Statistical Methods for Evaluating the Relationships Among Social, Spatial and Genetic Networks

Statistical Methods for Evaluating the Relationships Among Social, Spatial and Genetic Networks

a dissertation presented
by
Fei Li
to
The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts
May 2018

Thesis advisor: Professor Victor De Gruttola, Jukka-Pekka Onnela Fei Li

Statistical Methods for Evaluating the Relationships Among Social, Spatial and Genetic Networks

Abstract

Empirical findings suggest close relationships among social, spatial and genetic networks. Information learned from one network would provide beneficial insight on the others. These relationships offer the flexibility to obtain understanding of networks that are hard to observe, such as ones for disease transmission.

In Chapter 1, we start from theoretical models and empirical findings which suggest that the intensity of communication among groups of people declines with their degree of geographical separation. Based on the evidence that rather than decaying uniformly with distance, the intensity of communication might decline at different rates for shorter and longer distances, we introduce a statistical model based on Bayesian LASSO for estimating the rate of communication decline with geographic distance that allows for discontinuities in this rate. We apply our method to an anonymized mobile phone communication dataset and discover some geographic patterns.

Based on the findings in Chapter 1, it becomes clear that methods which can provide statistical justification for association between communities based on different networks are important to infer connections in one network based on another network.

In Chapter 2, we justify the use of a permutation test focusing on testing the null that there is no association between pairs of nodes in the same community in one assignment and they are in the same community according to the other assignment. In simulation, we evaluate its performances and find that the permutation test preserves the type I error and has adequate power. The use of the test is then demonstrated on work and social relations data of a tailor shop in Zambia.

In Chapter 3, with the tools introduced in Chapter 2, we take one step closer to evaluate association between clusters based on geographic distances and on viral genetic distances. By applying the hypothesis testing framework to analyze this association among individuals infected by HIV viral strains, we find significant association between geographic pattern and viral genetic clusters, which is not greatly affected by different divergence threshold values used in the construction for viral genetic clusters. Our findings provide an alternative view and are consistent with previous research.

Contents

Acknowledgments

We acknowledge Xihong Lin for many useful discussions. We thank Amanda King for her help in editing the manuscripts, and members of the Onnela Lab and Professor De Gruttola's discussion group in Biostatistics Department at Harvard.

1

Bayesian method for inferring the impact of geographical distance on intensity of communication

1.1 Problems Statement and Motivation

Observations of one way of communication are generally informative about the use of others (Eagle et al., 2008; Wang et al., 2011; Hawelka et al., 2014). For example, people who speak one the phone frequently are also likely to interact in person (Eagle et al., 2009). For researchers studying infectious diseases, such as HIV or Malaria, the structure of social interactions in a population can provide valuable insights into how viruses are transmitted among members of that population (Gregson et al., 2002; Jones & Handcock, 2003; Helleringer & Kohler, 2007). Because traditional surveys are resource intensive and scale poorly, mobile phone data, or more specifically call detail records (CDRs), have emerged as an alternative for inferring the structure of underlying interpersonal interactions (Onnela et al., 2007; Buckee et al., 2013; Tatem et al., 2014).

Although user interactions on the mobile phone network are not limited by geography, users themselves are subject to spatial constraints that restrict the locations they may frequent and therefore influence their overall interpersonal and mobile phone communication patterns. Individual-level analysis in Sailer & McCulloh (2012) demonstrated a relationship between spatial configuration of offices and social connections among employees. Overlap of geographical space and information flow network is discussed in Ter Wal & Boschma (2009) from a perspective of the spread of innovation and knowledge. The effect of geographic restrictions may differ for locations in different regions. For example, Lambiotte et al. (2008) and Expert et al. (2011) found that in Belgium, cell phone users communicate mostly within language-specific network communities (Porter et al., 2009) of French and Flemish speakers. In another example, Wang et al. (2013) showed that contact patterns between individuals with respect to disease transmission are location-specific. Potential overlap of the geographical and social networks on the topological level has also been explored. The connection between local network topology and tie strength is found to be consistent with the so-called weak-ties hypothesis (Granovetter, 1973) in Onnela et al. (2007). However, geographical and community centrality were not found to be related in Onnela et al. (2011).

In this study, we investigate the impact of spatial distance on cell phone communication using a statistical approach. Our choice of model is guided by the observation that the intensity of communication among groups of people tends to decay with

geographical distance and, further, the rate of decay in intensity appears to differ between short and long distances. To incorporate this feature, we allow for the existence of a break-point in the relationship between communication intensity and spatial distance. As the structure of electronic communication, human mobility and travel, and in-person social interactions are all related, we make use of existing methods and models in these areas. Some of the most widely studied models in these fields are the gravity model (Krings et al., 2009; Lambiotte et al., 2008; Balcan et al., 2009a; Noulas et al., 2012; Csáji et al., 2013), the radiation model (Simini et al., 2012), and the rank-based friendship model (Liben-Nowell et al., 2005). Both the radiation model and the rank-based friendship model make explicit mechanistic assumptions regarding the effect of distance and population sizes, and these models focus on prediction. The gravity model is simpler and ignores the geographical distribution of the population, as it uses only the source and destination population sizes and the spatial distance between them.

We extend the gravity model by relaxing the assumption of homogeneity in distance effects; its unsatisfactory performance in prediction compared with the radiation model shown in Simini et al. (2012) is mainly due to the assumption of an identical decay rate for all distances. We explicitly incorporate the potential for heterogeneity of distance effects into our model, and we also provide an estimate and an interval for the break-point between short and long distances.

This paper is organized as follows. Section 2 and 3 introduce notation and the mod-

els, sampling schemes, diagnosis of convergence and computational complexity. Section 4 describes the design of the simulation study and the dataset we analyze. Section 5 provides the results of the simulation study, and Section 6 presents the data analyses. We conclude with a discussion in Section 7.

1.2 Notation and Theory

Our starting point to investigating the relationship between spatial distance and communication intensity is the so-called gravity model. Using the notation from Krings et al. (2009), the gravity model can be written as

$$G_{ij} = K\frac{m_i n_j}{r_{ij}^2}, \tag{1.1}$$

where G_{ij} specifies the communication intensity from source location i to destination location j, K is a constant, m_i is the population of the source location i, n_j is the population of the destination location j, r_{ij} is the distance between source i and destination j.

In Simini et al. (2012), the model is extended to the following:

$$G_{ij} = \frac{m_i^\alpha n_j^\beta}{f(r_{ij})}, \tag{1.2}$$

where $f(\cdot)$ is a function that specifies the decay of G_{ij} with distance r_{ij}, and it is usu-

ally specified as r_{ij}^{γ}. Here, we adopt the following form of the model:

$$G_{ij} = K\frac{m_i^{\alpha} n_j^{\beta}}{r_{ij}^{\gamma}}. \tag{1.3}$$

Taking the logarithm of this yields

$$\log(G_{ij}) = \log(K) + \alpha\log(m_i) + \beta\log(n_j) - \gamma\log(r_{ij}). \tag{1.4}$$

1.3 A Bayesian approach

1.3.1 Model

We extend the gravity model shown in Equation 1.4 in the following way:

$$\begin{aligned} Y_{ij} &= \mu + \beta_1\log(n_i) + \beta_2\log(n_j) + \beta_{3,i}\log(d_{ij}) + \beta_{4,i}(\log(d_{ij}) - \theta_i)_+ + \epsilon_{ij}, \\ & i, j = 1, ..., S, j \neq i, \end{aligned} \tag{1.5}$$

where $Y_{ij} = g(G_{ij})$ and $g(\cdot)$ is a transformation function, in the gravity model, $g(\cdot) = \log(\cdot)$; μ is the intercept; θ_i represents the location of the break point measured on the logarithmic scale for communication initiated from location i; $\beta_{3,i}$ represents the distance effect before break point θ_i; and $\beta_{4,i}$ specifies the difference of distance effect before and after the break point. When $\beta_{4,i} = 0$, the difference is 0, i.e. the rate of decay does not change over the observed range. We denote the size of the population

at location i as n_i and refer the model with $\beta_{4,i}$ as the *full model* and the model that sets $\beta_{4,i}$ to 0 as the *reduced model*. By definition, $(d_{ij} - \theta_i)_+ = (d_{ij} - \theta_i)I(d_{ij} > \theta_i)$, which takes value 0 before the break point θ_i and $d_{ij} - \theta_i$ after the break point. We assume that $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$; S denotes the number of locations, and i and j are indexes. This formulation provides a straightforward way to compare the two nested models with regard to the effect of distance effect; the reduced model has the constraint $\beta_{4,i} = 0$. In this formulation, model selection becomes a variable selection problem that can be achieved using a variety of methods, such as LASSO. We are also interested in estimating θ_i and quantifying its uncertainty. To achieve these goals, we employ a Bayesian approach with a Metropolis sampling block for $\boldsymbol{\theta} \equiv (\theta_1, \theta_2, ..., \theta_S)^T$ and a Bayesian LASSO block dealing with $\boldsymbol{\beta_4} \equiv (\beta_{4,1}, \beta_{4,2}, ..., \beta_{4,S})^T$.

We note that the above model assumes that the full and nested models share the same intercept and population size effects — an assumption that might not hold in practice. To address this concern, we consider two distinct settings, or cases. In what follows, *case I* refers to the setting where the assumption holds, and *case II*, to the setting it does not. For the latter, we extend the model by making use of the Reversible Jump MCMC (RJMCMC) option in the `blasso` function in `R` package `monomvn`. This approach allows for statistical inference using Bayesian LASSO. Briefly, RJMCMC sampling procedure permits a change in the model matrix based on the variable selection results from the previous iteration; the intercept and population size effects are modeled separately for the two models. We provide details in the next

section.

1.3.2 Sampling algorithm

Initial values

To speed up convergence and prevent the algorithm from converging to a local mode, we calculate a set of crude initial values for all the parameters as follows:

1. Search through a grid over the distance range of location i for θ_i and choose the grid point that maximizes the likelihood function of the crude full model $\boldsymbol{\theta^{(0)}}$.

2. For case I, the preliminary values for the parameters are obtained by linear regression treating the break points as known. Substituting in the value of $\boldsymbol{\theta^{(0)}}$ from Step 1 leads to crude parameter estimates $\mu^{(0)}$, $\boldsymbol{\beta^{(0)}} \equiv (\beta_1^{(0)}, \beta_2^{(0)}, \boldsymbol{\beta_3^{(0)}}^T, \boldsymbol{\beta_4^{(0)}}^T)^T$ and $\sigma^2_{(0)}$. For case II, we fit two models for each source location: Model 1 has a break point at $\boldsymbol{\theta^{(0)}}$ estimated in Step 1 and Model 2 has no break point. We then assign $\eta_i^{(0)} = 1$ if Model 1 has a lower BIC than Model 2, and assign $\eta_i^{(0)} = 0$ otherwise. We use BIC to account for the fact that Model 1 has more parameters than Model 2. Based on $\boldsymbol{\eta^{(0)}} \equiv (\eta_1^{(0)}, \eta_2^{(0)}, ..., \eta_S^{(0)})^T$, we create a new corresponding model matrix, removing the column of $\beta_{4,i}$ if $\eta_i^{(0)} = 0$, and obtain the crude parameter estimates $\boldsymbol{\mu^{(0)}}, \boldsymbol{\beta^{(0)}}$ and $\sigma^2_{(0)}$ from linear regression. For cases where $\eta_i^{(0)} = 0$, we assign $\beta_{4,i} = 0$.

Metropolis block and Bayesian LASSO

Case I: Assuming same intercept and population size effects across all source locations With Bayesian LASSO, the model is specified as

$$\begin{aligned} Y_{ij} &= \mu + \beta_1 \log(n_i) + \beta_2 \log(n_j) + \beta_{3,i} \log(d_{ij}) + \beta_{4,i}(\log(d_{ij}) - \theta_i)_+ + \epsilon_{ij}, \\ \theta_i &\in (\min_j \log(d_{ij}), \max_j \log(d_{ij})), i, j = 1, ..., S, j \neq i, \end{aligned} \tag{1.6}$$

which can be written as $\boldsymbol{Y} = \mu\mathbf{1} + \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ using matrix notation. μ is not included in the Bayesian LASSO penalty term (Park & Casella, 2008); $\mathbf{1}$ is the vector of 1s; $\boldsymbol{X}$ is the model matrix consisting of logarithmic population sizes and distances, and $\boldsymbol{\beta}$ is the vector of βs.

In general, LASSO (Tibshirani, 1996) solves an unconstrained optimization problem subject to a given bound on the L_1 norm of the parameter vector that is equivalent to

$$\min_{\boldsymbol{\beta}} (\tilde{\boldsymbol{Y}} - \boldsymbol{X\beta})^{\boldsymbol{T}} (\tilde{\boldsymbol{Y}} - \boldsymbol{X\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{1.7}$$

where $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} - \mu\mathbf{1}$ is the centered outcome vector; p is the number of parameters after excluding the intercept. In the Bayesian setting, solution to Equation 1.7 provides the posterior mode estimates when β_j has i.i.d. double exponential priors. As explained in Park & Casella (2008), conditional double exponential priors are used in

the formulation to avoid multiple modes. They can be expressed hierarchically as

$$
\begin{aligned}
&\boldsymbol{Y}|\mu, \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 \sim N(\mu\mathbf{1} + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}), \\
&\boldsymbol{\beta}|\tau_1^2, ..., \tau_p^2, \sigma^2 \sim N(\mathbf{0}, \sigma^2\boldsymbol{D_r}), \text{ where } \boldsymbol{D_r} = \text{diag}(\tau_1^2, ..., \tau_p^2), \\
&\sigma^2, \tau_1^2, ..., \tau_p^2 \sim \pi(\sigma^2)d\sigma^2 \prod_{j=1}^{p} \frac{\lambda^2}{2} e^{-\lambda^2\tau_j^2/2} d\tau_j^2 \ , \ \sigma^2, \tau_1^2, ..., \tau_p^2 > 0.
\end{aligned}
\tag{1.8}
$$

The entire sampling procedure is available using function `blasso` in `R` package `monomvn` with the option for RJMCMC specified as False. To incorporate a Metropolis block for break point estimation, we alternate between the Metropolis and Bayesian LASSO blocks. Validity of this approach is established by regarding it as two components of a Gibbs sampling algorithm. In summary, conditional on break points, our problem is one of a variable selection; conditional on other parameters, break point sampling is a straightforward application of a Metropolis algorithm.

Thus after obtaining the initial values $\boldsymbol{\mu^{(0)}}, \boldsymbol{\beta^{(0)}}, \boldsymbol{\theta^{(0)}}$ and $\sigma^2_{(0)}$, we proceed as follows:

1. At iteration t for each source location i, update break point $\theta_i^{(t+1)}$ using Metropolis algorithm with a normal proposal $N(\theta_i^{(t)}, \sigma_\theta^2)$. The range of θ_i is determined empirically from data, i.e., the posterior likelihood of θ_i has an indicator function term in the product that is 0 if the proposed $\theta_i^{(t+1)}$ is out of the observed empirical log-distance range, thereby assuring that any out-of-range proposal will be rejected.

2. For each location i, if there are fewer than 5% of data points on either side of

$\theta_i^{(t+1)}$ for the subset of data, i.e., $\boldsymbol{Y_i}$, we consider it to be on the boundary, specify $\beta_{4,i}^{(t+1)} = 0$, and remove it from the model in the next estimation step. We denote the number of locations belonging to the boundary sets as $b^{(t+1)}$.

3. Create the corresponding $s(s-1) \times (2 + 2s - b^{(t+1)})$ covariate matrix (intercept column is not included) based on $\boldsymbol{\theta^{(t+1)}}$. Together with the data, $\boldsymbol{\beta^{(t)}}$ (after $\beta_{4,i}^{(t+1)} = 0$ are removed), $\sigma^{(t)2}$ and $\lambda^{(t)}$, input the covariate matrix into the `blasso` function for h iterations (2 or more). The output intercept is $\mu^{(t+1)}$. From the output we also get $\boldsymbol{\beta^{(t+1)}}$ ($\beta_{4,i}^{(t+1)} = 0$ are put back), $\sigma^{(t+1)2}$ and $\lambda^{(t+1)}$.

4. Repeat steps 1-3 until convergence (see below).

Case II: Allowing different intercepts and population size effects for models with and without break-points When there is evidence of the presence of break-points, we estimate these parameters separately in two different models. In this case, estimates of intercepts and population size effects depend on the set of source locations whose data contribute to the estimation in any given iteration. We denote the mean model as $\boldsymbol{\eta^{(t)}}$ for iteration t to maintain consistency with the notation we introduced earlier.

Estimation can be done using the Reversible Jump MCMC option in the `blasso` function, which allows sampling from different models. In our case, different models imply different specification of zeros in $\boldsymbol{\beta_4^{(t)}}$, and are characterized by $\boldsymbol{\eta^{(t)}}$, where $\eta_i^{(t)} = I(\beta_{4,i}^{(t)} > 0)$.

RJMCMC is a general version of the Metropolis-Hastings algorithm introduced by Green (1995), which allows transitions between different states or models of different dimensions. In RJMCMC, trans-dimensional moves are possible through dimension matching by augmenting the parameter vector with a random component. The difference compared to the usual Metropolis-Hastings procedure is the addition of a Jacobian term in the acceptance probability. A thorough review of RJMCMC with more recent comments can be found in Green & Hastie (2009).

Use of RJMCMC yields the following sampling scheme:

1. The first two steps are the same as in case I: At iteration t, for each source location i, update break point $\theta_i^{(t+1)}$ using Metropolis algorithm with a normal proposal $N(\theta_i^{(t)}, \sigma_\theta^2)$. For each location i, if there are fewer than 5% of data points on either side of $\theta_i^{(t+1)}$ for $\boldsymbol{Y_i}$, we specify $\beta_{4,i}^{(t+1)} = 0$ and remove it from the model in the next estimation step.

2. Conditional on $\boldsymbol{\theta^{(t+1)}}$, create the $s(s-1) \times (5 + 2s - b^{(t+1)})$ covariate matrix (intercept column is not included). Data from each source location contribute to their own group's estimation of intercept and population size effects, which depends on $\boldsymbol{\eta_i^{(t)}}$. All data and parameter values from the previous iteration t (including $\sigma^{(t)2}$ and $\lambda^{(t)}$) are used in the `blasso` function with RJMCMC for 3 iterations. 3 is the minimum number of iterations to avoid the situation in which zeros in the previous iteration are carried forward.

