Seeing the Unseen Network: Inferring Hidden Social Ties from Respondent-Driven Sampling

Lin Chen[1,2] and **Forrest W. Crawford**[2,3] and **Amin Karbasi**[1,2]
[1]Department of Electrical Engineering, [2]Yale Institute for Network Science, [3]Department of Biostatistics
Yale University, New Haven, CT 06520
{lin.chen, forrest.crawford, amin.karbasi}@yale.edu

Abstract

Learning about the social structure of hidden and hard-to-reach populations — such as drug users and sex workers — is a major goal of epidemiological and public health research on risk behaviors and disease prevention. Respondent-driven sampling (RDS) is a peer-referral process widely used by many health organizations, where research subjects recruit other subjects from their social network. In such surveys, researchers observe who recruited whom, along with the time of recruitment and the total number of acquaintances (network degree) of respondents. However, due to privacy concerns, the identities of acquaintances are not disclosed. In this work, we show how to reconstruct the underlying network structure through which the subjects are recruited. We formulate the dynamics of RDS as a continuous-time diffusion process over the underlying graph and derive the likelihood of the recruitment time series under an arbitrary inter-recruitment time distribution. We develop an efficient stochastic optimization algorithm called RENDER (REspoNdent-Driven nEtwork Reconstruction) that finds the network that best explains the collected data. We support our analytical results through an exhaustive set of experiments on both synthetic and real data.

Introduction

Random sampling is an effective way for researchers to learn about a population of interest. Characteristics of the sample can be generalized to the population of interest using standard statistical theory. However, traditional random sampling may be impossible when the target population is hidden, e.g. drug users, men who have sex with men, sex workers and homeless people. Due to concerns like privacy, stigmatization, discrimination, and criminalization, members of hidden populations may be reluctant to participate in a survey. In addition to the hidden populations for which no sampling "frame" exists, there are rare populations of great research interest, marked by their tiny fractions in the entire population. It is unlikely that random sampling from the general population would accrue a reasonable sample from a very rare population. However, it is often the case that members of hidden or rare populations know each other socially. This suggests that social referral would be an effective method for accruing a large sample. To this end, researchers have developed a variety of social link-tracing survey designs. The most popular is *respondent-driven sampling* (RDS) (Heckathorn 1997).

In RDS, a small set of initial subjects, known as "seeds" are selected, not necessarily simultaneously, from the target population. Subjects are given a few coupons, each tagged with a unique identifier, which they can use to recruit other eligible subjects. Participants are given a reward for being interviewed and for each eligible subject they recruit using their coupons. Each subject reports their degree, the number of others whom they know in the target population. No subject is permitted to enter the study twice and the date and time of each interview is recorded.

While RDS can be an effective recruitment method, it reveals only incomplete social network data to researchers. Any ties between recruited subjects along which no recruitment took place remain unobserved. The social network of recruited subjects is of great interest to sociologists, epidemiologists and public health researchers since it may induce dependencies in the outcomes of sampled individuals. Fortunately, RDS reveals information about the underlying social network that can be used to (approximately) reconstruct it. By leveraging the time series of recruitments, the degrees of recruited subjects, coupon information, and who recruited whom, it is possible to interpret the induced subgraph of RDS respondents as a simple random graph model (Crawford 2016).

In this paper, we introduce a flexible and expressive stochastic model of RDS recruitment on a partially observed network structure. We derive the likelihood of the observed time series; the model admits any edgewise inter-recruitment time distribution. We propose a stochastic optimization algorithm RENDER (REspoNdent-Driven nEtwork Reconstruction) to estimate unknown parameters and the underlying social network. We conduct extensive experiments, on synthetic and real data, to confirm the accuracy and the reconstruction performance of RENDER. In particular, we apply RENDER to reconstruct the network of injection drug users from an RDS study in St. Petersburg, Russia.

Related Work

RDS has been modeled as a random walk with replacement on a graph at its equilibrium distribution (Heckathorn 1997; Salganik and Heckathorn 2004; Volz and Heckathorn 2008;

Goel and Salganik 2009; Gile and Handcock 2010), and under this argument the sampling probability is proportional to degree, which is the basis for an estimator of the population mean (Heckathorn 2002; Salganik 2006). In this paper, we adopt an approach that focuses on the network structure estimable from the RDS sample using recruitment information (Crawford 2016). Crawford (2016) assumes that edgewise inter-recruitment times follow the exponential distribution, but this approach is relatively inflexible. In many other contexts, dynamic or random processes can reveal structural information of an underlying, partially observed, network (Kramer et al. 2009; Shandilya and Timme 2011; Linderman and Adams 2014). Network reconstruction and the edge prediction problem have been studied for diffusion processes where multiple realizations of the process on the same network are available (Liben-Nowell and Kleinberg 2007; Gomez Rodriguez, Leskovec, and Krause 2010; Gomez Rodriguez et al. 2011). In the case of RDS, reconstruction is particularly challenging because only a single realization of the diffusion process is observed. Furthermore, we must account for the role of coupons in recruitment as they intricately introduce bias. However, in contrast to general diffusion processes over graphs, some important network topological information is revealed by RDS. In this study, we leverage all the available data routinely collected during real-world RDS studies.

Problem Formulation

We conform to the following notation throughout the paper. Suppose that f is a real-valued function and that $\mathbf{v}$ is a vector. Then $f(\mathbf{v})$ is a vector of the same size as vector $\mathbf{v}$ and the i-th entry of $f(\mathbf{v})$ is denoted by $f(\mathbf{v})_i$, whose value is given by $f(\mathbf{v}_i)$. The transposes of matrix $\mathbf{A}$ and vector $\mathbf{v}$ are written as $\mathbf{A}'$ and $\mathbf{v}'$, respectively. And let $\mathbf{1}$ be the all-ones column vector.

