Day-3 Practical Session, 27 May 2021

Part 1: BIGS-IWE strategy for T-wave Snowball Sampling (T SBS)

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

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

In this illustration, we will apply BIGS-IWE strategy to T-wave snowball sampling (TSBS). Several types of *motifs* will be of interest. We will consider BIGS with the restricted-type of ancestors, β^*_κ, of the observed motifs. The efficiency of the HTE and the IWEs with equal and uneqal weights under T-wave SBS will be compared.

Description of the population and sampling strategies

- Population graph: $G = (U, A)$
- T-wave SBS: T-wave *incident observation procedure* (OP) applied to s_0
- The sample graph: $G_s = (U_s, A_s)$ formed based on $s_{ref} = s \times U \cup U \times s$ (for undireted graphs), where $s = \bigcup_{t=0}^{T-1} s_t$ is the *seed sample*, where $s_t = \alpha(s_{t-1}) \setminus \bigcup_{r=0}^{t-1} s_r$ is the t-th wave sample
- Let φ_{ij} be the geodesic distance from node i to node j in G
- The *geodesic distance* from node i to motif κ: $\varphi_{i,\kappa} = 0$ for all $i \in M(\kappa)$ and $\varphi_{i,\kappa} = \min_{j\in M(\kappa)} \varphi_{ij}$ if $i \notin M(\kappa)$
- The *radius distance* from node i to motif κ: $\lambda_{i,\kappa} = \max_{j\in M(\kappa)} \varphi_{ij}$
- The *observation distance* from node i to motif κ: $d_{i,\kappa} \leq 1 + \lambda_{i,\kappa}$
- The *diameter* of the motif κ: $\varphi_\kappa = \max_{i,j\in M(\kappa)} \varphi_{ij}$
- The *observation diameter* of the motif κ: $\zeta_\kappa = \max_{i\in M(\kappa)} d_{i,\kappa}$
- TSBS *ancestors* of κ: $\beta_\kappa = \{i : d_{i,\kappa} \leq T\}$;
- The population BIG: $\mathcal{B} = (F, \Omega; H)$, where H contains edges from β_κ to $\kappa \in \Omega$; however, additional waves of OP required generally
- The population BIG with *restricted ancestors*: $\mathcal{B}^* = (F, \Omega; H^*)$, where H^* contains edges from $\beta^*_\kappa \subseteq \beta_\kappa$ to $\kappa \in \Omega$; no additional waves required
- The sample BIG with β^*_κ: $\mathcal{B}^*_s = (s_0, \Omega_s; H^*_s)$, where H^* consists of edges from $\beta^*_\kappa \cap s_0$ to $\kappa \in \Omega_s$
- BIGS-IWE strategy, ($\mathcal{B}^*$,IWE): a sample motif $\kappa \in \Omega_s$ under T-wave incident OP becomes *eligible* for IWE $\iff \beta^*_\kappa \cap s_0 \neq \emptyset$
- By Lemma 5.4 (see lecture notes): the strategy BIGS-IWE with $\beta^*_\kappa = M(\kappa)$ unbiased for θ if $T = \zeta_\kappa$
- By Lemma 5.5 (see lecture notes): the strategy BIGS-IWE with $\beta^*_\kappa = M(\kappa) \cup \beta^t_\kappa(M)$, with $\beta^t_\kappa(M) = \{i \notin M(\kappa) : \varphi_{i,\kappa} \leq t\}$, unbiased for θ if $T = \varphi_\kappa + 2t$, for $t \geq 1$
- For the strategy BIGS-IWE with PIDA weights: additional waves up to φ_κ or $\zeta_\kappa + t$ maybe needed for $\beta^*_\kappa = M(\kappa)$ and $\beta^*_\kappa = M(\kappa) \cup \beta^t_\kappa(M)$, respectively

---

Formula sheet

- The parameter of interest: the total number of motifs in the population BIG

  $\theta = \sum_{\kappa\in\Omega} y_\kappa$, where $y_\kappa = 1$ for all $\kappa \in \Omega$

- Horvitz-Thompson estimator (HTE)

  $\hat{\theta}_{HT} = \sum_{\kappa\in\Omega_s} \frac{y_\kappa}{\pi_{(\kappa)}}$, where $\pi_{(\kappa)} = \Pr(\kappa \in \Omega_s)$ are the *first-order inclusion probabilies* calculated by, given SRS of s_0,