3. From Step 2 we get the updated $\boldsymbol{\beta^{(t+1)}}, \sigma^{(t+1)2}, \mu^{(t+1)}$ and $\lambda^{(t+1)}$. Now update

the $\boldsymbol{\eta}^{(\boldsymbol{t+1})}$: $\eta_i^{(t+1)} = 1$ if $\beta_{4,i}^{(t+1)} > 0$; otherwise 0.

4. Repeat steps 1-3 until convergence.

1.3.3 Diagnostics of convergence

The usual diagnostic framework for Bayesian LASSO (Gelman & Rubin, 1992; Brooks & Gelman, 1998; Gelman et al., 2014) includes trace plots for different chains and calculation of the *Potential Scale Reduction Factor* (PSRF). Diagnostics for RJMCMC can be developed by extending that framework to include within model and between model variations of the parameters.

We make use of the work of Castelloe & Zimmerman (2002) who define two PSRFs in the assessment. For a chosen parameter, PSRF_1 is the ratio between total variations $\widehat{V}$ and variation within chains W_c; PSRF_2 is the ratio between variations within models W_m and variations within models and chains $W_m W_c$. $\widehat{V}, W_c, W_m$ and $W_m W_c$ are defined as follows:

$$
\begin{aligned}
\widehat{V}(\theta) &= \frac{1}{CT-1}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}(\theta_{cm}^{r}-\overline{\theta_{..}}^{\cdot})^2,\\
W_c(\theta) &= \frac{1}{C(T-1)}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}(\theta_{cm}^{r}-\overline{\theta_{c.}}^{\cdot})^2,\\
W_m(\theta) &= \frac{1}{CT-M}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}(\theta_{cm}^{r}-\overline{\theta_{.m}}^{\cdot})^2,\\
W_mW_c(\theta) &= \frac{1}{C(T-M)}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}(\theta_{cm}^{r}-\overline{\theta_{cm}}^{\cdot})^2,
\end{aligned}
\tag{1.9}
$$

where $\theta_{cm}^{r}, \overline{\theta_{..}}^{\cdot}, \overline{\theta_{c.}}^{\cdot}, \overline{\theta_{.m}}^{\cdot}$ and $\overline{\theta_{cm}}^{\cdot}$ are the r^{th} appearance of θ in model m chain c, mean θ across all models and chains, mean θ within chain c across all models in that chain, mean θ within model m across all chains, mean θ within chain c and model m respectively. R_{cm} is number of θ in chain c model m. C and M are the number of chains and distinct models, respectively.

We follow the strategy given in Castelloe & Zimmerman (2002) to assess convergence and, for simplicity, illustrate this approach by considering a scalar. We choose σ^2, the variance of the error terms, for this illustration, as its interpretation remains the same across the models. Each chain is divided into batches of equal length. A sequence of PSRF_1 and PSRF_2 is calculated for each batch. A desirable result is that the two quantities move toward 1 as the iteration proceeds. In the simulation study below, we illustrate the use of diagnostic graphs for evaluating convergence; further details on this subject can be found in Brooks & Giudici (1998).

1.3.4 Interpretation

Under the assumption that intercept and population size effects are identical across source locations, we obtain a sample of $\beta_{4,i}$ as well as its 95% credible interval rather than an estimate of the probability that each source locations has a break point. Intervals that do not cover 0 imply the presence of a break point by providing evidence against the null hypothesis that the difference of the two slopes is zero. The interpretation of other parameters is straightforward. Approaches that allow variability in intercepts and population size effects yield a sample of models and their corresponding parameter estimates. For prediction, we make use of the collection of models; the estimated mean for predicted outcomes is a weighted average of the predicted outcomes of all models.

1.3.5 Computational complexity

Because of the computational burden of these methods, we consider an analysis of a subset of data. Simulation studies (Figure A.1 in Appendix A) show that computation time for the Bayesian LASSO function `blasso` increases sharply as the number of locations increases. We note that the size of the covariate matrix increases at $O(s^3)$ where s specifies the number of locations. Efron et al. (2004) showed that for the least angle regression formulation of the problem, the computational complexity is $O(m^3+m^2n)$, where m is the number of features and n is the number of the outcomes.

In our setting, the situation is even more challenging in that the number of outcomes grows quadratically with s, which renders the computational complexity to be $O(s^4)$.

1.4 Application: Analysis of call detail records

We apply our method to call detail records (CDRs) for a 3-month period to study the impact of geographical distance on communication intensity. The dataset consists of daily number of calls between distinct pairs of users. Each user is represented by a unique identifier created by the operator that made the dataset available for research. The actual phone numbers were available or recoverable from the dataset. Three covariates were made available for each person: billing zip code, sex, and age, though we only use zip code in our analysis here. We aggregated the dataset in two ways. First, we aggregated the daily call counts over the 3-month period, resulting in a single call count for each distinct pair of users. We distinguish between the caller and the receiver, so the count for each pair is directed. Second, we aggregated the data from the level of individuals to the level of counties. The resulting dataset describes communication intensity for calls among the counties. There were records for a total of 2,511,035 users; 359,759 of them resided in the largest county and 136 in the smallest one. The number of calls from one county to another ranged from 0 to 266,199, with 21,016,548 calls in total. There were 2,646 distinct zip codes nested within 427 counties. The geographical location of each county was calculated by first identifying the

latitude and longitude of each zip code and then taking the mean of the these coordinates over all zip codes that were nested within a given county. For each county we thus obtained the number of users residing in that county, and for each pair of counties we obtained the spatial distance between them and the number of calls made and received by users of those counties over the 3-month period. To reduce computational burden, we selected a subset of data that arose from 65 counties with the greatest number of users. The number of users in this subgroup of counties ranged from 7,879 to 359,759. The corresponding number of calls among pairs of counties ranged from 2 to 266,226.

1.5 Simulation

We conducted the following simulations to access the performance of our models comparing with naive approaches as well as to check the effects of different tuning parameter σ_θ^2. The values of the parameters in the data generation process are selected to be the estimates from the preliminary data analysis using $\sigma_\theta^2 = 0.03$. Actual geographical distances between counties are used. We assess the performance of the gravity model, the naive fit based on BIC and grid search, and the Bayesian LASSO model on scenarios with low (0.30), medium (0.38) and high (0.45) error variances (σ^2). This division is selected such that the medium scenario matches the estimates from the preliminary analyses. For each scenario, we simulate 2 data sets and apply our algorithm

with 4 chains. We also evaluate the effect of the tuning parameter for the Metropolis algorithm by specifying a series of different values for it: 0.015, 0.02, 0.025, 0.03, 0.04, 0.05, 0.06, 0.08, 0.1, 0.12, 0.15, 0.2, 0.25, 0.3, 0.4, 0.6. The diagnostic graphs in Appendix show that convergence is generally achieved. We assess the model fit and the effect of the tuning parameter based on the prediction error. One hundred new datasets were generated using the same covariates and parameters for each variance category. The findings are shown in Table 1.1.

As expected, estimates based both on BIC and Bayesian LASSO model perform better than those of the gravity model with respect to prediction error in low, medium and high error variances. The choice of tuning parameter had little effect; use of 0.2 in data analysis appears reasonable as this choice leads to a mean acceptance rate for the Metropolis algorithm on break-points in the range of 20% to 25% (Gelman et al., 2014), as shown in Table 1.2. The 95% credible interval coverages for break points also reach high values at tuning parameter 0.2. The crude model based on BIC and Bayesian LASSO estimates are comparable. An advantage of the latter is its ability to provide interval estimates on the break points and its smaller number of required parameters. These results imply that that we did not compromise predictive power because of the estimation of location of breakpoints, though Bayesian LASSO requires greater computational time. Computation time for 15,000 iterations takes around 9 to 10 hours, whereas the BIC approach requires only a few minutes.

Table 1.1: Prediction error of 3 models (2 trials each).

	Variance of error term σ^2					
	0.30		0.38		0.45	
Gravity model						
	0.807	0.807	0.887	0.887	0.956	0.956
Crude model based on BIC						
	0.331	0.327	0.412	0.435	0.485	0.486
Bayesian LASSO with breakpoints						
σ_θ^2						
0.015	0.321	0.329	0.403	0.411	0.479	0.486
0.020	0.322	0.329	0.405	0.413	0.479	0.487
0.025	0.319	0.329	0.404	0.411	0.479	0.486
0.030	0.321	0.326	0.407	0.411	0.479	0.485
0.040	0.318	0.323	0.409	0.409	0.481	0.486
0.050	0.317	0.323	0.411	0.411	0.480	0.487
0.060	0.318	0.321	0.411	0.411	0.481	0.487
0.080	0.318	0.321	0.411	0.410	0.482	0.487
0.100	0.318	0.320	0.411	0.409	0.482	0.488
0.120	0.320	0.320	0.410	0.412	0.481	0.485
0.150	0.319	0.320	0.411	0.414	0.486	0.487
0.200	0.321	0.319	0.413	0.414	0.486	0.490
0.250	0.321	0.320	0.415	0.416	0.489	0.488
0.300	0.320	0.321	0.413	0.417	0.490	0.489
0.400	0.321	0.321	0.417	0.420	0.496	0.489
0.600	0.325	0.322	0.414	0.419	0.494	0.489

Table 1.2: Mean acceptance rate for Metropolis algorithm on break points.

	Variance of the error terms σ^2					
σ_θ^2	0.30		0.38		0.45	
0.015	0.544	0.543	0.550	0.546	0.561	0.561
0.020	0.518	0.516	0.522	0.521	0.527	0.534
0.025	0.493	0.495	0.501	0.498	0.509	0.509
0.030	0.471	0.471	0.482	0.470	0.495	0.486
0.040	0.442	0.433	0.443	0.447	0.463	0.455
0.050	0.410	0.412	0.420	0.417	0.439	0.433
0.060	0.388	0.388	0.394	0.395	0.410	0.415
0.080	0.341	0.346	0.359	0.354	0.373	0.370
0.100	0.300	0.313	0.322	0.321	0.338	0.333
0.120	0.277	0.283	0.292	0.292	0.308	0.304
0.150	0.243	0.245	0.254	0.260	0.273	0.270
0.200	0.194	0.198	0.208	0.207	0.222	0.227
0.250	0.166	0.167	0.179	0.173	0.185	0.189
0.300	0.143	0.144	0.154	0.156	0.161	0.162
0.400	0.110	0.110	0.122	0.118	0.127	0.129
0.600	0.075	0.073	0.083	0.084	0.090	0.088

Table 1.3: 95% credible interval coverage for break points.

	Variance of the error terms σ^2					
σ_θ^2	0.30		0.38		0.45	
0.015	0.585	0.585	0.523	0.492	0.462	0.492
0.020	0.631	0.615	0.554	0.523	0.508	0.569
0.025	0.662	0.631	0.554	0.523	0.585	0.600
0.030	0.677	0.677	0.615	0.554	0.600	0.631
0.040	0.723	0.723	0.662	0.631	0.615	0.615
0.050	0.754	0.692	0.692	0.692	0.677	0.646
0.060	0.754	0.708	0.692	0.692	0.646	0.646
0.080	0.754	0.754	0.677	0.738	0.692	0.677
0.100	0.785	0.785	0.692	0.754	0.723	0.692
0.120	0.769	0.815	0.708	0.738	0.769	0.708
0.150	0.769	0.815	0.723	0.738	0.738	0.692
0.200	0.785	0.877	0.723	0.738	0.769	0.677
0.250	0.785	0.862	0.662	0.738	0.754	0.708
0.300	0.769	0.862	0.708	0.738	0.738	0.692
0.400	0.769	0.862	0.677	0.708	0.708	0.708
0.600	0.785	0.831	0.708	0.738	0.692	0.708

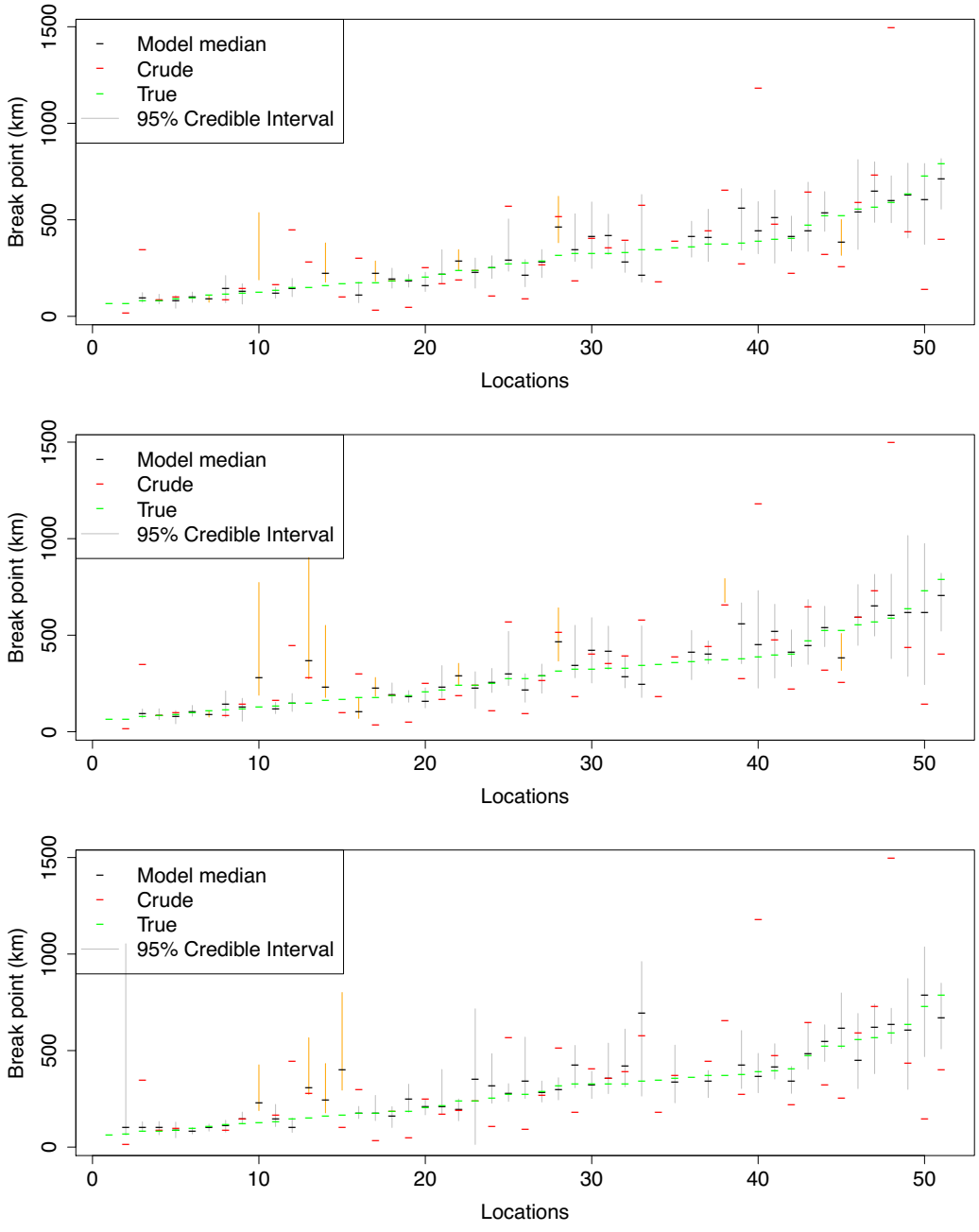


Figure 1.1: Estimated 95% credible intervals of break point θ_i (when true break points exist) under low (top), medium (middle) and high (bottom) error variance σ^2 with tuning parameter $\sigma_\theta^2 = 0.2$; orange color of the 95% credible interval indicates that the true value is not covered; no 95% credible interval showing indicates none available, i.e. estimates are from model without break points.

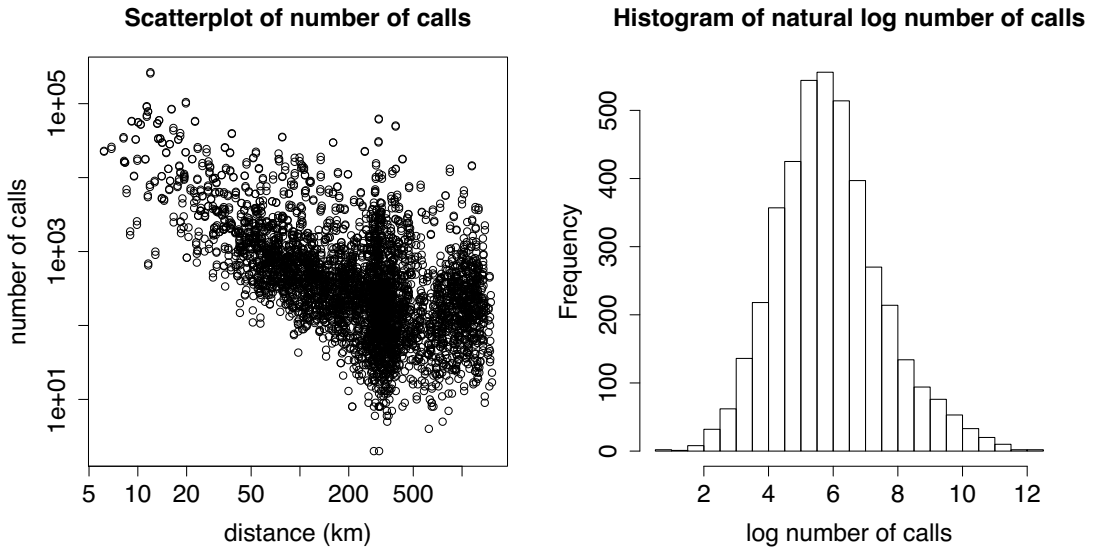


Figure 1.2: Left: scatter plot of natural log number of calls v.s. distances; Right: histogram of natural log number of calls.

1.6 Analysis of call records data

First, we observe that Figure 1.2 is consistent with our assumptions of continuous calling intensity and normality of natural log of the number of calls.