Dynamics of Respondent-Driven Sampling

We characterize the social network of the hidden population as a finite undirected simple graph $G = (V, E)$ with no parallel edges or self-loops. Members of the hidden population are vertices; $\{i, j\} \in E$ implies that $i \in V$ and $j \in V$ know each other. Using RDS, researchers recruit members from the hidden population into the study. The time-varying recruitment-induced subgraph $\{G_S(t) = (V_S(t), E_S(t)) : 0 \leq t \leq t_F\}$ is a nested collection of subgraphs of G, where t_F is the termination time of the study. For all $0 \leq t \leq t_F$, $G_S(t)$ is a subgraph of G. Here, $G_S(0)$ is the null graph since there are no subjects in the study initially. For simplicity, we write $G_S = (V_S, E_S)$ for $G_S(t_F) = (V_S(t_F), E_S(t_F))$, and call this the *recruitment-induced subgraph* or *induced subgraph* unless we explicitly specify the time t. The vertex set of the time-varying recruitment-induced subgraph at time t (i.e., $G_S(t)$) denotes the members in the hidden population that are known to the study by time t. The subgraph $G_S(t)$ is induced by the vertex set $V_S(t)$; i.e., $E_S(t) = \{\{i, j\} | i, j \in V_S(t), \{i, j\} \in E\}$. The time-varying recruitment-induced subgraph evolves in the following way (Crawford 2016).

1. At time $\tilde{t}$, researchers recruit a subject in the population as a seed. This subject is included in the vertex set $V_S(t)$ of $G_S(t)$ for all $t \geq \tilde{t}$. Researchers may provide this subject with coupons to recruit its neighbors.

2. Once recruited into this study (either by researchers or its neighbors) at time $\tilde{t}$, subjects currently holding coupons will attempt to recruit their yet-unrecruited neighbors. The inter-recruitment time along each edge connecting a recruiter with an unrecruited neighbor is i.i.d. with cdf $F(t)$. Recruitment happens when a neighbor is recruited into the study and is provided with a number of coupons. A successful recruitment costs the recruiter one coupon.

The directed recruitment graph is $G_R = (V_R, E_R)$, where $V_R = V_S(t_F)$ is the set of members in the study at the final stage. For two subjects $i, j \in V_R$, $(i, j) \in E_R$ if and only if i recruits j. Note that the subjects recruited by researchers (i.e., the seeds) have zero in-degree in G_R. Let n denote the cardinality of V_R (equivalently $V_S(t_F)$). For simplicity, we label the subject recruited in the ith recruitment event by i. The labels of the subjects in the study are $1, 2, 3, \ldots, n$. The vector of recruitment times is $\mathbf{t} = (t_1, t_2, t_3, \ldots, t_n)$, where t_i is the recruitment time of subject i. In shorthand, we write $\tau(i; j) = t_{j-1} - t_i$ for $i < j$. Let $\mathbf{C}$ be the $n \times n$ coupon matrix whose element $\mathbf{C}_{ij} = 1$ if subject i has at least one coupon just before the jth recruitment event, and zero otherwise. The rows and columns of the coupon matrix are ordered by subjects' recruitment time. The degree vector is $\mathbf{d} = (d_1, d_2, d_3, \ldots, d_n)'$, where d_i is the degree of i in G. At time t (where $t \neq t_i$ for $i = 1, 2, 3, \ldots, n$), if a subject has at least one coupon and at least one neighbor not in the study, we term it a *recruiter* at time t; if a subject has not entered the study and has at least one neighbor with at least one coupon, we term it a *potential recruitee* at time t. We assume that the observed data from a RDS recruitment process consist of $\mathbf{Y} = (G_R, \mathbf{d}, \mathbf{t}, \mathbf{C})$.

Likelihood of Recruitment Time Series

We assume that the inter-recruitment time along an edge connecting a recruiter and potential recruitee is i.i.d. with cdf $F(t)$. Let $F_s(t) = \Pr[W \leq t \mid W > s]$, where W is the random inter-recruitment time and $\Pr[W \leq t] = F(t)$. We write $f_s(t) = F_s'(t)$ for the corresponding conditional pdf. Let $S_s(t) = 1 - F_s(t)$ be the conditional survival function and $H_s(t) = f_s(t)/S_s(t)$ be the conditional hazard function.

We now derive a closed-form expression for the likelihood of the recruitment time series $L(\mathbf{t}|G_S, \theta)$. In what follows, let M be the set of seeds, and let $\mathbf{A}$ denote the adjacency matrix representation of the recruitment-induced subgraph at the final stage G_S; thus we use G_S and $\mathbf{A}$ interchangeably throughout this paper.

Theorem 1. (Proof in (Chen, Crawford, and Karbasi 2015)). *Let $R(i)$ and $I(i)$ be the recruiter set and potential recruitee set just before time t_i, respectively, and M be the set of seeds. The following statements with respect to the likelihood of the recruitment time series hold.*

1. *The likelihood of the recruitment time series is given by*

$$\begin{aligned} & L(\mathbf{t}|G_S,\theta) \\ = & \prod_{i=1}^{n}\left(\sum_{u\in R(i)} |I_u(i)| H_{\tau(u;i)}(t_i - t_u)\right)^{1\{i\notin M\}} \\ \times & \prod_{j\in R(i)} S^{|I_j|}_{\tau(j;i)}(t_i - t_j), \end{aligned}$$

   where $\tau(u;i) = t_{i-1} - t_u$.

