UNIVERSITY OF CALIFORNIA SAN DIEGO

Analyses of Viral Genetic Networks in the Presence of Missing Data

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Biostatistics

by

Tyler Vu

Committee in charge:

Professor Xin Tu, Chair
Professor Victor De Grutolla
Professor Natasha Martin
Professor Florin Vaida
Professor Joel Weirtheim
Professor Jingjing Zou

2022

The Dissertation of Tyler Vu is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

EPIGRAPH

You are both the screenwriter and the actor in your life.
If you don't like your movie, write a different scene
and/or act your part better. Some of us have written a
better script full of wealth, great health, social success
but you are still acting out your old role.

Tej Dosa

The only authority over us is God.

Tyler Vu

The downside is never failure. View failure as
another form of "success". There's tangible
success - growth in the external world. And
there's intangible success ("failure") - growth
in the internal world. As long as you're not stagnant,
physically or mentally - you're progressing.

Ian Du

TABLE OF CONTENTS

Dissertation Approval Page iii

Epigraph iv

Table of Contents v

List of Figures vii

List of Tables viii

Acknowledgements ix

Vita xi

Abstract of the Dissertation xii

Chapter 1 Introduction 1

Chapter 2 Estimating Viral Genetic Linkage Rates in the Presence of Missing Data 5
- 2.1 Introduction 5
- 2.2 Notation and Setting 6
- 2.3 Bias Arising from Incomplete Data 7
- 2.4 Methods 8
  - 2.4.1 An Unbiased Estimator of Probability of Linkage 9
  - 2.4.2 A Feasible and Consistent Estimator for Linkage Rate 10
  - 2.4.3 Inference on Linkage Rate: A U-Statistics Framework 12
  - 2.4.4 Diagnostic 17
- 2.5 Applications 17
  - 2.5.1 Simulation Study 17
  - 2.5.2 Botswana Combination Prevention Project 20
- 2.6 Discussion 22
- 2.7 Acknowledgements 23

Chapter 3 Estimating Probabilities of Linkage in the Presence of Missing Data 24
- 3.1 Introduction 24
- 3.2 Botswana Combination Prevention Project (BCPP) 26
  - 3.2.1 Study Introduction 26
  - 3.2.2 Missing Data 26
- 3.3 Notation and Setting 27
- 3.4 Methodology 28
  - 3.4.1 Estimating Probability of Linkage 29

3.4.2 Estimating Participation Probability . . . 30
3.4.3 Addressing Missing HIV+ Nonparticipants . . . 31
3.4.4 Inference on Probability of Linkage . . . 32
3.5 Application . . . 36
3.5.1 Simulation Study . . . 36
3.5.2 BCPP Application . . . 37
3.6 Discussion . . . 38
3.7 Acknowledgements . . . 40

Chapter 4 Nonparametric Estimation of Network Properties in the Presence of Missing Data . . . 41
4.1 Introduction . . . 41
4.2 Notation and Setting . . . 42
4.3 Methodology . . . 43
4.3.1 Calculating ϕ_m . . . 43
4.3.2 Consistency of $\widehat{\phi}$ . . . 44
4.3.3 Diagnostic . . . 46
4.4 Application . . . 48
4.4.1 Simulation . . . 48
4.4.2 BCPP . . . 49
4.5 Discussion . . . 53
4.6 Acknowledgements . . . 53

Chapter 5 Conclusions and Future Work . . . 54

Appendix A Estimating Viral Genetic Linkage Rates in the Presence of Missing Data . . . 56
A.1 Proof of Theorem 1 . . . 56
A.2 Derivation of Consistent Estimators for $\mathbf{h}_{rs1}(\mathbf{y}_{ki})$ . . . 59
A.2.1 A consistent estimator for $h^1_{rs1}(\mathbf{y}_{ki})$ . . . 59
A.2.2 A consistent estimator for $h^2_{rs1}(\mathbf{y}_{ki})$ . . . 61
A.2.3 A consistent estimator for $h^3_{rs1}(\mathbf{y}_{ki})$ . . . 62
A.3 A Consistent Estimator for the Variance of $\widetilde{\theta}_{rs}$ . . . 64
A.4 Comparison of Adjusted and Unadjusted Probabilities of Linkage . . . 65
A.4.1 Estimates for Probabilities of Linkage . . . 65
A.4.2 Standard Error of Probabilities of Linkage . . . 65

Bibliography . . . 66

LIST OF FIGURES

Figure 2.1. Plot of a network to show differences between u^s_{ri} and v^s_{ri}. For simplicity, we consider only a single group, r. Nodes that are colored in red are selected in $S_{n(r)}$. Note that u^s_{ri} and v^s_{ri} are only defined for $\mathbf{y}_{ri} \in S_{n(r)}$. 8

Figure 2.2. Distribution of adjusted estimators for various values of p when applying the proposed approach to a scale-free network with $\alpha = 3$. Here, the x-axis represents p and the y-axis represents values for the adjusted estimators where the red line indicates the true value. 18

Figure 2.3. Distribution of adjusted and unadjusted estimators for linkage rates from the simulation in Section 5.1. The red horizontal line is the true value. . . . 19

Figure 2.4. Adjusted estimators for linkage rates between the communities from the BCPP and coefficients of variation (CV) indicated by colors of circle and cell, respectively. 21

Figure 3.1. Distribution of adjusted and unadjusted estimators for Probabilities of Linkage from the simulation. The red horizontal line is the true value 37

Figure 3.2. Adjusted estimators for probabilities of linkage between the communities from the BCPP and coefficients of variation (CV) indicated by colors of circle and cell, respectively. 39

Figure 4.1. The proposed approach was applied to a power law network with $\alpha = 3$ and $N = 500$ for the following proportions: $0.20, 0.30, 0.40, 0.50, 0.60, 0.70$. The x axis represents the degree value and the y axis represents the proportion of nodes with the specified degree value. For values between 0 and 3, the degree distribution of the proposed approach diverges from the true degree distribution for proportions of 0.40 and less. For values between 4 and 7, the proposed degree distributions deviate considerably from the true values for all proportions. 50

Figure 4.2. Estimated α for each sample performed in the simulation study.The red horizontal line represents the true value for α. 51

Figure 4.3. Proposed degree distributions for the communities Maunatlala, Mmathethe, Ramokgonami, Shakawe. 52

Figure 4.4. Estimated values for α for the communities Maunatlala, Mmathethe, Ramokgonami, Shakawe. 52

LIST OF TABLES

Table 2.1. Coverage probabilities of adjusted estimators from simulation in Section 5.1. 20

Table 2.2. The proportion (p) and number (n) of HIV+ individuals in each community that participated in the BCPP. 21

Table 3.1. Each of the scenarios considered in the simulation study performed in Section 5.1. 37

Table 3.2. The coverage probabilities for the adjusted estimators for probabilities of linkage from the simulation. 37

Table 3.3. The proportion (p) and number (n) of HIV+ individuals in each community that participated in the BCPP. 38

Table A.1. Unadjusted (U) and adjusted (A) linkage rates for the communities investigated from the BCPP. 65

Table A.2. Standard errors of unadjusted (U) and adjusted (A) linkage rates for the communities investigated from the BCPP. 65

ACKNOWLEDGEMENTS

I will start off by saying my PhD thesis was honestly very easy and smooth and no, this is not me saying how smart I am (if anything, I'm quite dumb). This was a result of having the best group of people helping me along the way. Specifically, this group of people would be Xin Tu, Victor De Grutolla, Jingjing Zou, Ravi Goyal, Morris Wu and Tuo Lin. The fact, you all were there for me to help out with my research at any point in time seriously means so much and something I will always be grateful for.