Second, We use the preliminary binary assignments of break points group based on BIC in a simple linear regression to assess whether there is variability in intercepts and population size effects. Both models with only main effects (indicator variable of group assignments, log population sizes, log distance-before/after break point) and those with main effects and interaction terms show evidence (p-value$<$0.05) of variability. Hence we apply the method shown in the simulation study for the analysis of the cell phone data. The difference in intercepts and population size effects is true

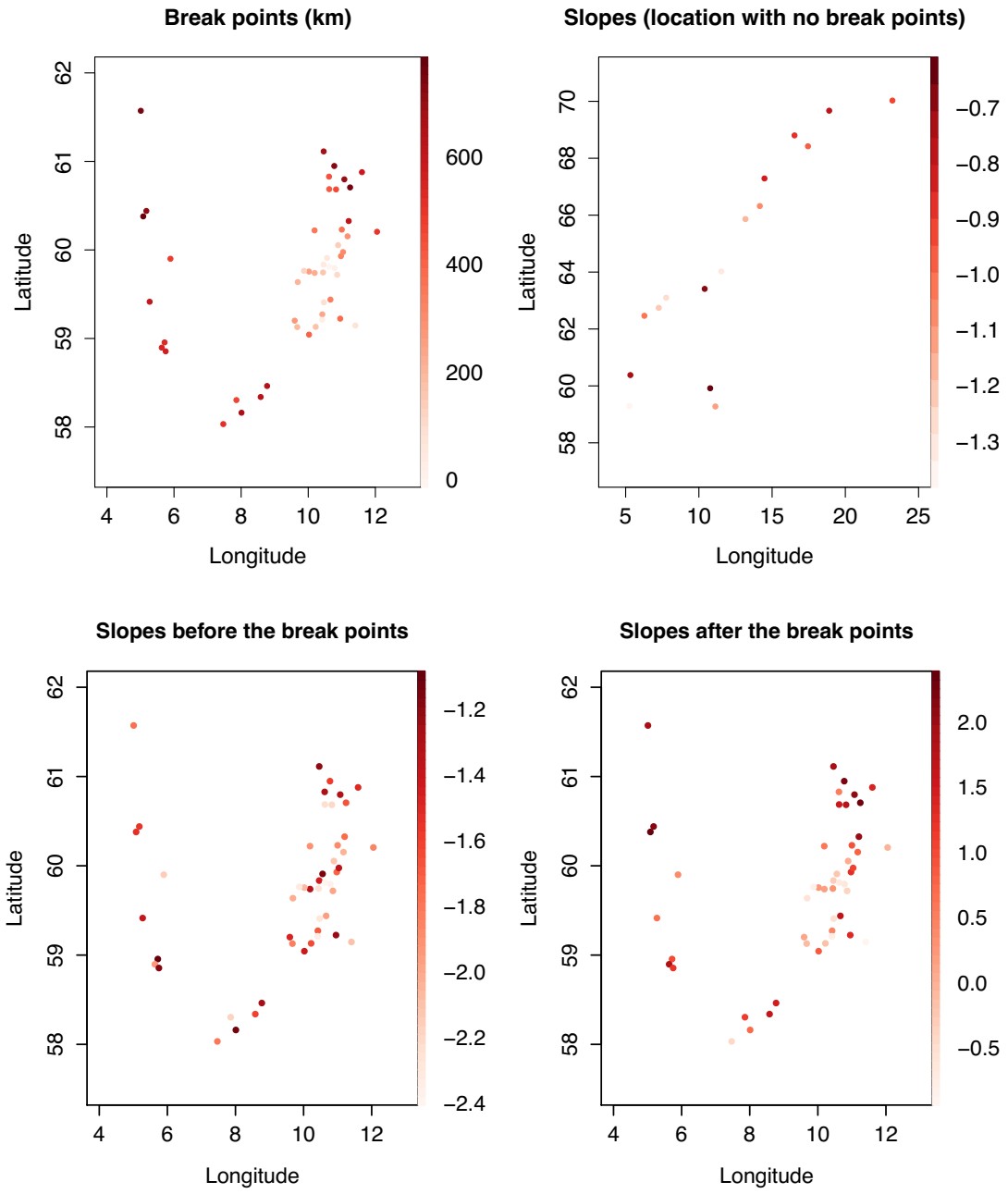


Figure 1.3: Top left: slope estimates for locations without break points; Top right: log distance of the estimated break points for locations with break points; Bottom: Slopes estimates before and after break points for different locations.

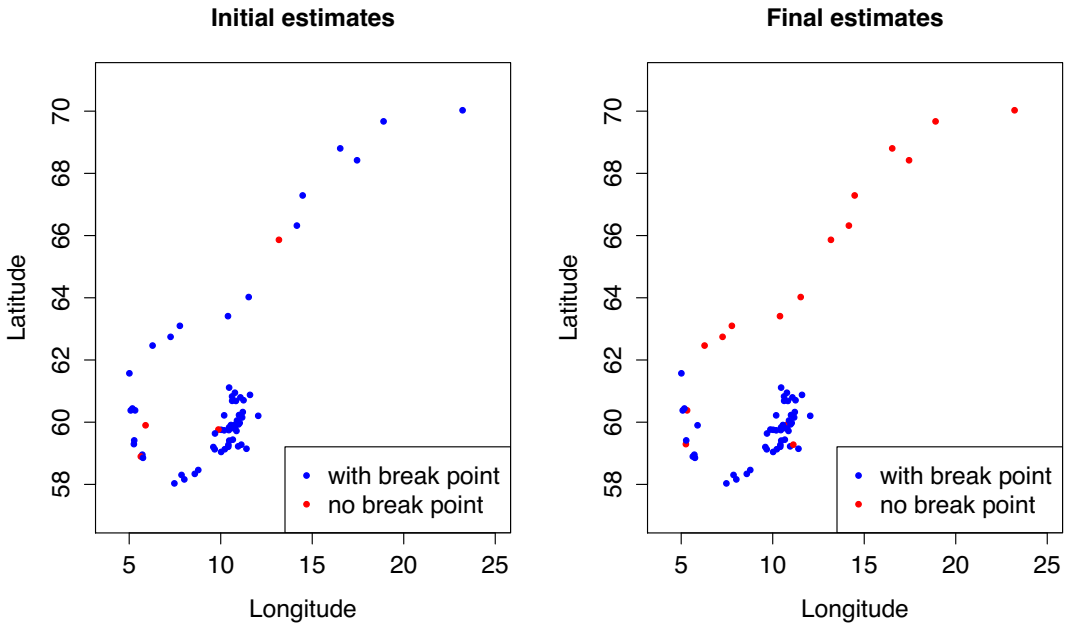


Figure 1.4: Initial and final estimates of the existence of the break points.

both for the general population from all 427 counties, and for the user subpopulation we described above.

In the analysis of call records (CDRs) data (Figure 1.3 and 1.4), we observe that the slopes for source locations in the northeast appear to be less steep. Locations near the capital city, where the population is dense, are more likely to have breakpoints in the relationship of communication and distance . No such patterns were observed for slopes of other locations, both before and after the break points. Model estimates revealed that locations with no break point tend to be in the north while those with breakpoints are concentrated in the south. For diagnosis on convergence, Figure 1.5 shows a trend of PSRF_2 approaching 1 very quickly and a PSRF_1 fluctuating below

1.5, which is acceptable.

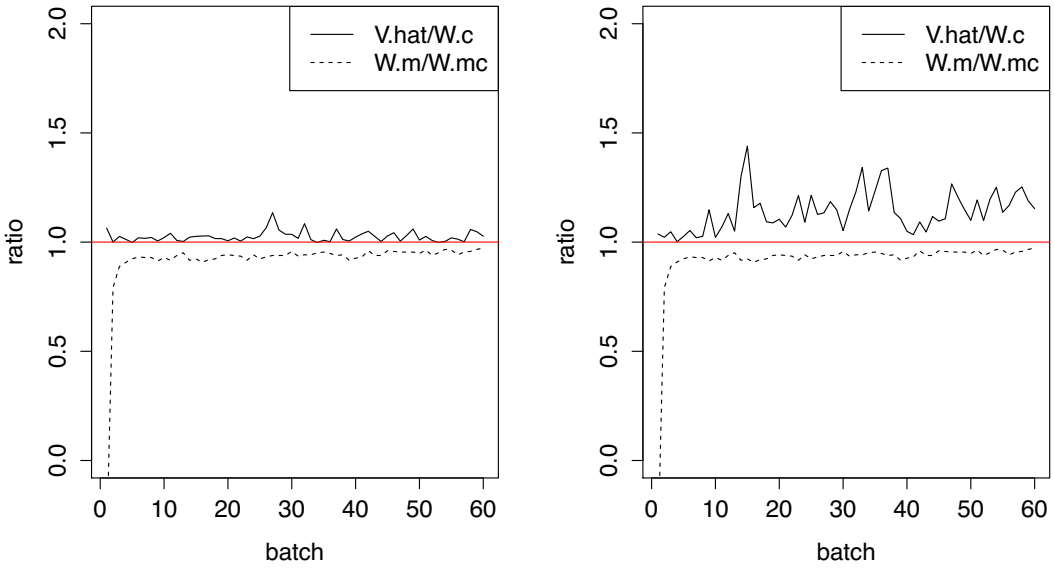


Figure 1.5: Left: diagnostic graph based on intercept estimates; Right: diagnostic graph based on σ^2; solid line: PSRF_1, dashed line: PSRF_2.

1.7 Discussion

To analyze the decline in communication intensity with geographical distance, we extended the gravity model by allowing for break-points in this relationship. We addressed the issue of the existence of break-points for each source location and quantify associated uncertainty using a Bayesian model. We also provided estimates of the slopes before and after each breakpoint. We investigated the geographical pattern of the existence of break-points and noted differences in these patterns between rural

and urban areas.

As an application of our method, we made use of an anonymized dataset of call detail records, using the number of mobile phone calls in our analysis as the measure of communication intensity between a pair of counties. The range of the outcomes is a count $\mathbb{N}$ before log transformation, and the regression model we specify treats the transformed outcomes as continuous, which is most appropriate when the number of calls between two locations is large (Figure 1.2). In settings were there may be 0 or very low counts, one could consider alternative models (e.g. negative binomial) or the addition of an arbitrary small positive number to 0, although the latter approach can add bias (Flowerdew & Aitkin, 1982; Burger et al., 2009). In this setting, a negative binomial model might be a better fit, though the interpretation of the parameters is less straightforward. Using Bayesian methods in a setting where the data are assumed to be negative binomial distributed requires non-standard approaches even without inclusion of break-points into models. Zhou et al. (2012); Pillow & Scott (2012), and Polson et al. (2013) provide some useful tools for sequentially updating the parameters using Gibbs sampler by augmenting the posterior distribution with auxiliary parameters. When the number of counts is large, the negative binomial approach may not be computationally feasible; fitting negative binomial outcomes in Bayesian LASSO needs further investigation. One possible direction is to extend the methods based on the conditional normal distribution in Polson et al. (2013) by transforming the variance matrix so that normal-distribution based LASSO method can be

employed.

Another extension of our methods would allow for aggregation of results across different subsamples; currently the number of locations we can analyze is limited by computational concerns. Developing a method to obtain consistent results from different overlapping sets of nodes-perhaps in a meta-analysis framework– would alleviate the computational concerns, but is challenging. Some potentially useful approaches are provided by Politis & Romano (1994); Politis et al. (2001); Geyer (2006) and Fitzenberger (1998). In particular, the stability selection in Meinshausen & Bühlmann (2010) may be used to assess the properties of the meta-analytic results. An example of the use of LASSO in analyses that combine across subsamples arose from analyses intended to discover adverse drug reactions provided by Ahmed et al. (2016). Another potentially useful approach is the use of path of partial posteriors in Strathmann et al. (2015). In this approach, the resampling procedure resembles the bootstrap, but with smaller resampling sizes. Because standard bootstrapping the LASSO estimator of the regression parameter for variance inference is known to yield inconsistent estimates (Knight & Fu, 2000; Chatterjee & Lahiri, 2010), modified bootstrapping must be used (Chatterjee & Lahiri, 2011). Nonetheless, Bayesian LASSO procedures provide straightforward and valid estimates for standard errors.

The findings from our analysis of mobile phone communication illustrate how such information might be used, should such communication networks prove to be accurate proxies for contact networks along which infectious diseases or other communicable

processes spread. If so, such analyses might help guide designs of cluster randomized trials. Randomized trials ideally enroll participants in a way that minimizes the extent to which treatment assignment of one subject affects outcome of another. For interventions in which such interference occurs at the individual but not cluster level (e.g. through contacts among randomized subjects), cluster randomization can be useful (Campbell et al., 2007). Clusters may be comprised of participants in the same geographical location, institution (e.g. school) or administrative unit (village). Cell phone data could potentially aid in the identification of appropriate clusters by providing information about the probability of interference. When mixing across clusters cannot be eliminated, identification of treatment effects requires models of the mixing process (Carnegie et al., 2016). Staples et al. (2015) and Wang et al. (2014) investigated the impact of interference across randomized units on power of a clinical trial to detect effects of an intervention in preventing spread of infectious disease. As geographical distance is likely to affect contact networks, knowing the relationship between communication and distance may be useful not only for identification of clusters, but also to aid in development of appropriate mixing models.

CONFLICT OF INTEREST

We declare no conflict of interest.

2

Investigating associations among network structures

2.1 Problems Statement and Motivation

In analyses of network community detection, three questions may arise. The first question is whether there is a community structure in the network; the null hypoth-

esis is that there is no community structure. This question has been explored by focusing on automatically determining the number of clusters in community detection (Zhao et al., 2011; Bickel & Sarkar, 2016).The second question arises in settings where there are two ways to define community membership for the same set of nodes. The question is whether the existence of a pair of nodes in the same community in one assignment implies that they are in the same community according to the other assignment. A corresponding null hypothesis is that it does not imply they are in the same community. The third question arises after we reject the second hypothesis. It asks whether the two assignments are the same, with the null being that they are the same. The third question has been explored in Tang et al. (2014) using random dot product graphs based on spectral embedding of the adjacency matrix. Some theoretical properties of the random dot product graphs and the asymptotic property of the corresponding eigenvectors are described in Tang & Priebe (2016).

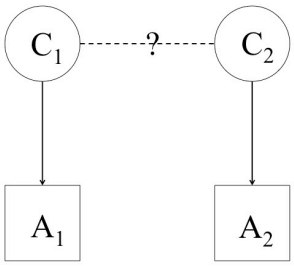


Figure 2.1: Latent community structure C_1 and C_2 generate adjacency matrice A_1 and A_2, question 2 asks whether C_1 and C_2 are informative of each other; question 3 asks that if C_1 and C_2 are informative of each other, whether they are the same.

In this paper, we focus on the second question, i.e. testing whether two community detection results imply each other. We specify the null as above. In our setting, community detection may result from two different algorithms applied to the same network, or from the same community detection methods applied on different networks that share the same set of nodes. One example of the latter situation is demonstrated by community detection results obtained from two networks, one based on cell phone communication and second, on viral genetic linkage. To use cell phone communities as proxies for viral genetic clusters, it would be useful to show that inference on community structure from one network is informative regarding the structure of the other. The hypothesis testing framework we propose is relevant for this purpose.

Our proposed approach uses existing community detection methods to assign community memberships to nodes. In general, a network or a graph can be represented by its node set and its edge set. The node set consists of all the nodes in the network, and the edge set specifies those pairs of nodes that are connected with an edge. A community can be defined as a set of densely connected nodes (Porter et al., 2009; Fortunato, 2010). Community detection methods aim at discovering communities in a network. There is large literature on various community detection methods(Clauset et al., 2004; Newman, 2006b; Richardson et al., 2009), and their performances and comparisons (Fortunato & Barthélemy, 2007; Bader & McCloskey, 2009; Lancichinetti & Fortunato, 2009; Fortunato, 2010; Good et al., 2010; Leskovec et al., 2010). We emphasize that our testing framework applies to most community detection settings re-

gardless of which specific community detection method is employed as long as the detection method is (1) appropriate for the graph or network considering its properties (e.g. weighted, directed, etc.) and (2) satisfies the assumption introduced in the paper, which indicates that the possibility of obtaining one specific detection result only depends on the topological properties of the graph and the detection method. We also note the differences and connections between our methods and those on multislice networks in Mucha et al. (2010), where community detection methods are extended to time-dependent settings with network slices of different types of links and scales. Our analysis framework complements the estimation model and provides insights from a hypothesis testing perspective.

The sections of this paper are organized as follows: section 2.2 introduces the setup and notation; section 2.3 describes the distributions of test statistics under the null hypothesis; section 2.4 describes the simulation study and provides results; section 2.5 provides results of an analysis on data with two types of connections among a group of workers at two different time points; section 2.6 is the conclusion and discussion.

2.2 Notation

Following the notation in Bondy et al. (1976) and Abbe (2017), let $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ be two undirected graphs generated from stochastic block model ($SBM(n, \mathbf{p}, W)$) with node or vertex sets V_1, V_2 and edge sets E_1, E_2, where $V_1 =$

$V_2 = V$, $|V| = n$, $\mathbf{p}$ is a probability vector on integers from 1 to k, W is the $k \times k$ symmetric matrix of the connectivity probabilities. Let A_1, A_2 be the corresponding adjacency matrices. Denote $\mathbf{C_1}, \mathbf{C_2}$ as the vectors of community assignments in G_1 and G_2. We note that $(\mathbf{C_1}, G_1)$ pair and $(\mathbf{C_2}, G_2)$ pair are drawn under the model we specified. Both $\mathbf{C_1}$ and $\mathbf{C_2}$ are vectors of random variables.

2.3 Distribution of test statistics under the null

The null hypothesis is described in terms of network community memberships of pairs of nodes. This approach can be used regardless of the number of communities for each detection method or network. The null hypothesis is that pairs of nodes being in the same communities in network 1 is not associated with their being in the same communities in network 2. Formally,

$$P(\delta(C_{2i}, C_{2j}) = k | \delta(C_{1s}, C_{1t}) = l) = P(\delta(C_{2i}, C_{2j}) = k), \forall i, j, s, t \in \{1, ..., n\} \quad (2.1)$$

where C_{2i} and C_{2j} are the elements in $\mathbf{C_2}$ indicating community assignments for node i and j in network 2; C_{1s} and C_{1t} are the elements in $\mathbf{C_1}$ indicating community assignments for node i and j in network 1; $\delta(\cdot, \cdot)$ is Kronecker delta; k and l take values in $\{0, 1\}$.

