

Supplementary Material: A Bayesian Framework for Modeling the Preferential Selection Process in Respondent-Driven Sampling

Katherine R. McLaughlin¹

¹ Department of Statistics, Oregon State University, Corvallis, OR, USA

Address for correspondence: Katherine R. McLaughlin, Department of Statistics, Oregon State University, 239 Weniger Hall, Corvallis, OR, 97331, USA.

E-mail: mclaugka@oregonstate.edu.

Phone: (+1) 541 737 3366.

Fax: (+1) 541 737 3489.

Abstract: In sampling designs that utilize peer recruitment, the sampling process is partially unknown and must be modeled to make inference about the population and estimate standard outcomes like prevalence. We develop a Bayesian model for the recruitment process for respondent-driven sampling (RDS), a network sampling methodology used worldwide to sample hidden populations that are not reachable by conventional sampling techniques, including those at high risk for HIV/AIDS. Current models for the RDS sampling process typically assume that recruitment occurs randomly given the population social network, but this is likely untrue in practice. To model preferential selection on covariates, we develop a sequential two-sided rational choice framework, which allows generative probabilistic network models to be created for the RDS sampling process. In the rational choice framework, members of the population make recruitment and participation choices based on observable nodal and dyadic covariates to maximize their utility given constraints. Inference is made about recruitment preferences given the observed recruitment chain in a Bayesian framework by incorporating the latent utilities and sampling from the joint posterior distribution via Markov chain Monte Carlo. We present simulation results and apply the model to an RDS study of Francophone migrants in Rabat, Morocco.

Key words: hidden population; Markov chain Monte Carlo; peer recruitment; rational choice model; respondent-driven sampling; sampling process model

S.1 Stable Matching Example

To aid understanding of stability, we provide several examples. Figure S1 gives an example of some possible preferences that would make a matching stable or unstable. In the simplest example of a coalition blocking a matching, there are two recruiter-peer dyads matched in

Scenario	Matching	Recruiter Preference	Peer Preference	Stable?
1				Yes
2				Yes
3				No

Figure S1: Stability example with recruiters A , B , and C shown as blue circles; and peers 1, 2, and 3 shown as red squares. In scenario 1, all recruiters and peers most prefer their current match, so the matching is stable. In scenario 2, all recruiters prefer their current partner even though none of the peers do, so the matching is still stable. In scenario 3, all members of the dyads $(A, 1)$ and $(B, 2)$ would prefer to swap partners so that the new pairing would be $(A, 2)$ and $(B, 1)$, making the matching unstable.

m such that $1 \in m(A)$ and $2 \in m(B)$. If recruiter A actually prefers peer 2 over peer 1, recruiter B prefers peer 1, peer 1 prefers recruiter B , and peer 2 prefers recruiter A , then the alternate matching m' where $2 \in m'(A)$ and $1 \in m'(B)$ is preferable to all parties. In this case, the coalition of $\{A, B, 1, 2\}$ blocks matching m , so it is not stable.

S.2 Matching Process Example

We provide an example to demonstrate the matching process. Consider a scenario with three recruiters $\mathcal{R} = \{r_1, r_2, r_3\}$ and five peers $\mathcal{P} = \{p_1, p_2, p_3, p_4, p_5\}$, where each recruiter has $n_c = 2$ coupons. Therefore $\mathcal{A} = \{a_1, a_2\}$ and $\mathcal{D} = \{d_1\}$. Assume the utility matrices \mathbf{U} and \mathbf{V} are known and can be ranked across each row to produce the row-ranked utility














































Iteration	Recruiter Choice	Peer Options	Peer Choice	Failed Attempts
1	$\mathcal{C}_{r_1} = $  $\mathcal{C}_{r_2} = $  $\mathcal{C}_{r_3} = $ 	$\mathcal{J}_{p_1} = $  $\mathcal{J}_{p_2} = $  $\mathcal{J}_{p_3} = $  $\mathcal{J}_{p_4} = $  $\mathcal{J}_{p_5} = $ 	$\tilde{\mathcal{J}}_{p_1} = $  $\tilde{\mathcal{J}}_{p_2} = $  $\tilde{\mathcal{J}}_{p_3} = $  $\tilde{\mathcal{J}}_{p_4} = $  $\tilde{\mathcal{J}}_{p_5} = $ 	$\mathcal{F}_{r_1} = $ $\mathcal{F}_{r_2} = $  $\mathcal{F}_{r_3} = $ 
2	$\mathcal{C}_{r_1} = $  $\mathcal{C}_{r_2} = $  $\mathcal{C}_{r_3} = $ 	$\mathcal{J}_{p_1} = $  $\mathcal{J}_{p_2} = $  $\mathcal{J}_{p_3} = $  $\mathcal{J}_{p_4} = $  $\mathcal{J}_{p_5} = $ 	$\tilde{\mathcal{J}}_{p_1} = $  $\tilde{\mathcal{J}}_{p_2} = $  $\tilde{\mathcal{J}}_{p_3} = $  $\tilde{\mathcal{J}}_{p_4} = $  $\tilde{\mathcal{J}}_{p_5} = $ 	$\mathcal{F}_{r_1} = $ $\mathcal{F}_{r_2} = $  $\mathcal{F}_{r_3} = $ 
3	$\mathcal{C}_{r_1} = $  $\mathcal{C}_{r_2} = $  $\mathcal{C}_{r_3} = $ 	$\mathcal{J}_{p_1} = $  $\mathcal{J}_{p_2} = $  $\mathcal{J}_{p_3} = $  $\mathcal{J}_{p_4} = $  $\mathcal{J}_{p_5} = $ 	$\tilde{\mathcal{J}}_{p_1} = $  $\tilde{\mathcal{J}}_{p_2} = $  $\tilde{\mathcal{J}}_{p_3} = $  $\tilde{\mathcal{J}}_{p_4} = $  $\tilde{\mathcal{J}}_{p_5} = $ 	$\mathcal{F}_{r_1} = $ $\mathcal{F}_{r_2} = $  $\mathcal{F}_{r_3} = $ 

Figure S2: Matching algorithm example where the row-ranked utility matrices \mathbf{U}^* and \mathbf{V}^* are given in Equation S2.1. Recruiters are indicated by blue circles, peers by red squares, and abstentions and declensions by gray hexagons. Shaded shapes indicate the first time a choice is possible. The final matching is achieved in three iterations (shown in bottom left cell), with r_1 recruiting p_1 , r_2 not recruiting anyone, and r_3 recruiting p_5 .

matrices \mathbf{U}^* and \mathbf{V}^* :

$$\mathbf{U}^* = \begin{matrix} & p_1 & p_2 & p_3 & p_4 & p_5 & a_1 & a_2 \\ \begin{matrix} r_1 \\ r_2 \\ r_3 \end{matrix} & \begin{pmatrix} 1 & 3 & 6 & 5 & 7 & 2 & 4 \\ 6 & 3 & 5 & 7 & 2 & 1 & 4 \\ 6 & 7 & 2 & 4 & 1 & 3 & 5 \end{pmatrix} & & & & & & \end{matrix} \quad \mathbf{V}^* = \begin{matrix} & r_1 & r_2 & r_3 & d_1 \\ \begin{matrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \end{matrix} & \begin{pmatrix} 1 & 2 & 4 & 3 \\ 3 & 4 & 2 & 1 \\ 4 & 1 & 3 & 2 \\ 2 & 4 & 3 & 1 \\ 3 & 4 & 1 & 2 \end{pmatrix} & & & \end{matrix} \quad (\text{S2.1})$$

where ‘1’ indicates the highest preference. Figure S2 shows how the matching algorithm applies in this example, which requires three iterations to complete. At each iteration, new choices are shown as filled in circles, while previously tentatively accepted choices are shown as open circles. The final outcome is that r_1 selects p_1 , r_2 does not recruit anyone, and r_3 selects p_5 .