2. *Let* $\mathbf{m}$ *and* $\mathbf{u}$ *be column vectors of size* n *defined as* $\mathbf{m}_i = 1\{i \notin M\}$ *and* $\mathbf{u} = \mathbf{d} - \mathbf{A} \cdot \mathbf{1}$*, and let* $\mathbf{H}$ *and* $\mathbf{S}$ *be* $n \times n$ *matrices, defined as* $\mathbf{H}_{ui} = H_{\tau(u;i)}(t_i - t_u)$ *and* $\mathbf{S}_{ji} = \log S_{\tau(j;i)}(t_i - t_j)$*. Furthermore, we form matrices* $\mathbf{B} = (\mathbf{C} \circ \mathbf{H})$ *and* $\mathbf{D} = (\mathbf{C} \circ \mathbf{S})$*, where* $\circ$ *denotes the Hadamard (entrywise) product. We let*

$$\begin{aligned} \beta &= \log(\mathbf{B}'\mathbf{u} + \mathrm{LowerTri}(\mathbf{AB})' \cdot \mathbf{1}) \\ \delta &= \mathbf{D}'\mathbf{u} + \mathrm{LowerTri}(\mathbf{AD})' \cdot \mathbf{1}, \end{aligned}$$

   where $\mathrm{LowerTri}(\cdot)$ *denotes the lower triangular part (diagonal elements inclusive) of a square matrix. Then, the log-likelihood of the recruitment time series can be written as*

$$l(\mathbf{t}|G_S,\theta) = \mathbf{m}'\beta + \mathbf{1}'\delta.$$

Network Reconstruction Problem

Given the observed time series, we seek to reconstruct the $n \times n$ binary symmetric, zero-diagonal adjacency matrix $\mathbf{A}$ of G_S and the parameter $\theta \in \Theta$ (Θ is the parameter space) that maximizes $\Pr(\mathbf{A},\theta|\mathbf{t})$. Recall that we use G_S and $\mathbf{A}$ interchangeably throughout this paper. We have

$$\Pr(\mathbf{A},\theta|\mathbf{t}) \propto L(\mathbf{t}|\mathbf{A},\theta)\Pr(\mathbf{A},\theta),$$

where $\Pr(\mathbf{A},\theta)$ is the prior distribution for $(\mathbf{A},\theta)$. The constraint for the parameter θ is obvious: θ must reside in the parameter space Θ; i.e., $\theta \in \Theta$. Now we discuss the constraint for the adjacency matrix $\mathbf{A}$—we require that the adjacency matrix $\mathbf{A}$ must be *compatible*.

We seek the matrix $\mathbf{A}$ that maximizes the probability $\Pr(\mathbf{A},\theta|\mathbf{t})$. However, we know that the directed recruitment subgraph, if viewed as an undirected graph, must be a subgraph of the true recruitment-induced subgraph. Let $\mathbf{A}_R$ be the adjacency matrix of the recruitment subgraph when it is viewed as an undirected graph; i.e., the (i,j) entry of $\mathbf{A}_R$ is 1 if a direct recruitment event occurs between i and j (either i recruits j or j recruits i), and is 0 otherwise. We require that $\mathbf{A}$ be greater than or equal to $\mathbf{A}_R$ entrywise. Recall that every subject in the study reports its degree; thus the adjacency matrix should also comply with the degree constraints. Following (Crawford 2016), we say that a symmetric, binary and zero-diagonal matrix $\mathbf{A}$ is a *compatible adjacency matrix* if $\mathbf{A} \geq \mathbf{A}_R$ entrywise, and $\mathbf{A} \cdot \mathbf{1} \leq \mathbf{d}$ entrywise.

Problem Statement

Now we formulate this problem as an optimization problem.

Algorithm 1 RENDER: Alternating inference of G_S and θ

Input: the observed data $\mathbf{Y} = (G_R, \mathbf{d}, \mathbf{t}, \mathbf{C})$; the initial guess for the distribution parameter θ, denoted by $\hat{\theta}_0$; the maximum number of iterations, $\iota_{\max}$.

Output: the estimated adjacency matrix $\mathbf{A}$ (denoted by $\hat{\mathbf{A}}$) and the estimated parameter θ (denoted by $\hat{\theta}$)

```
$\iota \leftarrow 0$
while $\iota < \iota_{\max}$
    $\hat{\mathbf{A}}_\iota \leftarrow \arg\max_{\mathbf{A}\text{ is compatible}} L(\mathbf{t}|\mathbf{A},\hat{\theta}_\iota)\Pr(\mathbf{A},\hat{\theta}_\iota)$ (A-step, we use Algorithm 2 here.)
    $\hat{\theta}_{\iota+1} \leftarrow \arg\max_{\theta\in\Theta} L(\mathbf{t}|\hat{\mathbf{A}}_\iota,\theta)\Pr(\hat{\mathbf{A}}_\iota,\theta)$    ($\theta$-step)
    $\iota \leftarrow \iota + 1$
end while
$\hat{\mathbf{A}} \leftarrow \hat{\mathbf{A}}_{\iota_{\max}-1}$
$\hat{\theta} \leftarrow \hat{\theta}_{\iota_{\max}}$
```