For simplicity, we use $\mathbf{\Delta_1}$ to represent the $\frac{n(n-1)}{2}$ vector of all $\delta(C_{1s}, C_{1t})$ for $s < t$ and $s, t \in \{1, ..., n\}$. $\mathbf{\Delta_2}$ to represents comparable quantities in network 2. By clas-

sifying pairs of nodes to be in the same community (or not) in network 1 and 2, we construct the 2 by 2 contingency table shown in Table 2.1.

Table 2.1: Contingency table with counts of pairs, where O_{ij} denotes number of counts in corresponding row and column.

		$\mathbf{\Delta_2}$		
		same	different	
$\mathbf{\Delta_1}$	same	O_{11}	O_{12}	$O_{1\cdot}$
	different	O_{21}	O_{22}	$O_{2\cdot}$
		$O_{\cdot 1}$	$O_{\cdot 2}$	O

Assume the true community membership vectors $\mathbf{C_1}$ and $\mathbf{C_2}$ are known. Table 2.1 reminds us of two statistics: Pearson's Chi-squared test statistic and likelihood ratio test statistic. The former is written as

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{2.2}$$

The latter is

$$Dev = \sum_{i,j} O_{ij} \log\left(\frac{O_{ij}}{E_{ij}}\right) \tag{2.3}$$

where E_{ij} is the expected counts in cell row i, column j, π_{ij} is the probability of the corresponding cell in Table 2.1. We note the connection of the latter with mutual in-

formation as defined in information theory, which can be written:

$$
\begin{aligned}
MutI(\Delta_1, \Delta_2) &= \sum_{i,j} \pi_{ij} \log\left(\frac{\pi_{ij}}{\pi_{i\cdot}\pi_{\cdot j}}\right) = \sum_{i,j} \frac{O_{ij}}{N} \log\left(\frac{O_{ij}}{E_{ij}}\right) \\
&= \frac{1}{N} \sum_{i,j} O_{ij} \log\left(\frac{O_{ij}}{E_{ij}}\right)
\end{aligned}
\tag{2.4}
$$

Equation 2.4 is the deviance of a Poisson (or multinomial) sampling scheme (Agresti, 2013) divided by $2N$. The deviance corresponds to a likelihood ratio test statistic comparing the saturated model and the nested null model. Under the null, the test statistics has an asymptotic χ^2 distribution with $df = N - p$ for large $\pi_{ij}N = \mu_{ij}$ when observations are independent.

Observations are not independent in our setting. Elements in $\mathbf{\Delta_1}$ are not randomly assigned and are subject to constraints. It means that there must be a community membership assignment $\mathbf{C_1}$ to which $\mathbf{\Delta_1}$ corresponds. Consider a triad of 3 nodes, (r, s, t), where $r, s, t \in V$, from which we create 3 pairs: $(r, s), (r, t), (s, t)$. Without loss of generality, let us assume in detection 1, i.e. the row of Table 2.1, nodes r and s are classified to the same community while t is in a different community. Hence, pair (r, s) will be in row "same" and pairs $(r, t), (s, t)$ are in row "different". However, knowing the status of 2 out of 3 pairs in this triad specifies the category of the third as well. Because of this complex deterministic mechanism, the observations in two rows are not independent. From another viewpoint, the effective number of observations in the samples is smaller than the observed number. This will tend to produce a

Pearson's Chi-square statistic with a liberal p-value if treated as independent, which is confirmed in the simulation.

We claim that the permutation is still valid under the null. In the following two theorems, we state and prove that by permuting the community labels, we are able to obtain a test which preserves type I error. The null hypothesis in both theorems is defined as above, i.e. pairs of nodes being in the same or in different communities in network (or detection method) 1 is not associated with their being in the same or different communities in network 2.

Theorem 1. *Assume* $\mathbf{C_1}$ *and* $\mathbf{C_2}$ *are observed, and marginally each element is independently and identically distributed from multinomial distribution* $Multinom(\mathbf{p_1})$ *and* $Multinom(\mathbf{p_2})$ *respectively, where* $\mathbf{p_1}$ *and* $\mathbf{p_2}$ *are the vectors of probability which sum to 1.* $\mathbf{\Delta_1}$ *is the* $\frac{n(n-1)}{2}$ *vector of all* $\delta(C_{1s}, C_{1t})$ *for* $s < t$ *and* $s, t \in \{1, ..., n\}$. $\mathbf{\Delta_2}$ *corresponds to that of* $\mathbf{C_2}$.

The null hypothesis H_0: $P(\mathbf{\Delta_1}, \mathbf{\Delta_2}) = P(\mathbf{\Delta_1})P(\mathbf{\Delta_2})$.

Let $\mathbf{C_{1(\pi)}}$ *be a permutation of* $\mathbf{C_1}$. $\mathbf{\Delta_{1(\pi)}}$ *corresponds to the vector of pairs for* $\mathbf{C_{1(\pi)}}$. Ω_c *denotes the set of all permutations of* $\mathbf{C_{1(\pi)}}$. Ω_Δ *denotes the set of all corresponding* $\mathbf{\Delta_{1(\pi)}}$. *Construct test statistic* $T(\mathbf{\Delta_1}, \mathbf{\Delta_2})$. *Keep* $\mathbf{\Delta_2}$ *fixed, calculate* T *for all elements in* Ω_Δ *and rank them.*

$$t^{(1)} \leqslant t^{(2)} \leqslant ... \leqslant t^{(|\Omega_\Delta|)}$$

For fixed level $\alpha > 0$*, let* $t^* = \min\{t^{(k)} : P(T \geqslant t^{(k)}) \leqslant \alpha, k \in \{1, ..., |\Omega_\Delta|\}$*. Reject* H_0 *if* $T(\mathbf{\Delta_1}, \mathbf{\Delta_2}) \geqslant t^*$*. Then*

$$P(\textit{reject } H_0 | H_0) \leqslant \alpha$$

PROOF: The proof is straightforward once it is proved that under H_0, the joint distribution $f(\mathbf{\Delta_{1(\pi)}}, \mathbf{\Delta_2})$ remains the same under permutations of $\mathbf{C_1}$.

First, we note the correspondence between $\mathbf{C_1}$ and $\mathbf{\Delta_1}$, which implies $\forall \mathbf{\Delta_{1(\pi)}} \in \Omega_\Delta, \exists \mathbf{C_{1(\pi)}} \in \Omega_c$ such that $\mathbf{C_{1(\pi)}}$ generates the $\mathbf{\Delta_{1(\pi)}}$. For $\mathbf{C_1}$ with distinct number of members in each community, the relationship between $\mathbf{C_1}$ and $\mathbf{\Delta_1}$ is one-to-one. When there are communities which have the same number of nodes, the correspondence is c-to-one, where c is a constant which only depends on $\mathbf{C_1}$. Specifically, c is the product of the factorial of the number of communities which have the same number of nodes.

Given fixed $\mathbf{C_1}$, we have $P(\mathbf{\Delta_{1(\pi)}} = \mathbf{x}) = cP(\mathbf{C_{1(\pi)}} = \mathbf{z})$. Under H_0, we have $P(\mathbf{\Delta_1}, \mathbf{\Delta_2}) = P(\mathbf{\Delta_1})P(\mathbf{\Delta_2}) = cP(\mathbf{C_1})P(\mathbf{\Delta_2})$.

For any permutation of $\mathbf{C_1}$, we have

$$\begin{aligned} P(\mathbf{\Delta_{1(\pi)}}, \mathbf{\Delta_2}) &= P(\mathbf{\Delta_{1(\pi)}})P(\mathbf{\Delta_2}) = cP(\mathbf{C_{1(\pi)}})P(\mathbf{\Delta_2}) = cP(\mathbf{C_1})P(\mathbf{\Delta_2}) \\ &= P(\mathbf{\Delta_1})P(\mathbf{\Delta_2}) = P(\mathbf{\Delta_1}, \mathbf{\Delta_2}) \end{aligned} \tag{2.5}$$

Hence, under H_0, the joint distribution $f(\mathbf{\Delta_{1(\pi)}}, \mathbf{\Delta_2})$ remains the same under the permutation of $\mathbf{C_1}$. It follows that for a chosen test statistic $T(\mathbf{\Delta_1}, \mathbf{\Delta_2})$, which is a

function of $(\mathbf{\Delta_1}, \mathbf{\Delta_2})$, the distribution is the same under the permutation. The permutation distribution is a discrete uniform distribution with equal probability mass on each $t^{(k)}$ for $k \in \{1, ..., |\Omega_\Delta|\}$.

$$P(\text{reject } H_0|H_0) = P(T(\mathbf{\Delta_1}, \mathbf{\Delta_2}) \geqslant t^*|H_0) \leqslant \alpha \tag{2.6}$$

Therefore the permutation test is valid and exact when $\mathbf{C_1}$ and $\mathbf{C_2}$ are observed. □

We note that the assumption in Theorem 1 is consistent with the network generating process for the Stochastic Block Model shown in Abbe (2017).

When $\mathbf{C_1}$ and $\mathbf{C_2}$ need to be estimated, the following assumption is required. We assume that the probability of getting a detection result depends only on the topological properties of the graph and the detection method. It is possible for one detection method used on the same graph to produce different detection results when the algorithm is run multiple times.

We note that the above assumption is very weak and should hold for any reasonable community detection methods.

Theorem 2. *Following the scenario in Theorem 1, now assume instead of being known,* $\mathbf{C_1}$ *and* $\mathbf{C_2}$ *need to be estimated. Denote them as* $\widehat{\mathbf{C_1}}$ *and* $\widehat{\mathbf{C_2}}$*. Correspondingly,* $\widehat{\mathbf{\Delta_1}}$ *and* $\widehat{\mathbf{\Delta_2}}$ *are the vectors of indicator function which indicates whether pairs of nodes are in the same community.* $\widehat{\mathbf{C_1}}$ *and* $\widehat{\mathbf{C_2}}$ *are estimated from adjacency matrices* A_1 *and* A_2*, i.e. if detection method is denoted as function* $\psi(\cdot)$*, it follows that*

$\widehat{\mathbf{C_1}} = \psi_1(A_1)$ *and* $\widehat{\mathbf{C_2}} = \psi_2(A_2)$. *We assume that the detection method satisfies the above mentioned assumption. Also* A_1 *and* A_2 *are both generated independently from true membership* $\mathbf{C_1}$ *and* $\mathbf{C_2}$ *following a Stochastic Block Model, we have* $A_1 = \phi_1(C_1)$ *and* $A_2 = \phi_2(C_2)$.

The null hypothesis H_0: $P(\mathbf{\Delta_1}, \mathbf{\Delta_2}) = P(\mathbf{\Delta_1})P(\mathbf{\Delta_2})$.

Let $\widehat{\mathbf{C_1}}_{(\pi)}$ *be a permutation of* $\widehat{\mathbf{C_1}}$. $\widehat{\mathbf{\Delta_1}}_{(\pi)}$ *corresponds to the vector of pairs for* $\widehat{\mathbf{C_1}}_{(\pi)}$. $\Omega_{\widehat{c}}$ *denotes the set of all permutations of* $\widehat{\mathbf{C_1}}_{(\pi)}$. $\Omega_{\widehat{\Delta}}$ *denotes the set of all corresponding* $\widehat{\mathbf{\Delta_1}}_{(\pi)}$. *Construct test statistic* $T(\widehat{\mathbf{\Delta_1}}, \widehat{\mathbf{\Delta_2}})$. *Keep* $\widehat{\mathbf{\Delta_2}}$ *fixed, calculate* T *for all elements in* $\Omega_{\widehat{\Delta}}$ *and rank them.*

$$t^{(1)} \leqslant t^{(2)} \leqslant ... \leqslant t^{(|\Omega_{\widehat{\Delta}}|)}$$

For fixed level $\alpha > 0$, *let* $t^* = \min\{t^{(k)} : P(T \geqslant t^{(k)}) \leqslant \alpha, k \in \{1, ..., |\Omega_\Delta|\}$. *Reject* H_0 *if* $T(\widehat{\mathbf{\Delta_1}}, \widehat{\mathbf{\Delta_2}}) \geqslant t^*$. *Then*

$$P(\textit{reject } H_0 | H_0) \leqslant \alpha$$

PROOF: Since A_1 is generated from true membership $\mathbf{C_1}$ from a Stochastic Block Model, for a permutation matrix Q, it follows that

$$P(A_1|\mathbf{C_1}) = P(QA_1|Q\mathbf{C_1}) = P(A_{1(\pi)}|\mathbf{C_1}_{(\pi)}) \tag{2.7}$$

Marginalized over membership, we have

$$P(A_1) = \sum_{\mathbf{C_1}} P(A_1|\mathbf{C_1})P(\mathbf{C_1}) = \sum_{\mathbf{C_{1(\pi)}}} P(A_{1(\pi)}|\mathbf{C_{1(\pi)}})P(\mathbf{C_{1(\pi)}}) = P(A_{1(\pi)}) \tag{2.8}$$

as $P(\mathbf{C_1}) = P(\mathbf{C_{1(\pi)}})$ and $\mathbf{C_1}$, $\mathbf{C_{1(\pi)}}$ are from the same sample space.

We note that for the detection method satisfying the previously mentioned assumption, the following must hold

$$P(\widehat{\mathbf{\Delta_1}}|A_1) = P(\widehat{\mathbf{\Delta_1}}_{(\pi)}|A_{1(\pi)}) \tag{2.9}$$

as the network structure remains the same while the node labels are permuted. Marginalized over A_1 and $A_{1(\pi)}$ respectively, we have

$$P(\widehat{\mathbf{\Delta_1}}) = P(\widehat{\mathbf{\Delta_1}}_{(\pi)}) \tag{2.10}$$

Under the null, we have H_0: $P(\mathbf{\Delta_1}, \mathbf{\Delta_2}) = P(\mathbf{\Delta_1})P(\mathbf{\Delta_2})$. It follows that as functions of $\mathbf{\Delta_1}$ and $\mathbf{\Delta_2}$ respectively, $\widehat{\mathbf{\Delta_1}}$ and $\widehat{\mathbf{\Delta_2}}$ are also independent, i.e., we have under the null:

$$P(\widehat{\mathbf{\Delta_1}}, \widehat{\mathbf{\Delta_2}}) = P(\widehat{\mathbf{\Delta_1}})P(\widehat{\mathbf{\Delta_2}}) \tag{2.11}$$

Together with Equation 2.10, we have

$$P(\widehat{\mathbf{\Delta_1}}, \widehat{\mathbf{\Delta_2}}) = P(\widehat{\mathbf{\Delta_1}})P(\widehat{\mathbf{\Delta_2}}) = P(\widehat{\mathbf{\Delta_1}_{(\pi)}})P(\widehat{\mathbf{\Delta_2}}) = P(\widehat{\mathbf{\Delta_1}_{(\pi)}}, \widehat{\mathbf{\Delta_2}}) \tag{2.12}$$

Thus, under the null the joint probability of $(\widehat{\mathbf{\Delta_1}}, \widehat{\mathbf{\Delta_2}})$ is also invariant under permutations. Following the same logic in Theorem 1, the permutation test is also valid and exact. □

The permutation test yields an exact p-value, by which we mean the actual probability of a Type I error is less than or equal to the α level of the test. As part of our exploration in the simulation section, we also exam the performance of another widely used resampling methods—bootstrap—to compare with that of the permutation test. In general, the bootstrap test is not exact, but should be asymptotically valid, given what is demonstrated by Equation 2.8.

The null hypothesis that there is no association between two sets of community assignments is different from the assertion that there is no community structure in the graph. As has been pointed out in both network science (Fortunato, 2010) and statistical literature (Zhao et al., 2011; Bickel & Sarkar, 2016), a precise definition of graphs with no community structure is yet to be established. When the graph is an Erdös-Rényi (ER) graph, no community structure are expected. In Lancichinetti & Fortunato (2009), the ER graph is selected as one type of random graph to evaluate the performance of different community detection methods. Theoretically, the ER

graph has only one community and a good detection method should ideally discover only one. However, in practice, certain pseudocommunities may still be identified due to specific realizations of random graphs(Lancichinetti & Fortunato, 2009); this is also illustrated in our simulation. However, detection results from two ER graphs are not associated. For graphs with community structure, such as two graphs generated from SBM with unassociated community membership assignments, detection results are not expected to be associated with each other. In Section 2.4, we simulate the null hypothesis by generating ER graphs and graphs from two unrelated community structures using SBM, and evaluate the performance of our tests in both situations.

Test Evaluation and Handling of the Resampling Risk

The bootstrap and permutation test introduced above are both Monte Carlo tests, which have the issue of resampling risk. Following the notations in Gandy (2009), the resampling risk $RR_p(\hat{p})$ is the probability that $\hat{p}$ and the true p are on different sides of the given threshold α.

$$RR_p(\hat{p}) \equiv \begin{cases} P_p(\hat{p} > \alpha) & \text{if } p \leqslant \alpha \\ P_p(\hat{p} \leqslant \alpha) & \text{if } p > \alpha \end{cases} \tag{2.13}$$

To address this issue and provide a better evaluation, we adopt the methods of Gandy (2009) and Gandy et al. (2013) and their `R` package `simctest`, which puts an upper bound on the resampling risk by a sequential implementation. Specifically, they con-

struct lower and upper bounds for the sum of the binary outcomes (reject the null or not) in the next step based on previous steps, and stop sampling once the lower or upper bound is hit. They prove that when the actual p-value is not 0.05, it takes finite steps for the algorithm to stop.

We iteratively use the implementation in two loops to evaluate the level of the tests. Adapted from Gandy (2009), we put an upper bound on the number of steps in the inner loop at M and use the interim estimated p-value if the maximum number of steps M is reached. According to their research, the results are (almost) the same with different M values. We briefly assess the effects of different M values in the simulation. We note that setting M will render the theoretical results of the upper bound not applicable, however, it is expedient to avoid infinite steps in the inner loop. Also less calculation is needed compared with fixed steps in the inner loop.

2.4 Simulation

We evaluate the test with a variety of number of nodes (50 to 500 by 50) in two aspects: the null case, in which there is no association, and the alternative case when there is some association. All graphs generated in the simulation section are undirected. For community detections, we employ a modularity-based method using greedy techniques (Blondel et al., 2008; Fortunato, 2010) which is implemented in `R` package `igraph`. This technique is usually referred as the Louvain method. We choose the

Louvain method due to the fact that modularity-based detection algorithms are well studied and widely used in the literature, and the Louvain method has demonstrated excellent accuracy and faster computation time. This choice makes our simulation more reasonable and realistic, even though the validity of the hypothesis testing does not depend greatly on the detection method.