S.3 RCPS Algorithm

We can make draws from the joint posterior of $\pi(\alpha, \beta, \xi, \zeta, \mathbf{U}, \mathbf{V} | m \in \mathcal{M})$ using Gibbs sampling because we can specify the full conditional distributions for α , β , ξ , ζ , \mathbf{U} , and \mathbf{V} . The full conditional distributions for α and \mathbf{U} are given in the text; those for β , ξ , ζ , and \mathbf{V} are provided here.

S.3.1 Full Conditional Distributions

The full conditional distribution for β is

$$\begin{aligned} p(\beta | \alpha, \xi, \zeta, \mathbf{U}, \mathbf{V}, m \in \mathcal{M}) &= p(\beta | \mathbf{V}, \mathbf{Y}) \\ &\propto \pi(\beta) p(\mathbf{V} | \beta, \mathbf{Y}) \end{aligned} \quad (\text{S3.1})$$

Using the conjugate prior $\pi(\beta) = \mathcal{MVN}(\mu_\beta, \Sigma_\beta)$, the update step in the Gibbs sampler is $\beta^{(k+1)} | \mathbf{V}^{(k)}, \mathbf{Y} \sim \mathcal{MVN}(\tilde{\mu}_\beta, \tilde{\Sigma}_\beta)$ where $\tilde{\Sigma}_\beta^{-1} = \sum_{j=1}^{n_p} \sum_{i=1}^{n_r} \mathbf{Y}_{ji} \mathbf{Y}'_{ji} + \Sigma_\beta^{-1}$ and $\tilde{\mu}_\beta = \tilde{\Sigma}_\beta \left(\sum_{j=1}^{n_p} \sum_{i=1}^{n_r} \mathbf{Y}_{ji} v_{ji}^{(k)} + \Sigma_\beta^{-1} \mu_\beta \right)$.

The same holds for ξ and ζ . The conditional distribution for ξ is

$$\begin{aligned} p(\xi | \alpha, \beta, \zeta, \mathbf{U}, \mathbf{V}, m \in \mathcal{M}) &= p(\xi | \mathbf{U}, \mathbf{A}) \\ &\propto \pi(\xi) p(\mathbf{U} | \xi, \mathbf{A}) \end{aligned} \quad (\text{S3.2})$$

Using the conjugate prior $\pi(\xi) = \mathcal{MVN}(\mu_\xi, \Sigma_\xi)$, the update step in the Gibbs sampler is $\xi^{(k+1)} | \mathbf{U}^{(k)}, \mathbf{A} \sim \mathcal{MVN}(\tilde{\mu}_\xi, \tilde{\Sigma}_\xi)$ where $\tilde{\Sigma}_\xi^{-1} = \sum_{i=1}^{n_r} \sum_{j=n_p+1}^{n_p+n_c} \mathbf{A}_{ij} \mathbf{A}'_{ij} + \Sigma_\xi^{-1}$ and $\tilde{\mu}_\xi = \tilde{\Sigma}_\xi \left(\sum_{i=1}^{n_r} \sum_{j=n_p+1}^{n_p+n_c} \mathbf{A}_{ij} u_{ij}^{(k)} + \Sigma_\xi^{-1} \mu_\xi \right)$. The conditional distribution for ζ is

$$\begin{aligned} p(\zeta | \alpha, \beta, \xi, \mathbf{U}, \mathbf{V}, m \in \mathcal{M}) &= p(\zeta | \mathbf{V}, \mathbf{D}) \\ &\propto \pi(\zeta) p(\mathbf{V} | \zeta, \mathbf{D}) \end{aligned} \quad (\text{S3.3})$$

Using the conjugate prior $\pi(\zeta) = \mathcal{MVN}(\mu_\zeta, \Sigma_\zeta)$, the update step in the Gibbs sampler is $\zeta^{(k+1)} | \mathbf{V}^{(k)}, \mathbf{D} \sim \mathcal{MVN}(\tilde{\mu}_\zeta, \tilde{\Sigma}_\zeta)$ where $\tilde{\Sigma}_\zeta^{-1} = \sum_{j=1}^{n_p} \sum_{i=n_r+1}^{n_r+1} \mathbf{D}_{ji} \mathbf{D}'_{ji} + \Sigma_\zeta^{-1}$ and $\tilde{\mu}_\zeta = \tilde{\Sigma}_\zeta \left(\sum_{j=1}^{n_p} \sum_{i=n_r+1}^{n_r+1} \mathbf{D}_{ji} v_{ji}^{(k)} + \Sigma_\zeta^{-1} \mu_\zeta \right)$. Note the change of the indices on the summation for abstentions and declensions.

Unlike the preference parameters α , β , ξ , and ζ , the conditional distributions of \mathbf{U} and \mathbf{V} depend on each other. The conditional distribution for \mathbf{U} is given in the main text in Equation (2.11), and the bounds for the truncated normal distribution are provided in Table 1. Additional detail about the bounds is provided here, as well as the full conditional distribution for \mathbf{V} .

Consider a recruiter $i \in \mathcal{R}$ and peer $j \in \mathcal{P}$ who are tied. We know that i must have a higher utility for j than for any abstention they do not select: $u_{ij} \geq \max(\mathcal{U}_{i, \mathcal{A}_i^-})$. Additionally,

i must have a higher utility for j than they do for any peers who prefer i over their own match. These are the peers $h \neq j$, $h \notin m(i)$ that i has in their opportunity set $\mathcal{O}(i)$:

$$\mathcal{O}(i) = \{h : v_{hi} \geq v_{h,m(h)}; h \in \mathcal{P}, h \notin m(i)\}. \quad (\text{S3.4})$$

Therefore, $u_{ij} \geq \max(\mathcal{U}_{i,h \in \mathcal{O}(i)})$ where $\mathcal{U}_{i,h \in \mathcal{O}(i)} = \{u_{ih} : h \in \mathcal{O}(i)\}$. These two bounds, required to satisfy stability constraints, result in $a_{\mathbf{U};ij} = \max(\mathcal{U}_{i,\mathcal{A}_i^-} \cup \mathcal{U}_{i,h \in \mathcal{O}(i)})$. Since i and j are matched, u_{ij} has no upper bound so $b_{\mathbf{U};ij} = \infty$. This is the scenario in the first row of Table 1.

The constraints are the same if $j \in \mathcal{A}$ and there is a match (j is abstention that is selected). This is the scenario in the third row of Table 1.

When i and j are not matched and $j \in \mathcal{P}$ (j is a peer), the utility that i has for j must be less than the utility i has for any of their matches if j prefers i over their own match. This is the scenario in the second row of Table 1. In particular, $u_{ij} \leq \min(\mathcal{U}_{i,m(i)})$ if $v_{ji} > v_{j,m(j)}$. Therefore, $a_{\mathbf{U};ij} = -\infty$ and $b_{\mathbf{U};ij}$ is either $\min(\mathcal{U}_{i,m(i)})$ or ∞ . If peer j prefers their own match over recruiter i , the utility that i has for j is irrelevant, so the draw is unconstrained.

Finally, when i and j are not matched and $j \in \mathcal{A}$ (j is an abstention), the utility that i has for j must be less than the utility i has for any of their matches. This is the scenario in the fourth row of Table 1. In particular, $u_{ij} \leq \min(\mathcal{U}_{i,m(i)})$. Thus, $a_{\mathbf{U};ij} = -\infty$ and $b_{\mathbf{U};ij} = \min(\mathcal{U}_{i,m(i)})$.

Based on these constraints, $u_{ij}^{(k+1)} \sim \psi(\alpha^{(k)} \mathbf{X}_{ij} + \xi^{(k)} \mathbf{A}_{ij}, \sigma_\epsilon^2, a_{\mathbf{U};ij}, b_{\mathbf{U};ij})$.

The derivation for the conditional distribution of \mathbf{V} is similar, differing only due to the many-to-one nature of the matching. For \mathbf{V} , again explicitly stating the observed covariates \mathbf{Y} and declension structure \mathbf{D} , and assuming $\gamma_{ji} \stackrel{iid}{\sim} \text{N}(0, \sigma_\gamma^2)$:

$$p(v_{ji} | \mathbf{V}_{-\{ji\}}, \beta, \zeta, \mathbf{U}, \mathbf{Y}_{ji}, \mathbf{D}_{ji}, m \in \mathcal{M}) = \text{N}(\beta \mathbf{Y}_{ji} + \zeta \mathbf{D}_{ji}, \sigma_\gamma^2) \quad (\text{S3.5})$$

constrained such that \mathbf{U} and the updated \mathbf{V} reproduce the observed matching via the matching mechanism in Algorithm 1.