Problem. Given the observed data $\mathbf{Y} = (G_S, \mathbf{d}, \mathbf{t}, \mathbf{C})$, we seek an $n \times n$ adjacency matrix $\mathbf{A}$ (symmetric, binary and zero-diagonal) and a parameter value $\theta \in \Theta$ that

$$\begin{aligned} \text{maximizes} \quad & L(\mathbf{t}|\mathbf{A},\theta)\Pr(\mathbf{A},\theta) \\ \text{subject to} \quad & \mathbf{A} \geq \mathbf{A}_R \text{ (entrywise)} \\ & \mathbf{A} \cdot \mathbf{1} \leq \mathbf{d} \text{ (entrywise)} \end{aligned}$$

Alternating Inference of A and θ

Given the observed data $\mathbf{Y}$, we wish to infer the adjacency matrix $\mathbf{A}$ of the recruitment-induced graph and the distribution parameter θ. Given $\mathbf{A}$, the maximum likelihood estimator (MLE) for θ is

$$\hat{\theta} = \arg\max_{\theta\in\Theta} L(\mathbf{t}|\mathbf{A},\theta)\Pr(\mathbf{A},\theta).$$

Similarly, given the true parameter θ, the MLE for the adjacency matrix $\mathbf{A}$ is given by

$$\hat{\mathbf{A}} = \arg\max_{\mathbf{A}\text{ is compatible}} L(\mathbf{t}|\mathbf{A},\theta)\Pr(\mathbf{A},\theta).$$

In practice, both the parameter θ and the true recruitment-induced subgraph G_S are unknown and need estimation. However, we can alternately estimate $\mathbf{A}$ and θ. This is what RENDER (presented in Algorithm 1) does. Each iteration is divided into two steps: $\mathbf{A}$-step and θ-step.

Estimation of A using simulated annealing The $\mathbf{A}$-step of RENDER solves

$$\max_{\mathbf{A}\text{ is compatible}} L(\mathbf{t}|\mathbf{A},\hat{\theta}_\iota)\Pr(\mathbf{A},\hat{\theta}_\iota);$$

equivalently, it suffices to solve

$$\max_{\mathbf{A}\text{ is compatible}} (l(\mathbf{t}|\mathbf{A},\hat{\theta}_\iota) + \log\Pr(\mathbf{A},\hat{\theta}_\iota)).$$

We employ a simulated-annealing-based method to estimate the adjacency matrix $\mathbf{A}$ (presented in Algorithm 2). Let the energy function be

$$\Lambda_\gamma(\mathbf{A};\mathbf{t},\hat{\theta}_\iota) \triangleq \exp\left[-\left(l(\mathbf{t}|\mathbf{A},\hat{\theta}_\iota) + \log\Pr(\mathbf{A},\hat{\theta}_\iota)\right)/\gamma\right],$$

where γ is the *temperature*. We specify a cooling schedule, which is a sequence of positive numbers $\{\gamma_\jmath\}_{\jmath\geq1}$ that satisfy

$\lim_{\jmath \to \infty} \gamma_\jmath = 0$, where $\gamma_\jmath$ is the temperature in the $\jmath$-th iteration. Note that the $\jmath$-th iteration of the simulated annealing procedure has a compatible adjacency matrix $\mathbf{A}(\jmath)$ as its state. Algorithm 3 specifies which state (compatible adjacency matrix) the algorithm should transition into in the next iteration. Concretely, in each iteration of Algorithm 2, it randomly proposes an edge that connects vertices i and j. If the edge does not appear in $\mathbf{A}(\jmath)$ and it still conforms to the degree constraint if we add the edge, then we simply add it to $\tilde{\mathbf{A}}(\jmath+1)$. In contrast, if the edge appears in $\mathbf{A}(\jmath)$ and it still conforms to the subgraph constraint if we remove the edge, then we simply remove it from $\tilde{\mathbf{A}}(\jmath+1)$. If neither condition is satisfied, the algorithm tries again with a new proposal. We prove in (Chen, Crawford, and Karbasi 2015) that the space of compatible adjacency matrices is connected by the proposal method.

Algorithm 2 Simulated-annealing-based optimization

Input: the number of iterations, $\jmath_{\max}$; the cooling schedule $\{\gamma_\jmath\}_{\jmath \geq 1}$; initial compatible adjacency matrix $\mathbf{A}(1)$; estimated parameter $\hat{\theta}_\iota$.

Output: the estimated adjacency matrix $\hat{\mathbf{A}}_\iota$

```
for $\jmath = 1$ to $\jmath_{\max}$ do
    Use Algorithm 3 to propose a compatible adjacency matrix $\tilde{\mathbf{A}}(\jmath+1)$ based on $\mathbf{A}(\jmath)$.
    $\psi \leftarrow \min\left\{1, \frac{\Lambda_{\gamma_\jmath}(\tilde{\mathbf{A}}(\jmath+1);\mathbf{t},\hat{\theta}_\iota)}{\Lambda_{\gamma_\jmath}(\mathbf{A}(\jmath);\mathbf{t},\hat{\theta}_\iota)} \cdot \frac{\Pr(\mathbf{A}(\jmath)|\tilde{\mathbf{A}}(\jmath+1))}{\Pr(\tilde{\mathbf{A}}(\jmath+1)|\mathbf{A}(\jmath))}\right\}$.
    $\mathbf{A}(\jmath+1) \leftarrow \begin{cases} \tilde{\mathbf{A}}(\jmath+1) & \text{with probability } \psi; \\ \mathbf{A}(\jmath) & \text{with probability } 1-\psi. \end{cases}$
end for
$\hat{\mathbf{A}}_\iota \leftarrow \mathbf{A}(\jmath\text{max}+1)$.
```