2.4.1 Null case

Two Erdös-Rényi graph

We simulate two ER graphs with edge probability $p = 0.2$ and apply the community detection method to obtain two detected assignments $\widehat{\mathbf{C_1}}$ and $\widehat{\mathbf{C_2}}$.

Two networks from SBM with random community membership assignments

We evaluate and compare the performance of tests in settings with different $\mathbf{p}$ and same W in graph generation using SBM: (i) $\mathbf{p} = (0.7, 0.3)$; (ii) $\mathbf{p} = (0.6, 0.4)$; (iii) $\mathbf{p} = (0.5, 0.5)$; (iv) $\mathbf{p} = (0.3, 0.3, 0.4)$. W has diagonal elements 0.15 and off diagonal elements 0.05 for all settings.

1. To evaluate the test under accurate recovery, we generate two random labels $\mathbf{C_1}$ and $\mathbf{C_2}$ by permuting community memberships sampled from $multinomial(\mathbf{p})$ and check the performance of the tests assuming the detection method accurately recovers

the community structure.

2. We simulate two graphs using SBM with $\mathbf{C_1}$ and $\mathbf{C_2}$ generated by permuting community memberships sampled from $multinomial(\mathbf{p})$. Then we apply community detection method to obtain two detected assignments $\widehat{\mathbf{C_1}}$ and $\widehat{\mathbf{C_2}}$.

2.4.2 Alternative case

Expecting the tests to have high power when community structures in two graphs are the same, we simulate settings with weaker association. Specifically, we assess the performance of the tests in settings with (i) 20% and 30% structural overlap (ii) a smaller difference in within-community and between-community edge probability:

1. We simulate two graphs using SBM (same $\mathbf{p}$ and W as in the null case (i)and (iii)) with graph 2 matching ρ percentage (e.g. 20%, 30%) of community memberships in graph 1. The remaining community memberships are random permutations of those in graph 1. Then we obtain $\widehat{\mathbf{C_1}}$ and $\widehat{\mathbf{C_2}}$ assuming accurate recovery, or using community detection.

2. We simulate two graphs using SBM with the same community membership labels (within-community edge probability of 0.1 and between-community edge probability of 0.05). Then we obtain $\widehat{\mathbf{C_1}}$ and $\widehat{\mathbf{C_2}}$ using community detection.

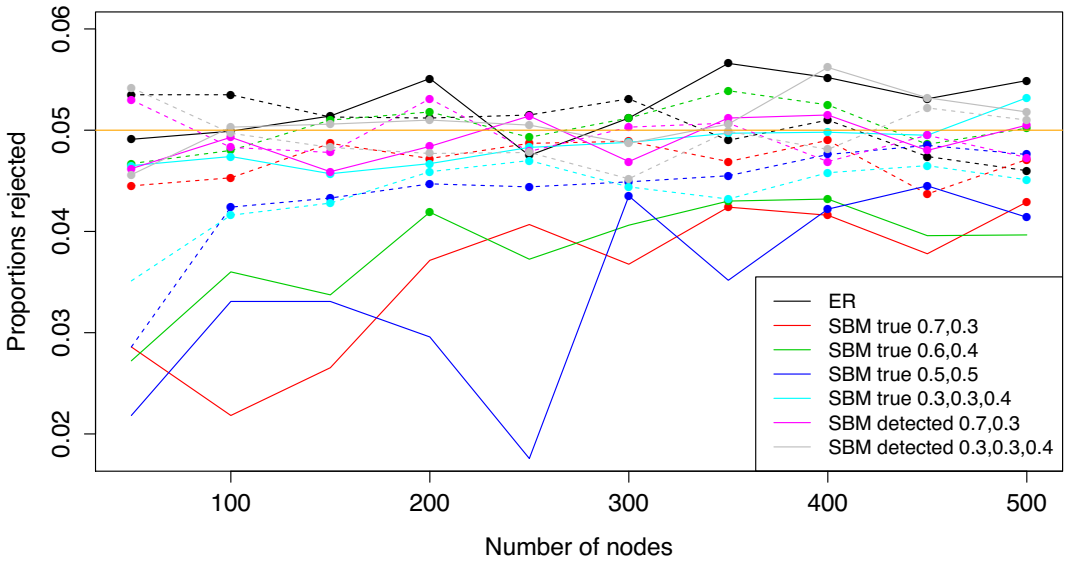


Figure 2.2: Size of the permutation and bootstrap tests under the null. The solid line: the permutation test; the dashed line: the bootstrap test. Dots indicate the estimated p-value is not significantly different from 0.05.

2.4.3 Simulation results

We found that in both permutation and bootstrap tests, test statistics based on mutual information and Pearson's Chi-square produce the same p-value. In the figures below, only the results with Pearson's Chi-square are shown, as they require slightly less time to compute.

In Figure 2.2, we observe that both permutation and bootstrap tests have the right level in different null cases: when the two graphs are ER graphs, and when the two graphs are from SBM assuming accurate recovery, or from SBM with detection.

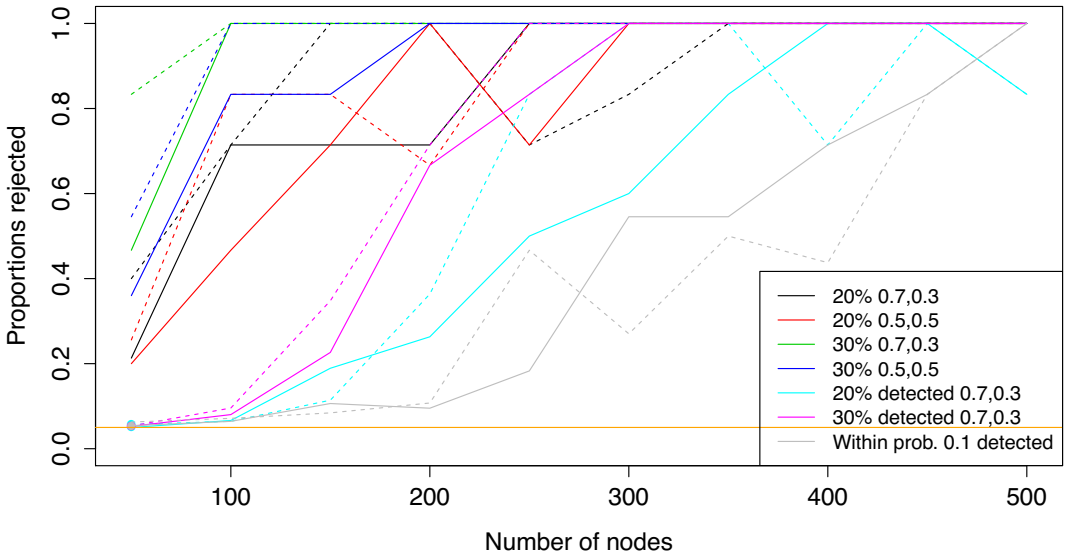


Figure 2.3: Power of the permutation and bootstrap tests under the alternative. The solid line: the permutation test; the dashed line: the bootstrap test. Dots indicate the estimated p-value is not significantly different from 0.05.

When the number of nodes is small, the permutation test assuming accurate recovery shows conservativeness. As the number of nodes increases, both permutation and bootstrap tests approach the right size of the test, with fluctuations under all scenarios we specified in the simulation. We note that this trend is observed regardless of community size differences, number of communities detected, as well as whether detection is used or assuming accurate recovery. Unsurprisingly, we also observe that as detection introduces uncertainties, the permutation test becomes less conservative using detection compared with cases assuming accurate recovery. The Chi-squared test and the LRT have an inflated level in general as the number of nodes increase towards 500 (Figure B.1 in Appendix), which is consistent with our discussion in the theory section. We also find that a different maximum number of steps M in the inner loop in the evaluation algorithm does not greatly affect the results.

Figure 2.3 shows the performance of permutation and bootstrap tests under different alternative scenarios. We observe that the power of the tests rapidly approaches 1 as the number of nodes increases for all cases. As expected, cases with weaker associations have flatter slopes for the trend. The power of both permutation and bootstrap tests are comparable to Chi-squared tests and LRT as shown in Figure B.2 in Appendix B.

2.5 Data analysis-Kapferer's tailor shop

We demonstrate our method using the data of interactions in a tailor shop in Zambia (then Northern Rhodesia) over a period of 10 months from Kapferer (1972), available in *Ucinet 5* (Borgatti et al., 1998). The dataset comes from the sociology literature and it records two types of interaction between 39 workers at two different time points (T1 and T2, 7 months apart) over a period of 1 month. One type is "instrumental" (work and assistance related), the other type is "sociational" (friendship, socioemotional). Work interactions are directed. Social interactions are undirected. An abortive strike occurred after T1, and a successful strike took place after T2. In our analysis, we convert all networks to be undirected considering the fact that complex interactions between individuals are sometimes hard to identify as reciprocal or unilateral, even though they may be distinguished with careful consideration.

There has been research on identifying and estimating the block structure in the data (Nowicki & Snijders, 2001) in order to discover the changing patterns over time. Kapferer (1972) approaches the problem by rearranging columns and rows of the adjacency matrix based on a sociometric method by Beum & Brundage (1950) to achieve greatest clustering. Nowicki & Snijders (2001) employ a statistical method focusing on the poststrike relations. Both approaches make inferences from the estimation perspective. Kapferer (1972) finds more inter-community interactions at T2 compared with T1 and identifies one individual who plays the crucial role of connecting commu-

nities, which is also revealed in the statistical analysis in Nowicki & Snijders (2001).

Our perspective distinguishes from both. Considering the close relationship between strike organizing in work and influences passing through social connections, one would suspect that there might be changes in the association between the two types of interactions across time which contribute to the difference in strike results. The hypothesis testing framework we proposed fits in the setting and is able to provide some statistical insight on the issue. We demonstrate our analysis using two modularity-based community detection methods, the Louvain method introduced in the simulation section, and the spectral optimization from Newman (2006b).

Table 2.2: Permutation test results for the tailor shop data in Zambia

	P-value		Proportion agreed	
Networks compared	Louvain	Spectral	Louvain	Spectral
T1 work v.s. T2 work	0.004	0.444	0.787	0.757
T1 work v.s. T1 social	0	0.152	0.771	0.702
T1 work v.s. T2 social	0	0	0.748	0.734
T2 work v.s. T1 social	0	0.6	0.673	0.605
T2 work v.s. T2 social	0	0	0.750	0.686
T1 social v.s. T2 social	0	0	0.745	0.703

Figure 2.4 shows the community detection results using the Louvain method. Table 2.2 and Figure 2.5 represent pairwise statistical associations between T1 and T2 social and work relations. We find that all pairwise associations are statistically significant based on the Louvain method while the test results based on spectral optimization are insignificant for all pairwise associations involving T2 work. These findings seems to suggest some structural changes in the work interactions at T2 that affects its re-

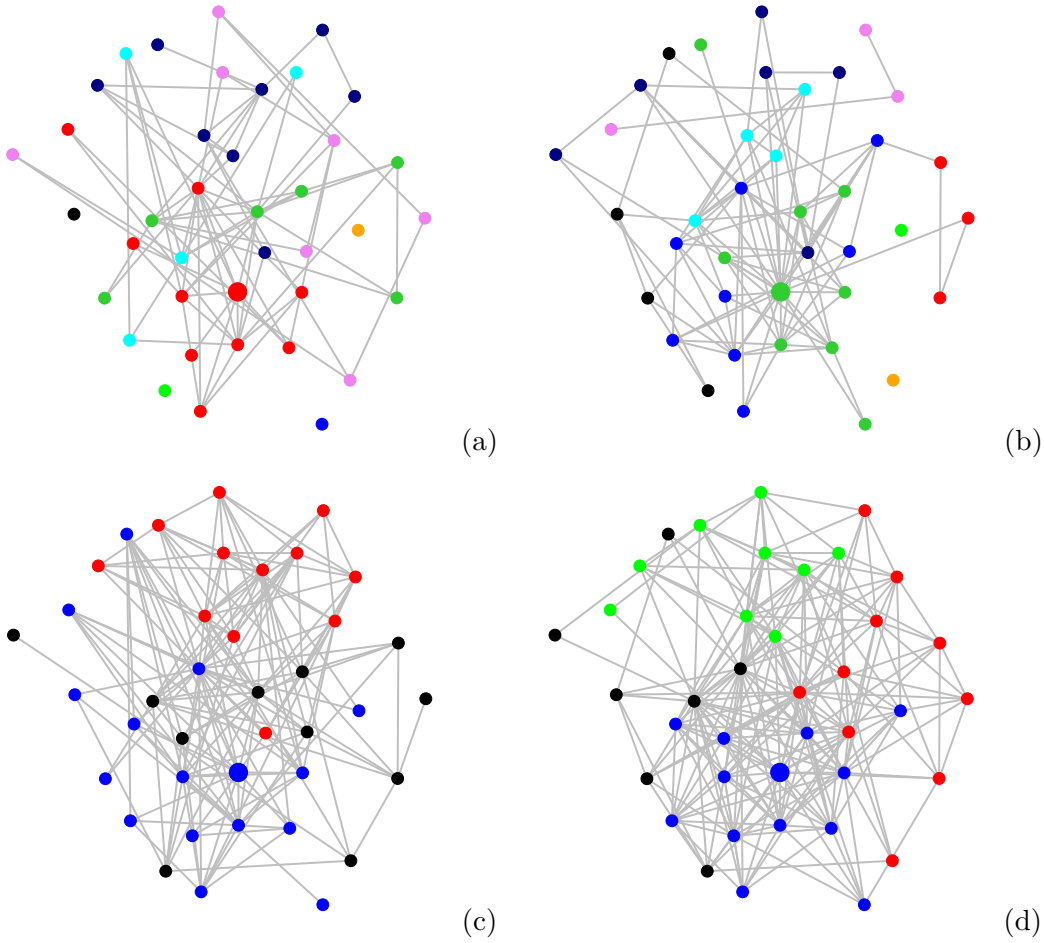


Figure 2.4: Community detection results using the Louvain method. Nodes of the same color belong to the same community in each subgraph. (a): T1 work interactions; (b): T2 work interactions; (c): T1 friend interactions; (d): T2 friend interactions. Vertex 11 is shown in larger node size.

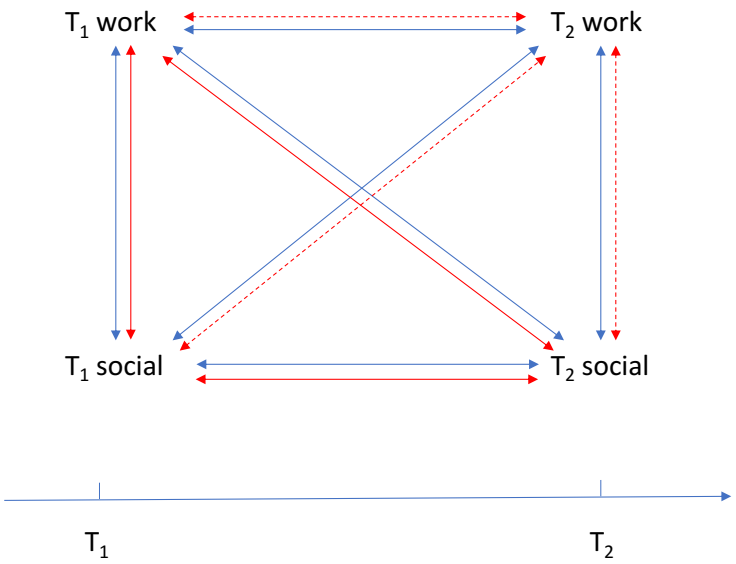


Figure 2.5: Graphical illustration of the permutation test results for the tailor shop data in Zambia. The dashed line indicates insignificant association; the solid line indicates significant association. Red: spectral optimization; Blue: the Louvain method.

lations with other observed networks. Those insignificant results are consistent with the speculations in Kapferer (1972) and Nowicki & Snijders (2001) where both find that the emergence of vertex 11 has altered the intra-cluster connections, and there are more interactions between individuals from different clusters.

2.6 Discussion

In this study, we first introduced a series of interrelated scientific questions that are of interest: whether a network has no community structure; whether underlying community structures in two networks are related; whether underlying community structures in two networks are the same. We then provided some theoretic explorations on per-

mutation tests focusing on the second issue. We compared the performance of the test with that of the bootstrap test, the Chi-squared test and the LRT in the simulation section and conclude that the permutation test preserves the type I error and has good power. We demonstrated the use of the test on work and social relations data of a tailor shop in Zambia from the sociology literature and compare results with previous research, where our method provides an alternative view and consistent results.

We note the connection between our permutation test and the discussion on the randomization test in the presence of interference in Rosenbaum (2007). In the article, two nulls are described: the null hypothesis of no primary effect H_0 and the null hypothesis of no effect $\widetilde{H_0}$. H_0 states that with repeating the same randomization procedure, the outcomes do not change. $\widetilde{H_0}$ states that the outcomes under the same randomization procedure do not change and equal those of the "uniformity trial" of a double blind placebo assignment to all units. Similar nulls can be defined in our setting as well, though the interpretations are different. The "uniformity trial" of a double blind placebo assignment to all units in our setting would be the case that there is no underlying community structure in network 1. The subsequently defined null of no effect $\widetilde{H_0}$ would then mean the outcome under every randomization assignment equals the corresponding outcome when network 1 is an Erdös-Rényi graph. $\widetilde{H_0} \Rightarrow H_0$ still holds. Fisher's randomization test has the correct level for testing either null, though it cannot differentiate the two.

3

Viral genetic linkage and geographic structure: analysis of data from Mexico

3.1 Problem Statement and Motivation

Contagious diseases, such as HIV infection, are often transmitted through close interpersonal interactions. As individuals live and interact within spatial and social

networks, it is interesting and important to investigate the association between disease transmission and geographical distance. Furthermore, at the group level (village, town, institution, etc.), investigating this relationship may provide information about chains of transmission that cannot be inferred from knowledge obtained at the individual level (e.g. regarding sexual or other forms of behavior). Studies of the effect of human mobility and geographic connections on viral transmission have been reported in the literature (Lagarde et al., 2003; Mangili & Gendreau, 2005; Balcan et al., 2009b; Faria et al., 2014; Bogoch et al., 2015). Demographics, socioeconomics and other biological, social, and cultural factors also play a role in affecting the probability of transmission among different categories of people (Buvé et al., 2002; Gregson et al., 2002; Helleringer & Kohler, 2007).