Therefore, given $\beta^{(k)}$ and $\zeta^{(k)}$, new draws of $\mathbf{V}^{(k+1)}$ are obtained from truncated normal distributions: $v_{ji}^{(k+1)} \sim \psi(\beta^{(k)} \mathbf{Y}_{ji} + \zeta^{(k)} \mathbf{D}_{ji}, \sigma_\gamma^2, a_{\mathbf{V};ji}, b_{\mathbf{V};ji})$ with constraints shown in Table S1. Note that these constraints are similar to those for draws of $\mathbf{U}^{(k+1)}$, but differ because of the many-to-one nature of matching.

S.3.2 Implementation

Implementation of the RCPS method is performed in R (R Core Team, 2017) via the `rcps` package, which is available from the authors upon request and will be posted on Github at the time of publication. This package is designed to function in conjunction with from the RDS package (Handcock et al., 2016). For a desired number of iterations K the Markov

Dyad Type	Match	$a_{\mathbf{V};ji}$	$b_{\mathbf{V};ji}$
Peer-Rec	Yes	$\max(v_{j,\mathcal{D}_j} \cup \mathcal{V}_{j,\ell \in \mathcal{O}(j)})$	∞
Peer-Rec	No	$-\infty$	$\begin{cases} v_{j,m(j)} & \text{if } u_{ij} > \min(\mathcal{U}_{i,m(i)}) \\ \infty & \text{otherwise} \end{cases}$
Peer Decline	Yes	$\max(\mathcal{V}_{j,\ell \in \mathcal{O}(j)})$	∞
Peer Decline	No	$-\infty$	$v_{j,m(j)}$

Table S1: Upper and lower bounds for truncated draws of v_{ji} where $\mathcal{V}_{j,\ell \in \mathcal{O}(j)}$ is the set of utilities j has for recruiters $\ell \neq i$ such that ℓ is in j 's opportunity set $\mathcal{O}(j)$. $\mathcal{O}(j)$ is the set of recruiters $\ell \in \mathcal{R}$, $\ell \neq m(j)$ such that $u_{\ell j} > \min(\{u_{\ell,m(\ell)}\})$. v_{j,\mathcal{D}_j} is j 's declension utility. $m(i)$ is the set of peers that i is matched to and $m(j)$ is the recruiter of j .

chain is updated as described in Algorithm S1. Note that the procedure described above in Section S.3 is only for one wave (that is, for one bipartite network between one set of recruiters and one set of peers). In practice, we observe multiple waves of recruitment, and thus we sequentially apply the method where the α , β , ξ , and ζ from wave $w - 1$ during iteration (k) are used as the starting values of α , β , ξ , and ζ during wave w of iteration (k). Values of α , β , ξ , and ζ for wave 1 of iteration ($k + 1$) are informed by wave n_w of iteration (k). Algorithm S1 includes this sequential application.

Inputs to the inference function are: the observed recruitment chain m ; the observed covariate array \mathbf{Z} ; priors μ_α , Σ_α , μ_β , Σ_β , μ_ξ , Σ_ξ , μ_ζ , and Σ_ζ for the preference coefficients; and the desired number of iterations K .

S.3.3 Model Selection

The RCPS model framework can include numerous covariates Z . Values of α and β for each covariate can be examined to determine whether it seems to impact recruitment, and thus candidate models can be compared. In particular, the posterior probability that the coefficient is above zero for homophilous recruitment will be close to 1. Although there are no set cut-off points for when a covariate is ‘important’ for recruitment, potential comparison values are 0.9 and 0.95. Thus, for example, if the posterior probability that α is greater than zero is 0.97, we would say that this covariate seems to impact recruitment. On the other hand, if the posterior probability that α is greater than zero is 0.5, we would say that this covariate does not seem to affect recruitment. For heterophilous recruitment, we would consider the posterior probability that α and β are less than zero. In order to select a final model (i.e., a final set of covariates upon which participants recruit preferentially), a larger set of covariates can be pruned by considering their posterior probability above (or below) zero.

To compare whether sets of covariates impact preferential recruitment, we could compare them to the null model where $\alpha = 0$ and $\beta = 0$ for all covariates. Then *Bayes factors* can be used to compare the models (Gelman et al., 2004).

Algorithm S1: Gibbs sampler for joint posterior $\pi(\alpha, \beta, \xi, \zeta, \mathbf{U}, \mathbf{V} \mid m \in \mathcal{M})$

Input: Hyperparameters $\mu_\alpha, \mu_\beta, \mu_\xi, \mu_\zeta, \mu_\epsilon, \mu_\gamma, \Sigma_\alpha, \Sigma_\beta, \Sigma_\xi, \Sigma_\zeta, \sigma_\epsilon, \sigma_\gamma$; covariates \mathbf{Z} ; matching m ; number of iterations K