The proposed compatible adjacency matrix $\tilde{\mathbf{A}}(\jmath+1)$ is accepted as the state for the next iteration with probability ψ, where ψ equals

$$\min\left\{1, \frac{\Lambda_{\gamma_\jmath}(\tilde{\mathbf{A}}(\jmath+1);\mathbf{t},\hat{\theta}_\iota)}{\Lambda_{\gamma_\jmath}(\mathbf{A}(\jmath);\mathbf{t},\hat{\theta}_\iota)} \cdot \frac{\Pr(\mathbf{A}(\jmath)|\tilde{\mathbf{A}}(\jmath+1))}{\Pr(\tilde{\mathbf{A}}(\jmath+1)|\mathbf{A}(\jmath))}\right\};$$

otherwise, the matrix $\mathbf{A}(\jmath)$ remains the state for the next iteration. The term $\frac{\Lambda_{\gamma_\jmath}(\tilde{\mathbf{A}}(\jmath+1);\mathbf{t},\hat{\theta}_\iota)}{\Lambda_{\gamma_\jmath}(\mathbf{A}(\jmath);\mathbf{t},\hat{\theta}_\iota)}$ is called the *likelihood ratio* and the term $\frac{\Pr(\mathbf{A}(\jmath)|\tilde{\mathbf{A}}(\jmath+1))}{\Pr(\tilde{\mathbf{A}}(\jmath+1)|\mathbf{A}(\jmath))}$ is called the *proposal ratio*. To implement Algorithm 2, we have to compute the likelihood ratio and the proposal ratio efficiently. In fact, they can be evaluated efficiently in a recursive manner.

Theorem 2. (Proof in (Chen, Crawford, and Karbasi 2015)). *The proposal ratio* $\frac{\Pr(\mathbf{A}(\jmath)|\tilde{\mathbf{A}}(\jmath+1))}{\Pr(\tilde{\mathbf{A}}(\jmath+1)|\mathbf{A}(\jmath))}$ *is given by*

$$\frac{\text{Add}(\mathbf{A}(\jmath)) + \text{Remove}(\mathbf{A}(\jmath))}{\text{Add}(\tilde{\mathbf{A}}(\jmath+1)) + \text{Remove}(\tilde{\mathbf{A}}(\jmath+1))},$$

where $\text{Add}(\mathbf{A}) = \sum_{1 \leq i < j \leq n} 1\{\mathbf{A}_{ij} = 0, \sum_{k=1}^n \mathbf{A}_{ik} < \mathbf{d}_i, \sum_{k=1}^n \mathbf{A}_{jk} < \mathbf{d}_j\}$, *and* $\text{Remove}(\mathbf{A}) = \sum_{1 \leq i < j \leq n} 1\{\mathbf{A}_{ij} = 1$ *and* $\mathbf{A}_R^{(ij)} = 0\}$. *Here, we let* $\mathbf{A}_R^{(ij)}$ *denote the* (i,j)*-entry of the matrix* $\mathbf{A}_R$.

Algorithm 3 Proposal of compatible adjacency matrix

Input: the compatible adjacency matrix $\mathbf{A}(\jmath)$ in the $\jmath$-th iteration.

Output: a compatible adjacency matrix $\tilde{\mathbf{A}}(\jmath+1)$, which will be a candidate for the state in the $(\jmath+1)$-th iteration.

```
loop
    $i, j \leftarrow$ two distinct random integers in $[1, n] \cap \mathbb{N}$
    if $\mathbf{A}_{ij}(\jmath) = 0$ and $\sum_{k=1}^n \mathbf{A}_{ik}(\jmath) < \mathbf{d}_i$ and $\sum_{k=1}^n \mathbf{A}_{jk}(\jmath) < \mathbf{d}_j$ then
        $\tilde{\mathbf{A}}^+(\jmath+1) \leftarrow \mathbf{A}(\jmath)$
        $\tilde{\mathbf{A}}^+_{ij}(\jmath+1) \leftarrow 1$ and $\tilde{\mathbf{A}}^+_{ji}(\jmath+1) \leftarrow 1$
        $\tilde{\mathbf{A}}(\jmath+1) \leftarrow \tilde{\mathbf{A}}^+(\jmath+1)$ and exit loop
    else
        if $\mathbf{A}_{ij}(\jmath) = 1$ and $\mathbf{A}_R^{(ij)} = 0$ ($\mathbf{A}_R^{(ij)}$ is the $(i,j)$-entry of the matrix $\mathbf{A}_R$) then
            $\tilde{\mathbf{A}}^-(\jmath+1) \leftarrow \mathbf{A}(\jmath)$
            $\tilde{\mathbf{A}}^-_{ij}(\jmath+1) \leftarrow 0$ and $\tilde{\mathbf{A}}^-_{ji}(\jmath+1) \leftarrow 0$
            $\tilde{\mathbf{A}}(\jmath+1) \leftarrow \tilde{\mathbf{A}}^-(\jmath+1)$ and exit loop
        end if
    end if
end loop
return $\tilde{\mathbf{A}}(\jmath+1)$
```

The same way, we can find the likelihood ratio in a recursive manner.

