrow[d, no head] & 5 & \cdots \\
& \cdots & 3 & \cdots & 6 \arrow[r, no head] & 7 & \cdots
\end{tikzcd}$$

The *geodesic distance from node i to motif κ* is the number of waves it takes from i to reach *any* nodes in $M(\kappa)$:

$$\varphi_{i,\kappa} = 0 \text{ if } i \in M(\kappa) \quad \text{and} \quad \varphi_{i,\kappa} = \min_{j \in M(\kappa)} \varphi_{ij} \text{ if } i \notin M(\kappa)$$

The *radius distance from node i to motif κ* is the number of waves it takes from i to reach *all* the nodes in $M(\kappa)$:

$$\lambda_{i,\kappa} = \max_{j \in M(\kappa)} \varphi_{ij}$$

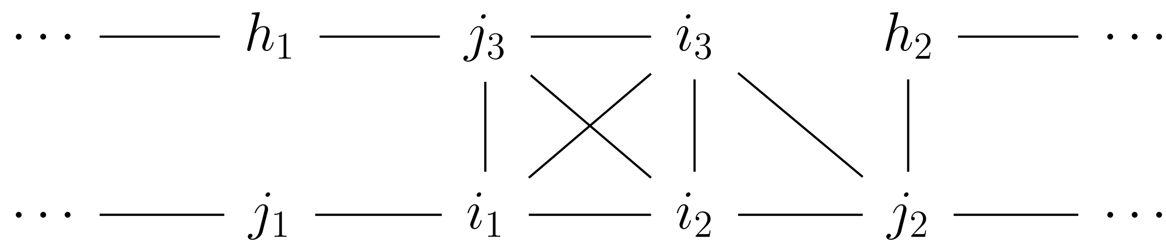
The *observation distance from node i to motif κ* is the number of waves it takes from i to observe the motif κ, denoted by $d_{i,\kappa}$:

$d_{i,\kappa} \leq 1 + \lambda_{i,\kappa}$ if connected $M(\kappa)$ – Zhang and Oguz-Alper (2020)

The node i in G is a TSBS *ancestor* of κ if $d_{i,\kappa} \leq T$

Strategy: Using all TSBS ancestors β_κ

Requires additional waves of OP generally

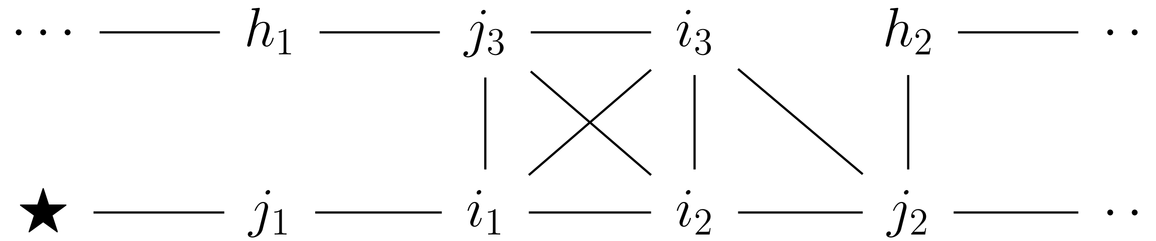


Triangle κ of $M = \{i_1, i_2, i_3\}$ under 3SBS

- 3SBS ancestors: $\beta_\kappa = M \cup \{j_1, j_2, j_3\} \cup \{h_1, h_2\}$
- if only $i_1 \in s_0$, then all β_κ identified in G_s
  similarly for i_2, i_3, j_2 and j_3, which have $d_{i,\kappa} = 2$
- if only $j_1 \in s_0$, then h_2 unidentified as $(j_2 h_2) \notin A_s$
  if only $h_1 \in s_0$, then h_2 unidentified as $(j_2 h_2) \notin A_s$
- if only $h_2 \in s_0$, then $\{j_1, h_1\}$ unidentified since neither $(j_1 h_1) \notin A_s$ nor $(h_1 j_3) \notin A_s$

Strategy: Using subset $\beta_\kappa^* \subseteq \beta_\kappa$

No additional waves of OP, using $\beta_\kappa^* \subseteq \beta_\kappa$ identifiable given *any* G_s by TSBS (Zhang and Oguz-Alper, 2020)