I specifically would like to acknowledge Professors Xin Tu, Jingjing Zou and Victor De Grutolla. Although on paper, Xin is the only chair of my committee. From the very beginning of my thesis I have always viewed it as the three of you all being co-chairs.

I would like to then thank all my colleagues and mentors that have shaped my entire experience in school: Laura Smith (California State University, Fullerton), Adam Glesser (California State University, Fullerton), Mikhail Gofman(California State University, Fullerton), Sherri Rose (Stanford University), Savannah Bergquist (UC Berkeley), Heather Mattie (Harvard T.H. Chan School of Public Health).

I would like to thank my friends and family. Specifically, I'd like to thank my Mom and Dad, my brothers Westin and Hunter. My grandma and grandpa, Ba and Ong.

I'd also like to thank my friend whether it be through home, college or online biz: Leeroy, Andy, Cristian, Tiltnam, Colean, Trevy, Brona, Handsome Man, Dog, Alex, Mannyman, Elijah, Luqui, Tenzin, Chris, Amir, Mattman, Jon, Patty, Ryan, Jeanelle, Tiff, Jenna, Mamà, Logan, Colby, Micah, Aiden, Jonathan, David, Vicky and Yng Morris.

Chapter 2, in full, is a reprint of the materials as it appears in Estimating Viral Genetic Linkage Rates in the Presence of Missing Data 2022. Victor De Gruttola, Tu, Xin, Vladimir Novitsky, Jingjing Zou, Tuo Lin. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reprint of the materials as it appears in Estimating Probabilities of Linkage in the Presence of Missing Data 2022. Victor De Gruttola, Tu, Xin, Jingjing Zou, Tuo

Lin. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, is a reprint of the materials as it appears in Nonparametric Estimation of Network Properties in the Presence of Missing Data 2022. Victor De Gruttola, Ravi Goyal, Tu, Xin, Jingjing Zou, Tuo Lin. The dissertation author was the primary investigator and author of this paper.

VITA

2018	Bachelor of Arts, California State University, Fullerton
2022	Doctor of Philosophy, University of California, San Diego

ABSTRACT OF THE DISSERTATION

Analyses of Viral Genetic Networks in the Presence of Missing Data

by

Tyler Vu

Doctor of Philosophy in Biostatistics

University of California San Diego, 2022

Professor Xin Tu, Chair

Molecular epidemiology is increasingly used to investigate patterns of HIV transmission. To do so, many analyses consider investigating properties of a sexual or transmission network. The use of sampled data to estimate such properties is a common practice; however, in the presence of missing data, even missing completely at random, networks based on sampled data do not represent their population counterparts. As a result, inferences on sampled networks become unreliable. To address this challenge, we propose statistical approaches to accommodating missing data in the analysis of sampled networks.

Chapter 1

Introduction

Molecular epidemiology is increasingly used to investigate patterns of disease transmission dynamics and the effect of interventions on them [2, 10, 29]; a useful analysis in this regard is of viral genetic linkage (VGL) where linkage is defined on the genetic distance between viral sequences taken from pairs of infected individuals. Pairs with a genetic distance below a specified threshold would be considered linked and demonstrate transmission between the two individuals [17]. Such analyses can reveal which viral strains are propagating within and between-communities, the characteristics of people infected with such strains, and the effects of interventions designed to control rates at which viral genetic clusters grow.

Both the CDC and NIH have proposed that viral genetic analyses be used to guide resources intended to end the AIDS epidemic where there are a million deaths annually as a result of HIV [6, 8]. The development of antiretroviral therapy (ART) has shown to significantly reduce the mortality rate of HIV-infected individuals [11]. Even with such advances, over 69% of the individuals with HIV are located in Sub-Saharan Africa [18]. Mah et. al. showed that concurrent sexual relationships played an important role in the AIDS epidemic in this region [20]. The use of viral genetic sequencing has been of interest to model such sexual relationships to further investigate HIV transmission dynamics and the impact of prevention interventions [2, 10, 29]. Yerly et al. used viral genetic sequencing to demonstrate that drug resistance transmission decreased considerably in Switzerland between 1996 and 1999 by using phylogenetics trees and epidemiological linkages to identify clusters between treatment groups in the study [29]. Similarly, Aldous et al. showed that HIV transmission was associated with lack of antiretroviral drug use and higher viral load [1].

Currently most studies regarding VGL make inferences about a population by estimating network properties from a set of observed viral genetic sequences while disregarding any potential effects from sampling and missing data [1, 4]. However, in the presence of missing data, even if missing completely at random, observed VGL networks do not represent their population counterparts, rendering inferences based on sampled networks unreliable without adjustment for missing data. Specifically, estimates of network properties from observed VGL networks will generally be biased unless networks are completely observed, which generally is unfeasible. As a result, inference pertaining to transmission may not accurately represent transmission networks.

Liu, et al addressed this bias by developing a multiple imputation framework in which the missing sequences are imputed [17]. Carnegie et al. made use of a subsampling approach to generate estimators that adjust for this bias [3]. Note that both of these approaches focus on estimating specific network properties; both papers investigate properties of the proposed estimators only through simulation; neither provide a theoretical foundation for properties like unbiasedness and consistency, i.e. convergence of a network property to its true value as the size of the network sample and the network itself increases. To address this challenge, this dissertation presents novel methods to generate consistent and asymptotically normally estimators for various network properties pertinent to VGL linkage.

To guide development of methods, we will analyze data from a large pair-matched cluster-randomized trial, the Botswana Combination Prevention Project (BCPP) [9]. The overarching goal of the BCPP was to estimate the impact of a package of combination prevention interventions on reducing population-level cumulative three year HIV incidence in Botswana, a country in Sub-Saharan Africa. The BCPP randomized 15 matched pairs of communities (30 communities in all) Botswana to intervention vs control; matching was based on size of community, pre-existing health services, population age structure, and geographical location.

The intervention consisted of home-based and mobile HIV testing and counselling, point-of-care CD4 testing, linkage to care support, expanded antiretroviral therapy. The study generated viral sequences for all infected members of a 20% sample of households at baseline and on all participants who became infected during the study. Additionally, 3 pairs of intervention-control communities participated in the End of Survey Study (ESS), in which eligible members of all households were tested for HIV at the end of the study.

