Online Appendix

Appendix A.

Estimating G_S and λ

A.1. Adding and Removing Edges

To estimate G_S, it is helpful to have a recipe for generating compatible subgraphs. Let $\mathcal{C}(G_R, \mathbf{d})$ denote the set of all recruitment-induced subgraphs that are compatible with the observed data G_R and $\mathbf{d}$ under Definition 4. To obtain a new compatible subgraph $\widehat{G}_S$ from a current compatible subgraph $G_S = (V_S, E_S)$, we randomly choose two vertices i and j, where $i \neq j$. If $\{i, j\} \notin E_S$, $\mathbf{u}_i > 0$, and $\mathbf{u}_j > 0$, then we propose to add the edge $\{i, j\}$ to E_S. Alternatively, if $\{i, j\} \in E_S$ and $\{i, j\} \notin E_R$, then we propose to remove the edge $\{i, j\}$ from E_S. If neither of these conditions hold, we pick another pair $\{i, j\}$ and try again. This procedure is described formally in the following algorithm.

```
loop
    Choose two vertices $i$ and $j$ at random, with $i < j$.
    if $\{i, j\} \notin E_S$ and $\mathbf{u}_i \geq 1$ and $\mathbf{u}_j \geq 1$ then
        let $E_S^+ = \{i, j\} \cup E_S$ and $G_S^+ = (V_S, E_S^+)$
        let $\mathbf{u}_k^+ = \mathbf{u}_k$ for all $k \neq i, j$ and $\mathbf{u}_i^+ = \mathbf{u}_i - 1$, $\mathbf{u}_j^+ = \mathbf{u}_j - 1$.
        return $G_S^+$ and $\mathbf{u}^+$
    else if $\{i, j\} \in E_S$ and $\{i, j\} \notin E_R$ then
        let $E_S^- = E_S \setminus \{i, j\}$ and $G_S^- = (V_S, E_S^-)$
        $\mathbf{u}_i^- = \mathbf{u}_i + 1$ and $\mathbf{u}_j^- = \mathbf{u}_j + 1$
        return $G_S^-$ and $\mathbf{u}^-$
    end if
end loop
```

This procedure chooses a vertex pair $\{i, j\}$ uniformly at random from all pairs whose change (addition or removal of the edge between i and j) would result in a compatible graph. The space of compatible subgraphs $\mathcal{C}(G_R, \mathbf{d})$ is connected via proposals of this type. To see why this is so, consider two compatible graphs G_S^1 and G_S^2 in $\mathcal{C}(G_R, \mathbf{d})$. Let $G_R^r = (V_S, E_R^r)$ be the undirected recruitment graph obtained by making each edge in the directed recruitment graph G_R reciprocal. By definition, G_R^r is a subgraph of every graph in

$\mathcal{C}(G_R, \mathbf{d})$. From G_S^1, we can obtain G_R^r by successively removing non recruitment edges, one at a time and each of these steps occurs with positive probability. From G_R^r we can obtain G_S^2 by adding non recruitment edges, one at a time. Since we can reach G_S^2 from G_S^1 in a similar manner, $\mathcal{C}(G_R, \mathbf{d})$ is connected via the given proposal algorithm.

A.2. Monte Carlo Sampling

To decide whether to accept a proposal G_S^* as a sample from the conditional distribution $p(G_S|\lambda, \mathbf{Y})$, we form the Metropolis-Hastings acceptance probability

$$\rho = \min\left\{1, \frac{L(\mathbf{w}|G_S^*, \lambda)}{L(\mathbf{w}|G_S, \lambda)} \cdot \frac{\Pr(G_S|G_S^*)}{\Pr(G_S^*|G_S)}\right\}, \tag{A.1}$$

and we accept the proposed graph G_S^* with probability ρ. Section A.3 below gives a simple and computationally efficient expression for the likelihood ratio, and Section A.3.1 gives a derivation of the proposal ratio $\Pr(G_S|G_S^*)/\Pr(G_S^*|G_S)$.

To sample λ conditional on G_S, we employ a Metropolis-Hastings step based on an approximation to the conditional distribution of λ. From the likelihood, we can easily find the maximum likelihood estimator of λ,

$$\hat{\lambda} = \frac{n - |M|}{\mathbf{s}'\mathbf{w}} \tag{A.2}$$

with asymptotic variance $\sigma^2 = \lambda^2/(n - |M|)$. Let

$$g(\lambda|G_S) = \frac{1}{\sqrt{2\pi}\sigma} \exp[-(\lambda - \hat{\lambda})^2/\sigma^2] \tag{A.3}$$

be a normal approximation to the conditional distribution of λ given G_S. Suppose λ is the current value and we propose a new value λ^* from $g(\lambda|G_S)$. We accept the proposal with probability

$$\rho = \min\left\{\frac{L(\mathbf{w}|G_S, \lambda^*)\, \pi(\lambda^*)}{L(\mathbf{w}|G_S, \lambda)\, \pi(\lambda)} \cdot \frac{g(\lambda|G_S)}{g(\lambda^*|G_S)}\right\}. \tag{A.4}$$

A.3. Computing the Likelihood Ratio

The ratio of likelihoods for two different compatible subgraphs can be computed very efficiently. Suppose we have generated a new subgraph by either adding or removing an edge between vertices i and j, where $t_i < t_j$. Since i was recruited before j, we have $t_i < t_j$. Let t_i^* be the minimum of the time that vertex i used its last coupon and t_n, the end of the study. Let $\mathrm{s} = \mathrm{lt}(\mathbf{AC})'\mathbf{1} + \mathbf{C}'\mathbf{u}$ where $\mathbf{A}$ is the adjacency matrix of a particular realization of G_S and let s^+ be the susceptible vector after addition of an edge between i and j, where $i < j$. Likewise let s^- be the susceptible vector after removal of an edge between i and j. The following result will be useful in computing the likelihood ratio in a simple way.

Lemma 1. *Given* $\mathbf{s}$*, i, and j, where $t_i < t_j$, the vectors* $\mathbf{s}^+$ *and* $\mathbf{s}^-$ *are given by*

$$\mathbf{s}_k^+ = \mathbf{s}_k - \mathbb{1}\{k > j\}C_{ik} - C_{jk} \tag{A.5}$$

$$\mathbf{s}_k^- = \mathbf{s}_k + \mathbb{1}\{k > j\}C_{ik} + C_{jk} \tag{A.6}$$

for $k = 1, \dots, n$.