Theorem 3. (Proof in (Chen, Crawford, and Karbasi 2015)). *If we view* β *in Theorem 1 as a function of the adjacency matrix* $\mathbf{A}$*, denoted by* $\beta(\mathbf{A})$*, then the recurrence relation between* $\beta(\tilde{\mathbf{A}}(\jmath+1))$ *and* $\beta(\mathbf{A}(\jmath))$ *is as follows:*

$$(e^{\beta(\tilde{\mathbf{A}}(\jmath+1))} - e^{\beta(\mathbf{A}(\jmath))})_j = \pm\left(\mathbf{B}_{bj} 1\{a < j\} - \mathbf{B}_{aj} 1\{b < j\}\right). \quad (1)$$

Here, we assign the minus sign "$-$" in "$\pm$" if $\tilde{\mathbf{A}}(\jmath+1) = \tilde{\mathbf{A}}^+(\jmath+1)$*, and assign the plus sign "$+$" in "$\pm$" if* $\tilde{\mathbf{A}}(\jmath+1) = \tilde{\mathbf{A}}^-(\jmath+1)$*. By the same convention, the likelihood ratio* $\Lambda_{\gamma_\jmath}(\tilde{\mathbf{A}}(\jmath+1)|\mathbf{t},\hat{\theta}_\iota)/\Lambda_{\gamma_\jmath}(\mathbf{A}(\jmath)|\mathbf{t},\hat{\theta}_\iota)$ *is given by*

$$\exp\left\{-\gamma_\jmath^{-1}\left[-\log\frac{\Pr(\tilde{\mathbf{A}}(\jmath+1))}{\Pr(\mathbf{A}(\jmath))} + \mathbf{m}'\left(\beta(\tilde{\mathbf{A}}(\jmath+1)) - \beta(\mathbf{A}(\jmath))\right) \pm \sum_{j=1}^n \left(\mathbf{D}_{bj} 1\{a < j\} - \mathbf{D}_{aj} 1\{b < j\}\right)\right]\right\},$$

Estimation of distribution parameter In a θ-step in Algorithm 1, we have to solve the optimization problem $\hat{\theta}_{\iota+1} \leftarrow \arg\max_{\theta \in \Theta} L(\theta|\hat{\mathbf{A}}_\iota, \mathbf{t}) \Pr(\hat{\mathbf{A}}_\iota, \theta)$. If the parameter space Θ is a subset of the p-dimensional Euclidean space $\mathbb{R}^p$, this problem can be solved using off-the-shelf solvers, e.g., FMINSEARCH in MATLAB.

Experiments

We evaluated the proposed method in two aspects, the reconstruction performance of the recruitment-induced sub-

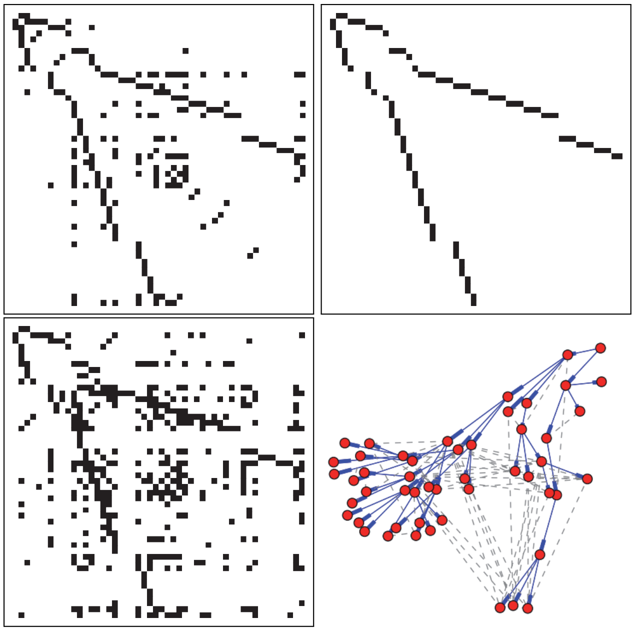

Figure 1: Example reconstruction procedure. Clockwise from top left: the true recruitment-induced subgraph G_S, the observed recruitment graph G_R, the estimated network with recruitments as blue arrows and dashed lines as inferred edges, and the estimated recruitment-induced subgraph $\hat{G}_S$.

graph and the parameter estimation of the edgewise inter-recruitment time model. Let $\mathbf{A}$ be the adjacency matrix of the true recruitment-induced subgraph G_S and $\hat{\mathbf{A}}$ be the estimate. We define the true and false positive rates (TPR and FPR) as $\mathrm{TPR}(\hat{\mathbf{A}}, \mathbf{A}) = \sum_{i<j} 1\{\hat{\mathbf{A}}_{ij} = 1 \text{ and } \mathbf{A}_{ij} = 1\}/\binom{n}{2}$ and $\mathrm{FPR}(\hat{\mathbf{A}}, \mathbf{A}) = \sum_{i<j} 1\{\hat{\mathbf{A}}_{ij} = 1 \text{ and } \mathbf{A}_{ij} = 0\}/\binom{n}{2}$. Fig. 1 illustrates an example of the reconstruction procedure. We simulated a RDS process over the Project 90 graph (Woodhouse et al. 1994) with power-law edgewise inter-recruitment time distribution, whose shape parameter $\alpha = 2$ and scale parameter $x_{\min} = 0.5$. Fifty subjects are recruited in this process. We show clockwise from top left: the true recruitment-induced subgraph G_S, the observed recruitment graph G_R, the estimated network with recruitments as blue arrows and dashed lines as inferred edges, and the estimated recruitment-induced subgraph $\hat{G}_S$. The TPR equals 0.769 and the FPR equals 0.106. The estimated parameter $\hat{\alpha} = 1.95$ and $\hat{x}_{\min} = 0.49$.

