Web Appendix to "Identification of homophily and preferential recruitment in respondent-driven sampling"

Web Appendix 1

Unfortunately there are no general closed-form expressions for the extrema of h and p on $\mathcal{C}(G_R, \mathbf{d}_R, \mathbf{Z}_R)$. The space of compatible subgraphs can be very large, but straightforward optimization techniques permit finding these bounds quickly. We describe a stochastic optimization algorithm for finding the global optimum of an arbitrary function J of h and p, based on simulated annealing (1–4). The approach is similar to a quadratic programming framework introduced by De Paula, Richards-Shubik, and Tamer (5) for finding the identification set for certain functionals of graphs and vertex attributes. The optimization routine described here is constructive: it returns the (possibly not unique) pair $(G_{SU}, \mathbf{Z}_{SU}) \in \mathcal{C}(G_R, \mathbf{d}_R, \mathbf{Z}_R)$ that maximizes a carefully chosen objective function $\pi(\cdot)$.

Let $J : [-1, 1]^2 \to \mathbb{R}$ be a function taking arguments $h(G_{SU}, \mathbf{Z}_{SU})$ and $p(G_{SU}, G_R, \mathbf{t}_R, \mathbf{Z}_{SU})$ for $(G_{SU}, \mathbf{Z}_{SU}) \in \mathcal{C}(G_R, \mathbf{d}_R, \mathbf{Z}_R)$. We choose this function, abbreviated $J(h, p)$, so that a desired feature of $\mathcal{C}(G_R, \mathbf{d}_R, \mathbf{Z}_R)$ coincides with the maximum of J. For example, the maximum of the function

$$J(h, p) = \frac{1}{1 + \epsilon + h}$$

on $\mathcal{C}(G_R, \mathbf{d}_R, \mathbf{Z}_R)$ where $\epsilon > 0$, coincides with the lower identification bound of h. For concreteness in what follows, we will assume $J(h, p)$ has this form; similar definitions can be formulated individually to find the maximum of h, and the minimum and maximum of p.

For $T > 0$, define the objective function $\pi(h, p) \propto \exp[J(h, p)/T]$. Our goal is to find $(G_{SU}, \mathbf{Z}_{SU}) \in \mathcal{C}(G_R, \mathbf{d}_R, \mathbf{Z}_R)$ such that $\pi\big(h(G_{SU}, \mathbf{Z}_{SU}), p(G_{SU}, G_R, \mathbf{t}_R, \mathbf{Z}_{SU})\big)$ is maximized. Let

$$K\big((G_{SU}, \mathbf{Z}_{SU}), (G^*_{SU}, \mathbf{Z}^*_{SU})\big)$$

be a transition kernel that describes the probability of moving from a state $(G_{SU}, \mathbf{Z}_{SU}) \in \mathcal{C}(G_R, \mathbf{d}_R, \mathbf{Z}_R)$ to another state $(G^*_{SU}, \mathbf{Z}^*_{SU}) \in \mathcal{C}(G_R, \mathbf{d}_R, \mathbf{Z}_R)$. Let T_t be a positive non-decreasing sequence indexed by t, with $\lim_{t\to\infty} T_t = 0$. We construct an inhomogeneous Markov chain on $\mathcal{C}(G_R, \mathbf{d}_R, \mathbf{Z}_R)$. At step t, where the current state is $(G_{SU}, \mathbf{Z}_{SU})$, we accept the proposed state $(G^*_{SU}, \mathbf{Z}^*_{SU}) \sim K\big((G_{SU}, \mathbf{Z}_{SU}), \cdot\big)$ with probability

$$\rho_t = \min\left\{1, \exp\left[\frac{J\big(h(G^*_{SU}, \mathbf{Z}^*_{SU}), p(G^*_{SU}, G_R, \mathbf{t}_R, \mathbf{Z}^*_{SU})\big) - J\big(h(G_{SU}, \mathbf{Z}_{SU}), p(G_{SU}, G_R, \mathbf{t}_R, \mathbf{Z}_{SU})\big)}{T_t}\right]\right\}.$$

The proposal function is described formally below.