A proof of Lemma 1 is given in Section D.5. The following Proposition establishes likelihood ratios for addition and removal of edges in G_S.

Proposition 1. *Suppose* $G_S = (V_S, E_S)$ *has* $\{i,j\} \notin E_S$, $\mathbf{u}_i \geq 1$*, and* $\mathbf{u}_j \geq 1$*. For a proposal* $G_S^+ = (V_S, E_S^+)$ *identical to* G_S *except that* $\{i,j\} \in E_S^+$*, the likelihood ratio is*

$$\frac{L(\mathbf{w}|G_S^+,\lambda)}{L(\mathbf{w}|G_S,\lambda)} = \left(\prod_{k \notin M} \frac{\mathbf{s}_k^+}{\mathbf{s}_k}\right) e^{\lambda(t_i^* - \min\{t_j, t_i^*\} + t_j^* - t_j)}. \tag{A.7}$$

Now suppose $G_S = (V_S, E_S)$ *has* $\{i,j\} \in E_S$ *and* $\{i,j\} \notin E_R$*. For a proposal* $G_S^- = (V_S, E_S^-)$ *identical to* G_S *except that* $\{i,j\} \notin E_S^-$*, the likelihood ratio is*

$$\frac{L(\mathbf{w}|G_S^-,\lambda)}{L(\mathbf{w}|G_S,\lambda)} = \left(\prod_{k \notin M} \frac{\mathbf{s}_k^-}{\mathbf{s}_k}\right) e^{-\lambda(t_i^* - \min\{t_j, t_i^*\} + t_j^* - t_j)}. \tag{A.8}$$

A proof of Proposition 1 is given in Section D.6. Note that the change in susceptible edge time, $t_i^* - \min\{t_j, t_i^*\} + t_j^* - t_j$, does not depend on any unknown parameters and can be computed in advance for every i and $j > i$. Likewise, $\mathbf{s}_k$ can be updated at each step using (A.5) and (A.6) without computing $\mathbf{s} = \mathrm{lt}(\mathbf{AC})'\mathbf{1} + \mathbf{C}'\mathbf{u}$ explicitly.

A.3.1 Proposal Ratio

To define the subgraph proposal ratio in (A.1), consider a given subgraph G_S. The number of possible vertex pairs between which an edge can be added is

$$\text{Add}(G_S) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{1}\{\{i,j\} \notin E_S \text{ and } \mathbf{u}_i \geq 1 \text{ and } \mathbf{u}_j \geq 1\}. \tag{A.9}$$

Likewise, for a proposed removal of an edge, the number of possible vertex pairs is

$$\text{Remove}(G_S) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{1}\{\{i,j\} \in E_S \text{ and } \{i,j\} \notin E_R\}. \tag{A.10}$$

Then the proposal probability for obtaining G_S^* from G_S is

$$\Pr(G_S^*|G_S) = 1/(\text{Add}(G_S) + \text{Remove}(G_S)), \tag{A.11}$$

and the ratio of backward and forward proposal probabilities for an addition is

$$\frac{\Pr(G_S|G_S^*)}{\Pr(G_S^*|G_S)} = \frac{1/(\text{Add}(G_S^*) + \text{Remove}(G_S^*))}{1/(\text{Add}(G_S) + \text{Remove}(G_S))} = \frac{\text{Add}(G_S) + \text{Remove}(G_S)}{\text{Add}(G_S^*) + \text{Remove}(G_S^*)}. \tag{A.12}$$

A.4. Maximum Likelihood and Maximum A Posteriori *Estimation*

We have described a Markov chain Monte Carlo algorithm for drawing from the posterior distribution of G_S. It is often faster to find a single "most likely" pair (G_S, λ) by a stochastic optimization algorithm known as "simulated annealing." To illustrate, let $\gamma > 0$ be a scale factor and let

$$L_\gamma(\mathbf{w}|G_S, \lambda) = \exp\left[-\left(\lambda \mathbf{s}'\mathbf{w} + \sum_{k \notin M} \log(\lambda \mathbf{s}_k)\right)/\gamma\right] \tag{A.13}$$

Define a sequence $\gamma_1, \gamma_2, \dots$ such that $\lim_{i\to\infty} \gamma_i = 0$. At each iteration i, we propose G_S^* and compute (A.1) with $L_{\gamma_i}(\mathbf{w}|G_S^*, \lambda)$ to accept the proposal with probability ρ. Once G_S is sampled, the most likely λ is obtained by (A.2). The joint sequence of sampled subgraphs and λ's tends toward the maximum likelihood estimates very rapidly.

A simple illustrative example of MAP estimation with $\lambda = 1$, $n = 50$, $|M| = 1$ seed, and three coupons per subject is shown in Figure A.1. The prior distribution of λ has mean 1 and SD 0.1. The population network is derived from the Project 90 network data. The top row shows the true subgraph G_S with the recruitment graph G_R overlaid (arrows indicate recruitment edges), adjacency matrices of G_R, G_S, and an estimate $\widehat{G}_S$. The bottom row shows the number of edges $|\widehat{E}_S|$ in the estimated subgraph at each iteration, the trace of λ, log posterior values, and accuracy.

Appendix B.

B.1. Validation Using Simulations

We analyze the performance of reconstruction on simulated Erdős-Rényi networks and a real-world network derived from a network study heterosexuals at high risk of contracting HIV in Colorado Springs, CO, from 1988-1990 called Project 90 [Woodhouse et al., 1994, Klovdahl et al., 1994, Rothenberg et al., 1995, Potterat et al., 2004]. We also evaluate reconstruction under mis-specification of the waiting time model, in which Assumption 5 is violated.