Reconstruction Performance

Impact of distribution parameter We simulated 50 RDS over the Project 90 graph with inter-recruitment time distribution $\mathrm{Gamma}(\alpha, \alpha)$ (parametrized by the shape and scale) for each $\alpha = 0.01$, 0.1, 1, 10, and 100. Thus a total number of 250 RDS processes are simulated and the mean inter-recruitment time is fixed to be 1. The reconstruction performance is illustrated in Fig. 2. Each point on the receiver operating characteristic (ROC) plane represents a reconstruction accuracy performance of a simulated RDS process.

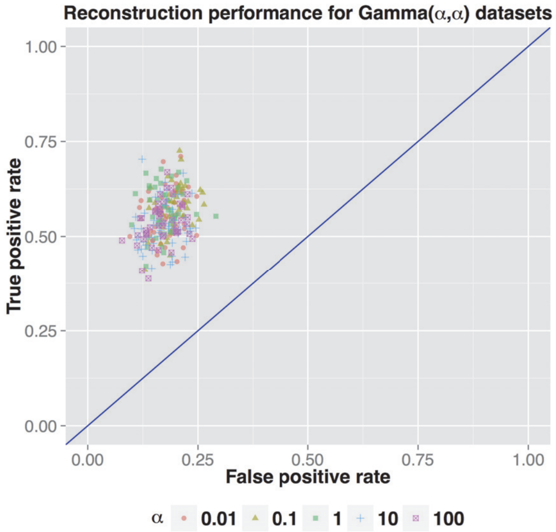

Figure 2: True and false positive rates of the reconstruction results for $\mathrm{Gamma}(\alpha, \alpha)$ datasets. Each point corresponds to the reconstruction result of one dataset. The different shapes of the points indicate different shape parameters.

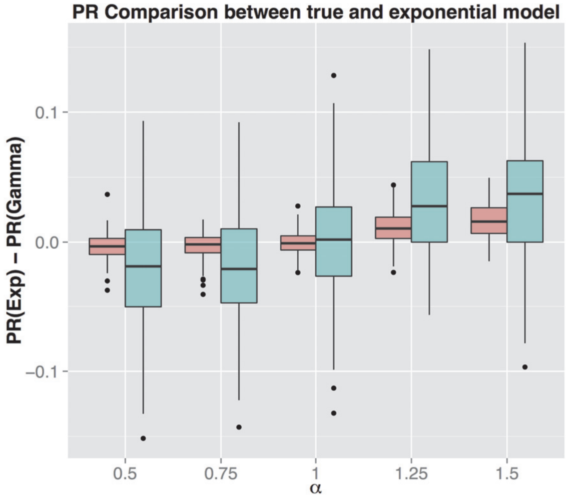

Figure 3: Difference between true and false positive rates (green and red boxplots, respectively) under the exponential model and the true Gamma model (the vertical axis) with the true shape parameter α (the horizontal axis).

Points with the same marker have the same inter-recruitment time distribution parameter. From the figure, we can observe that there is no significant sign of separation of points with different inter-recruitment time distribution parameters. Reconstruction accuracy is robust to the distribution parameter.

Impact of distribution model We simulated 50 RDS process over the Project 90 graph with inter-recruitment time distribution $\mathrm{Gamma}(0.5, 0.5)$. The recruitment-induced subgraph is reconstructed via the model of the true inter-

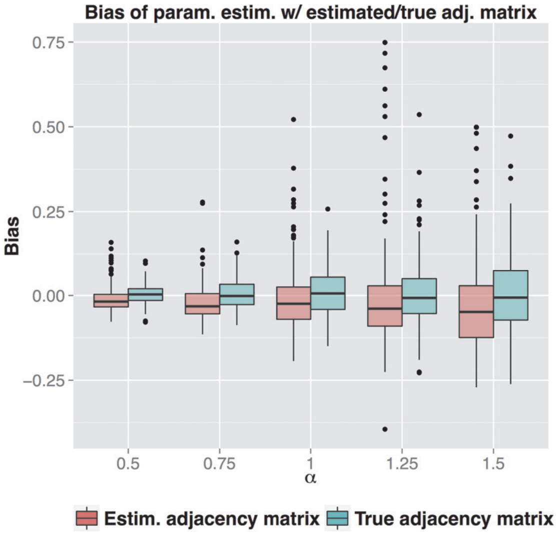

Figure 4: Distribution of bias in the estimated shape parameter $\hat{\alpha}$ given the estimated adjacency matrix (red boxes) and the true adjacency matrix (green boxes). The horizontal axis represents the values of the true shape parameter α. Whisker lengths are 1.5 times the inter-quartile range.

recruitment distribution $\text{Gamma}(0.5, 0.5)$ and the exponential distribution $\text{Exp}(1)$, respectively. The TPR and FPR of each dataset are presented in Fig. 3. The TPR is always higher than the FPR, which reaffirms the effectiveness of our proposed reconstruction method. For each dataset, the TPR and FPR under the true and the exponential models are very close to each other. We observe that there is no significant reconstruction skewness incurred by mis-specification of the inter-recruitment time distribution model.

Parameter Estimation

We simulated 200 RDS processes over the Project 90 graph with edgewise inter-recruitment time distribution $\text{Gamma}(\alpha, \alpha)$ (parametrized by the shape and scale parameters) for each $\alpha = 0.5$, 0.75, 1, 1.25, and 1.5. We used the method in the "Estimation of distribution parameter" section. We assess the bias of the estimated shape parameter $\hat{\alpha}$, which is given by $\hat{\alpha} - \alpha$. Fig. 4 shows the distribution of the bias using Tukey boxplots. In Fig. 4, the red boxes depict the distribution of the biases of the estimated shape parameters inferred through the estimated adjacency matrix, while the green boxes illustrate the distribution of those inferred given the true adjacency matrix. The horizontal axis shows the value of the true shape parameter.