However, issues arise in assessing the intervention effect, because of the presence of mixing between randomized clusters, i.e., some individuals have sexual partners outside communities randomized to their assigned treatment group, and hence violating the Stable Unit Treatment Value Assumption (SUTVA). As a result, one can obtain unbiased estimates of the randomized treatment effect (including the presence of mixing) but not of causal estimands of public health interest, such as the difference in outcome between rolling out the intervention everywhere versus nowhere. I intend for the developments made in this dissertation to be used in the development of approaches to address this issue in the future.

To accurately estimate the causal effect of the BCPP intervention described above (not just the randomized effect) we must first develop knowledge about mixing. To investigate this issue, BCPP investigators made use of the viral sequences obtained from HIV-infected individuals, with a goal of learning about patterns of HIV transmission dynamics in Botswana by estimating network properties pertinent to VGL linkage [19]. However, with currently available methods, one cannot be assured of the conditions under which estimators have good asymptotic properties. This dissertation develops these methods and applies them to the investigation of transmission flows across three pairs of communities in Botswana. The results from these analyses reflects transmission flows between three sets. Such information is useful for estimating a counterfactual estimator of interest—the difference between implementing the intervention in all communities versus implementing it in none. Although we use the BCPP to guide development of our methods, the developments made in this dissertation extend to general networks beyond HIV linkage. The methods developed will prove useful in revealing features of transmission patterns within and across communities and assessing the effect of interventions for infectious diseases such as COVID-19.

Chapter 2

Estimating Viral Genetic Linkage Rates in the Presence of Missing Data

2.1 Introduction

While networks have become widely used to analyze elements in a system and how these elements interconnect, the challenge of sampling complete network data remains a prevalent issue. In most instances, we only sample a portion of the nodes and hence don't observe the edges corresponding to the missing nodes. As a result, estimators for linkage rates that ignore the impact of missing data will be biased downwards.

Additionally, obtaining asymptotic properties for estimators of linkage rates is challenging, because linkage indicators across pairs of individuals are correlated; hence, the central limit theorem and law of large numbers cannot be directly applied to such estimators.

For our parameter of interest, past work has been explored to accommodate missing data in the case of viral genetic linkage networks (which in turn would apply to networks general). In Liu et. al., a multiple imputation framework in which the missing sequences are imputed is used to adjust for the bias in estimation of linkage rates across individuals that results from the missing data [17]. Carnegie et. al. consider a subsampling approach to develop such an adjustment [3]. Neither of these papers demonstrate desirable asymptotic properties such as consistency and asymptotic normality.

In this paper, the overall goal is to develop estimators for linkage rates under the assumption that unobserved nodes are missing completely at random (MCAR). First, we show that the bias can, under the MCAR assumption, be represented as a multiplicative factor equal to the probability that we observe a node's edge in the sample. From estimates of this factor, we construct an improved estimator for this multiplicative factor using a subsampling approach. A U-Statistics approach facilitates development of an improved estimator for linkage probabilities that are asymptotically normal. We refer to the proposed estimator as the adjusted estimator. Lastly, we propose a diagnostic approach for assessing the reliability of the method.

We apply these methods to analyses of HIV viral genetic linkage network in Botswana where the data is from a large cluster-randomized trial of a combination HIV prevention intervention - the Botswana Combination Prevention Project (BCPP) [21]. The interest in these analyses is to investigate the patterns of HIV transmission between communities in Botswana.

The paper is organized as follows. Section 2 introduces the notation and setting along with our parameter of interest. Section 3 illustrates the bias of an estimator for θ_{rs} that arises when we do not adjust for incomplete data. Section 4 shows our proposed approach to adjust for incomplete data. Section 5 demonstrates the proposed approach applied to a simulation setting and the HIV viral genetic network from the BCPP. In Section 6, we discuss the overall findings from the proposed approach.

2.2 Notation and Setting

Consider a population of nodes, Ω, of finite size N partitioned into w disjoint groups, $\Omega_1, \ldots, \Omega_w$, with N_r being the number of subjects in group r. Let $\mathbf{y}_{ri}$ denote the ith individual in group r. We consider two nodes to be linked if the pairwise distance between their viral sequences is less than some given threshold. Let the network be represented by $G = (\Omega, E)$ where E is the set of links between individuals, $E \subset \Omega \times \Omega$. Note that G is an undirected network.

Let D_{ri}^{s} be the number of individuals in Ω_s that are linked to $\mathbf{y}_{ri}$ (excluding $\mathbf{y}_{ri}$ if $r = s$). Let $N_{rs} = N_r + N_s$ if $r \neq s$. Otherwise, let $N_{rs} = N_r$. We assume that the nodes and edges in Ω come from some network generating process and that N_{rs} is sufficiently large such that the network structure of Ω is that of its network generating process. Also, we assume that $\frac{\max(D^s_{r1},\ldots,D^s_{rN_r})}{\sqrt{N_{rs}}} \to 0$ as $N_{rs} \to \infty$.

We are interested in inference about the probability that a randomly selected individual in group r links to at least one individual in group s (excluding itself if $r = s$), and we refer to it as the linkage rate. We denote the linkage rate as the following:

$$\theta_{rs} = \Pr(D^s_{ri} \geq 1)$$

where Pr() is defined by the superpopulation of infinite size. In practice, we need to estimate θ_{rs} based on a sample from Ω. Consider a random sample of subjects from the individuals in Ω of size n, which we denote by S_n, such that the proportion of sampled subjects in group r is p_r (known). We denote the sample from Ω_r as $S_{n(r)}$ and the size of $S_{n(r)}$ as n_r. Then $n = \sum_{r=1}^{w} p_r N_r$.

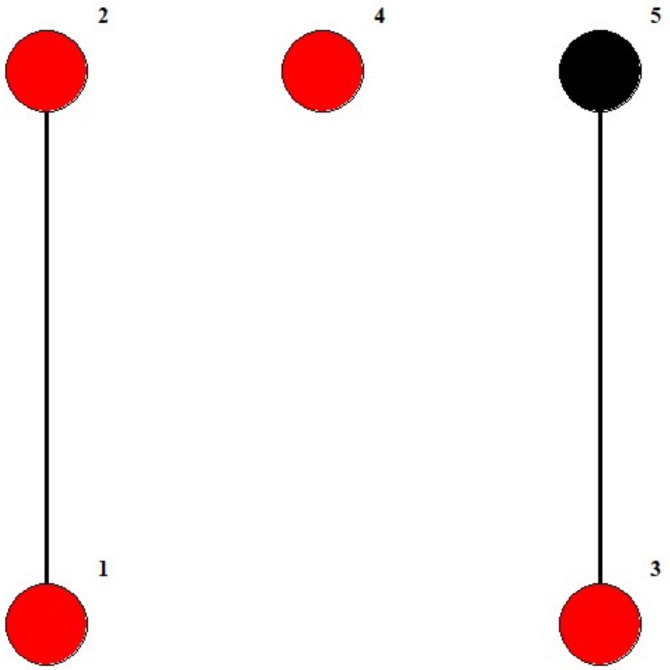
2.3 Bias Arising from Incomplete Data

Let $\widetilde{D}^s_{ri}$ be the number of edges that $\mathbf{y}_{ri}$ has in $S_{n(s)}$ (excluding $\mathbf{y}_{ri}$ if $r = s$). For $\mathbf{y}_{ri} \in S_{n(r)}$, we define