Let $\hat{\mathbf{A}}$ be the adjacency matrix of the estimated subgraph $\widehat{G}_S$ and let $\mathbf{A}$ be the adjacency matrix of the true subgraph G_S. We measure the accuracy, true positive rate (TPR), and true negative rate (TNR) of each estimated subgraph. These measures are defined as follows:

$$\begin{aligned}
\text{Accuracy}(\hat{\mathbf{A}}, \mathbf{A}) &= \sum_{i<j} \mathbb{1}\{\hat{\mathbf{A}}_{ij} = \mathbf{A}_{ij}\} \Big/ \binom{n}{2} \\
\text{TPR}(\hat{\mathbf{A}}, \mathbf{A}) &= \sum_{i<j} \mathbb{1}\{\hat{\mathbf{A}}_{ij} = 1 \text{ and } \mathbf{A}_{ij} = 1\} \Big/ \sum_{i<j} \mathbb{1}\{\hat{\mathbf{A}}_{ij} = 1\} \\
\text{TNR}(\hat{\mathbf{A}}, \mathbf{A}) &= \sum_{i<j} \mathbb{1}\{\hat{\mathbf{A}}_{ij} = 0 \text{ and } \mathbf{A}_{ij} = 0\} \Big/ \sum_{i<j} \mathbb{1}\{\hat{\mathbf{A}}_{ij} = 0\}\ .
\end{aligned} \tag{B.1}$$

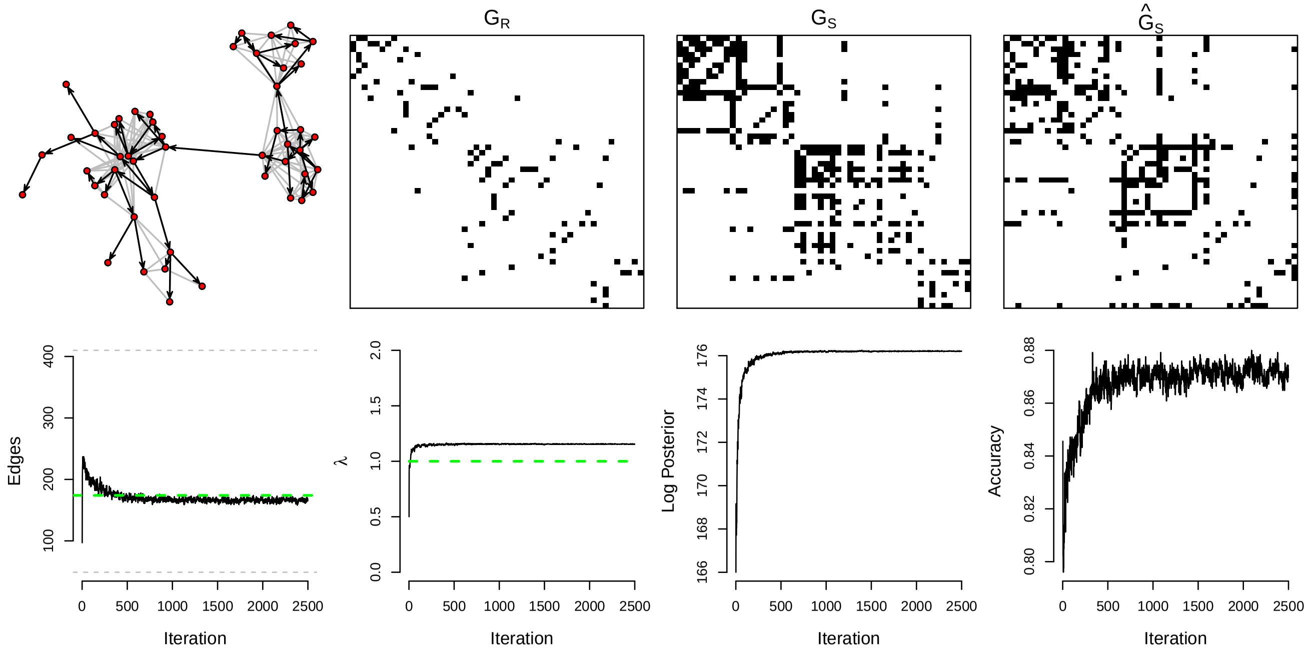

Figure A.1: Example maximum *a posteriori* reconstruction of G_S and estimation of λ from the Project 90 network and simulated RDS with $\lambda = 1$, $n = 50$, $|M| = 1$ seed, and three coupons per subject. The Gamma prior for λ has mean 1 and SD 0.1. The top row shows the recruitment-induced subgraph G_S with the directed recruitment graph G_R overlaid, and the adjacency matrices of G_R, G_S, and a sample reconstruction $\widehat{G}_S$ of G_S. The bottom row shows the traces of the estimated number of edges $|E_S|$, λ, the unnormalized log-posterior value, and the accuracy of $\widehat{G}_S$ as a predictor of G_S at each iteration. Accuracy is defined as the number of correct entries in the adjacency matrix of G_S divided by the total number of entries. The true number of edges is $|E_S| = 174$ and the number of edges in the estimated graph is 169. The estimated λ is 1.15.

In every simulation, we generate an RDS sample of $n = 500$ subjects, starting from $|M| = 10$ seeds with three coupons per recruit. Since the choice of time scale is arbitrary, we set $\lambda = 1$ for every simulation. We first evaluate reconstruction accuracy by simulating RDS on random undirected networks $G = (V, E)$ generated according to the Erdős-Rényi random graph model with with total population size $N = 1000$, 5000, and 10000 vertices and densities $p = 5/N$, $10/N$, and $15/N$. In addition, we evaluate the performance of reconstruction on a real-world network: Project 90 surveyed networks of heterosexuals at high risk of contracting HIV in Colorado Springs, CO, from 1988-1990 [Woodhouse et al., 1994, Klovdahl et al., 1994, Rothenberg et al., 1995, Potterat et al., 2004]. The network data G in the Project 90 data consist of $|V| = 5492$ individuals and $|E| = 43288$ edges. Network edges represent social, sexual, or drug use links between individuals. The Project 90 data have been used in other simulation studies to evaluate the performance of RDS estimators [Goel and Salganik, 2010].