  $\pi_{(\kappa)} = 1 - \bar{\pi}_{\beta^*_\kappa} = 1 - \binom{N-|\beta^*_\kappa|}{n} / \binom{N}{n}$, where $|\beta^*_\kappa|$ is the size of the ancestor set of κ in $\mathcal{B}^*$

- Hansen-Hurwitz (HH) type estimator

  $\hat{\theta}_{HH} = \sum_{i\in s_0} \frac{z_i}{\pi_i}$, where $z_i = \sum_{\kappa\in\alpha^*_i} w_{i\kappa} y_\kappa$, α^*_i the *restricted* successor set of i in $\mathcal{B}^*$

  - *Multiplicity* estimator; equal weights

    $w_{i\kappa} \equiv \frac{1}{|\beta^*_\kappa|}$

  - HH-type estimator with *unequal* weights given SRS of s_0: *probability and inverse degree-adjusted (PIDA) weights*

    $w_{i\kappa} = \frac{1}{|\alpha^*_i|^\gamma} \left(\sum_{i\in\beta^*_\kappa} \frac{1}{|\alpha^*_i|^\gamma}\right)^{-1}, \gamma \geq 0$

NB. When $\gamma = 0$, PIDA weights reduce to the eqaul-share weights: $w_{i\kappa} = 1/ \mid \beta^*_\kappa|$ for $i \in \beta^*_\kappa$

- The sampling variance of the HTE of θ in $\mathcal{B}^*$ by T-wave *incident* OP is calculated by

  $\mathrm{V}(\hat{\theta}_{HT}) = \sum_{\kappa\in\Omega} \sum_{\ell\in\Omega} \left(\frac{\pi_{(\kappa\ell)}}{\pi_{(\kappa)}\pi_{(\ell)}} - 1\right)$, where $\pi_{(\kappa\ell)} = \Pr(\kappa, \ell \in \Omega_s)$ are the *second-order inclusion probabilies* calculated by, given SRS of s_0,

  $\pi_{(\kappa\ell)} = \pi_{(\kappa)} + \pi_{(\ell)} - (1 - \bar{\pi}_{\beta^*_\kappa \cup \beta^*_\ell}) = 1 - \binom{N-|\beta^*_\kappa \cup \beta^*_\ell|}{n} / \binom{N}{n}$

- The sampling variance of the HH-type estimator of θ in $\mathcal{B}^*$ by T-wave *incident* OP is calculated by

  $\mathrm{V}(\hat{\theta}_{HH}) = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{\sum_{i\in F}(z_i - \bar{Z})^2}{n-1}$, where $N = \mid F|$, $n = \mid s_0|$, and $\bar{Z} = \sum_{i\in F} z_i/N$

  $\pi_{(\kappa\ell)} = \pi_{(\kappa)} + \pi_{(\ell)} - (1 - \bar{\pi}_{\beta^*_\kappa \cup \beta^*_\ell}) = 1 - \binom{N-|\beta^*_\kappa \cup \beta^*_\ell|}{n} / \binom{N}{n}$

- The sampling variance of the HTE of θ under *induced* OP is calculated by

  $\mathrm{V}(\hat{\theta}^{ind}_{HT}) = \sum_{\kappa\in\Omega} \sum_{\ell\in\Omega} \left(\frac{\pi_{(M(\kappa)M(\ell))}}{\pi_{(M(\kappa))}\pi_{(M(\ell))}} - 1\right)$, where $\pi_{(M(\kappa))}$ and $\pi_{(M(\ell))}$ are the $|M(\kappa)|$-th and the $|M(\ell)|$-th *order joint inclusion probabilities*, respectively, such that $\pi_{(M(\kappa))} = \Pr(M(\kappa) \in s_0)$ and $\pi_{(M(\ell))} = \Pr(M(\ell) \in s_0)$. For 3-th *order* motifs, e.g. triangles, 2-stars, etc., we have, under SRS of s_0,

  $\pi_{(M(\kappa))} = \pi_{(M(\ell))} = 1 - \binom{N-3}{n} / \binom{N}{n}$

  The $\pi_{(M(\kappa)M(\ell))}$ are the $|M(\kappa) \cup M(\ell)|$-th *order joint inclusion probabilies*, such that, $\pi_{(M(\kappa)M(\ell))} = \Pr(M(\kappa) \cup M(\ell) \in s_0)$, and calculated by, given SRS of s_0,

  $\pi_{(M(\kappa)M(\ell))} = 1 - \binom{N-|M(\kappa)\cup M(\ell)|}{n} / \binom{N}{n}$