$$u^s_{ri} = I(D^s_{ri} \geq 1)$$
$$v^s_{ri} = I(\widetilde{D}^s_{ri} \geq 1)$$

so u^s_{ri} is the indicator for an edge between $\mathbf{y}_{ri} \in S_{n(r)}$ with at least one node in Ω_s and v^s_{ri} is a "sample version" of u^s_{ri} with respect to $S_{n(s)}$. The differences between u^s_{ri} and v^s_{ri} is shown in Figure 2.1 . Note that $E(u^s_{ri}) = \theta_{rs}$.

Nodes in $S_{n(r)}$, who do not link to any nodes in $S_{n(s)}$, may in fact be linked to nodes in Ω_s but were not observed in $S_{n(s)}$. Thus, $v^s_{ri} \leq u^s_{ri}$ for all $1 \leq i \leq n_r$ and the estimator that ignores the



i	v^r_{ri}	u^r_{ri}
1	1	1
2	1	1
3	0	1
4	0	0

Figure 2.1. Plot of a network to show differences between u^s_{ri} and v^s_{ri}. For simplicity, we consider only a single group, r. Nodes that are colored in red are selected in $S_{n(r)}$. Note that u^s_{ri} and v^s_{ri} are only defined for $\mathbf{y}_{ri} \in S_{n(r)}$.

impact of incomplete data, $\widetilde{\theta}_{rs} = \frac{1}{n_r}\sum_{i=1}^{n_r} v^s_{ri}$, is biased downward (unless all of Ω is sampled):

$$E\left(\widetilde{\theta}_{rs}\right) = \frac{1}{n_r}\sum_{i=1}^{n_r} E\left(v_i^{rs}\right) = E\left(v_i^{rs}\right) \leq E\left(u_i^{rs}\right) = \theta_{rs}. \tag{2.1}$$

We refer to $\widetilde{\theta}_{rs}$ as the unadjusted estimator.

2.4 Methods

As shown in Section 3, the unadjusted estimator for the linkage rate is biased downwards for θ_{rs}. Additionally, even if $\widetilde{\theta}_{rs}$ were an unbiased estimator,the Central Limit Theorem and Law of Large Numbers cannot be directly applied, because the independence assumption is violated. Hence, we use a U-Statistics framework to derive an estimator for θ_{rs} that is asymptotically normal and consistent for θ_{rs}.

2.4.1 An Unbiased Estimator of Probability of Linkage

First, we note that $v_{ri}^{s} = 1$ implies that $u_{ri}^{s} = 1$, because if a link is observed in S_n, then it must exist in Ω. It follows that

$$\begin{aligned} E(\widetilde{\theta}_{rs}) &= \Pr(v_{ri}^{s} = 1) \\ &= \Pr(v_{ri}^{s} = 1, u_{ri}^{s} = 1) \\ &= \Pr(v_{ri}^{s} = 1 \mid u_{ri}^{s} = 1) \Pr(u_{ri}^{s} = 1) \\ &= \pi_{rs}\theta_{rs} \end{aligned}$$

where $\pi_{rs} = \Pr(v_{ri}^{s} = 1 \mid u_{ri}^{s} = 1)$, the probability of observing the link of an individual in S_{n_r} with some individual in $S_{n(s)}$ given that this individual is linked to at least one individual in Ω_s. We have that

$$\begin{aligned} \pi_{rs} &= \Pr\left(v_{ri}^{s} = 1 \mid u_{ri}^{s} = 1\right) \\ &= \frac{\Pr\left(v_{ri}^{s} = 1, u_{ri}^{s} = 1\right)}{\Pr\left(u_{ri}^{s} = 1\right)} \\ &= \frac{\Pr\left(v_{ri}^{s} = 1\right)}{\Pr\left(u_{ri}^{s} = 1\right)} \end{aligned}$$

Therefore, the following is an unbiased estimator for θ_{rs}:

$$\frac{1}{n_r \pi_{rs}} \sum_{i=1}^{n_r} v_{ri}^{s}, \tag{2.2}$$

However, in practice, π_{rs} is unknown, because the event $\{u_i^{rs} = 1\}$ is not observed. Thus, the above is not a feasible estimator for θ_{rs}.

2.4.2 A Feasible and Consistent Estimator for Linkage Rate

Consider a subsample from the individuals in S_n of size $m = \sum_{r=1}^{w} p_r n_r$, which we denote by S_m, such that for each group r we randomly sample a proportion p_r of the individuals in $S_{n(r)}$. We define $m_r = p_r n_r$. We denote the subsample from group r as $S_{m(r)}$. Note that $n_r = p_r N_r$ as well. Thus, in the subsample, we recapitulate the sampling of the observed data from the entire population. Let $\widetilde{\widetilde{D}}^s_{ri}$ be the number of individuals in $S_{m(s)}$ that are linked to $\mathbf{y}_{ri}$ (excluding $\mathbf{y}_{ri}$ is $r = s$). For any $\mathbf{y}_{ri} \in S_{m(r)}$, we define

$$\widetilde{u}^s_{ri} = I(\widetilde{D}^s_{ri} \geq 1)$$
$$\widetilde{v}^s_{ri} = I(\widetilde{\widetilde{D}}^s_{ri} \geq 1)$$

We then denote $\widetilde{\pi}_{rs}$ as the following:

$$\begin{aligned} \widetilde{\pi}_{rs} &= \Pr(\widetilde{v}^s_{ri} = 1 \mid \widetilde{u}^s_{ri} = 1) \\ &= \frac{\Pr(\widetilde{v}^s_{ri} = 1)}{\Pr(\widetilde{u}^s_{ri} = 1)} \end{aligned}$$

We can then estimate $\widetilde{\pi}_{rs}$ by

$$\widehat{\widetilde{\pi}}_{rs} = \frac{\frac{1}{m_r} \sum_{i=1}^{m_r} \widetilde{v}^s_{ri}}{\frac{1}{m_r} \sum_{i=1}^{m_r} \widetilde{u}^s_{ri}}$$

which is well-defined based on S_n as $\widetilde{v}^s_{ri}$ and $\widetilde{u}^s_{ri}$ are observed.