Table B.1 shows estimate summaries; each row aggregates the results of 100 simulations on distinct networks. Conditional on each simulated network, we simulate the recruitment process and report the mean and SD of estimated quantities over the 100 repetitions. We report the parameters N and p used in the network simulation, the prior standard deviation (SD) of λ, the mean and standard deviation (SD) of accuracy, TPR, and TNR for reconstruction of G_S, and the mean and SD of estimates of λ. Accuracy and TNR are generally very high, with lower values of TPR. The high values of accuracy and TNR indicate that the reconstruction method recovers the true density of G_S fairly well on average. Assignment of the non-recruitment edges is more difficult, and TPR is lower. Figure A.1 shows an example in which the general structure of the adjacency matrix is recovered, but individual edges (shown as black entries in the adjacency matrix) may not always be correctly placed. The overall accuracy of edge inference depends on the pattern of coupon use and the structure of the recruitment graph. Accuracy is strongly affected by the proportion of recruitment edges in G_S: G_R is always a subgraph of G_S, so these edges are always present in estimates of G_S. Therefore simulated data sets with low edge density contain very few non-recruitment edges in G_S and hence reconstructions of G_S enjoy very high accuracy and high TNR, while TPR is lower. More dense subgraphs generally have higher TPR. Some estimates of λ exhibit small upward bias, which is reduced under more informative priors for λ.

In real-world RDS studies, Assumption 5 may be violated. It is therefore important to assess the performance of subgraph reconstruction and estimation of λ when the waiting time distribution is mis-specified and Assumption 5 does not hold. To this end, we draw random edge-wise waiting times T_{ij} from the Gamma distribution with density $f(t) = \lambda^\delta t^{\delta-1} e^{-\delta t}/\Gamma(\delta)$ with shape and rate parameters $\delta > 0$. Setting $\delta = 1$ recovers the Exponential$(\lambda = 1)$ distribution. When $0 < \delta < 1$, the Gamma density decays monotonically with the waiting time. When $\delta > 1$, the density has a nonzero mode. In simulations under this waiting time distribution, we fix the mean waiting time at $\mathbb{E}[T_{ij}] = 1$ and vary δ. In this way, δ provides a convenient continuous parameter to change the magnitude of mis-specification of the waiting time model.

Simulated Network		Prior	Accuracy		TPR		TNR		λ	
N	Np	SD$_\lambda$	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1000	5	1.00	0.994	2.9E-4	0.574	1.3E-2	0.997	2.1E-4	1.292	9.8E-2
		0.10	0.994	3.1E-4	0.589	1.5E-2	0.997	2.4E-4	1.093	2.6E-2
		0.01	0.994	2.8E-4	0.600	1.7E-2	0.997	2.1E-4	1.001	4.1E-4
	10	1.00	0.985	5.4E-4	0.429	1.4E-2	0.994	5.7E-4	1.401	1.0E-1
		0.10	0.986	4.9E-4	0.455	1.2E-2	0.994	4.8E-4	1.103	2.3E-2
		0.01	0.986	5.5E-4	0.468	1.5E-2	0.994	3.9E-4	1.001	3.8E-4
	15	1.00	0.977	7.2E-4	0.374	1.6E-2	0.991	8.2E-4	1.409	1.1E-1
		0.10	0.979	7.0E-4	0.403	1.4E-2	0.990	7.5E-4	1.104	2.6E-2
		0.01	0.980	8.3E-4	0.421	1.7E-2	0.990	6.3E-4	1.001	4.3E-4
5000	5	1.00	0.996	4.7E-4	0.542	3.2E-2	0.999	7.7E-5	1.401	9.7E-2
		0.10	0.997	5.0E-4	0.617	4.8E-2	0.999	8.2E-5	1.141	2.6E-2
		0.01	0.998	5.7E-4	0.717	7.3E-2	0.999	6.7E-5	1.002	7.4E-4
	10	1.00	0.990	1.0E-3	0.339	2.6E-2	0.999	1.3E-4	1.540	1.1E-1
		0.10	0.993	1.1E-3	0.440	4.8E-2	0.999	1.3E-4	1.149	2.3E-2
		0.01	0.995	1.1E-3	0.537	8.2E-2	0.999	1.1E-4	1.003	6.1E-4
	15	1.00	0.984	1.5E-3	0.258	2.0E-2	0.998	1.7E-4	1.576	1.4E-1
		0.10	0.989	1.7E-3	0.355	4.5E-2	0.998	1.7E-4	1.150	2.5E-2
		0.01	0.992	1.9E-3	0.472	9.0E-2	0.998	1.5E-4	1.003	8.0E-4
10000	5	1.00	0.996	4.4E-4	0.546	3.2E-2	1.000	5.5E-5	1.434	1.1E-1
		0.10	0.997	5.2E-4	0.635	5.0E-2	1.000	5.4E-5	1.146	2.9E-2
		0.01	0.998	5.9E-4	0.727	7.7E-2	1.000	5.7E-5	1.003	6.7E-4
	10	1.00	0.991	1.2E-3	0.337	3.8E-2	0.999	7.7E-5	1.541	1.4E-1
		0.10	0.994	1.0E-3	0.441	4.7E-2	0.999	7.9E-5	1.158	2.6E-2
		0.01	0.996	1.2E-3	0.569	9.4E-2	0.999	7.8E-5	1.003	7.3E-4
	15	1.00	0.986	1.6E-3	0.246	2.3E-2	0.999	1.1E-4	1.578	1.1E-1
		0.10	0.991	1.8E-3	0.354	5.7E-2	0.999	1.0E-4	1.161	2.7E-2
		0.01	0.994	1.8E-3	0.481	1.1E-1	0.999	9.0E-5	1.003	7.4E-4
Project 90		1.00	0.973	1.6E-3	0.376	2.6E-2	0.989	1.1E-3	1.263	1.0E-1
		0.10	0.974	1.3E-3	0.370	2.5E-2	0.988	9.5E-4	1.085	3.6E-2
		0.01	0.974	1.5E-3	0.374	2.3E-2	0.987	9.3E-4	1.001	4.1E-4

Table B.1: Simulation results for RDS on Erdős-Rényi random networks and the Project 90 network. The recruitment rate λ in every simulation is 1. Each row summarizes 100 simulations of RDS on different random networks under the given network parameters N and p. The prior SD of λ is given in the next column. The mean and SD of accuracy, TPR, and TNR for reconstruction of G_S and the mean MAP estimate and SD of λ are in the last three columns. Table B.2 shows simulation results for RDS on the Project 90 network.