As $T_t \to 0$, the samples $(G_{SU}, \mathbf{Z}_{SU})_t$ become more concentrated around local maxima of π. Convergence of the sequence $(G_{SU}, Z_{SU})_t$ to a global optimum depends on its ability to escape local maxima of J. The sequence T_t, called the "cooling schedule", controls the rate of convergence. Let

$\mathcal{M}$ denote the set of $(G_{SU}, \mathbf{Z}_{SU}) \in \mathcal{C}(G_R, \mathbf{d}_R, \mathbf{Z}_R)$ for which $J\big(h(G_{SU}, \mathbf{Z}_{SU}), p(G_{SU}, G_R, \mathbf{t}_R, \mathbf{Z}_{SU})\big)$ is equal to the global maximum. Careful choice of T_t ensures that the sequence of samples converges in probability to an element of $\mathcal{M}$.

Proposition 1. *Let the cooling schedule be given by* $T_t = \big(\epsilon \log(t)\big)^{-1}$ *where* $\epsilon > 0$ *is a constant. Then* $\lim_{t\to\infty} \Pr\big((G_{SU}, \mathbf{Z}_{SU})_t \in \mathcal{M}\big) = 1$.

Proof. Let $J(h,p) = 1/(1+\epsilon+h)$ for $0 < \epsilon < 1$ and let $\mathcal{M}$ be the set of $(G_{SU}, \mathbf{Z}_{SU})$ that achieve the global maximum of J on $\mathcal{C}(G_R, \mathbf{d}_R, \mathbf{Z}_R)$. Let the cooling schedule be given by $T_t = (\epsilon \log(t))^{-1}$. Following Hajek (3), we say that a state $(G_{SU}, \mathbf{Z}_{SU}) \in \mathcal{C}(G_R, \mathbf{d}_R, \mathbf{Z}_R)$ communicates with $\mathcal{M}$ at depth D if there exists a path in $\mathcal{C}(G_R, \mathbf{d}_R, \mathbf{Z}_R)$ that starts at $(G_{SU}, \mathbf{Z}_{SU})$ and ends at an element of $\mathcal{M}$ such that the least value of J along the path is $J\big(h(G_{SU}, \mathbf{Z}_{SU}), p(G_{SU}, G_R, \mathbf{t}_R, \mathbf{Z}_{SU})\big) - D$. Let D^* be the smallest number such that every $(G_{SU}, \mathbf{Z}_{SU}) \in \mathcal{C}(G_R, \mathbf{d}_R, \mathbf{Z}_R)$ communicates with $\mathcal{M}$ at depth D^*. Theorem 1 of Hajek (3) states that if $T_t \to 0$ and $\sum_{t=1}^{\infty} \exp[-D^*/T_t]$ diverges, then the sequence $(G_{SU}, \mathbf{Z}_{SU})_t$ converges in probability to an element of $\mathcal{M}$.

First, note that since $J(h,p) > 0$ for all h, D^* is bounded above by the maximum of J on $\mathcal{C}(G_R, \mathbf{d}_R, \mathbf{Z}_R)$, and so

$$\begin{aligned} D^* &\leq \max_{(G_{SU},\mathbf{Z}_{SU})\in\mathcal{C}(G_R,\mathbf{d}_R,\mathbf{Z}_R)} J\big(h(G_{SU}, \mathbf{Z}_{SU}), p(G_{SU}, G_R, \mathbf{t}_R, \mathbf{Z}_{SU})\big) \\ &\leq \max_{(h,p)\in[-1,1]^2} J(h,p) \\ &= \max_{(h,p)\in[-1,1]^2} 1/(1+\epsilon+h) \\ &= 1/\epsilon. \end{aligned} \tag{1}$$

Now examining the divergence criterion,

$$\begin{aligned} \sum_{t=1}^{\infty} \exp[-D^*/T_t] &= \sum_{t=1}^{\infty} \exp\left[-D^*\epsilon \log(t)\right] \\ &= \sum_{t=1}^{\infty} \frac{1}{t^{D^*\epsilon}} \\ &\geq \sum_{t=1}^{\infty} \frac{1}{t} = \infty \end{aligned} \tag{2}$$

where the inequality is a consequence of $D^*\epsilon \leq 1$. Therefore $\lim_{t\to\infty} \Pr\big((G_{SU}, \mathbf{Z}_{SU})_t \in \mathcal{M}\big) = 1$, as claimed. □