With respect to the red boxes (those based on the estimated adjacency matrices), The middle line of each box is very close to zero and thus the estimation is highly accurate. The interquartile range (IQR) , which measures the deviation of the biases, declines as the shape parameter decreases. Even for the box with largest deviation (i.e., the box with $\alpha = 1.5$), the IQR is approximately $[-0.125, 0.02]$ and 99.3% of the biases reside in the interval $[-0.27, 0.25]$. Compared with the parameter estimation via the true adjacency matrix, this estimator based on the estimated adjacency matrix is biased to some degree. In Fig. 4, we can observe that it underestimates α.

Then consider the green boxes (those based on the true adjacency matrix). Similar to those based on the estimated adjacency matrix, the deviation of this estimator declines as the value of the shape parameter α decreases. The middle line of each box is noted to coincide perfectly with zero bias line, which suggests that this estimator is unbiased given the true adjacency matrix.

Experiments on Real Data

We also apply RENDER to data from an RDS study of $n = 813$ drug users in St. Petersburg, Russian Federation. We use RENDER to infer the underlying social network structure of the drug users in this study. Since it could be confusing to visualize the whole inferred network, we only show the inferred subgraph of the largest community of the network, as presented in Fig. 5. The blue arrows represent the edges in the recruitment subgraph that indicates the recruiter and recruitee. Gray dashed edges are inferred from the data.

Conclusion

In this paper, we precisely formulated the dynamics of RDS as a continuous-time diffusion process over the underlying graph. We derived the likelihood for the recruitment time series under an arbitrarily recruitment time distribution. As a result, we develop an efficient stochastic optimization algorithm, RENDER, that identifies the optimum network that best explains the collected data. We then supported the performance of RENDER through an exhaustive set of experiments on both synthetic and real data.

Acknowledgements

FWC was supported by NIH/NCATS grant KL2 TR000140 and NIMH grant P30MH062294. LC would like to thank Zeya Zhang and Yukun Sun for their encouragement.

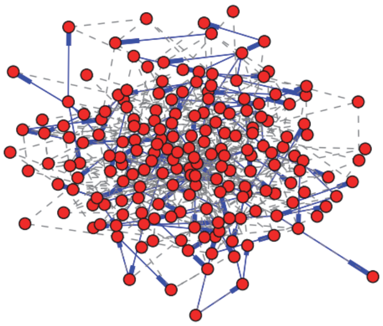

Figure 5: Illustration of the largest community of the inferred underlying social network of the St. Petersburg dataset. The blue arrows represent the edges in the recruitment subgraph that indicates the recruiter and recruitee. Gray dashed edges are inferred from the data.

References

Chen, L.; Crawford, F. W.; and Karbasi, A. 2015. Seeing the unseen network: Inferring hidden social ties from respondent-driven sampling. *arXiv preprint arXiv:1511.04137*.

Crawford, F. W. 2016. The graphical structure of respondent-driven sampling. *Sociological Methodology* to appear.

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

Goel, S., and Salganik, M. J. 2009. Respondent-driven sampling as Markov chain Monte Carlo. *Statistics in Medicine* 28(17):2202–2229.

Gomez Rodriguez, M.; Balduzzi, D.; Schölkopf, B.; Scheffer, G. T.; et al. 2011. Uncovering the temporal dynamics of diffusion networks. In *28th International Conference on Machine Learning (ICML 2011)*, 561–568. International Machine Learning Society.

Gomez Rodriguez, M.; Leskovec, J.; and Krause, A. 2010. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1019–1028. ACM.

Heckathorn, D. D. 1997. Respondent-driven sampling: a new approach to the study of hidden populations. *SOCIAL PROBLEMS-NEW YORK-* 44:174–199.

Heckathorn, D. D. 2002. Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems* 49(1):11–34.

Kramer, M. A.; Eden, U. T.; Cash, S. S.; and Kolaczyk, E. D. 2009. Network inference with confidence from multivariate time series. *Physical Review E* 79(6):061916.

Liben-Nowell, D., and Kleinberg, J. 2007. The link-prediction problem for social networks. *Journal of the American society for information science and technology* 58(7):1019–1031.

Linderman, S., and Adams, R. 2014. Discovering latent network structure in point process data. In *Proceedings of The 31st International Conference on Machine Learning*, 1413–1421.

Salganik, M. J., and Heckathorn, D. D. 2004. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology* 34(1):193–240.

Salganik, M. J. 2006. Variance estimation, design effects, and sample size calculations for respondent-driven sampling. *Journal of Urban Health* 83(1):98–112.

Shandilya, S. G., and Timme, M. 2011. Inferring network topology from complex dynamics. *New Journal of Physics* 13(1):013004.

Volz, E., and Heckathorn, D. D. 2008. Probability based estimation theory for respondent driven sampling. *Journal of Official Statistics* 24(1):79.

Woodhouse, D. E.; Rothenberg, R. B.; Potterat, J. J.; Darrow, W. W.; Muth, S. Q.; Klovdahl, A. S.; Zimmerman, H. P.; Rogers, H. L.; Maldonado, T. S.; Muth, J. B.; et al. 1994. Mapping a social network of heterosexuals at high risk for HIV infection. *Aids* 8(9):1331–1336.