Output: Joint posterior $\pi(\alpha, \beta, \xi, \zeta, \mathbf{U}, \mathbf{V} \mid m \in \mathcal{M})$

```

1 initialize  $\alpha_0^{(0)}, \beta_0^{(0)}, \xi_0^{(0)}$ , and  $\zeta_0^{(0)}$ ;
2  $\alpha_0^{(0)} \sim \mathcal{MVN}(\mu_\alpha, \Sigma_\alpha)$ ;
3  $\beta_0^{(0)} \sim \mathcal{MVN}(\mu_\beta, \Sigma_\beta)$ ;
4  $\xi_0^{(0)} \sim \mathcal{MVN}(\mu_\xi, \Sigma_\xi)$ ;
5  $\zeta_0^{(0)} \sim \mathcal{MVN}(\mu_\zeta, \Sigma_\zeta)$ ;
6 for  $w = 1 \dots, n_w$  do
7   initialize  $\epsilon_w$  and  $\gamma_w$ ;
8    $\epsilon_{w;ij} \sim N(\mu_\epsilon, \sigma_\epsilon^2)$ ;
9    $\gamma_{w;ji} \sim N(\mu_\gamma, \sigma_\gamma^2)$ ;
10  calculate initial values  $\mathbf{U}_w^{(0)}$  and  $\mathbf{V}_w^{(0)}$ ;
11   $u_{w;ij}^{(0)} \mid \alpha^{(0)}, \xi^{(0)}, \mathbf{U}_{w;-\{ij\}}, \mathbf{V}_w, m \in \mathcal{M} = \alpha_{w-1}^{(0)} \mathbf{X}_{w;ij} + \xi_{w-1}^{(0)} \mathbf{A}_{w;ij} + \epsilon_{w;ij}$ ;
12   $v_{w;ji}^{(0)} \mid \beta^{(0)}, \zeta^{(0)}, \mathbf{U}_w, \mathbf{V}_{w;-\{ji\}}, m \in \mathcal{M} = \beta_{w-1}^{(0)} \mathbf{Y}_{w;ji} + \zeta_{w-1}^{(0)} \mathbf{D}_{w;ji} + \gamma_{w;ji}$ ;
13  draw initial values  $\alpha_w^{(1)}, \beta_w^{(1)}, \xi_w^{(1)}$ , and  $\zeta_w^{(1)}$ ;
14   $\alpha_w^{(1)} \mid \mathbf{U}_w^{(0)}, m \in \mathcal{M} \sim \mathcal{MVN}(\tilde{\mu}_\alpha^{(1)}, \tilde{\Sigma}_{\alpha;w})$ ;
15   $\beta_w^{(1)} \mid \mathbf{V}_w^{(0)}, m \in \mathcal{M} \sim \mathcal{MVN}(\tilde{\mu}_\beta^{(1)}, \tilde{\Sigma}_{\beta;w})$ ;
16   $\xi_w^{(1)} \mid \mathbf{U}_w^{(0)}, m \in \mathcal{M} \sim \mathcal{MVN}(\tilde{\mu}_\xi^{(1)}, \tilde{\Sigma}_{\xi;w})$ ;
17   $\zeta_w^{(1)} \mid \mathbf{V}_w^{(0)}, m \in \mathcal{M} \sim \mathcal{MVN}(\tilde{\mu}_\zeta^{(1)}, \tilde{\Sigma}_{\zeta;w})$ ;
18   $k = 1$ 
19 end
20 while  $k < K$  do
21   for  $w = 1, \dots, n_w$  do
22      $check = 0$ ;
23     while  $check = 0$  do
24       update  $\mathbf{U}_w^{(k+1)}$  and  $\mathbf{V}_w^{(k+1)}$ ;
25        $u_{w;ij}^{(k+1)} \mid \alpha^{(k)}, \xi^{(k)}, \mathbf{U}_{w;-\{ij\}}, \mathbf{V}_w, m \in \mathcal{M}$ ;
26          $\sim \psi(\alpha_{w-1}^{(k)} \mathbf{X}_{w;ij} + \xi_{w-1}^{(k)} \mathbf{A}_{w;ij}, \sigma_\epsilon^2, a_{\mathbf{U};w;ij}^{(k+1)}, b_{\mathbf{U};w;ij}^{(k+1)})$ ;
27        $v_{w;ji}^{(k+1)} \mid \beta^{(k)}, \zeta^{(k)}, \mathbf{U}_w, \mathbf{V}_{w;-\{ji\}}, m \in \mathcal{M}$ ;
28          $\sim \psi(\beta_{w-1}^{(k)} \mathbf{Y}_{w;ji} + \zeta_{w-1}^{(k)} \mathbf{D}_{w;ji}, \sigma_\gamma^2, a_{\mathbf{V};w;ji}^{(k+1)}, b_{\mathbf{V};w;ji}^{(k+1)})$ ;
29       if  $\mathbf{U}_w^{(k+1)}$  and  $\mathbf{V}_w^{(k+1)}$  reproduce the observed matching then
30          $check = 1$ ;
31       end
32     end
33     update  $\alpha_w^{(k+1)}$  and  $\beta_w^{(k+1)}$ ;
34      $\alpha_w^{(k+1)} \mid \mathbf{U}_w^{(k)}, m \in \mathcal{M} \sim \mathcal{MVN}(\tilde{\mu}_\alpha^{(k+1)}, \tilde{\Sigma}_{\alpha;w})$ ;
35      $\beta_w^{(k+1)} \mid \mathbf{V}_w^{(k)}, m \in \mathcal{M} \sim \mathcal{MVN}(\tilde{\mu}_\beta^{(k+1)}, \tilde{\Sigma}_{\beta;w})$ ;
36      $\xi_w^{(k+1)} \mid \mathbf{U}_w^{(k)}, m \in \mathcal{M} \sim \mathcal{MVN}(\tilde{\mu}_\xi^{(k+1)}, \tilde{\Sigma}_{\xi;w})$ ;
37      $\zeta_w^{(k+1)} \mid \mathbf{V}_w^{(k)}, m \in \mathcal{M} \sim \mathcal{MVN}(\tilde{\mu}_\zeta^{(k+1)}, \tilde{\Sigma}_{\zeta;w})$ ;
38   end
39    $k = k + 1$ ;
40 end

```

Variable	Description	a	b	c	d	e	f
N	Population size	2000					
n_v	Number of covariates	1					
\mathbf{Z}	$N \times n_v$ matrix of covariates	$z_i \stackrel{iid}{\sim} \text{Bernoulli}(0.5)$					
n_c	Number of coupons	3					
n_s	Number of seeds	2	2	2	12	4	3
n_w	Number of waves	4	4	4	4	6	10
α	Pref coef for recruiters to peers	2	-2	0	0	0.5	0.01
β	Pref coef for peers to recruiters	2	-2	0	0	0.5	0.01
ξ	Pref coef for recruiter abstention	(-2,-2,-2)	(-2,-2,-2)	(-2,-2,-2)	(6,6,-2)	(2,2,2)	(1.5,1.5,1.5)
ζ	Pref coef for peer declension	-2	-2	-2	-2	3.5	3.5

Table S2: Simulation input parameters producing the recruitment trees shown in Figure S3.

S.4 Simulation of Recruitment Chains

This section demonstrates a variety of recruitment chains that can be produced by altering the simulation input parameters. Input parameters are shown in the first two columns of Table S2. These parameters can be varied to obtain a wide range of recruitment chains, encompassing many different types we observe in real data. Six examples are shown in Figure S3, demonstrating homophilous, heterophilous, and non-preferential recruitment with a variety of recruitment chain shapes. The specific inputs for these recruitment chains are given in the last six columns of Table S2. The population size, number of type of covariates, and number of coupons are kept the same in these simulations to highlight the effect of varying the values of α , β , ξ , and ζ .

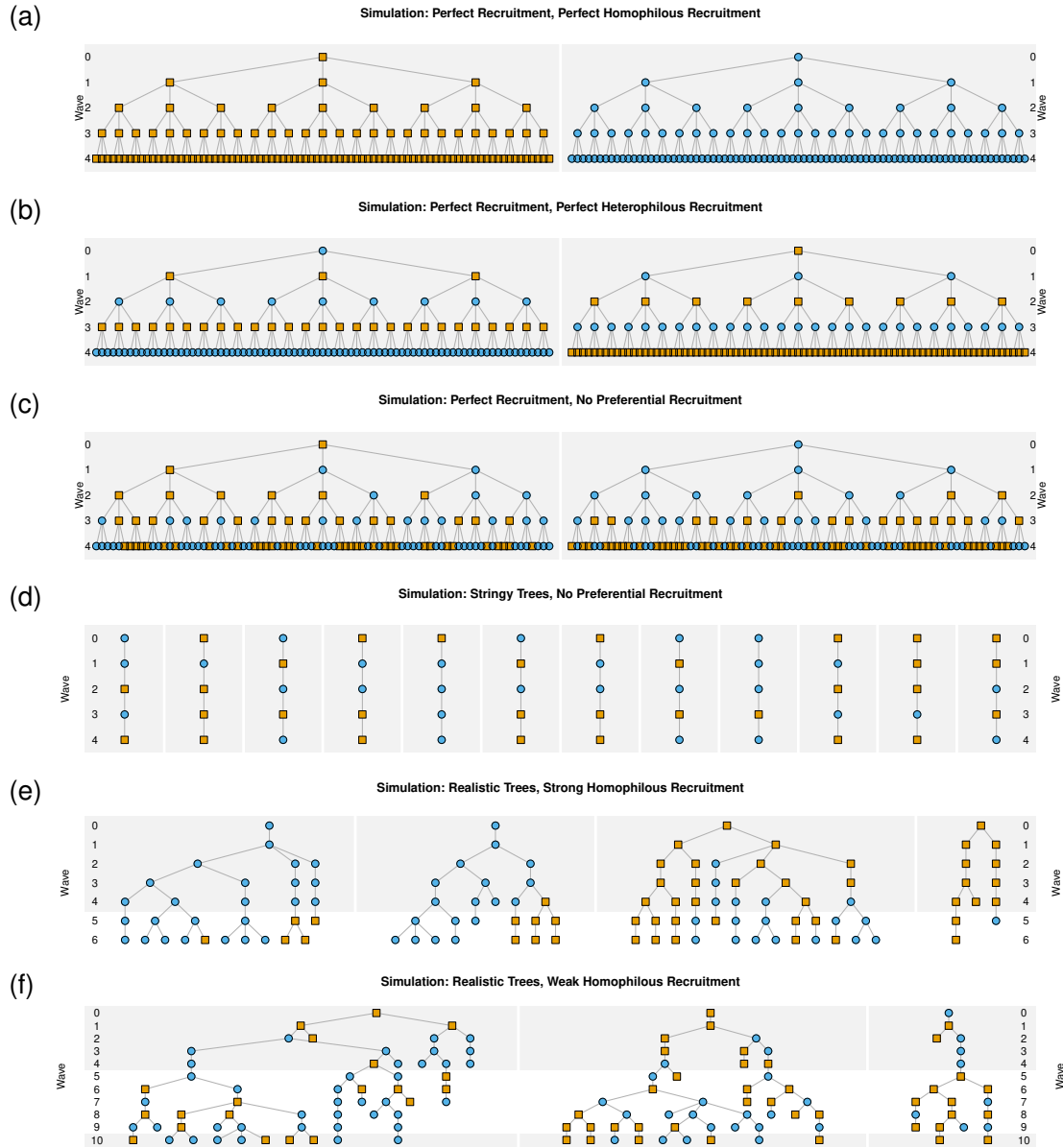


Figure S3: The RCPS framework can be used to create a variety of recruitment chains, such as: bushy (a,b,c), stringy (d), and realistic (e,f) trees; homophilous (a-perfect, e-strong, f-weak), heterophilous (b), and non-preferential (c,d) recruitment. Points colored and shaped uniquely for a covariate with 2 categories. Input parameters are given in Table S2.