Triangle κ of $M = \{i_1, i_2, i_3\}$ under TSBS

- Can let $\beta_\kappa^* = M$ for 2SBS, which excludes $\{j_2, j_3\}$
- Can let $\beta_\kappa^* = M(\kappa) \cup \beta_\kappa^1(M)$ for 4SBS, where

$$\beta_\kappa^t(M) = \{i \notin M(\kappa) : \varphi_{i,\kappa} \leq t\}$$
$$\beta_\kappa^1(M) = \{j_1, j_2, j_3\}$$

which excludes $\{h_1, h_2\}$ that are 3SBS ancestors, e.g. h_2 not identified as 4SBS ancestor if only ★ $\in s_0$

Strategy: Sample-dependent $\beta_{\kappa|s}$ for TSBS

No additional waves of OP, using TSBS ancestors given *actual* G_s and sample observation distance $d_{i,\kappa}(G_s)$:

$$\beta_{\kappa|s} = \{i : d_{i,\kappa}(G_s) \leq T\}$$

$$\begin{tikzcd}
\cdots \arrow[r, dash] & h_1 \arrow[r, dash] & j_3 \arrow[r, dash] \arrow[d, dash] \arrow[dr, dash] & i_3 \arrow[d, dash] \arrow[dl, dash] \arrow[dr, dash] & h_2 \arrow[r, dash] \arrow[d, dash] & \cdots \\
\cdots \arrow[r, dash] & j_1 \arrow[r, dash] & i_1 \arrow[r, dash] & i_2 \arrow[r, dash] & j_2 \arrow[r, dash] & \cdots
\end{tikzcd}$$

Triangle κ of $M = \{i_1, i_2, i_3\}$ under TSBS

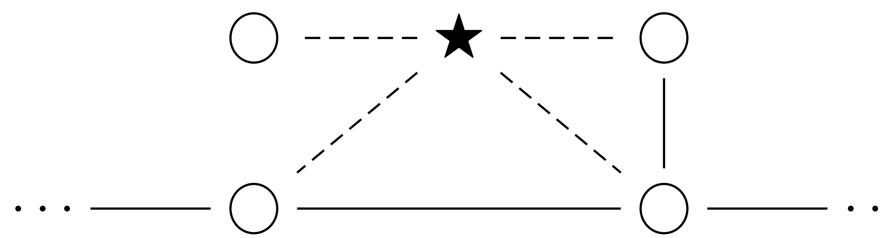
- For 2SBS: $\beta_{\kappa|s} = M \cup \{j_2, j_3\} = \beta_\kappa$
- For 3SBS:
  if only $h_2 \in s_0$, then $\beta_{\kappa|s} = M \cup \{j_2, j_3\} \cup \{h_2\}$;
  if only $\{j_1, h_1\} \cap s_0 \neq \emptyset$, then $\beta_{\kappa|s} = M \cup \{j_2, j_3\} \cup \{j_1, h_1\}$;
  otherwise, $\beta_{\kappa|s} = M \cup \{j_2, j_3\} \cup \{j_1, h_1, h_2\} = \beta_\kappa$

Random walk in graphs

Relevant in many fields (Masuda et al. 2017), e.g.

- PageRank (Brin and Page, 1998)
- Betweenness (Newman, 2010)

Need random jumps, e.g. Avrachenkov et al. (2010):



Control probability of moving via ★ by r, such that

$$\text{stationary probability} \quad \pi_i \propto d_i + r$$

Generalised ratio estimator (e.g. Thompson 2006):

$$\hat{\mu} = \Big(\sum_{i=1}^{n} y_i/\pi_i\Big) / \Big(\sum_{i=1}^{n} 1/\pi_i\Big)$$

Targeted random walk (TRW)