	Prior	Accuracy		TPR		TNR		λ	
δ	SD$_\lambda$	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.5	1.00	0.974	1.6E-3	0.394	2.9E-2	0.990	1.0E-3	19.816	3.4E-0
	0.10	0.986	1.1E-3	0.969	6.0E-2	0.986	9.1E-4	3.872	2.3E-1
	0.01	0.986	9.3E-4	1.000	3.3E-4	0.986	9.4E-4	1.043	8.8E-4
0.6	1.00	0.974	1.5E-3	0.387	3.0E-2	0.990	1.0E-3	8.748	1.4E-0
	0.10	0.985	1.9E-3	0.814	1.2E-1	0.986	9.2E-4	2.698	1.7E-1
	0.01	0.986	9.6E-4	0.998	8.2E-3	0.986	9.6E-4	1.037	1.3E-3
0.7	1.00	0.973	1.5E-3	0.385	2.6E-2	0.989	1.0E-3	4.390	5.9E-1
	0.10	0.981	2.2E-3	0.596	9.1E-2	0.986	1.0E-3	1.885	9.4E-2
	0.01	0.986	1.1E-3	0.977	4.6E-2	0.986	1.0E-3	1.026	2.4E-3
0.8	1.00	0.973	1.4E-3	0.383	2.5E-2	0.989	9.4E-4	2.710	2.6E-1
	0.10	0.978	2.0E-3	0.464	4.3E-2	0.987	1.0E-3	1.441	7.7E-2
	0.01	0.984	2.3E-3	0.794	1.4E-1	0.986	1.1E-3	1.013	2.6E-3
0.9	1.00	0.974	1.6E-3	0.380	2.5E-2	0.989	1.0E-3	1.736	1.6E-1
	0.10	0.975	1.6E-3	0.399	2.3E-2	0.987	9.3E-4	1.197	3.9E-2
	0.01	0.978	2.3E-3	0.458	6.2E-2	0.987	1.0E-3	1.004	1.1E-3
1.1	1.00	0.974	1.7E-3	0.379	2.7E-2	0.989	1.1E-3	0.929	7.2E-2
	0.10	0.973	1.5E-3	0.380	2.7E-2	0.989	1.0E-3	0.970	3.6E-2
	0.01	0.973	1.6E-3	0.374	2.5E-2	0.989	1.0E-3	0.999	8.2E-4
1.2	1.00	0.973	1.5E-3	0.373	2.4E-2	0.988	1.0E-3	0.731	5.3E-2
	0.10	0.974	1.7E-3	0.394	2.8E-2	0.990	9.4E-4	0.834	3.3E-2
	0.01	0.974	2.0E-3	0.405	3.3E-2	0.991	1.0E-3	0.995	1.6E-3
1.3	1.00	0.973	1.5E-3	0.370	2.3E-2	0.988	1.2E-3	0.587	4.4E-2
	0.10	0.974	1.9E-3	0.401	3.2E-2	0.990	1.0E-3	0.724	3.0E-2
	0.01	0.976	2.1E-3	0.436	3.4E-2	0.993	1.0E-3	0.987	2.4E-3
1.4	1.00	0.973	1.5E-3	0.366	2.4E-2	0.988	1.1E-3	0.488	3.6E-2
	0.10	0.974	1.8E-3	0.404	2.9E-2	0.990	1.0E-3	0.626	2.4E-2
	0.01	0.977	1.9E-3	0.460	3.3E-2	0.994	9.8E-4	0.978	3.1E-3
1.5	1.00	0.973	1.6E-3	0.352	2.2E-2	0.988	1.1E-3	0.412	3.0E-2
	0.10	0.974	2.1E-3	0.400	3.4E-2	0.990	1.0E-3	0.542	2.1E-2
	0.01	0.978	2.1E-3	0.481	3.6E-2	0.995	9.1E-4	0.968	3.4E-3

Table B.2: Simulation results for RDS on the Project 90 data under mis-specification of the waiting time distribution. The mean waiting time to recruitment is distributed as Gamma(δ, δ), for the given values of δ. When $\delta = 1$, the waiting time distribution is correctly specified; these results are given in the last three lines of Table B.1.

Table B.2 gives the results for Gamma-distributed waiting times, where δ is specified. The prior SD of λ, accuracy, TPR, TNR, and the mean and SD of estimates of λ are given. In general, accuracy, TPR, and TNR are roughly the same as in the Exponential($\lambda = 1$) case. But as expected, estimates of λ under a mis-specified waiting time distribution appear to be subject to bias. When $\delta < 1$, λ is typically overestimated; when $\delta > 1$, λ is usually underestimated. We can explain the relative robustness of reconstruction by recalling two features of the proposed framework. First, the compatibility conditions (Definition 4) impose strong constraints on the topology of G_S when G_R and $\mathbf{d}$ are observed. These constraints are in effect regardless of whether the waiting time model is correctly specified, and serve to ensure that all edges in G_R are correctly estimated. Second, under any model in which the recruitment times across edges are independent, the rate of new recruitments is positively associated with the number of susceptible edges. Under the exponential model this relationship is linear, and under other waiting time models the relationship may be non-linear (and in general depends on time waited along each susceptible edge up to the current time).

Appendix C.

C.1. Prior for λ in the St. Petersburg Application

To obtain a sensible prior distribution for the edge-wise recruitment rate λ, we adopt an empirical Bayesian approach based on bounding the maximum likelihood estimate of λ given by (A.2). First, observe that the maximum number of susceptible edges that can be added by a vertex i with degree d_i is d_i minus 1 if i was recruited by another subject, $d_i - \mathbb{1}\{i \notin M\}$. Then the maximum number of susceptible edges at step k in the recruitment process is $\mathrm{s}_k = \sum_{i=1}^{k-1}(d_i - \mathbb{1}\{i \notin M\})$. Therefore a minimum estimate of λ is

$$\lambda_{\text{lo}} = \frac{n - |M|}{\sum_{k=1}^{n} \mathbf{w}_i \sum_{i=1}^{k-1}(d_i - \mathbb{1}\{i \notin M\})}.$$

Now let n_i be the number of subjects recruited by subject i. The minimum number of susceptible edges at step k is $\mathrm{s}_k = \sum_{i=1}^{k-1}(n_i - \mathbb{1}\{i \notin M\})$, and the maximum estimate of λ is

$$\lambda_{\text{hi}} = \frac{n - |M|}{\sum_{k=1}^{n} \mathbf{w}_i \sum_{i=1}^{k-1}(n_i - \mathbb{1}\{i \notin M\})}.$$

Applying these bounds to the St. Petersburg data yields $\lambda_{\text{lo}} = 9.8 \times 10^{-4}$ and $\lambda_{\text{hi}} = 4.2 \times 10^{-2}$. We therefore specify a prior distribution for λ that takes most of its mass in the interval $[\lambda_{\text{lo}}, \lambda_{\text{hi}}]$. Let $\lambda_{\text{mean}} = (\lambda_{\text{lo}} + \lambda_{\text{hi}})/2 = 0.022$. Suppose $\eta > 0$ is given and let $\xi = \eta/\lambda_{\text{mean}}$. Now by varying η, we obtain a family of Gamma prior distributions with mean λ_{mean}.