Chaillon et al. (2017) identified clusters of HIV sequences based on genetic distance among the viral sequences. This study was based on partial HIV *pol* sequences and found significant associations between the probability of individuals being clustered genetically and their various demographic and geographic characteristics. There is a large degree of variability in the cluster sizes, and the majority of clusters are small, with only 2 or 3 members. As individuals within a genetic cluster are likely to be closer to each other in a transmission chain than to those not in the cluster, it is of interest to assess the factors that might affect the probability of clustering, such as geographical distance. Large divergence between genetic sequences often implies a long distance of separation in the phylogenetic tree (Korber et al., 2000; Junqueira

et al., 2011; Worobey et al., 2016). In Chaillon et al. (2017), a threshold of genetic distance is chosen to define transmission links (Wertheim et al., 2013, 2016); pairs of sequences with divergence below the threshold are regarded as implying direct or indirect transmissions within the 5-year period of their study. This research suggests that transmissions did not likely occur within a given period among people with sequences in different transmission clusters. These finding support the idea that viral genetic linkage analysis combined with geographic information may be able to provide insight into the effect of geographical distance on viral transmission.

In this study, we make use of the hypothesis testing framework discussed in Chapter 2 to analyze the association between clusters based on pairwise viral genetic and geographical distances among people infected by these viral strains. As a sensitivity analysis, we demonstrate the test results for various threshold values.

3.2 Genetic Distance Data from Mexico

We use the viral genetic sequences from participants in Mexico discussed in Chaillon et al. (2017), from which the authors obtained pairwise viral genetic distances (GD). Chaillon et al. (2017) drew up data from the multi-center Mesoamerican Project, in which 4,192 HIV-1 subtype B *pol* sequences were sampled from unique Antiretroviral treatment (ART)-naive HIV-infected individuals from 23 states across Mexico between 2001 and 2016. Individuals were enrolled at participating clinics prior to expo-

sure to ART. Participants donated a single blood sample which was processed at the Center for Research in Infectious Diseases (CIENI) of the National Institute of Respiratory Diseases (INER) in Mexico City, a WHO-designated laboratory for HIV genotyping within 72 hours of collection. Partial HIV *pol* (HXB2 positions 2253–3554) was bulk sequenced and assembled following the procedures described in Chaillon et al. (2017). Pairwise genetic distances (GDs) were computed with all sequences aligned to the HXB2 reference sequence based on a nucleotide substitution model (Tamura & Nei, 1993; Oster et al., 2015).

3.3 Methods

To infer viral genetic clusters, Oster et al. (2015) and Chaillon et al. (2017) define links between subjects as occurring when the Genetic Distance (GD) is $\leqslant 0.015$ substitutions/site. The resulting viral genetic linkage analysis defines a network with edge weight

$$W_{ij} = I(\text{GD}_{ij} \leqslant d). \tag{3.1}$$

where GD_{ij} is the Genetic Distance between sequence i and j, $I(\cdot)$ is the indicator function, and threshold d is chosen to be 0.015.

Based on Equation 3.1, we note that there are other potential ways to define weight W. To make more complete use of the genetic distance information rather than simply dichotomizing it, one can alternatively define an inverse relationship between the

two which has been studied in other distance matrices, such as the geographical distances matrix(Lambiotte et al., 2008; Krings et al., 2009), using

$$W_{ij} = \frac{1}{\mathrm{GD}_{ij}}. \tag{3.2}$$

It is also possible to combine the two to yield

$$W_{ij} = \frac{1}{\mathrm{GD}_{ij}} \cdot I(\mathrm{GD}_{ij} \leqslant d). \tag{3.3}$$

With some modifications on the exponent of GD, we can also use

$$W_{ij} = \frac{1}{\mathrm{GD}_{ij}^2} \cdot I(\mathrm{GD}_{ij} \leqslant d). \tag{3.4}$$

Network constructions with different threshold values d in Equation 3.1 have been used for analyses of HIV (Wertheim et al., 2013, 2016; Chaillon et al., 2017). Wertheim et al. (2013) set the threshold value at 1%. Wertheim et al. (2016) used the value of 1.5% and explored the analysis using 1% and 2% without finding much difference. Chaillon et al. (2017) selected 1.5% as the threshold. These thresholds are chosen based on the findings in Wertheim et al. (2016) that the divergence between baseline and subsequent HIV *pol* sequences often do not exceed 1% over a 10-year period. These empirical threshold values for *pol* gene are similar to the estimates obtained in Korber et al. (2000) for *env*(gp160) and *gag*. In Korber et al. (2000), the evolution

rate for *env*(gp160) and *gag* is estimated to be 0.0024 (0.0018 to 0.0028) and 0.0019 (0.0009 to 0.0027) substitutions per base pair per year. These findings provide evidence for the interpretation of a linear relation between time and divergence at least for a short period of one to two decades on the evolutionary scale, which justifies the threshold value up to 2.5%.

For a given threshold, we can construct a network of viral genetic clusters, denoted as G. Without loss of generality, we choose weight Equation 3.1 to demonstrate the connection between the network and transmission clusters. Following the definition and notation in Bondy et al. (1976), we use G to denote an undirected graph and we use V for its vertex set and E for its edge set. As in Bondy et al. (1976), a *walk* in G is defined as a finite non-null sequence W of alternating vertices and edges such that each edge is adjacent to the vertices in the sequence that occur immediately before it and immediately after it. When the edges and vertices are all distinct, W is called a *path*. Two vertices are *connected* if there is a path in G between them. The vertex set V of G can be partitioned into nonempty subsets such that two vertices are connected if and only if they belong to the same subset. The subgraph of G generated from the partition are the *connected component*s of G. As mentioned above, for our G defined by viral genetic distance, connected components of G signify transmission clusters of potential transmission partners (Wertheim et al., 2013).

To test whether there is an association between clusters based on viral genetic distances and those based on geographical distances among the infected individuals, we

make use of the hypothesis testing methods developed in Chapter 2, which demonstrated the use of a permutation test in similar settings. The geographic network is constructed as a fully connected weighted graph. We use the inverse of the squared geographical distances between nodes as edge weights in the graph, where the geographical distances are calculated as the spherical distance in kilometers using the latitude and longitude of the centroids of the states associated with the geographic information of the individuals. The weight function is formulated based on previous research on human mobility (Lambiotte et al., 2008; Balcan et al., 2009a; Krings et al., 2009). To obtain clusters within the largest connected component of the viral genetic distance network, we use community detection methods to detect groups of individuals with viral sequences that are more closely related, which we refer to as communities within transmission clusters. We choose to use a modularity-based greedy algorithm—the Louvain method, which has been shown to have good performance in practice (Blondel et al., 2008; Fortunato, 2010) and is available in `R` package `igraph`.

3.4 Viral genetic and geographical distances among HIV infected participants in Mexico

Figure 3.1 shows the relationship between number of connected components in the viral genetic network and threshold values. The size of the largest connected component increases with the threshold, while the size of the second largest connected

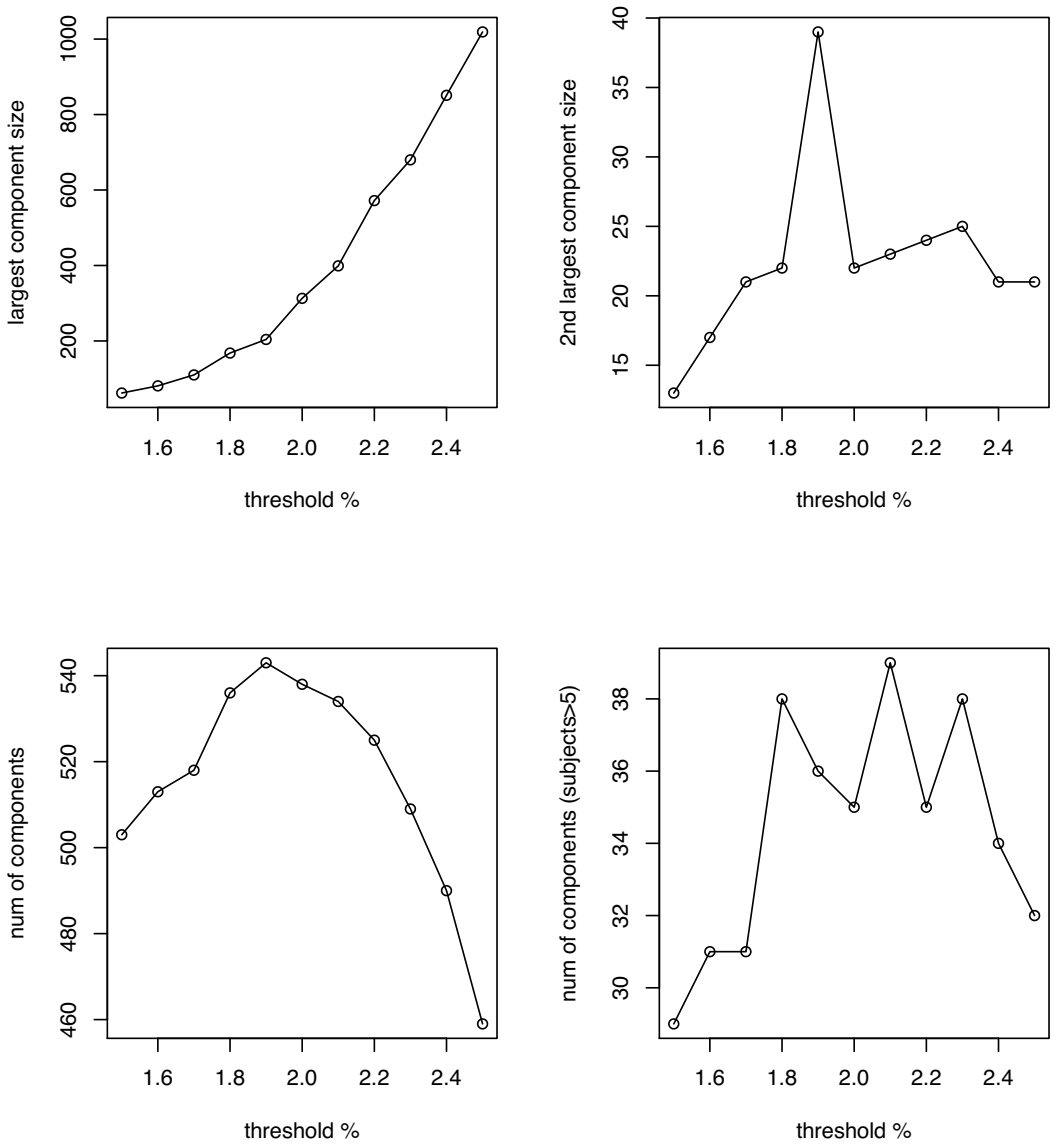
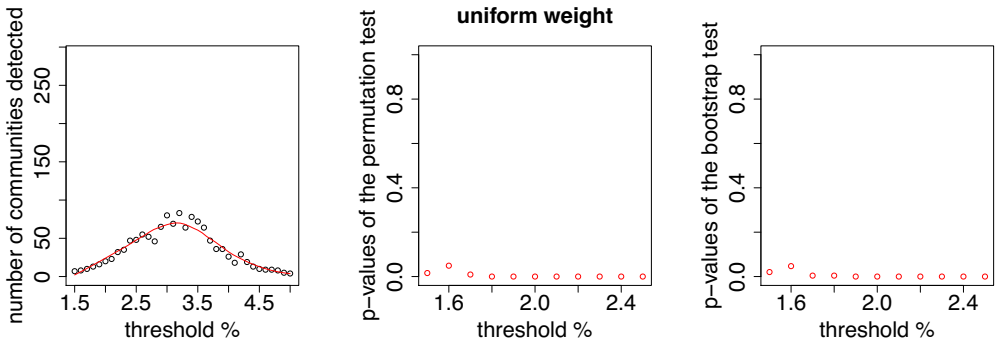


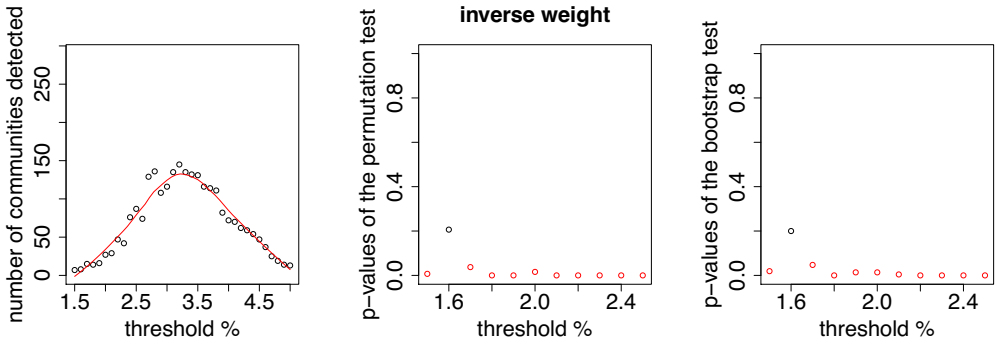
Figure 3.1: Descriptive statistics of the connected components for networks based on viral genetic distances. Upper left: size of the largest connected component; upper right: size of the second largest connected component; bottom left: number of connected components in the network; bottom right: number of connected components (size>5) in the network.

component increases as the threshold value increases from 1.5% to 1.8%, then fluctuates between 20 and 40 nodes with a peak at 1.9%. The number of connected components decreases steadily for threshold values larger than 2%; from about 500 at 2% to about 460 at 2.5%. These results show that as the value of the threshold increases, the largest connected component grows larger, as expected. Smaller connected components do not grow greatly before they are absorbed.

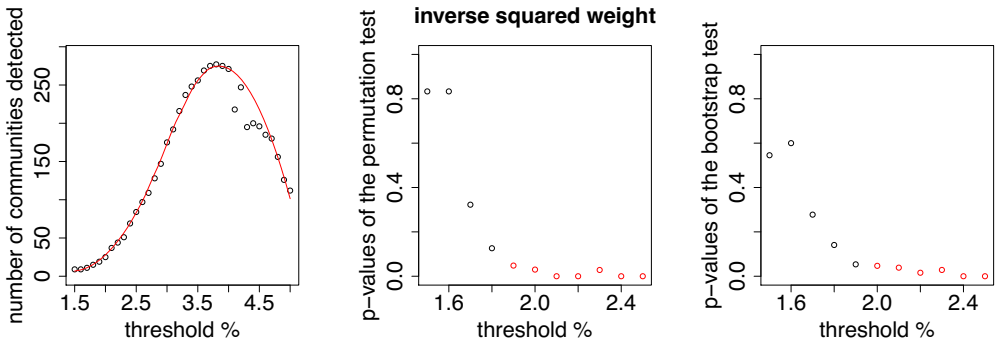
Figure 3.2 demonstrates the hypothesis testing results for different weight specifications and threshold values on the largest transmission cluster. Due to the fact that the test has low power shown in Chapter 2 when the number of individuals is around 20 and that all connected components are small except for the largest one, we first conduct the test on only the largest connected component. We explore three different weight specifications in the construction of genetic networks: (a) $W_{ij} = I(\mathrm{GD}_{ij} \leqslant d)$; (b) $W_{ij} = \frac{1}{\mathrm{GD}_{ij}} \cdot I(\mathrm{GD}_{ij} \leqslant d)$; (c) $W_{ij} = \frac{1}{\mathrm{GD}_{ij}^2} \cdot I(\mathrm{GD}_{ij} \leqslant d)$. In all three settings, we observe a bell-shaped curve for the number of detected communities in the largest connected component as the threshold varies. For weight specification in Equation 3.1, which dichotomizes genetic distances, the tests are significant at the 0.05 level for all thresholds between 1.5% and 2.5%. Settings with weights as inverse genetic distance in Equation 3.3 show similar trends with one exception at the threshold of 1.6%. Settings with weights chosen to be inverse squared genetic distance in Equation 3.4 show a decreasing p-value as the threshold varies from 1.5% to 2.5%. Test results are significant for all settings where the threshold exceeds 2%.



(a)



(b)



(c)

Figure 3.2: Hypothesis testing results for the association between clusters based on viral genetic distances and those based on geographical distances among individuals within the largest transmission cluster. Weight specifications in genetic networks are (a) $W_{ij} = I(\text{GD}_{ij} \leqslant d)$; (b) $W_{ij} = \frac{1}{\text{GD}_{ij}} \cdot I(\text{GD}_{ij} \leqslant d)$; (c) $W_{ij} = \frac{1}{\text{GD}_{ij}^2} \cdot I(\text{GD}_{ij} \leqslant d)$.