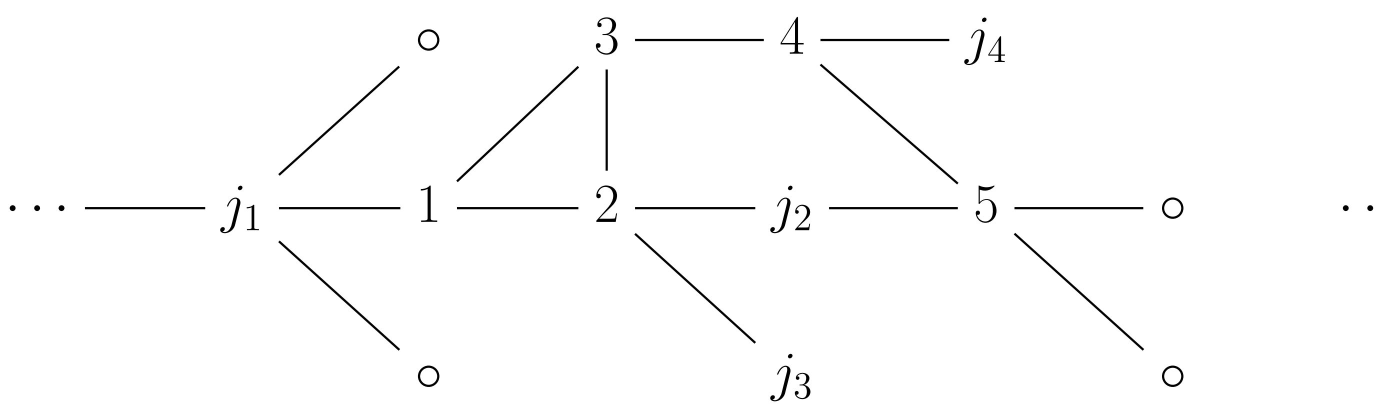
The OP of TRW is applied to the seed sample

$$s = \{X_0, X_1, ..., X_T\}$$

A node can appear more than once in s of TRW sampling, whereas the seed sample nodes are all distinct under TSBS by definition

At each $X_t = i$, observe $\{i\} \times \alpha_i$ and $\alpha_i \times \{i\}$, such that

$$s_{\mathrm{ref}} = s \times U \cup U \times s$$



Triangle κ of $M = \{1, 2, 3\}$ observed given one adjacent move $(X_t, X_{t+1}) = (1, 2), (2, 1), (1, 3), (3, 1), (2, 3)$, or $(3, 2)$

Basis of inference

Stationary probability at equilibrium $\pi_i = \Pr(X_t = i)$
Stationary sampling probability π_M may be unknown, e.g.

$$\pi_{X_t X_{t+2} X_{t+4}} = \pi_{135} = \pi_1 \Big(\sum_{i \in U} p_{1i} p_{i3} \Big) \Big(\sum_{i \in U} p_{3i} p_{i5} \Big)$$

Stationary successive sampling probability (S3P), e.g.

$$\pi_{X_t X_{t+1} X_{t+2} X_{t+3}} = \pi_{1234} = \pi_1 p_{12} p_{23} p_{34}$$

is known except for the proportionality constant in π_i
Generating states of T-step walk with seed sample s:

$$\mathcal{C}_s = \{M : M \subseteq s\}$$

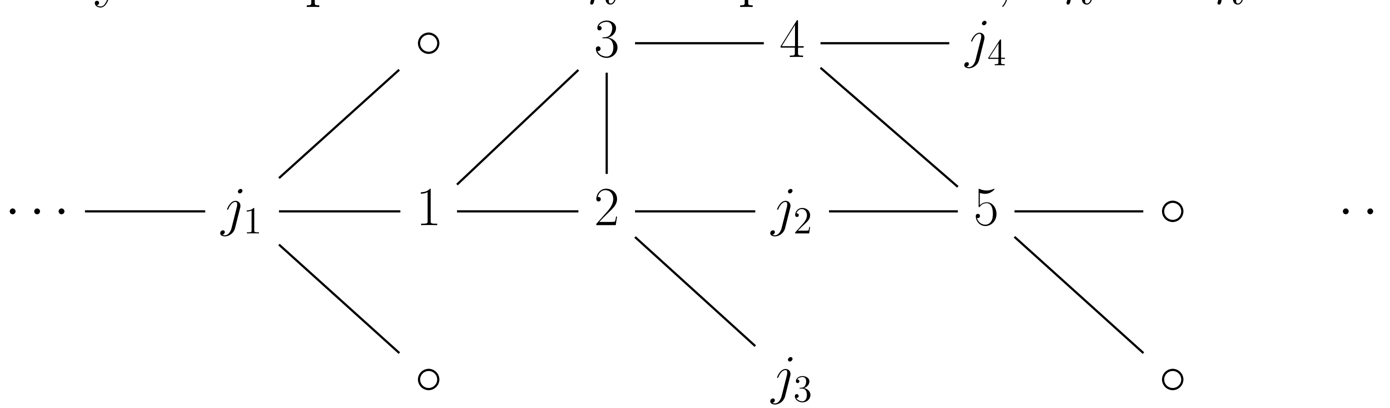
E.g. $(X_t, X_{t+1}, X_{t+2}) = (1, 2, 3)$ with S3P π_{123} actually...
can also calculate S3P $\pi_{132}, \pi_{213}, \pi_{231}$ etc. hypothetically