S.5 Simulation Study 1: No Preferential Recruitment

This section presents additional figures for the simulation study presented in the manuscript where $\alpha = \beta = 0$ (no preferential recruitment). Figure S4 shows MCMC posterior draws for all preference, abstention, and declension parameters for each of the 7 generated recruitment chain replicates. For each, a burn-in of 5000 iterations and a thinning interval of 20 iterations were used, and the sampler was run until 2250 posterior draws were attained. Figure S5 shows the posterior distributions for the preference parameters α and β for each of the 7 generated recruitment chain replicates.

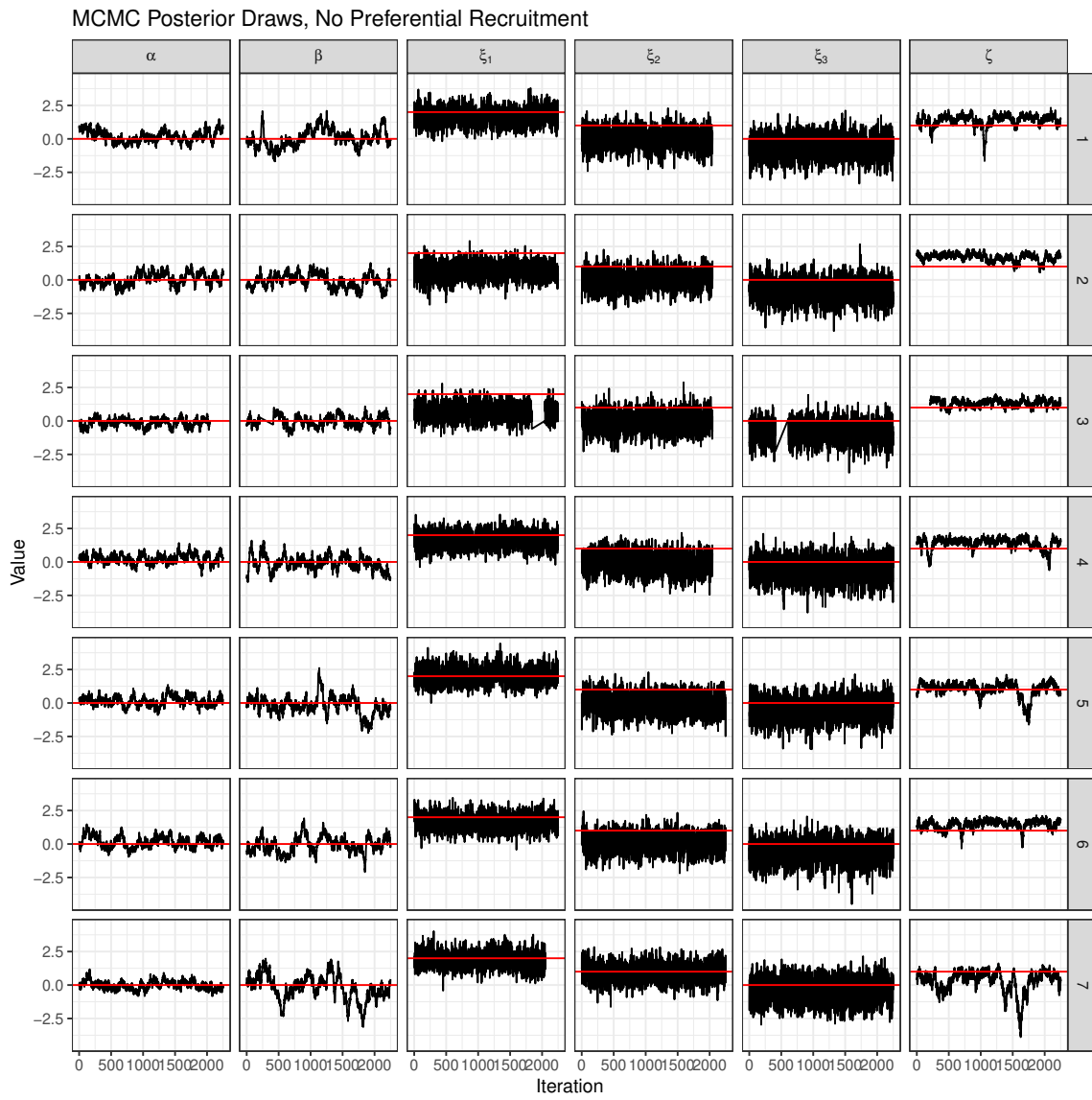


Figure S4: MCMC posterior draws for all preference, abstention, and declension parameters for the case of no preferential recruitment for each of the 7 generated recruitment chains. True values are shown as red horizontal lines.

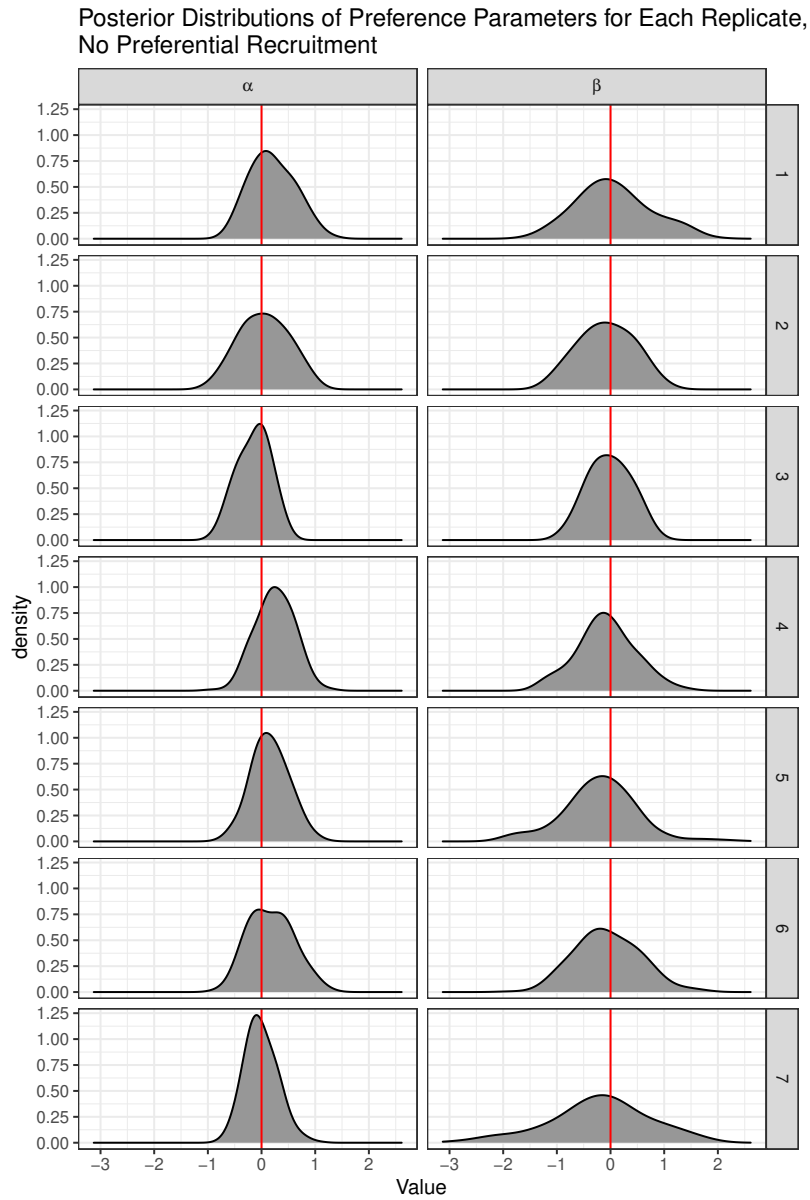


Figure S5: Posterior distributions for the preference parameters α and β for the case of no preferential recruitment for each of the 7 generated recruitment chains. True values are shown as red vertical lines.