Appendix D.

D.1. Proof of Proposition 1

Let $u \in R$ be a particular recruiter and let S_u be the set of susceptible vertices that are neighbors of u at a given time in the recruitment process. Let W_{ux} be the waiting time for u to recruit its susceptible neighbor $x \in S_u$. By Assumption 5, $W_{ux} \sim \text{Exponential}(\lambda)$ independently for each $x \in S_u$. Given that u recruits a random vertex X in S_u before any other recruiter, define the first recruited vertex to be

$$X = \operatorname*{argmin}_{x \in S_u} W_{ux}. \tag{D.1}$$

We follow the competing risks perspective of Lange [2010, 188] and consider the joint probability

$$\begin{aligned} \Pr(X = x, W_{ux} \geq t) &= \Pr(W_{ux} \geq t, W_{uk} > W_{ux} \text{ for all } k \neq x) \\ &= \int_t^\infty \lambda e^{-\lambda s} \Pr(W_{uk} > s \text{ for all } k \neq x)\, \mathrm{d}s \\ &= \int_t^\infty \lambda e^{-\lambda s} \prod_{\substack{k \in S_u \\ k \neq x}} e^{-\lambda s}\, \mathrm{d}s \\ &= \frac{1}{|S_u|} e^{-\lambda |S_u| t}. \end{aligned} \tag{D.2}$$

Therefore X is recruited uniformly at random from S_u, the waiting time to this recruitment has distribution Exponential$(\lambda|S_u|)$, and X and W_{uX} are independent. □

D.2. Proof of Proposition 2

Let $W_u = \min_{x \in S_u} W_{ux}$ be the waiting time to the first recruitment by recruiter $u \in R$ and let $W = \min_{u \in R} \min_{x \in S_u} W_{ux}$ be the waiting time to the first recruitment by any recruiter. The first recruiter is $U = \text{argmin}_{u \in R} W_u$ and the first recruited vertex is $X = \text{argmin}_{x \in S_U} W_{Ux}$. We again consider the joint probability of the recruited vertex $X = x$

and the waiting time W_{ux},

$$\begin{aligned}\Pr(X=x, W_{Ux}\geq t) &= \sum_{u\in R}\Pr(W_{ux}\geq t,\ W_{jk}>W_{ux} \text{ for all } k\in R, j\in S, \{u,x\}\neq\{j,k\})\,\mathbb{1}\{u\in R_x\}\\ &= \sum_{u\in R_x}\int_t^\infty \lambda e^{-\lambda s}\Pr(W_{jk}>s \text{ for all } k\in R, j\in S, \{u,x\}\neq\{j,k\})\,\mathrm{d}s\\ &= \sum_{u\in R_x}\int_t^\infty \lambda e^{-\lambda s}\prod_{j\in R}\prod_{\substack{k\in S_j\\ \{j,k\}\neq\{u,x\}}} e^{-\lambda s}\,\mathrm{d}s\\ &= \sum_{u\in R_x}\frac{1}{\sum_{j\in R}|S_j|}\exp\left[-\lambda t\sum_{j\in R}|S_j|\right]\\ &= \frac{|R_x|}{\sum_{k\in S}|R_k|}\exp\left[-\lambda t\sum_{j\in R}|S_j|\right]\end{aligned} \tag{D.3}$$

where the last line is obtained because $\sum_{j\in R}|S_j| = \sum_{k\in S}|R_k|$. Therefore $X\in S$ is the first recruit with probability proportional to the number of recruiters it has (equivalently, the number of susceptible edges incident to it), the waiting time W_{UX} to the first recruitment is Exponential($\lambda\sum_{j\in R}S_j$), and X and W_{UX} are independent. □

D.3. Proof of Corollary 1

Proposition 2 shows that the probability that a given vertex $x\in S$ is recruited at the next step in the sampling process under the model described in this paper is $|R_v|/\sum_{k\in S}|R_k|$. Gile and Handcock [2010] describe a recruitment process in which the recruiter u is chosen first, without regard to the number of susceptible vertices linked to it. Then, conditional on the identity of the chosen u, a susceptible neighbor $x\in S_u$ is recruited with uniform probability $1/|S_u|$. Marginalizing over the recruiter u, we find that the probability of recruiting vertex $x\in S$ in the model of Gile and Handcock [2010] is

$$\begin{aligned}\Pr(x\in S \text{ is recruited}) &= \sum_{u\in R}\Pr(u \text{ is recruiter})\Pr(x \text{ is recruited}|u \text{ is recruiter})\\ &= \sum_{u\in R}\frac{1}{|R|}\frac{\mathbb{1}\{x\in S_u\}}{|S_u|}\\ &= \frac{1}{|R|}\sum_{u\in R_x}\frac{1}{|S_u|},\end{aligned} \tag{D.4}$$

where the last line is obtained because $x\in S_u$ if and only if $u\in R_x$. In general, this probability distribution is not equal to $|R_v|/\sum_{k\in S}|R_k|$. □