We then want to show that as $N_r, N_s \to \infty$,

$$\widetilde{\pi}_{rs} \to \pi_{rs}$$

so that $\widehat{\widetilde{\pi}}_{rs}$ is a consistent estimator.

Theorem 1 *Suppose* $\frac{max(D^s_{r1},\ldots,D^s_{rN_r})}{\sqrt{N_{rs}}} \to 0$ *as* $N_{rs} \to \infty$*. We note that* p_r *and* p_s *are fixed so* $n_r \to \infty$ *and* $n_s \to \infty$ *as* $N_r \to \infty$ *and* $N_s \to \infty$*, respectively. Suppose also that* $\Pr(\widetilde{D}_{ri} = k \mid \widetilde{D}_{ri} \geq 1) \to \Pr(D^s_{ri} = k \mid D^s_{ri} \geq 1)$ *as* $N_r, N_s \to \infty$*. Then*

$$\widetilde{\pi}_{rs} = P(\widetilde{v}^s_{ri} = 1 \mid \widetilde{u}^s_{ri} = 1) \to P(v^s_{ri} = 1 \mid u^s_{ri} = 1) = \pi_{rs}, \quad \text{as } N_r, N_s \to \infty.$$

The proof of Theorem 1 is provided in the appendix. The first assumption is made in Section 2. Due to the assumption in Theorem 1 that $\Pr(\widetilde{D}_{ri} = k \mid \widetilde{D}_{ri} \geq 1) \to \Pr(D^s_{ri} = k \mid D^s_{ri} \geq 1)$ as $N_r, N_s \to \infty$, consistency requires further assumptions on the structure of our network. Hence, we assume that the degree distribution for linked individuals in the superpopulation follow a power-law distribution. For such distributions, there exists some k_0 such that for $k \geq k_0$ we have

$$\Pr(D^s_{rs} = k \mid D^s_{rs} \geq 1) = \beta k^{-\alpha}$$

where $2 \leq \alpha \leq 3$. Networks whose degree distributions follow a power law are referred to as scale-free networks. Several investigators have notes that HIV genetic linkage networks appear to have this property [14, 27, 28]. Although Stumpf et. al. showed that this assumption will not hold theoretically with networks of power law distributions, we show in Section 4.4 that this assumption approximately holds for large enough values of p_s resulting in consistent estimators. Theorem 1 shows that although $\{u^s_{ri} = 1\}$ is not observed, we can develop a subsample S_m of S_n and estimate π_{rs} by treating S_n as Ω and S_m as S_n.

Since $\widehat{\widetilde{\pi}}_{rs}$ applies only to a single subsample, the estimator can depend heavily on the specific subsample that was selected. Hence, we propose the following estimator for π_{rs} :

$$\widehat{\pi}_{rs} = \frac{\binom{n_r}{m_r}^{-1}\binom{n_s}{m_s}^{-1}\sum_{S_{m(r)}\in C_{m(r)}^{n(r)}}\sum_{S_{m(s)}\in C_{m(s)}^{n(s)}}\frac{1}{m_r}\sum_{i=1}^{m_r}\widetilde{v}_{ri}^{s}}{\binom{n_r}{m_r}^{-1}\binom{n_s}{m_s}^{-1}\sum_{S_{m(r)}\in C_{m(r)}^{n(r)}}\sum_{S_{m(s)}\in C_{m(s)}^{n(s)}}\frac{1}{m_r}\sum_{i=1}^{m_r}\widetilde{u}_{ri}^{s}},$$

where $C_{m(j)}^{n(j)}$ is the set of all possible combinations from sampling m_j individuals from $S_{n(j)}$. With such an estimator $\widehat{\pi}_{rs}$, we can consider a feasible estimator of θ_{rs} as:

$$\widehat{\theta}_{rs} = \frac{1}{n_r\widehat{\pi}_{rs}}\sum_{i=1}^{n_r} v_{ri}^{s}, \tag{2.3}$$

We refer to $\widehat{\theta}_{rs}$ as the adjusted estimator for θ_{rs}.

To establish consistency and asymptotic normality of the estimate in (2.3), standard asymptotic methods such as the law of large numbers and central limit theorem cannot be directly applied. This is because $\widetilde{u}_{ri}^{s}, v_{ri}^{s}$ and $\widetilde{v}_{ri}^{s}$ are not stochastically independent, thereby violating the required independence assumption. Below in Section 4.3, we describe an approach to establish such properties.

2.4.3 Inference on Linkage Rate: A U-Statistics Framework

First, we let

$$\gamma_{rs1} = \Pr\left(\widetilde{v}_{ri}^{s} = 1\right)$$
$$\gamma_{rs2} = \Pr\left(\widetilde{u}_{ri}^{s} = 1\right)$$
$$\gamma_{rs3} = \Pr\left(v_{ri}^{s} = 1\right).$$

Now, we denote γ_{rs} as the following:

$$\gamma_{rs} = \begin{pmatrix} \gamma_{rs1} \\ \gamma_{rs2} \\ \gamma_{rs3} \end{pmatrix}.$$

Then $\frac{\gamma_{rs1}\gamma_{rs3}}{\gamma_{rs2}} \to \theta_{rs}$ as $N_{rs} \to \infty$ and by Theorem 1, $\widetilde{\pi}_{rs} = \frac{\gamma_{rs1}}{\gamma_{rs2}}$. From Section 4.1 and 4.2, we propose the following estimator for γ:

$$\begin{aligned} \widehat{\gamma}_{rs}\left(\mathbf{y}_{r1}, \mathbf{y}_{r2}, \dots, \mathbf{y}_{rn_r}; \mathbf{y}_{s1}, \mathbf{y}_{s2}, \dots, \mathbf{y}_{sn_s}\right) &= \begin{pmatrix} \widehat{\gamma}_1 \\ \widehat{\gamma}_2 \\ \widehat{\gamma}_3 \end{pmatrix} \\ &= \begin{pmatrix} \binom{n_r}{m_r}^{-1}\binom{n_s}{m_s}^{-1} \sum_{S_{m(r)} \in C^{n(r)}_{m(r)}} \sum_{S_{m(s)} \in C^{n(s)}_{m(s)}} \frac{1}{m_r} \sum_{i=1}^{m_r} \widetilde{v}^s_{ri} \\ \binom{n_r}{m_r}^{-1}\binom{n_s}{m_s}^{-1} \sum_{S_{m(r)} \in C^{n(r)}_{m(r)}} \sum_{S_{m(s)} \in C^{n(s)}_{m(s)}} \frac{1}{m_r} \sum_{i=1}^{m_r} \widetilde{u}^s_{ri} \\ \frac{1}{n_r} \sum_{i=1}^{n_r} v^s_{ri} \end{pmatrix}. \end{aligned}$$

Thus, provided that we can establish consistency and asymptotic normality of $\widehat{\gamma}_{rs}$, by the Delta Method and Theorem 1, $\widehat{\theta}_{rs}$ is consistent and asymptotically normal. To this end, we adopt the following U-Statistics framework.