S.6 Simulation Study 2: Homophilous Recruitment

This section presents additional figures for the simulation study presented in the manuscript where $\alpha = \beta = 0.5$ (heterophilous recruitment). Figure S6 shows MCMC posterior draws for all preference, abstention, and declension parameters for each of the 7 generated recruitment chain replicates. For each, a burn-in of 5000 iterations and a thinning interval of 20 iterations were used, and the sampler was run until 1000 posterior draws were attained. Figure S7 shows the posterior distributions for the preference parameters α and β for each of the 7 generated recruitment chain replicates.

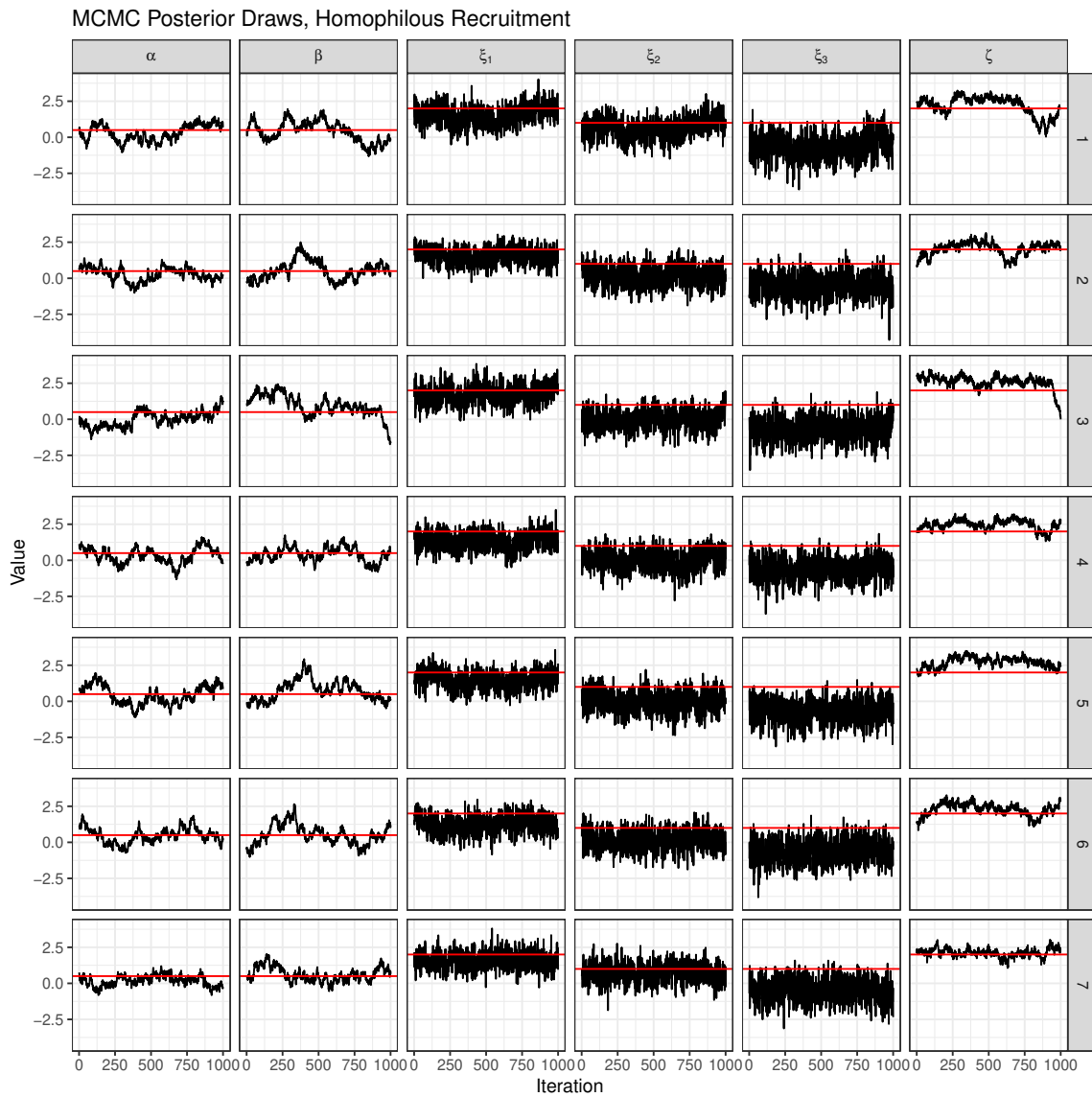


Figure S6: MCMC posterior draws for all preference, abstention, and declension parameters for the case of homophilous recruitment for each of the 7 generated recruitment chains. True values are shown as red horizontal lines.

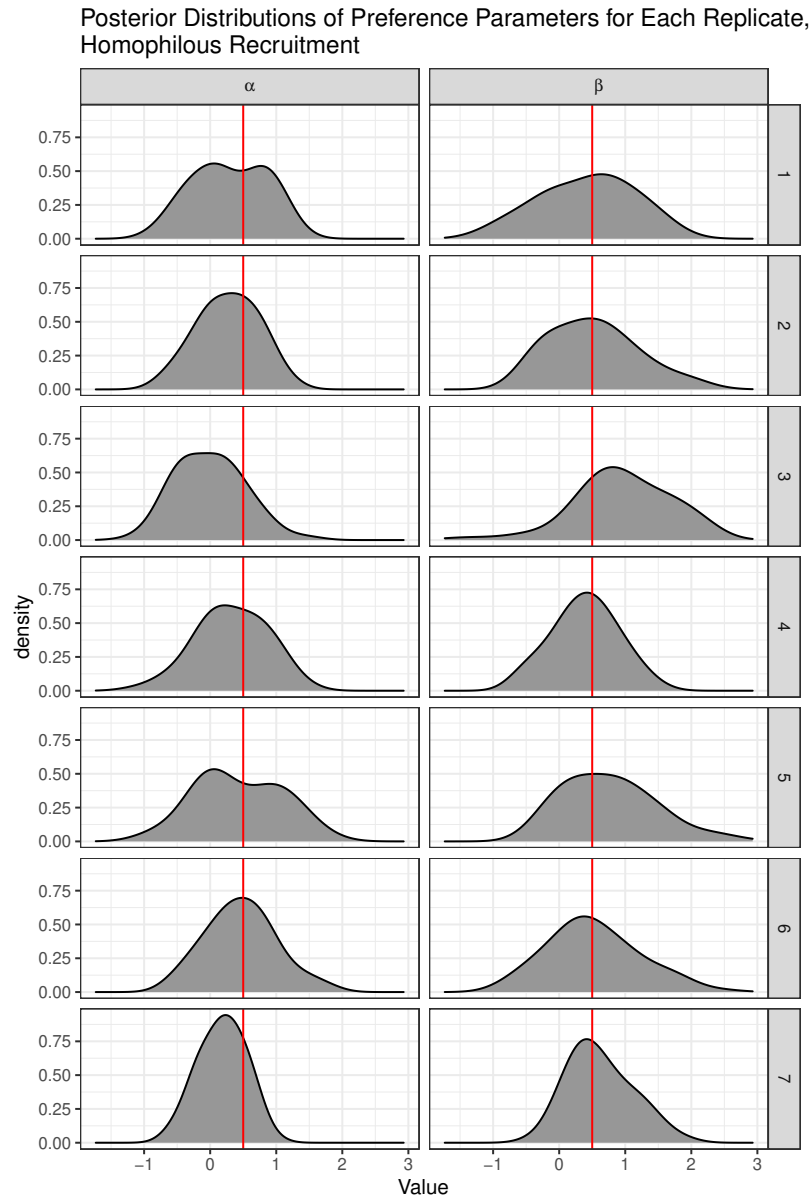


Figure S7: Posterior distributions for the preference parameters α and β for the case of homophilous recruitment for each of the 7 generated recruitment chains. True values are shown as red vertical lines.