D.4. Proof of Proposition 3

We first give a rigorous definition of the coupon matrix $\mathbf{C}$. Define the function $C_i(t)$ to be 1 if subject i has at least one coupon just before time t, and zero otherwise. The function $C_i(t)$ is left-continuous. Let t_j be the time of the jth recruitment event. Then define the i,jth element of $\mathbf{C}$ as

$$\mathbf{C}_{ij} = \lim_{t \to t_j^-} C_i(t). \tag{D.5}$$

where $t \to t_j-$ means that t approaches t_j from the left. Recall that $\mathbf{A}$ is the adjacency matrix of the recruitment-induced subgraph, with rows and columns in the order of vertices' recruitment into the study. The i,jth element of the matrix product $\mathbf{AC}$ is the number of recruiters connected to i just before the time t_j of the jth recruitment. Then

$$\{\mathbf{AC}\}_{ij} = \sum_{k=1}^{n} \mathbf{A}_{ik}\mathbf{C}_{kj} \tag{D.6}$$

is the number of possible recruiters of subject i at time t_j^-, and

$$\sum_{i=j}^{n}\sum_{k=1}^{n} \mathbf{A}_{ik}\mathbf{C}_{kj} \tag{D.7}$$

is the number of susceptible edges at time t_j^- connecting recruiters to vertices that will eventually be sampled. Recruiters may also have connections to vertices that are never recruited into the study. The ith element of the $n \times 1$ vector $\mathbf{u}$ is the number of pendant edges connecting vertex i to unknown/unsampled vertices. Each of these pendant edges is susceptible while i is a recruiter, so the number of susceptible pendant edges just before time t_j is

$$\sum_{i=j}^{n} \mathbf{C}_{ij}\mathbf{u}_i. \tag{D.8}$$

Finally, the total number of susceptible edges just before time t_j is the sum of (D.7) and (D.8),

$$\mathbf{s}_j = \sum_{i=j}^{n} \left(\sum_{k=1}^{n} \mathbf{A}_{ik}\mathbf{C}_{kj} \right) + \mathbf{C}_{ij}\mathbf{u}_i \tag{D.9}$$

and in vector form, $\mathbf{s} = \text{lt}(\mathbf{AC})'\mathbf{1} + \mathbf{C}'\mathbf{u}$. Now let $\mathbf{w} = (0, t_2 - t_1, \ldots, t_n - t_{n-1})$ be the $n \times 1$ vector of waiting times between recruitments. By Proposition 2, the random waiting time between recruitment of subject $j-1$ and j has distribution Exponential$(\lambda \mathbf{s}_j)$. For recruited vertices j, this waiting time $\mathbf{w}_j$ is fully observed and has density $\lambda \mathbf{s}_j \exp[-\lambda \mathbf{s}_j \mathbf{w}_j]$ where $j \notin M$, where M is the set of seeds. In contrast, seeds are recruited not by other vertices, but by another mechanism. If a seed j shares edges with any recruiters before it is chosen as a seed, we observe that the actual waiting time to its recruitment must be greater than the waiting time actually observed, so the density of the waiting time $\mathbf{w}_j$ of a seed is $\exp[-\lambda \mathbf{s}_j \mathbf{w}_j]$. Therefore, the full likelihood of the recruitment time series is

$$\begin{aligned} L(\mathbf{w}|G_S, \lambda) &= \prod_{i=1}^{n} (\lambda \mathbf{s}_i)^{\mathbb{1}\{i \notin M\}} \exp[-\lambda \mathbf{s}_i \mathbf{w}_i] \\ &= \left(\prod_{i \notin M} \lambda \mathbf{s}_i \right) \exp[-\lambda \mathbf{s}' \mathbf{w}], \end{aligned} \tag{D.10}$$

where M is the set of seeds, as claimed. □

D.5. Proof of Lemma 1

Consider the adjacency matrix $\mathbf{A}$ of the current estimate of the recruitment-induced subgraph G_S and suppose we would like to add an edge between i and j, where $t_i < t_j$, $\mathbf{A}_{ij} = \mathbf{A}_{ji} = 0$, $\mathbf{u}_i \geq 1$, and $\mathbf{u}_j \geq 1$. Define the proposal graph as G_S^+ with adjacency matrix $\mathbf{A}^+$ to be a matrix identical to $\mathbf{A}$, except that $\mathbf{A}^+_{ij} = \mathbf{A}^+_{ji} = 1$, with $\mathbf{u}^+$ identical to $\mathbf{u}$ except $\mathbf{u}^+_i = \mathbf{u}_i - 1$ and $\mathbf{u}^+_j = \mathbf{u}_j - 1$. By (D.9),

$$\begin{aligned} \mathbf{s}^+_k &= \sum_{x=k}^{n} \left(\sum_{y=1}^{n} \mathbf{A}^+_{xy} \mathbf{C}_{yk} \right) + \mathbf{C}_{xk} \mathbf{u}^+_x \\ &= \sum_{x=k}^{n} \left(\sum_{y=1}^{n} (\mathbf{A}_{xy} + \mathbb{1}\{x = i, y = j\} + \mathbb{1}\{x = j, y = i\}) \mathbf{C}_{yk} \right) + \mathbf{C}_{xk} \mathbf{u}_x \\ &\quad - \mathbf{C}_{xk} (\mathbb{1}\{x = i\} + \mathbb{1}\{x = j\}) \\ &= \mathbf{s}_k + \mathbb{1}\{i \geq k\} \mathbf{C}_{jk} + \mathbb{1}\{j \geq k\} \mathbf{C}_{ik} - \mathbb{1}\{i \geq k\} \mathbf{C}_{ik} - \mathbb{1}\{j \geq k\} \mathbf{C}_{jk} \\ &= \mathbf{s}_k + 0 + (1 - \mathbb{1}\{j < k\}) \mathbf{C}_{ik} - (1 - \mathbb{1}\{i < k\}) \mathbf{C}_{ik} - (1 - \mathbb{1}\{j < k\}) \mathbf{C}_{jk} \\ &= \mathbf{s}_k - \mathbf{C}_{ik} \mathbb{1}\{k > j\} - \mathbf{C}_{jk}, \end{aligned} \tag{D.11}$$

where the last line is obtained because $\mathbf{C}_{jk} = 0$ for $k < j$.