First, we note that

$$E(\widehat{\gamma}_{rs}) = \gamma_{rs}$$

and that the arguments of $\widehat{\gamma}_{rs}$ are invariant to permutations of individuals within each group, i.e.,

$$\widehat{\gamma}_{rs}\left(\mathbf{y}_{r1'}, \mathbf{y}_{r2'}, \dots, \mathbf{y}_{rn'_r}; \mathbf{y}_{s1''}, \mathbf{y}_{s2''}, \dots, \mathbf{y}_{sn''_s}\right) = \widehat{\gamma}_{rs}\left(\mathbf{y}_{r1}, \mathbf{y}_{r2}, \dots, \mathbf{y}_{rn_r}; \mathbf{y}_{s1}, \mathbf{y}_{s2}, \dots, \mathbf{y}_{sn_s}\right)$$

where $(1',2',\ldots,n_r')$ and $(1'',2'',\ldots,n_s'')$ are any permutations of $(1,2,\ldots,n_r)$ and $(1,2,\ldots,n_s)$, respectively. Thus, by [13], $\widehat{\gamma}_{rs}$ is a multivariate U-Statistic. Let

$$\begin{aligned} \mathbf{h}_{rs1}\left(\mathbf{y}_{ki}\right) = E\left(\widehat{\gamma}_{rs}\left(\mathbf{y}_{r1}\ldots,\mathbf{y}_{rn_r};\mathbf{y}_{s1}\ldots,\mathbf{y}_{sn_s}\right) \mid \mathbf{y}_{ki}\right), \quad \widetilde{\mathbf{h}}_{rs1}\left(\mathbf{y}_{ki}\right) = \mathbf{h}_{rs1}\left(\mathbf{y}_{ki}\right) - \gamma_{rs}, \\ \Sigma_k = \mathrm{Var}\left[\widetilde{\mathbf{h}}_{rs1}\left(\mathbf{y}_{ki}\right)\right] = E\left[\widetilde{\mathbf{h}}_{rs1}\left(\mathbf{y}_{ki}\right)\widetilde{\mathbf{h}}_{rs1}^{\top}\left(\mathbf{y}_{ki}\right)\right]. \end{aligned}$$

By [13, 24, 26], it follows that

$$\sqrt{n_{rs}}\left(\widehat{\gamma}_{rs} - \gamma_{rs}\right) \to_d N\left(\mathbf{0},\Sigma_{\gamma(rs)}\right).$$

where

$$n_{rs} = \begin{cases} n_r & r = s \\ n_r + n_s & r \neq s \end{cases}.$$

$$\Sigma_{\gamma(rs)} = \frac{N_r - n_r}{N_s}\rho_r^2 n_r^2 \Sigma_r + \frac{N_s - n_s}{N_s}\rho_s^2 n_s^2 \Sigma_s$$

and

$$\rho_k^2 = \lim_{n_{rs}\to\infty} \frac{n_{rs}}{n_k}.$$

A consistent estimate of $\Sigma_{\gamma_{(rs)}}$ is given by:

$$\begin{aligned} \widehat{\Sigma}_{\gamma(rs)} &= \frac{N_r - n_r}{N_r}\frac{n_{rs}}{n_r}n_r^2\widehat{\Sigma}_r + \frac{N_s - n_s}{N_s}\frac{n_{rs}}{n_r}n_s^2\widehat{\Sigma}_s = n_{rs}\left(\frac{N_r - n_r}{N_r}n_r\widehat{\Sigma}_r + \frac{N_s - n_s}{N_s}n_s\widehat{\Sigma}_s\right), \\ \widehat{\Sigma}_k &= \frac{1}{n_k - 1}\sum_{i=1}^{n_k}(\widehat{\mathbf{h}}_{rs1}(\mathbf{y}_{ki}) - \widehat{\gamma}_{rs})(\widehat{\mathbf{h}}_{rs1}(\mathbf{y}_{ki}) - \widehat{\gamma}_{rs})^T \end{aligned}$$

where $\widehat{\mathbf{h}}_{rs1}(\mathbf{y}_{ki})$ is a consistent estimate of $\mathbf{h}_{rs1}\left(\mathbf{y}_{ki}\right)$. In the appendix, it is shown that $\widehat{\mathbf{h}}_{rs1}(\mathbf{y}_{\mathbf{ri}})$

defined as follows is consistent:

$$\widehat{\mathbf{h}}_{rs1}(\mathbf{y}_{ri}) = \begin{pmatrix} \frac{1}{n_r}\binom{n_r-1}{m_r-1}^{-1}\sum_{S_{m(r)}\in C_{m(r)}^{n(r)}}\binom{n_s}{m_s}^{-1}\sum_{S_{m(s)}\in C_{m(s)}^{n(s)}}\widetilde{v}_{ri}^{s}+\frac{n_r-1}{n_r}\widehat{\gamma}_{rs1} \\ \frac{\widetilde{u}_{ri}^{s}}{n_r}+\frac{n_r-1}{n_r}\widehat{\gamma}_{rs2} \\ \frac{\widetilde{u}_{ri}^{s}}{n_r}+\frac{n_r-1}{n_r}\widehat{\gamma}_{rs3} \end{pmatrix}$$

and

$$\widehat{\mathbf{h}}_{rs1}(\mathbf{y}_{si}) = \begin{pmatrix} \widehat{\gamma}_{rs1} \\ \widehat{\gamma}_{rs2} \\ \widehat{\gamma}_{rs3} \end{pmatrix}$$

Thus, $\widehat{\Sigma}_s = 0$, which implies

$$\widehat{\Sigma}_{\gamma(rs)} = n_{rs}n_r\widehat{\Sigma}_r.$$

Theorem 2 *Let*

$$\mathbf{h}_{rs1}(\mathbf{y}_{ki}) = E\left(\widehat{\gamma}_{rs}\left(\mathbf{y}_{r1}\ldots,\mathbf{y}_{rn_r};\mathbf{y}_{s1}\ldots,\mathbf{y}_{sn_s}\right)\mid\mathbf{y}_{ki}\right), \quad \widetilde{\mathbf{h}}_{rs}(\mathbf{y}_{ki}) = \mathbf{h}_{rs1}(\mathbf{y}_{ki})-\gamma_{rs},$$

$$\Sigma_k = Var\left[\widetilde{\mathbf{h}}_{rs}(\mathbf{y}_{ki})\right] = E\left[\widetilde{\mathbf{h}}_{rs}(\mathbf{y}_{ki})\widetilde{\mathbf{h}}_{rs}^{\top}(\mathbf{y}_{ki})\right],$$

$$n_{rs} = \begin{cases} n_r & \textit{if } r=s \\ n_r+n_s & \textit{if } r\neq s \end{cases}, \quad \rho_k^2 = \lim_{n_{rs}\to\infty}\frac{n_{rs}}{n_k}<\infty, \quad k=r,s$$