S.7 Simulation Study 3: Heterophilous Recruitment

This section presents additional figures for the simulation study presented in the manuscript where $\alpha = \beta = -0.5$ (heterophilous recruitment). Figure S8 shows MCMC posterior draws for all preference, abstention, and declension parameters for each of the 7 generated recruitment chain replicates. For each, a burn-in of 5000 iterations and a thinning interval of 20 iterations were used, and the sampler was run until 4750 posterior draws were attained. Figure S9 shows the posterior distributions for the preference parameters α and β for each of the 7 generated recruitment chain replicates.

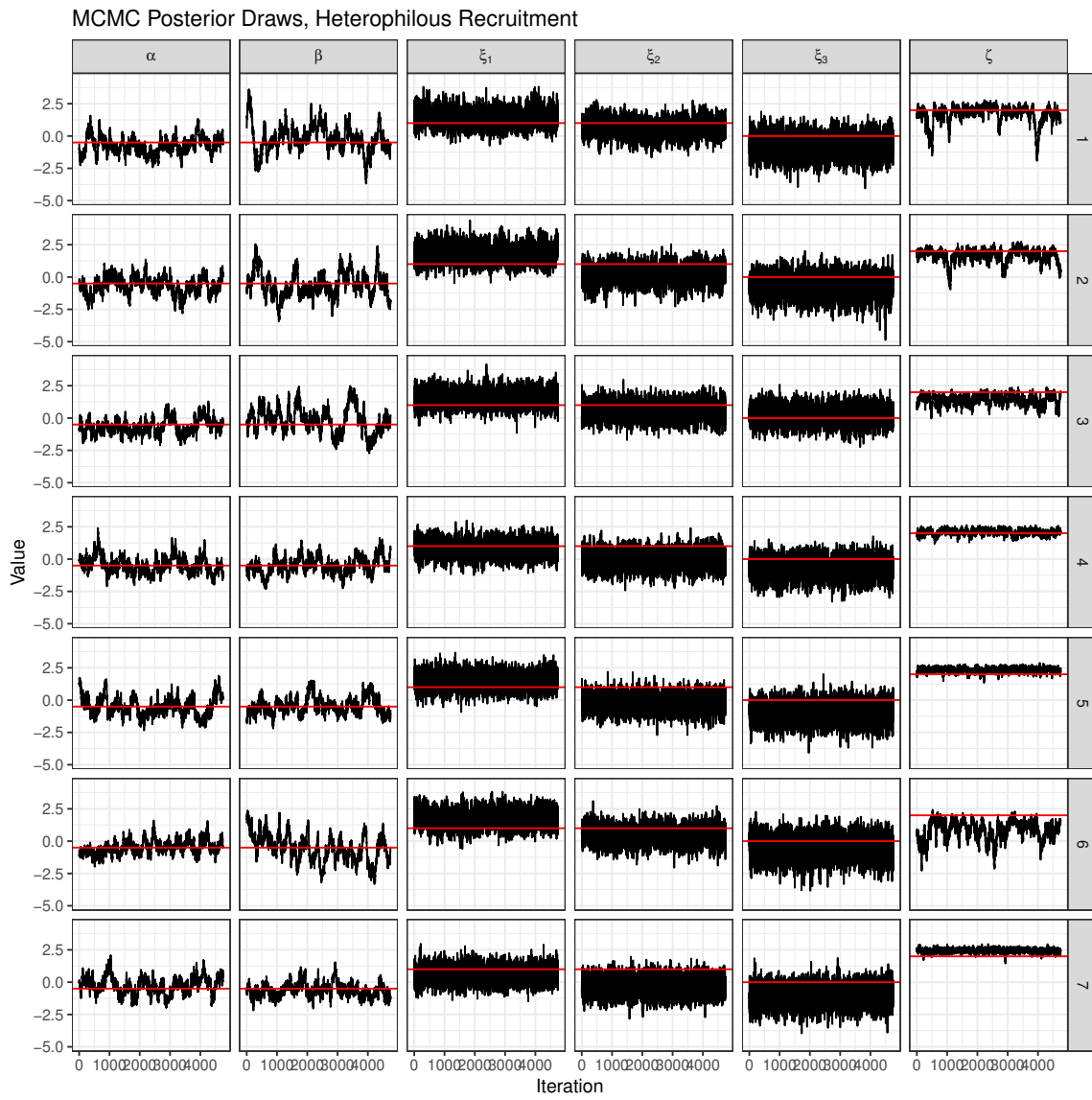


Figure S8: MCMC posterior draws for all preference, abstention, and declension parameters for the case of heterophilous recruitment for each of the 7 generated recruitment chains. True values are shown as red horizontal lines.

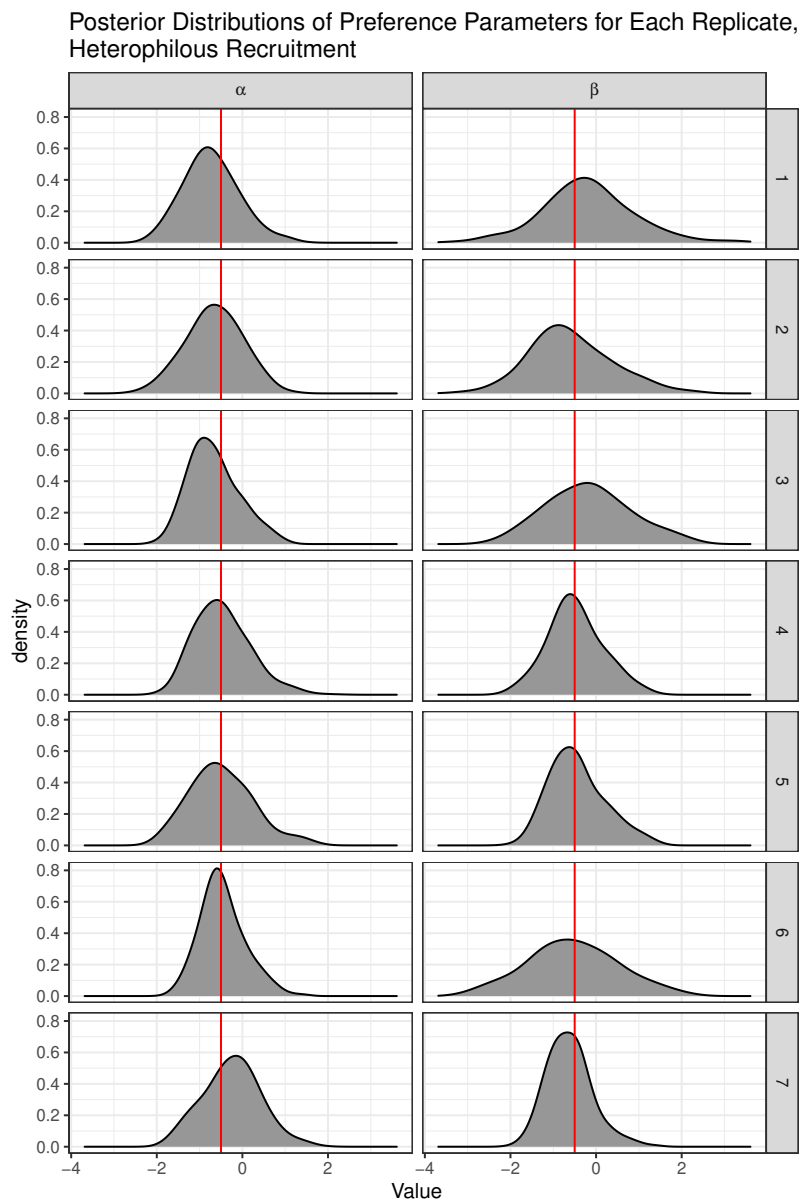


Figure S9: Posterior distributions for the preference parameters α and β for the case of heterophilous recruitment for each of the 7 generated recruitment chains. True values are shown as red vertical lines.