Now we consider removing an edge between i and j, where $t_i < t_j$. Suppose the current estimate of the recruitment-induced subgraph is $\mathbf{A}$ with $\mathbf{A}_{ij} = \mathbf{A}_{ji} = 1$ with no recruitment taking place across this edge, $\{i, j\} \notin E_R$. Define the proposal graph as G_S^- with adjacency matrix $\mathbf{A}^-$ to be a matrix identical to $\mathbf{A}$, except that $\mathbf{A}^-_{ij} = \mathbf{A}^-_{ji} = 0$, with $\mathbf{u}^-$ identical to $\mathbf{u}$ except $\mathbf{u}^-_i = \mathbf{u}_i + 1$ and $\mathbf{u}^-_j = \mathbf{u}_j + 1$. Then the number of susceptible edges just before the time t_k of the kth recruitment is

$$\begin{aligned}
\mathbf{s}^-_k &= \sum_{x=k}^{n} \left(\sum_{y=1}^{n} \mathbf{A}^-_{xy} \mathbf{C}_{yk} \right) + \mathbf{C}_{xk} \mathbf{u}^-_x \\
&= \sum_{x=k}^{n} \left(\sum_{y=1}^{n} (\mathbf{A}_{xy} - \mathbb{1}\{x = i, y = j\} - \mathbb{1}\{x = j, y = i\}) \mathbf{C}_{yk} \right) + \mathbf{C}_{xk} \mathbf{u}_x \\
&\quad + \mathbf{C}_{xk} (\mathbb{1}\{x = i\} + \mathbb{1}\{x = j\}) \\
&= \mathbf{s}_k - \mathbb{1}\{i \geq k\} \mathbf{C}_{jk} - \mathbb{1}\{j \geq k\} \mathbf{C}_{ik} + \mathbb{1}\{i \geq k\} \mathbf{C}_{ik} + \mathbb{1}\{j \geq k\} \mathbf{C}_{jk} \\
&= \mathbf{s}_k + \mathbb{1}\{k > j\} \mathbf{C}_{ik} + \mathbf{C}_{jk}. \quad \square
\end{aligned} \tag{D.12}$$

D.6. Proof of Proposition 1

For the addition of an edge between i and j with $t_i < t_j$, the likelihood ratio is

$$\begin{aligned}
\frac{L(\mathbf{w}|G^+_S, \lambda)}{L(\mathbf{w}|G_S, \lambda)} &= \left(\prod_{k \notin M} \frac{\mathbf{s}^+_k}{\mathbf{s}_k} \right) \exp\left[-\lambda (\mathbf{s}^+ - \mathbf{s})' \mathbf{w} \right] \\
&= \left(\prod_{k \notin M} \frac{\mathbf{s}^+_k}{\mathbf{s}_k} \right) \exp\left[\lambda \sum_{k=1}^{n} (\mathbf{C}_{ik} \mathbb{1}\{k > j\} + \mathbf{C}_{jk}) \mathbf{w}_k \right].
\end{aligned} \tag{D.13}$$

We have

$$\sum_{k=1}^{n} \mathbf{C}_{ik} \mathbb{1}\{k > j\} \mathbf{w}_k = t^*_i - \min\{t^*_i, t_j\} \tag{D.14}$$

and

$$\sum_{k=1}^{n} \mathbf{C}_{jk} \mathbf{w}_k = t^*_j - t_j. \tag{D.15}$$

Then the ratio becomes

$$\frac{L(\mathbf{w}|G^+_S, \lambda)}{L(\mathbf{w}|G_S, \lambda)} = \left(\prod_{k \notin M} \frac{\mathbf{s}^+_k}{\mathbf{s}_k} \right) \exp\left[\lambda (t^*_i - \min\{t_j, t^*_i\} + t^*_j - t_j) \right]. \tag{D.16}$$

For the removal of an edge between i and j with $t_i < t_j$, the same arguments apply. The likelihood ratio is

$$\begin{aligned}
\frac{L(\mathbf{w}|G^-_S, \lambda)}{L(\mathbf{w}|G_S, \lambda)} &= \left(\prod_{k \notin M} \frac{\mathbf{s}^-_k}{\mathbf{s}_k} \right) \exp\left[-\lambda (\mathbf{s}^- - \mathbf{s})' \mathbf{w} \right] \\
&= \left(\prod_{k \notin M} \frac{\mathbf{s}^-_k}{\mathbf{s}_k} \right) \exp\left[-\lambda \sum_{k=1}^{n} (\mathbf{C}_{ik} \mathbb{1}\{k > j\} + \mathbf{C}_{jk}) \mathbf{w}_k \right] \\
&= \left(\prod_{k \notin M} \frac{\mathbf{s}^-_k}{\mathbf{s}_k} \right) \exp\left[-\lambda (t^*_i - \min\{t_j, t^*_i\} + t^*_j - t_j) \right],
\end{aligned} \tag{D.17}$$

as claimed. □

References

Krista J Gile and Mark S Handcock. Respondent-driven sampling: an assessment of current methodology. *Sociological Methodology*, 40(1):285–327, 2010.

Sharad Goel and Matthew J Salganik. Assessing respondent-driven sampling. *Proceedings of the National Academy of Sciences*, 107(15):6743–6747, 2010.

Alden S Klovdahl, John J Potterat, Donald E Woodhouse, John B Muth, Stephen Q Muth, and William W Darrow. Social networks and infectious disease: The Colorado Springs study. *Social Science & Medicine*, 38(1):79–88, 1994.

Kenneth Lange. *Applied Probability*. Springer texts in statistics. Springer New York, 2nd edition, 2010.

John J Potterat, Donald E Woodhouse, Steven Q Muth, Richard B Rothenberg, William W Darrow, Alden S Klovdahl, and John B Muth. Network dynamism: History and lessons of the Colorado Springs study. In Martina Morris, editor, *Network Epidemiology: A Handbook for Survey Design and Data Collection*, pages 87–114. Oxford University Press, 2004.

Richard B Rothenberg, Donald E Woodhouse, John J Potterat, Stephen Q Muth, William W Darrow, and Alden S Klovdahl. Social networks in disease transmission: the Colorado Springs study. In Richard Needle, Sander G Genser, and Robert T Trotter, editors, *Social Networks, Drug Abuse, and HIV Transmission*. US Department of Health and Human Services, Public Health Service, National Institutes of Health, National Institute on Drug Abuse, 1995.

Donald E Woodhouse, Richard B Rothenberg, John J Potterat, William W Darrow, Stephen Q Muth, Alden S Klovdahl, Helen P Zimmerman, Helen L Rogers, Tammy S Maldonado, John B Muth, and Judith U Reynolds. Mapping a social network of heterosexuals at high risk for HIV infection. *AIDS*, 8(9):1331–1336, 1994.