Web Appendix 2

Suppose $(G_{SU}, \mathbf{Z}_{SU}) \in \mathcal{C}(G_R, \mathbf{d}_R, \mathbf{Z}_R)$ is a compatible augmented subgraph and trait set, and we wish to propose another compatible pair $(G^*_{SU}, \mathbf{Z}^*_{SU}) \in \mathcal{C}(G_R, \mathbf{d}_R, \mathbf{Z}_R)$. We outline two proposal mechanisms. The first removes or adds an edge in G_{SU}. If necessary, a new unsampled vertex u is invented, and assigned a trait value Z_u. Let $U = \{u \in V_{SU} :\ u \notin V_R\}$ be the set of unsampled vertices. Furthermore, let $U_{-k} = \{u \in V_{SU} \setminus V_R :\ \{k,u\} \notin E_{SU}\}$ be the set of unsampled vertices in U that are *not* connected to $k \in V_R$.

1: Let $G^*_{SU} = G_{SU}$ and $\mathbf{Z}^*_{SU} = \mathbf{Z}_{SU}$
2: Randomly choose $i \in V_R$ and $j \in V_{SU}$ with $i \neq j$.

if $\{i,j\} \in E_{SU}$ and $\{i,j\} \notin E_R$ **then**
  Remove $\{i,j\}$ from E^*_{SU}
  $B \sim \text{Bernoulli}(1/2)$
  if $B < 0.5$ and $U_{-i} \neq \emptyset$ **then**
    Randomly choose $u \in U_{-i}$
  else
    Add a new vertex u to V^*_{SU}
    Randomly choose a trait $Z^*_u \in \{0,1\}$
  end if
  Add $\{i,u\}$ to E^*_{SU}
  if $j \in V_R$ **then**
    $B \sim \text{Bernoulli}(1/2)$
    if $B < 0.5$ and $U_{-j} \neq \emptyset$ **then**
      Randomly choose $u \in U_{-j}$
    else
      Add a new vertex u to V^*_{SU}
      Randomly choose a trait $Z^*_u \in \{0,1\}$
    end if
  end if
  Add $\{j,u\}$ to E^*_{SU}
else if $\{i,j\} \notin E_{SU}$ and $\exists u_1, u_2 \in U : \{i,u_1\} \in E_{SU}$ and $\{j,u_2\} \in E_{SU}$ **then**
  Remove $\{i,u_1\}$ and $\{j,u_2\}$ from E^*_{SU}
  Add $\{i,j\}$ to E^*_{SU}
end if
Remove any isolated vertices from V^*_{SU}

The space $\mathcal{C}(G_R, \mathbf{d}_R, \mathbf{Z}_R)$ is connected via proposals of this type (see 6, for explanation). The second proposal mechanism accelerates exploration of $\mathcal{C}(G_R, \mathbf{d}_R, \mathbf{Z}_R)$ by switching the trait of an unsampled vertex:

Choose $u \in \{u \in V_{SU} : u \notin V_R\}$.
Set $Z^*_u = 1 - Z_u$.

Together, these proposal mechanisms result in a well-mixing sequence $(G_{SU}, \mathbf{Z}_{SU})_t$.

References

[1] Kirkpatrick C. D., Vecchi M. P.. Optimization by simmulated annealing *Science.* 1983;220:671–680.

[2] Černỳ Vladimír. Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm *Journal of Optimization Theory and Applications.* 1985;45:41–51.

[3] Hajek Bruce. Cooling schedules for optimal annealing *Mathematics of Operations Research.* 1988;13:311–329.

[4] Bertsimas Dimitris, Tsitsiklis John. Simulated annealing *Statistical Science.* 1993;8:10–15.

[5] De Paula Aureo, Richards-Shubik Seth, Tamer Elie T. Identification of Preferences in Network Formation Games *SSRN 2577410.* 2014.

[6] Crawford Forrest W.. The graphical structure of respondent-driven sampling *Sociol Methodol.* 2016;46:187–211.