- Let B be the number of Monte-Carlo replications. The emprical expectation and the variance of an estimator over B replications given, respectively, by

  $E_{MC}(\hat{\theta}) = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b, \quad \mathrm{V}_{MC}(\hat{\theta}) = \frac{1}{B} \sum_{b=1}^{B} \left(\hat{\theta}_b - E_{MC}(\hat{\theta})\right)^2$, where $\hat{\theta}_b$ is the estimate on the b-th replication

- The Monte-Carlo bias and the mean square error (MSE) of an estimator given, respectively, by

  $\mathrm{Bias}_{MC}(\hat{\theta}) = E_{MC}(\hat{\theta}) - \theta, \quad MSE_{MC}(\hat{\theta}) = \mathrm{Bias}_{MC}(\hat{\theta})^2 + \mathrm{V}_{MC}(\hat{\theta})$

N.B. 1. R-package **igraph** and **latex2exp** have to be installed.

N.B. 2. The R-functions **checkSum**, **cliqueFun**, **motifFun**, **varSRSCliqueFun**, **varSRSMotifFun**, **cliquezimkFun**, **motifzimkFun**, **cliqueziUnS** and **motifziUnS** will be called *subsidiary functions*, and only be used implicitly meaning that they will be called by the functions for the parameters of which you are allowed to choose values. Therefore, no explanations, except from one-two lines of information about what function does before each function, will be given in detail.

Description of R-function **skthG**

1. Function parameters

- **sizeF**: the number of nodes (vertices) in the population graph G; default value 40
- **p**: The probability of drawing an edge between any arbitrary nodes in G; default value 0.1 resulting in $E(\mid A \mid) = p* \mid U \mid *(\mid U \mid -1)/2 = 78$, for $|U \mid= 40$, for undirected graph
- **showplot**: Use **TRUE** to get the skecth of the population graph; default **FALSE**

2. Main steps of the function

- The population graph generated by the Erdos-Renyi model for chosen number of vertices and the probability of drawing edges. An undirected graph generated.

3. Main outputs of the function

- The population graph shown if **showplot**= **TRUE**
- The population graph generated is returned as a graph object. It shall be called via **$G**

Description of R-function **makeMotif**

1. Function parameters

- **orderM**: the number of nodes (vertices) in the motif; default value 2. The minimum number of nodes for stars and paths has to be 2.
- **motif**: the type of the motif: whether a *clique*, *cycle*, *star* or *path*; default value *clique*

NB. A 2-clique may also be called 2-cycle or 1-path or *dyad*. A 3-clique may also be called 3-cycle or *triangle*. A node may also be called 1-clique or 1- cycle.

2. Main steps of the function

- A subgraph with the chosen order and the type, *motif*, is generated

3. Main outputs of the function

- The motif is returned as a graph object.

Description of R-function **skthMotifs**

1. Function parameters

- No user-defined function parameters

2. Main steps of the function

- Well-known motifs up to order 4 generated
- A plot is drawn to illustrate the motifs

3. Main outputs of the function

- A sketch of the motifs returned

Description of R-function **countMotif**

1. Function parameters

- **popgraph**: The population graph as a graph object
- **orderM**: the number of nodes (vertices) in the motif; default value 2
- **motif**: the type of the motif: whether a *clique*, *cycle*, *star* or *path*; default value *clique*

2. Main steps of the function

- The number of motifs with chosen order and the type is counted in the population graph. One of the subsidiary functions, **cliqueFun** or **motifFun**, called depending on whether the motif is a clique or not. If it is a clique, the counting is based on the R-function **cliques**, otherwise, it is based on the function **subgraph_isomorphisms**. Both functions available in the **igraph** package. The latter may take significant amount of time, especially when the number of motifs is large

3. Main outputs of the function

- The number of the motifs with specified order and type in the population is returned

Description of R-function **varSRSinduced**

1. Function parameters

- **popgraph**: The population graph as a graph object
- **orderM**: the number of nodes (vertices) in the motif; default value 2
- **motif**: the type of the motif: whether a *clique*, *cycle*, *star* or *path*; default value *clique*
- **sizes0**: the sample size of the initial sample s_0; default value 2: The minimum sample size has to be equal to the order of the motif. Otherwise, the inclusion probability of the motif will be zero.