Then, we have: (1) $\widehat{\gamma}_{rs}$ is a consistent, unbiased and asymptotically normal estimator of γ_{rs}:

$$\sqrt{n_{rs}}\left(\widehat{\gamma}_{rs}-\gamma_{rs}\right)\to_d N\left(\mathbf{0},\Sigma_{\gamma(rs)} = \frac{N_r-n_r}{N_r}\rho_r^2n_r^2\Sigma_r+\frac{N_s-n_s}{N_s}\rho_s^2n_s^2\Sigma_s\right),$$

(2) A consistent estimator of the asymptotic variance is given by:

$$
\begin{aligned}
\widehat{\Sigma}_{\gamma(rs)} &= \frac{N_r - n_r}{N_r}\frac{n_{rs}}{n_r} n_r^2 \widehat{\Sigma}_r + \frac{N_s - n_s}{N_s}\frac{n_{rs}}{n_r} n_s^2 \widehat{\Sigma}_s = n_{rs}\left(\frac{N_r - n_r}{N_r} n_r \widehat{\Sigma}_r + \frac{N_s - n_s}{N_s} n_s \widehat{\Sigma}_s\right), \\
\widehat{\Sigma}_k &= \frac{1}{n_k - 1}\sum_{i=1}^{n_k}\left(\widehat{\mathbf{h}}_{rs1}\left(\mathbf{y}_{ki}\right) - \widehat{\gamma}_{rs}\right)\left(\widehat{\mathbf{h}}_{rs1}\left(\mathbf{y}_{ki}\right) - \widehat{\gamma}_{rs}\right)^{\top}, \\
\widehat{\mathbf{h}}_{rs1}\left(\mathbf{y}_{ri}\right) &= \begin{pmatrix} \frac{1}{n_r}\binom{n_r-1}{m_r-1}^{-1}\sum_{S_{m(r)}\in C_{m(r)}^{n(r)}}\binom{n_s}{m_s}^{-1}\sum_{S_{m(s)}\in C_{m(s)}^{n(s)}}\widetilde{v}_{ri}^{s} + \frac{n_r-1}{n_r}\widehat{\gamma}_{rs1} \\ \frac{\widetilde{u}_{ri}^{s}}{n_r} + \frac{n_r-1}{n_r}\widehat{\gamma}_{rs2} \\ \frac{\widetilde{u}_{ri}^{s}}{n_r} + \frac{n_r-1}{n_r}\widehat{\gamma}_{rs3} \end{pmatrix}, \\
\widehat{\mathbf{h}}_{rs1}(\mathbf{y}_{si}) &= \begin{pmatrix} \widehat{\gamma}_{rs1} \\ \widehat{\gamma}_{rs2} \\ \widehat{\gamma}_{rs3} \end{pmatrix}.
\end{aligned}
$$

With the asymptotic results in Theorem 2, we can readily obtain the consistency and asymptotic normality by the Delta method. Let

$$
f\left(\gamma_{rs}\right) = \frac{\gamma_{rs1}\gamma_{rs3}}{\gamma_{rs2}}.
$$

Then, $\widehat{\theta}_{rs} = f\left(\widehat{\gamma}_{rs}\right)$. By the Delta method and Theorem 2,

$$
\sqrt{n_{rs}}\left(\widehat{\theta}_{rs} - \theta_{rs}\right) \rightarrow_d N\left(\mathbf{0}, \sigma_{\theta(rs)}^2 = \phi^{\top}(\gamma_{rs})\Sigma_{\gamma(rs)}\phi(\gamma_{rs})\right),
$$

where

$$
\phi(\gamma_{rs}) = \frac{\partial}{\partial\gamma_{rs}} f\left(\gamma_{rs}\right) = \begin{pmatrix} \frac{\partial f(\gamma_{rs})}{\partial\gamma_{rs2}} \\ \frac{\partial f(\gamma_{rs})}{\partial\gamma_{rs1}} \\ \frac{\partial f(\gamma_{rs})}{\partial\gamma_{rs3}} \end{pmatrix} = \begin{pmatrix} -\frac{\gamma_{rs1}\gamma_{rs3}}{\gamma_{rs2}^2} \\ \frac{\gamma_{rs3}}{\gamma_{rs2}} \\ \frac{\gamma_{rs1}}{\gamma_{rs2}} \end{pmatrix}.
$$

A consistent estimator of $\sigma_{\theta(rs)}^2$ is given by:

$$
\widehat{\sigma}_{\theta(rs)}^2 = \phi^{\top}\left(\widehat{\gamma}_{rs}\right)\widehat{\Sigma}_{\gamma_{rs}}\phi(\widehat{\gamma}_{rs}),
$$

where $\phi^\top(\widehat{\gamma}_{rs})$ and $\widehat{\Sigma}_{\gamma_{rs}}$ denote the respective quantities by substituting $\widehat{\gamma}_{rs}$ in place of γ_{rs}.

2.4.4 Diagnostic

The consistency of the proposed estimator depends on the assumption that $\Pr(\widetilde{D}^s_{ri} = k \mid \widetilde{D}^s_{ri}) \approx \Pr(D^s_{ri} = k \mid D^s_{ri})$. Stumpf et. al. performed a simulation with sampling from scale-free networks [25]. They showed that the larger the value of α, the greater the sample degree distribution deviates from the true distribution. Since for many settings $2 \le \alpha \le 3$ [25], we apply our approach to a scale free network with $\alpha = 3$ to evaluate values of p such that

$$\Pr(\widetilde{D}^s_{ri} = k \mid \widetilde{D}^s_{ri}) \approx \Pr(D^s_{ri} = k \mid D^s_{ri})$$

From Figure 2.2, we find that the estimates deviate greatly from the true value when $p < .40$.

2.5 Applications

2.5.1 Simulation Study

We apply the proposed methods to a population with two communities. We denote them both as community 1 and 2. We let $N_1 = 1000$ and $N_2 = 1200$. For the degree distributions, we have

$$\Pr(D^r_{ri} = k) = \begin{cases} ak^{-2.5} & k \ge 1 \\ 0.50 & k = 0 \end{cases}$$

$$\Pr(D^s_{ri} = k) = \begin{cases} \beta k^{-2.6} & k \ge 1 \\ 0.60 & k = 0 \end{cases}$$