S.8 Re-Simulating From Posterior Medians

To assess simulation performance, we compare the original simulated recruitment chains to those generated from new simulations using the posterior medians of ξ and ζ (as well as α and β) as the new input values. We simulated 200 new recruitment chains for each scenario in this manner, and compare the number of recruits each person enrolls ($0, 1, \dots, n_c$) to the median from each of the 7 replicates. Figure S10 shows the distribution of the re-sampled number of recruits, with red vertical lines indicating the median from each of the replicates. For the most part, the RCPS model performs well, although it perhaps results in ξ and ζ values that return trees that are too stringy in a few cases.

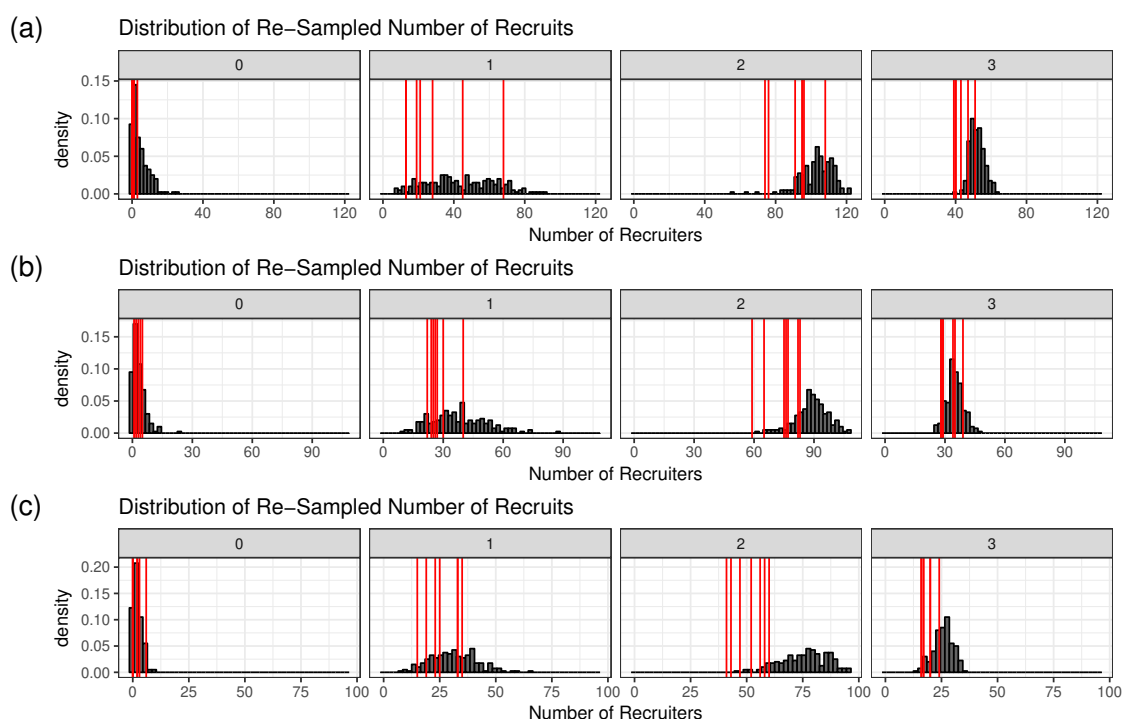


Figure S10: Distribution of number of recruits for new recruitment chains simulated from posterior medians of ξ and ζ for (a) non-preferential, (b) homophilous, and (c) heterophilous recruitment. Posterior medians from each of 7 replicates shown by red solid line.

S.9 Francophone Migrants: Additional Plots

We assess the structure of the recruitment chains simulated using the RCPS model fit for the Francophone migrants in Rabat, Morocco. Figure S11 shows histograms of the number of recruits per person in (a) and the number of people in each wave in (b). Despite the visual similarity of the recruitment chains in Figure 4, the simple model does not appear to

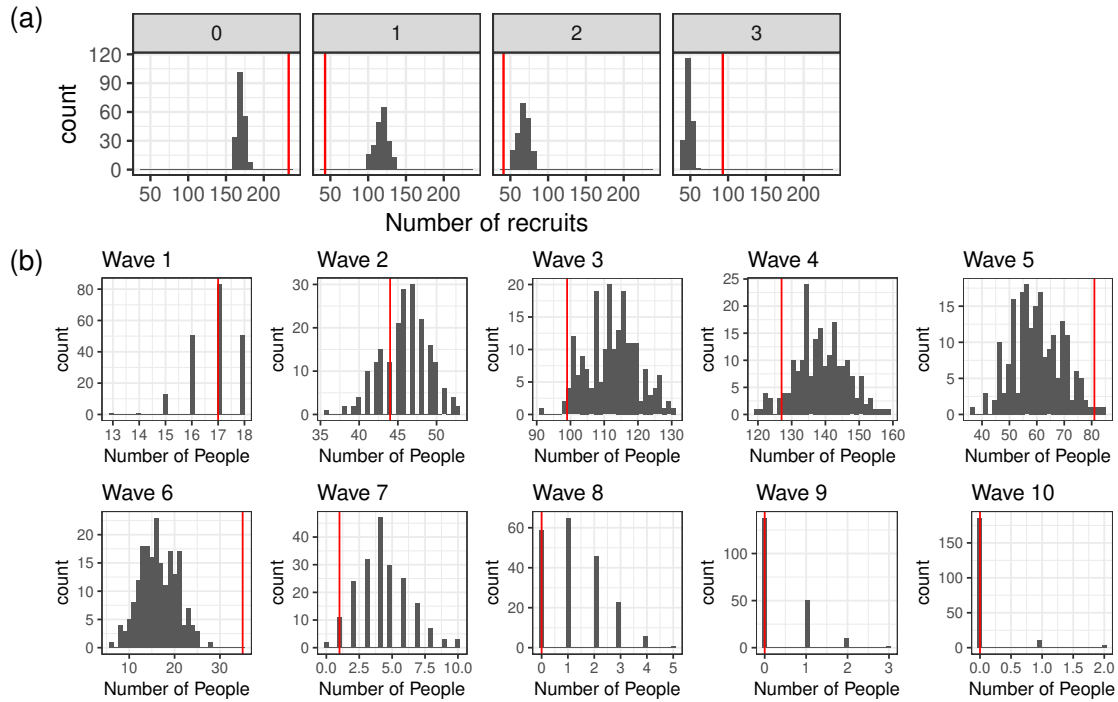


Figure S11: Histograms of (a) the number of recruits per person and (b) the number of people in each wave, in 200 simulations using the posterior medians from the RCPS model for the Francophone migrants. The red vertical lines show the true values.

capture all aspects of the recruit distribution, with too many people recruiting one or two peers, and too few recruiting zero or three peers. This suggests a more realistic extension of the RCPS model would be to have a separate parameter for whether a person will recruit at all, and then if they do recruit, model the number of peers they will select as in the RCPS model. The model does a good job at capturing the true values for small and large number waves, but a poorer job in waves 4-6. The abstention and declension parameters are not wave-specific (that is, we assume there is one overall parameter governing the recruitment process), so this suggests another extension would be to allow abstention and declension to vary based on wave to more realistically capture recruitment dynamics. Nevertheless, the RCPS model captures the preferential recruitment observed for both living location and network size.

References

- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, 2nd edition.
- Handcock, M. S., Fellows, I. E., and Gile, K. J. (2016). *RDS: Respondent-Driven Sampling*.

Los Angeles, CA. URL <https://CRAN.R-project.org/package=RDS>. R package version 0.7-8.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.