2. Main steps of the function

- The sampling variance of the HTE of θ under SRS of s_0 and with induced OP calculated. One of the subsidiary functions, **varSRSCliqueFun** or **varSRSMotifFun**, called depending on whether the motif is a clique or not. If it is a clique, the identification of the motifs in G is based on the R-function **cliques**, otherwise, it is based on the function **subgraph_isomorphisms**. Both functions available in the **igraph** package. The latter may take significant amount of time, especially when the number of motifs is large. Identification of the motifs required to calculate the joint probability of two different motifs to be selected in s_0.

3. Main outputs of the function

- The sampling variance returned

Description of R-function **diagMotif**

1. Function parameters

- **orderM**: the number of nodes (vertices) in the motif; default value 2
- **motif**: the type of the motif: whether a *clique*, *cycle*, *star* or *path*; default value *clique*

2. Main steps of the function

- The diameter, φ_κ, and the observation diameter, ζ_κ, of the motif calculated

3. Main outputs of the function

- A list two values that shall be called via $\mathbf{\$diamkappa}$ or $\mathbf{\$obsdiamkappa}$ for the diameter and the observation diameter, respectively

Description of R-function **mainTSBS**

1. Function parameters

- **popgraph**: The population graph as a graph object
- **n**: the sample size of s_0; default value 2
- **Twave**: the number of waves of the SBS; default value 1. The minimum number has to be equal to the observation diameter of the motif. Otherwise, an error message returned.
- **orderM**: the number of nodes (vertices) in the motif; default value 2
- **motif**: the type of the motif: whether a *clique*, *cycle*, *star* or *path*; default value *clique*
- **B**: the number of Monte-Carlo replications; dafeult value 50. If $\binom{N}{n} \leq 1000$, $N = \mid U|$, all possible samples selected and B becomes, regardless of the user-specified choice, equal to the number of all possible subets of size n from N. Otherwise, the user-specified value will be used.
- **include PIDA**: use **TRUE** to get the results for the IWE with PIDA weights; default value **FALSE**. Enabling **includePIDA** may increase the computational time significantly, especially when the motif is not a clique.

2. Main steps of the function

- Using the diagnostics of the motif, the maximum numbers for the T-waves required for restricted BIGS-IWE strategy for the multiplicity estimator and the IWE with PIDA weights are calculated. The calculations based on Lemma 5.5. Within the function, the former and the latter denoted by **gomaxgeo** and **gomaxgeoPIDA**, respectively.
- B random initial samples selected with SRS from U
- A list of sample graphs as graph objects constructed based on the incident OP for each wave, including the T-th wave
- Sample estimates are obtained from using the HTE and the multiplicity estimator under $\mathcal{B}^*$ with $\beta^*_\kappa = M(\kappa)$. The subsidiary functions **cliqueFun** or **motifFun** called for the HTE depending on whether the motif a clique or not, and the functions **cliquezimkFun** or **motifzimkFun** called for the multiplicity estimator depending on whether the motif a clique or not
- Sample estimates are obtained from using the IWE with PIDA weights if $\varphi_\kappa + \zeta_\kappa \leq T$ and **include PIDA**=**TRUE** under $\mathcal{B}^*$ with $\beta^*_\kappa = M(\kappa)$. The subsidiary functions **cliqueziUnSFun** and **motifziUnSFun** called depending on whether the motif a clique or not
- Sample estimates are obtained from using the HTE and the multiplicity estimator under $\mathcal{B}^*$ with $\beta^*_\kappa = M(\kappa) \cup \beta^t(M)$. An estimate for each t, which has a maximum value given by **gomaxgeo**, is obtained. The subsidiary functions **cliquezimkFun** or **motifzimkFun** called depending on whether the motif a clique or not
- Sample estimates are obtained from using the IWE with PIDA weights if $\varphi_\kappa + \zeta_\kappa + 3 * t \leq T$ and **include PIDA**=**TRUE** under $\mathcal{B}^*$ with $\beta^*_\kappa = M(\kappa) \cup \beta^t(M)$. An estimate for each t, which has a maximum value given by **gomaxgeoPIDA**, is obtained. The subsidiary functions **cliqueziUnSFun** and **motifziUnSFun** called depending on whether the motif a clique or not.
- The Monte-Carlo expectation of the sample sizes for each wave where the sample graph is constructed to identify the observed motifs is calculated. The maximum value for t-wave equivalent to **gomaxgeo**
- The Monte-Carlo expectations, variances and MSEs calculated for each of the estimators used. The first one is equivalent to the population value θ when the replications based on the all possible samples

3. Main outputs of the function

- Expected sample size for each wave if **gomaxgeo**> 0
- The MC expectations, variances and MSEs of the estimators used

---