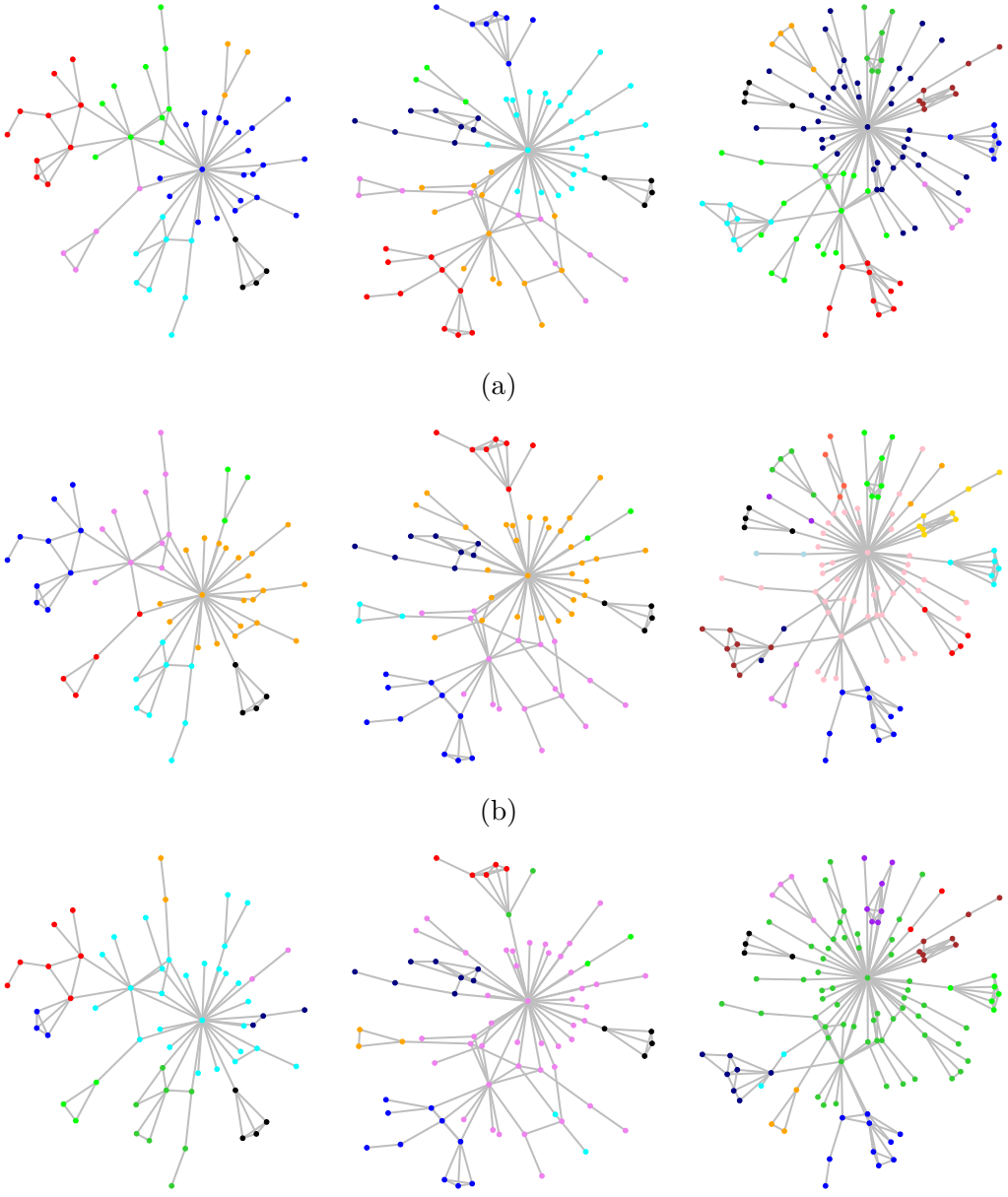


Figure 3.3: Detected communities based on viral genetic distances among individuals within the largest transmission cluster. Nodes in the same community in each subgraph in (a), (b), and (c) are shown in the same color. Weight specifications in genetic networks are (a) $W_{ij} = I(\text{GD}_{ij} \leqslant d)$; (b) $W_{ij} = \frac{1}{\text{GD}_{ij}} \cdot I(\text{GD}_{ij} \leqslant d)$; (c) $W_{ij} = \frac{1}{\text{GD}_{ij}^2} \cdot I(\text{GD}_{ij} \leqslant d)$. Threshold values are Left: $d = 1.5\%$; Middle: $d = 1.6\%$; Right: $d = 1.7\%$

Comparing these three weight specifications, Equation 3.1 assigns a uniform weight for all edges, which implies larger weights for distant connections among all three specifications. Equation 3.3 assigns more weight to shorter genetic distances and Equation 3.4 amplifies this effect. Setting (b) is between (a) and (c) with regard to the weighting. It is not surprising that the hypothesis testing results for (b) seem to show some characteristics from both (a) and (c), with an insignificant result at 1.6%.

Test results across (a), (b), and (c) suggest that tests of the association between clusters based on viral genetic and geographical distances among individuals within the largest transmission cluster is sensitive to the specification of weights. For uniform weights, test results show significant association, while for weights inversely related to squared genetic distances, the results are insignificant for smaller thresholds. The difference is explained in Figure 3.3 where communities which are detected based on different weights are shown for viral genetic networks obtained using thresholds 1.5%, 1.6%, and 1.7%. Nodes in the same community in each subgraph are shown in the same color, and different communities correspond to different colors. Use of uniform weights treats all edges equally and the community detection method in this setting focuses on finding groups of nodes which are more densely connected within each group. Use of heavier weights on smaller genetic distances tends to favor nodes that are more closely related even when there is an absence of an edge between them. As weights increases from top to bottom in Figure 3.3, detected communities extend further to include nodes that are close to them with respect to viral genetic distances.

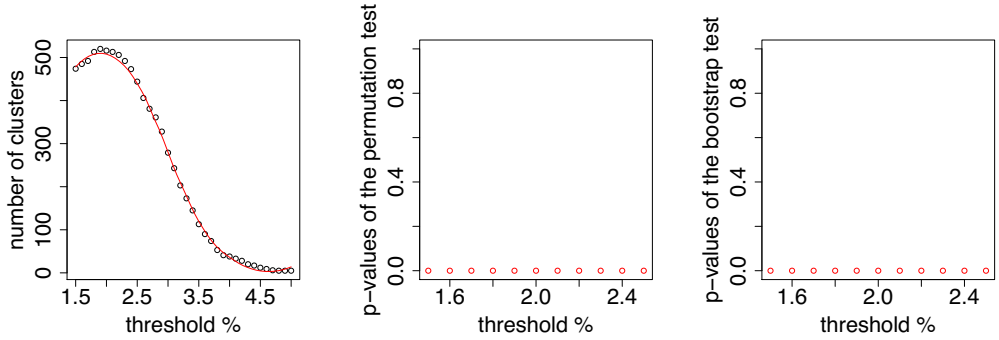


Figure 3.4: Hypothesis testing results for the association between clusters based on viral genetic and geographical distances among all participants where viral genetic clusters are defined as connected components in the viral genetic distance network.

For threshold 1.5%, no difference is observed between (a) and (b), while (c) has a well-spread blue community. For thresholds 1.6% and 1.7%, we see expanding purple and green communities in (c). Significances of the test results are similar to those which use other community detection methods such as spectral optimization (Newman, 2006a).

In Figure 3.4, we conduct tests for the association between clusters based on geographical distances and clusters defined as connected components in the viral genetic distance network. Both the permutation and the bootstrap tests are significant at the 0.05 level for all threshold values from 1.5% to 2.5%; this result suggests a strong association of HIV clustering and geographical distance regardless of the value of the threshold. Test results that combine across all connected components except for the largest one are similar to those in Figure 3.4.

3.5 Discussion

In this study, we use a hypothesis testing framework to analyze the association between clusters based on viral genetic and geographical distances among individuals infected by viral strains. Based on both within-cluster and cluster-wise analyses, significant associations between geographic pattern and viral genetic clusters are observed, which in general are not sensitive to the threshold used in the construction of the viral genetic clusters. Our findings are consistent with the results reported in the literature, though from a different perspective.

There are several aspects of our findings that might be of interest to pursue in future research. First, it would be interesting to investigate the role of different weight specifications for both geographic and genetic networks. In our study, we explored three different weight specifications for genetic distances. For clusters from networks based on geographical distances, because the gravity model tends to overemphasize model fit for shorter distances (Simini et al., 2012) in Chapter 1, the weights inversely related to geographical distance could be further modified, or could take a more complex form. For constructing viral genetic clusters from genetic distance networks, an empirical analysis comparing community detection results obtained using different weight specifications with structures shown in the phylogenetic tree would also be helpful. The second aspect of the study that could be expanded concerns a thorough phylogenetic analysis regarding the evolutionary timeline of the viral sequences and

its relation to different divergence thresholds. Original sequencing data are needed for this approach and to test for the validity of the molecular clock assumption (Tajima, 1993; Takezaki et al., 1995; Tamura et al., 2011). The third aspect that could be revisited in future research concerns cases when there are more than one large connected component of the transmission network. The stepdown procedure for multi-testing in Romano & Wolf (2005) provides some potentially useful tools. However, the monotonicity assumptions which guarantees preservation of type I error need to be checked and satisfied.

A

Supplemental Material 1

Computational complexity

Please see Figure A.1.

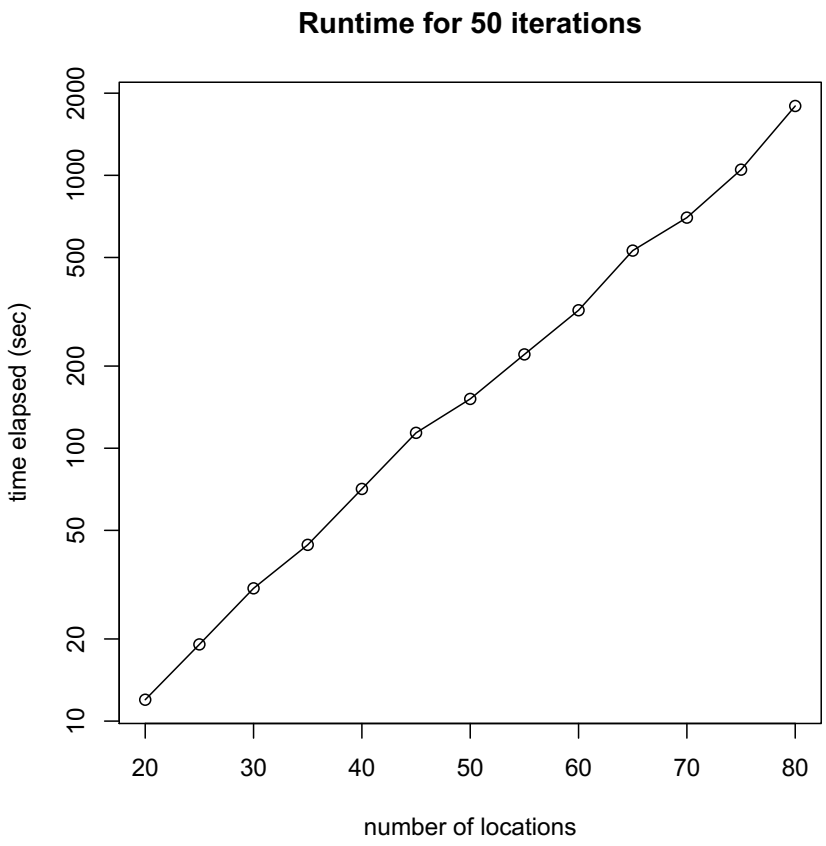


Figure A.1: Computation time (in seconds) versus number of locations in the simulation s. Note that the vertical axis is on logarithmic scale.

Discussion on two models

Other models to study the impact of spatial distance on communication intensity have been proposed in the literature, such as the radiation model (Simini et al., 2012), which predicts commuting fluxes between locations, and the rank-based friendship model proposed in Liben-Nowell et al. (2005), which ranks friendships based on the

geographical distance between them. Both models reduce to Equation 1.4 with certain constraints on their parameters or assumptions. The radiation model from Simini et al. (2012) uses the following specification

$$\langle T_{ij} \rangle = T_i \frac{m_i n_j}{(m_i + s_{ij})(m_i + n_j + s_{ij})}, \tag{A.1}$$

where $\langle T_{ij} \rangle$ is the average commuting or mobility flux from location i to j (for simplicity, we denote average flux as T_{ij} to keep the notation consistent), $T_i = \sum_{j \neq i} T_{ij}$ is the total number of commuters from i. s_{ij} is the population living in the circle centered at the source with a radius of r_{ij} (not including m_i). Adopting this notation,

$$T_{ij} = T_i \frac{m_i n_j}{(m_i + s_{ij})(m_i + n_j + s_{ij})}, \tag{A.2}$$

Taking the logarithm yields,

$$\log(T_{ij}) = \log(T_i) + \log(m_i) + \log(n_j) - \log(m_i + s_{ij}) - \log(m_i + n_j + s_{ij}). \tag{A.3}$$

As in Simini et al. (2012), we note that Equation A.3 reduces to Equation 1.4 with $\alpha + \beta = 1$ and $\gamma = 4$ when the population is uniformly distributed such that $m = n$ and $s_{ij} \approx m_i r_{ij}^2$. The model is mechanistic and has no parameter to fit.

The rank-based friendship model in Liben-Nowell et al. (2005) is formulated as follows. Let u and v be two individuals. They define $\text{rank}_u(v) = |\{w : d(u, w) <=$

$d(u,v)\}|$, where $d(u,w)$ is the distance between individual u and individual w. The probability of u and v being friends is modeled as

$$\Pr[u \to u] \propto \frac{1}{\text{rank}_u(v)}. \tag{A.4}$$

As $\text{rank}_u(v) \approx d(u,v)^2$ when the the population is uniformly distributed, Equation A.4 reduces to Equation 1.4 with $m = n = 1$ and $\gamma = 2$.

Both the gravity and radiation models are based on strict assumptions of the underlying mechanism, which are hard to validate. The gravity model, which uses the same parameters for each pair of locations, implicitly assumes a homogeneous effect of distance for the intensity function. The radiation model addresses this issue by modeling the intrinsic heterogeneity of the geographical distribution of population as indicated by the incorporation of s_{ij} in the model. However, subject to its strict assumption and 'parameter-free' property, it does not allow for fluctuations of other forms or from other sources. The rank-based model deals with the heterogeneity by substituting distance with rank, which seems to have a similar role as the s_{ij} in the radiation model. Thus the rank function in Equation A.4 can be regarded as an implicit function of distance and population distribution. We can make Equation A.4 a parametric model by putting a parameter at the power of the rank, which when assuming population is uniformly distributed across the area, would be equivalent to the gravity model with parameter γ for the distance r_{ij}.

We note here that even though the rank-based approach shed some lights on the question in which we are interested, to move from resolution at individual level to zip code or county level requires a completely different set of assumptions. Therefore, a rank-based gravity model cannot be seen as a simple extension from the rank-based friendship model.

B

Supplemental Material 2

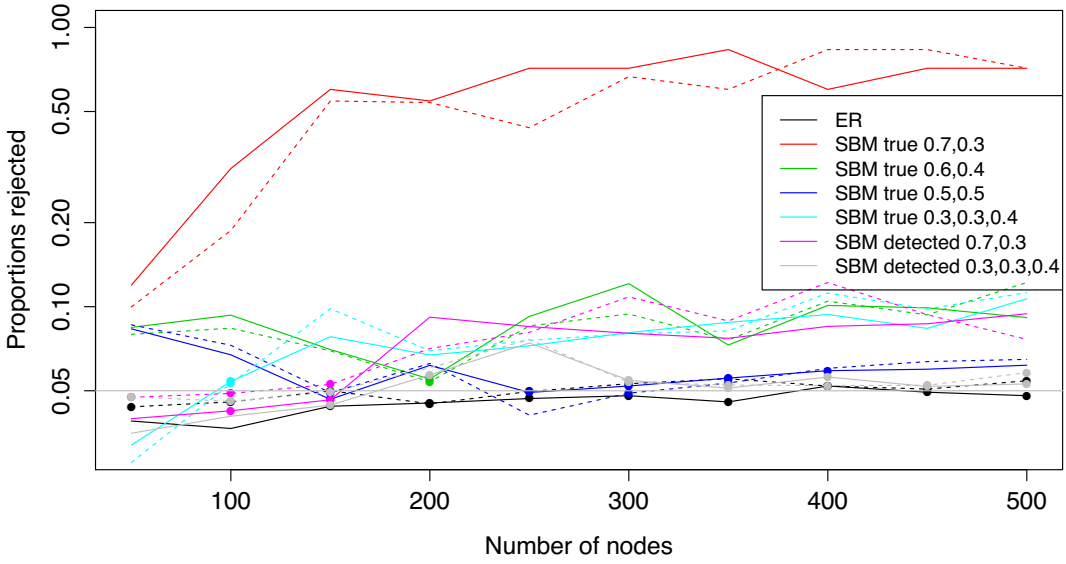


Figure B.1: Size of the Chi-squared test and LRT under the null. Solid line: the Chi-squared test; Dashed line: the LRT. Dots indicate the estimated p-value is not significantly different from 0.05.

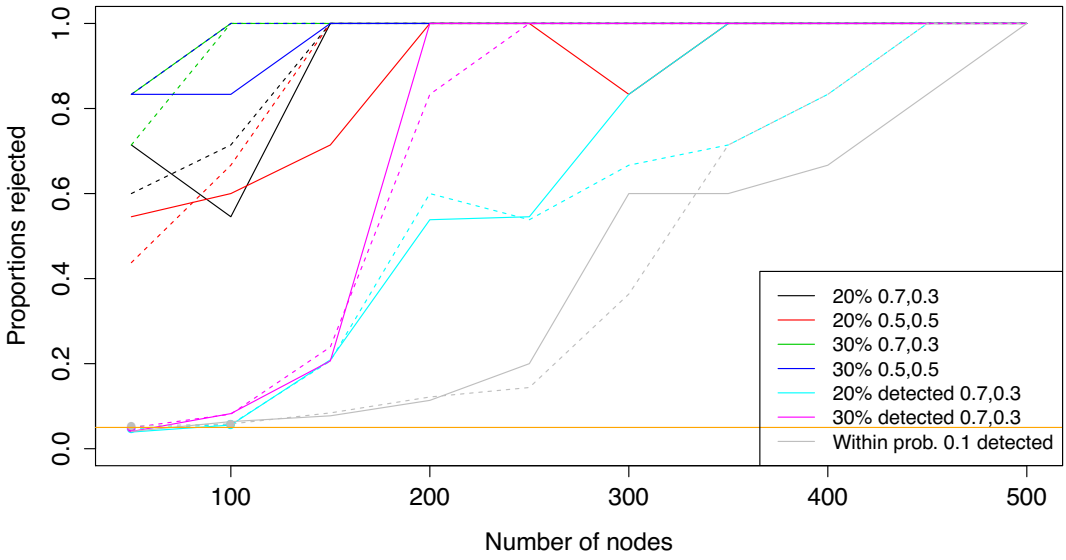


Figure B.2: Power of the Chi-squared test and LRT under the alternative. Solid line: the Chi-squared test; Dashed line: the LRT. Dots indicate the estimated p-value is not significantly different from 0.05.