Eligible sample motifs

Actual sampling sequence of states (AS3) of motif κ:

$$s_K = (X_t, ..., X_{t+q})$$

Equivalent sampling sequence of states (ES3) of s_K, $R_K = \{\tilde{s}_K : \tilde{s}_K \sim s_K\}$, contains any possible sequence of states with $|\tilde{s}_K| = |s_K|$, such that the motif κ would be observed given $(X_t, X_{t+1}, ..., X_{t+q}) = \tilde{s}_K$ but not based on any subsequence of $\tilde{s}_K$. In particular, $s_K \sim s_K$.



Triangle κ of $M = \{1, 2, 3\}$ observed given AS3 $(X_t, X_{t+1}) = (1, 2)$
ES3 are $R_\kappa = \{(1,2), (2,1), (1,3), (3,1), (2,3), (3,2)\}$

Generalised ratio estimator using IWE

Let AS3 $s_\kappa = (X_t, ..., X_{t+q})$ for eligible sample motif $\kappa \in \Omega_s$
Let M be a possible sequence of states $(X_t, ..., X_{t+q})$
Let $\delta_M = 1$ if M is realised and 0 otherwise, with associated S3P π_M
Let $I_\kappa(M) = 1$ if $M \in R_\kappa$, and 0 otherwise
Let $\{w_{M\kappa} : M \in R_\kappa\}$ be the *incidence weights*, $\sum_{M \in R_\kappa} w_{M\kappa} = 1$

$$\hat{\theta}(X_t, ..., X_{t+q}) = \sum_{\kappa \in \Omega} \sum_{M} \frac{\delta_M}{\pi_M} I_\kappa(M) w_{M\kappa} y_\kappa$$

Given the AS3 s_κ is of the order $|s_\kappa| = q + 1$ and $|s| = n$
Let $\mathbb{I}_t = 1$ if $\sum_{\kappa \in \Omega} I_\kappa\big(\{X_t, ..., X_{t+q}\}\big) > 0$, and 0 otherwise

$$\hat{\theta} = \Big(\sum_{t=1}^{n-q} \mathbb{I}_t \hat{\theta}_t\Big) / \Big(\sum_{t=1}^{n-q} \mathbb{I}_t\Big)$$

Given $\theta = \sum_{\kappa \in \Omega} y_\kappa$ and $\theta' = \sum_{\kappa \in \Omega'} y'_\kappa$, can use

$$\hat{\mu} = f(\hat{\theta}, \hat{\theta}')$$

if $\hat{\mu}$ invariant towards the unknown proportionality constant in S3P

[1] Avrachenkov, K., Ribeiro, B. and Towsley, D. (2010). Improving Random Walk Estimation Accuracy with Uniform Restarts. *Research report*, RR-7394, INRIA. inria-00520350

[2] Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30: 107–117.

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

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

[5] Masuda, N., Porter, M.A. and Lambiotte, R. (2017) Random walks and diffusion on networks. *Physics Reports*, 716-717: 1–58. `http://dx.doi.org/10.1016/j.physrep.2017.07.007`

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

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

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

[9] Zhang, L.-C. and Oguz-Alper, M. (2020) Bipartite incidence graph sampling. `https://arxiv.org/abs/2003.09467`