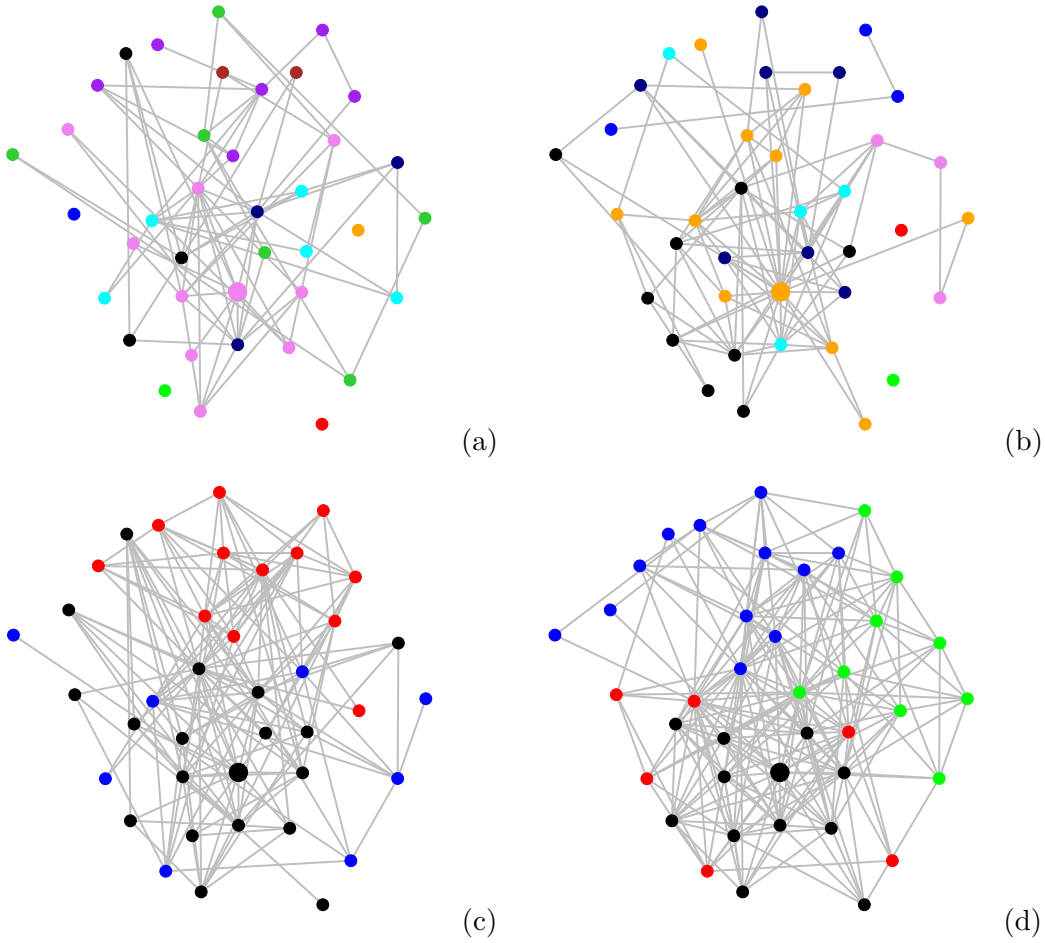


Figure B.3: Community detection results using the spectral optimization. Nodes of the same color belong to the same community in each subgraph. (a): T1 work interactions; (b): T2 work interactions; (c): T1 friend interactions; (d): T2 friend interactions. Vertex 11 is shown in larger node size.

References

Abbe, E. (2017). Community detection and stochastic block models: recent developments. *arXiv preprint arXiv:1703.10146*.

Agresti, A. (2013). *Categorical data analysis*. Wiley.

Ahmed, I., Pariente, A., & Tubert-Bitter, P. (2016). Class-imbalanced subsampling lasso algorithm for discovering adverse drug reactions. *Statistical Methods in Medical Research*.

Bader, D. & McCloskey, J. (2009). Modularity and graph algorithms. *SIAM AN10 Minisymposium on Analyzing Massive Real-World Graphs*, (pp. 12–16).

Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J. J., & Vespignani, A. (2009a). Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, 106(51), 21484–21489.

Balcan, D., Hu, H., Goncalves, B., Bajardi, P., Poletto, C., Ramasco, J. J., Paolotti, D., Perra, N., Tizzoni, M., Van den Broeck, W., et al. (2009b). Seasonal transmission potential and activity peaks of the new influenza a (h1n1): a monte carlo likelihood analysis based on human mobility. *BMC medicine*, 7(1), 45.

Beum, C. O. & Brundage, E. G. (1950). A method for analyzing the sociomatrix. *Sociometry*, 13(2), 141–145.

Bickel, P. J. & Sarkar, P. (2016). Hypothesis testing for automated community detection in networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(1), 253–273.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10), P10008.

Bogoch, I. I., Creatore, M. I., Cetron, M. S., Brownstein, J. S., Pesik, N., Miniota, J., Tam, T., Hu, W., Nicolucci, A., Ahmed, S., et al. (2015). Assessment of the potential for international dissemination of ebola virus via commercial air travel during the 2014 west african outbreak. *The Lancet*, 385(9962), 29–35.

Bondy, J. A., Murty, U. S. R., et al. (1976). *Graph theory with applications*, volume 290. Citeseer.

Borgatti, S., Everett, M., & Freeman, L. (1998). Ucinet v, reference manual. *Columbia, SC: Analytic Technologies*.

Brooks, S. & Giudici, P. (1998). Convergence assessment for reversible jump mcmc simulations.

Brooks, S. P. & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4), 434–455.

Buckee, C. O., Wesolowski, A., Eagle, N. N., Hansen, E., & Snow, R. W. (2013). Mobile phones and malaria: modeling human and parasite travel. *Travel Medicine and Infectious Disease*, 11(1), 15–22.

Burger, M., Van Oort, F., & Linders, G.-J. (2009). On the specification of the gravity model of trade: zeros, excess zeros and zero-inflated estimation. *Spatial Economic Analysis*, 4(2), 167–190.

Buvé, A., Bishikwabo-Nsarhaza, K., & Mutangadura, G. (2002). The spread and effect of hiv-1 infection in sub-saharan africa. *The Lancet*, 359(9322), 2011–2017.

Campbell, M., Donner, A., & Klar, N. (2007). Developments in cluster randomized trials and statistics in medicine. *Statistics in Medicine*, 26(1), 2–19.

Carnegie, N. B., Wang, R., & De Gruttola, V. (2016). Estimation of the overall treatment effect in the presence of interference in cluster-randomized trials of infectious disease prevention. *Epidemiologic Methods*, 5(1), 57–68.

Castelloe, J. M. & Zimmerman, D. L. (2002). Convergence assessment for reversible jump mcmc samplers. *Department of Statistics and Actuarial Science, University of Iowa, Technical Report*, 313.

Chaillon, A., Avila-Ríos, S., Wertheim, J. O., Dennis, A., García-Morales, C., Tapia-Trejo, D., Mejía-Villatoro, C., Pascale, J. M., Porras-Cortés, G., Quant-Durán, C. J., et al. (2017). Identification of major routes of hiv transmission throughout mesoamerica. *Infection, Genetics and Evolution*, 54, 98–107.

Chatterjee, A. & Lahiri, S. (2010). Asymptotic properties of the residual bootstrap for lasso estimators. *Proceedings of the American Mathematical Society*, 138(12), 4497–4509.

Chatterjee, A. & Lahiri, S. N. (2011). Bootstrapping lasso estimators. *Journal of the American Statistical Association*, 106(494), 608–625.

Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical review E*, 70(6), 066111.

Csáji, B. C., Browet, A., Traag, V. A., Delvenne, J.-C., Huens, E., Van Dooren, P., Smoreda, Z., & Blondel, V. D. (2013). Exploring the mobility of mobile phone users. *Physica A: Statistical Mechanics and its Applications*, 392(6), 1459–1473.

Eagle, N., Pentland, A. S., & Lazer, D. (2008). Mobile phone data for inferring social network structure. In *Social Computing, Behavioral Modeling, and Prediction* (pp. 79–88). Springer.

Eagle, N., Pentland, A. S., & Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36), 15274–15278.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of statistics*, 32(2), 407–499.

Expert, P., Evans, T. S., Blondel, V. D., & Lambiotte, R. (2011). Uncovering space-independent communities in spatial networks. *Proceedings of the National Academy of Sciences*, 108(19), 7663–7668.

Faria, N. R., Rambaut, A., Suchard, M. A., Baele, G., Bedford, T., Ward, M. J., Tatem, A. J., Sousa, J. D., Arinaminpathy, N., Pépin, J., et al. (2014). The early spread and epidemic ignition of hiv-1 in human populations. *science*, 346(6205), 56–61.

Fitzenberger, B. (1998). The moving blocks bootstrap and robust inference for linear least squares and quantile regressions. *Journal of Econometrics*, 82(2), 235–287.

Flowerdew, R. & Aitkin, M. (1982). A method of fitting the gravity model based on the poisson distribution. *Journal of Regional Science*, 22(2), 191–202.

Fortunato, S. (2010). Community detection in graphs. *Physics reports*, 486(3-5), 75–174.

Fortunato, S. & Barthélemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1), 36–41.

Gandy, A. (2009). Sequential implementation of monte carlo tests with uniformly bounded resampling risk. *Journal of the American Statistical Association*, 104(488), 1504–1511.

Gandy, A., Rubin-Delanchy, P., et al. (2013). An algorithm to compute the power of monte carlo tests with guaranteed precision. *The Annals of Statistics*, 41(1), 125–142.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis*, volume 2. Taylor & Francis.

Gelman, A. & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, (pp. 457–472).

Geyer, C. J. (2006). 5601 notes: The subsampling bootstrap. *Unpublished manuscript*.

Good, B. H., de Montjoye, Y.-A., & Clauset, A. (2010). Performance of modularity maximization in practical contexts. *Physical Review E*, 81(4), 046106.

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, 78(6), 1360–1380.

Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4), 711–732.

Green, P. J. & Hastie, D. I. (2009). Reversible jump mcmc. *Genetics*, 155(3), 1391–1403.

Gregson, S., Nyamukapa, C. A., Garnett, G. P., Mason, P. R., Zhuwau, T., Caraël, M., Chandiwana, S. K., & Anderson, R. M. (2002). Sexual mixing patterns and sex-differentials in teenage exposure to hiv infection in rural zimbabwe. *The Lancet*, 359(9321), 1896–1903.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geo-located twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(3), 260–271.

Helleringer, S. & Kohler, H.-P. (2007). Sexual network structure and the spread of hiv in africa: evidence from likoma island, malawi. *AIDS*, 21(17), 2323–2332.

Jones, J. H. & Handcock, M. S. (2003). An assessment of preferential attachment as a mechanism for human sexual network formation. *Proceedings of the Royal Society of London B: Biological Sciences*, 270(1520), 1123–1128.

Junqueira, D. M., De Medeiros, R. M., Matte, M. C. C., Araújo, L. A. L., Chies, J. A. B., Ashton-Prolla, P., & de Matos Almeida, S. E. (2011). Reviewing the history of hiv-1: spread of subtype b in the americas. *PLoS One*, 6(11), e27489.

Kapferer, B. (1972). *Strategy and transaction in an African factory: African workers and Indian management in a Zambian town*. Manchester University Press.

Knight, K. & Fu, W. (2000). Asymptotics for lasso-type estimators. *Annals of Statistics*, (pp. 1356–1378).

Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., Hahn, B., Wolinsky, S., & Bhattacharya, T. (2000). Timing the ancestor of the hiv-1 pandemic strains. *science*, 288(5472), 1789–1796.

Krings, G., Calabrese, F., Ratti, C., & Blondel, V. D. (2009). Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(07), L07003.

Lagarde, E., Schim Van Der Loeff, M., Enel, C., Holmgren, B., Dray-Spira, R., Pison, G., Piau, J.-P., Delaunay, V., M'Boup, S., Ndoye, I., et al. (2003). Mobility and the spread of human immunodeficiency virus into rural areas of west africa. *International journal of epidemiology*, 32(5), 744–752.

Lambiotte, R., Blondel, V. D., de Kerchove, C., Huens, E., Prieur, C., Smoreda, Z., & Van Dooren, P. (2008). Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387(21), 5317–5325.

Lancichinetti, A. & Fortunato, S. (2009). Community detection algorithms: a comparative analysis. *Physical review E*, 80(5), 056117.

Leskovec, J., Lang, K. J., & Mahoney, M. (2010). Empirical comparison of algorithms for network community detection. In *Proceedings of the 19th international conference on World wide web* (pp. 631–640).: ACM.

Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., & Tomkins, A. (2005). Geographic routing in social networks. *Proceedings of the National Academy of Sciences*, 102(33), 11623–11628.

Mangili, A. & Gendreau, M. A. (2005). Transmission of infectious diseases during commercial air travel. *The Lancet*, 365(9463), 989–996.

Meinshausen, N. & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4), 417–473.

Mucha, P. J., Richardson, T., Macon, K., Porter, M. A., & Onnela, J.-P. (2010). Community structure in time-dependent, multiscale, and multiplex networks. *science*, 328(5980), 876–878.

Newman, M. E. (2006a). Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3), 036104.

Newman, M. E. (2006b). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23), 8577–8582.

Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., & Mascolo, C. (2012). A tale of many cities: universal patterns in human urban mobility. *PloS one*, 7(5), e37027.

Nowicki, K. & Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455), 1077–1087.

Onnela, J.-P., Arbesman, S., González, M. C., Barabási, A.-L., & Christakis, N. A. (2011). Geographic constraints on social network groups. *PLoS one*, 6(4), e16939.

Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., & Barabási, A.-L. (2007). Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18), 7332–7336.

Oster, A. M., Wertheim, J. O., Hernandez, A. L., Ocfemia, M. C. B., Saduvala, N., & Hall, H. I. (2015). Using molecular hiv surveillance data to understand transmission between subpopulations in the united states. *Journal of acquired immune deficiency syndromes (1999)*, 70(4), 444.

Park, T. & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482), 681–686.

Pillow, J. W. & Scott, J. G. (2012). Fully bayesian inference for neural models with negative-binomial spiking. In *NIPS* (pp. 1907–1915).

Politis, D. N. & Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, (pp. 2031–2050).

Politis, D. N., Romano, J. P., & Wolf, M. (2001). On the asymptotic theory of subsampling. *Statistica Sinica*, (pp. 1105–1124).

Polson, N. G., Scott, J. G., & Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504), 1339–1349.

Porter, M. A., Onnela, J.-P., & Mucha, P. J. (2009). Communities in networks. *Notices of the AMS*, 56(9), 1082–1097.

Richardson, T., Mucha, P. J., & Porter, M. A. (2009). Spectral tripartitioning of networks. *Physical Review E*, 80(3), 036111.

Romano, J. P. & Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469), 94–108.

Rosenbaum, P. R. (2007). Interference between units in randomized experiments. *Journal of the American Statistical Association*.

Sailer, K. & McCulloh, I. (2012). Social networks and spatial configuration—how office layouts drive social interaction. *Social Networks*, 34(1), 47–58.

Simini, F., González, M. C., Maritan, A., & Barabási, A.-L. (2012). A universal model for mobility and migration patterns. *Nature*, 484(7392), 96–100.

Staples, P. C., Ogburn, E. L., & Onnela, J.-P. (2015). Incorporating contact network structure in cluster randomized trials. *Scientific Reports*, 5.

Strathmann, H., Sejdinovic, D., & Girolami, M. (2015). Unbiased bayes for big data: Paths of partial posteriors. *arXiv preprint arXiv:1501.03326*.

Tajima, F. (1993). Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics*, 135(2), 599–607.

Takezaki, N., Rzhetsky, A., & Nei, M. (1995). Phylogenetic test of the molecular clock and linearized trees. *Molecular biology and evolution*, 12(5), 823–833.

Tamura, K. & Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial dna in humans and chimpanzees. *Molecular biology and evolution*, 10(3), 512–526.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., & Kumar, S. (2011). Mega5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular biology and evolution*, 28(10), 2731–2739.

Tang, M., Athreya, A., Sussman, D. L., Lyzinski, V., & Priebe, C. E. (2014). A semiparametric two-sample hypothesis testing problem for random dot product graphs. *arXiv preprint arXiv:1403.7249*.

Tang, M. & Priebe, C. E. (2016). Limit theorems for eigenvectors of the normalized laplacian for random graphs. *arXiv preprint arXiv:1607.08601*.

Tatem, A. J., Huang, Z., Narib, C., Kumar, U., Kandula, D., Pindolia, D. K., Smith, D. L., Cohen, J. M., Graupe, B., Uusiku, P., et al. (2014). Integrating rapid risk mapping and mobile phone call record data for strategic malaria elimination planning. *Malaria journal*, 13(1), 52.

Ter Wal, A. L. & Boschma, R. A. (2009). Applying social network analysis in economic geography: framing some key analytic issues. *The Annals of Regional Science*, 43(3), 739–756.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, (pp. 267–288).

Wang, D., Pedreschi, D., Song, C., Giannotti, F., & Barabasi, A.-L. (2011). Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1100–1108).: ACM.

Wang, L., Wang, Z., Zhang, Y., & Li, X. (2013). How human location-specific contact patterns impact spatial transmission between populations? *Scientific Reports*, 3.

Wang, R., Goyal, R., Lei, Q., Essex, M., & De Gruttola, V. (2014). Sample size considerations in the design of cluster randomized trials of combination hiv prevention. *Clinical Trials*, 11(3), 309–318.

Wertheim, J. O., Leigh Brown, A. J., Hepler, N. L., Mehta, S. R., Richman, D. D., Smith, D. M., & Kosakovsky Pond, S. L. (2013). The global transmission network of hiv-1. *The Journal of infectious diseases*, 209(2), 304–313.

Wertheim, J. O., Oster, A. M., Hernandez, A. L., Saduvala, N., Bañez Ocfemia, M. C., & Hall, H. I. (2016). The international dimension of the us hiv transmission network and onward transmission of hiv recently imported into the united states. *AIDS research and human retroviruses*, 32(10-11), 1046–1053.

Worobey, M., Watts, T. D., McKay, R. A., Suchard, M. A., Granade, T., Teuwen, D. E., Koblin, B. A., Heneine, W., Lemey, P., & Jaffe, H. W. (2016). 1970s and 'patient 0'hiv-1 genomes illuminate early hiv/aids history in north america. *Nature*, 539(7627), 98.

Zhao, Y., Levina, E., & Zhu, J. (2011). Community extraction for social networks. *Proceedings of the National Academy of Sciences*, 108(18), 7321–7326.

Zhou, M., Li, L., Dunson, D., & Carin, L. (2012). Lognormal and gamma mixed negative binomial regression. In *Machine learning: proceedings of the International Conference. International Conference on Machine Learning*, volume 2012 (pp. 1343).: NIH Public Access.