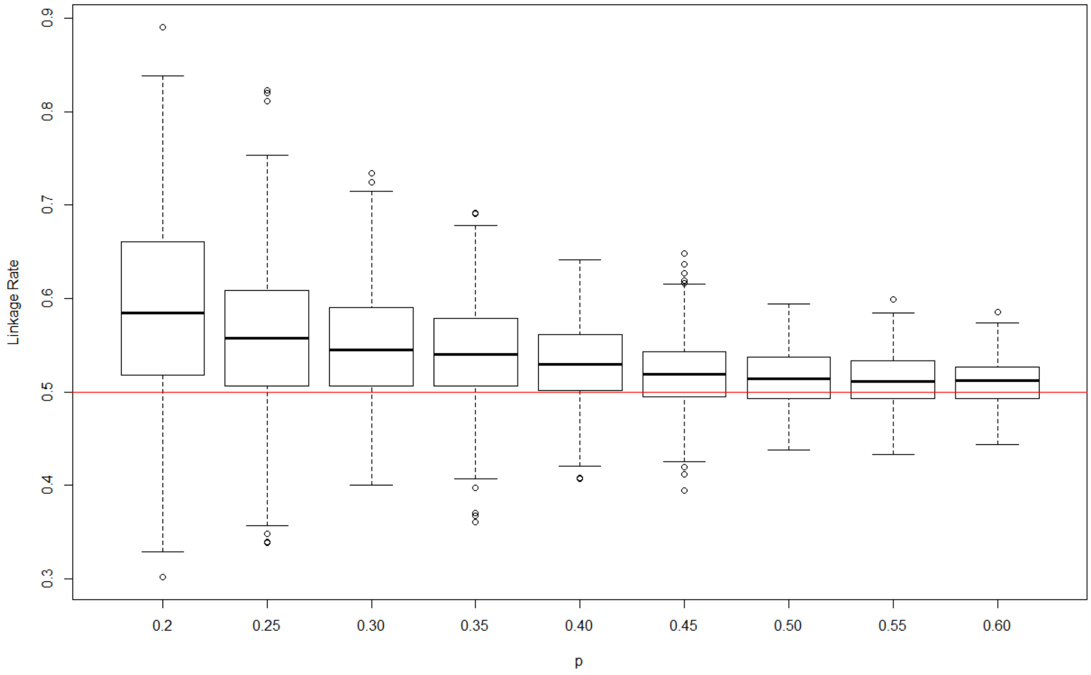


Figure 2.2. Distribution of adjusted estimators for various values of p when applying the proposed approach to a scale-free network with $\alpha = 3$. Here, the x-axis represents p and the y-axis represents values for the adjusted estimators where the red line indicates the true value.

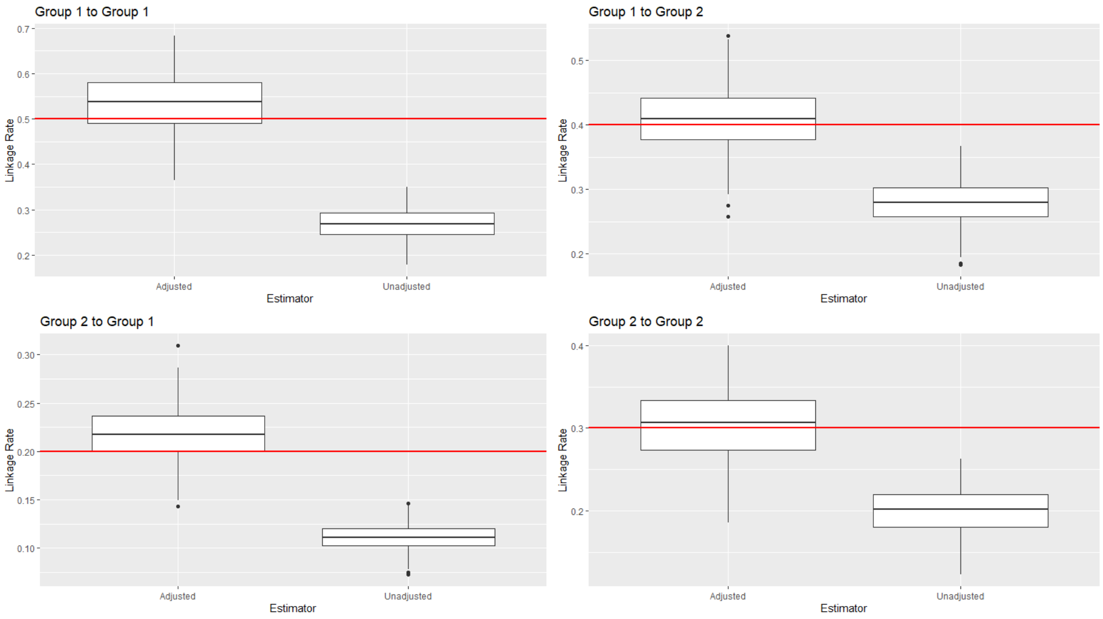


Figure 2.3. Distribution of adjusted and unadjusted estimators for linkage rates from the simulation in Section 5.1. The red horizontal line is the true value.

$$\Pr(D^r_{si} = k) = \begin{cases} \beta k^{-2.3} & k \geq 1 \\ 0.80 & k = 0 \end{cases}$$

$$\Pr(D^s_{si} = k) = \begin{cases} \beta k^{-3} & k \geq 1 \\ 0.70 & k = 0 \end{cases}$$

We have that

$$\theta = \begin{bmatrix} 0.50 & 0.40 \\ 0.20 & 0.30 \end{bmatrix}$$

We apply the proposed approach with letting $p_1 = 0.40$ and $p_2 = 0.60$.

Figure 2.3 demonstrates that the adjusted estimators are considerably less biases than are the unadjusted estimators. As expected, the sampling fraction of the group impacts the performance of the adjusted estimator. Figure 2.3 confirms the expectation from Theorem 1 that when estimating the probability of linkage from Group ”A” to Group ”B”, sampling additional nodes from Group ”B” rather than Group ”A” contributes more to improved performance. Most of the biases observed in the adjusted estimators is upwards.

Table 2.1 shows that the coverage probabilities for the between-community estimates are much greater than the estimates for within communities.

Table 2.1. Coverage probabilities of adjusted estimators from simulation in Section 5.1.

	Group 1	Group 2
Group 1	0.68	0.74
Group 2	0.64	0.47

2.5.2 Botswana Combination Prevention Project

As discussed earlier, the intent of developing this approach was to estimate linkage rates for the HIV viral genetic linkage networks Botswana. The data used comes from a large cluster-randomized trial of a combination HIV prevention intervention - the Botswana Combination Prevention Project (BCPP).

In the BCPP, all households were targeted for a survey in 6 of the 30 participating communities in Botswana. The communities that were selected are Gumare, Mauntalala, Mmankgodi, Mmathethe, Ramokgonami and Shakawe. For those that choose to participate in the survey, demographic and household data along with HIV status is ascertained. For those who are HIV+, the viral genetic sequences are obtained. Hence, missing data arises due to the fact we have individuals who choose not to participate in the BCPP. However, going forward we assume that all individuals in the BCPP are MCAR, but acknowledge the fact that this assumption may not entirely hold.

It follows that we define an edge between two individuals to exist if and only if the pairwise distance between their viral genetic sequences is below some threshold, c. Following Novitstky et. al, we set $c = 0.07$.

Table 2.2 provides the proportions of HIV+ in individuals that participated in the BCPP;

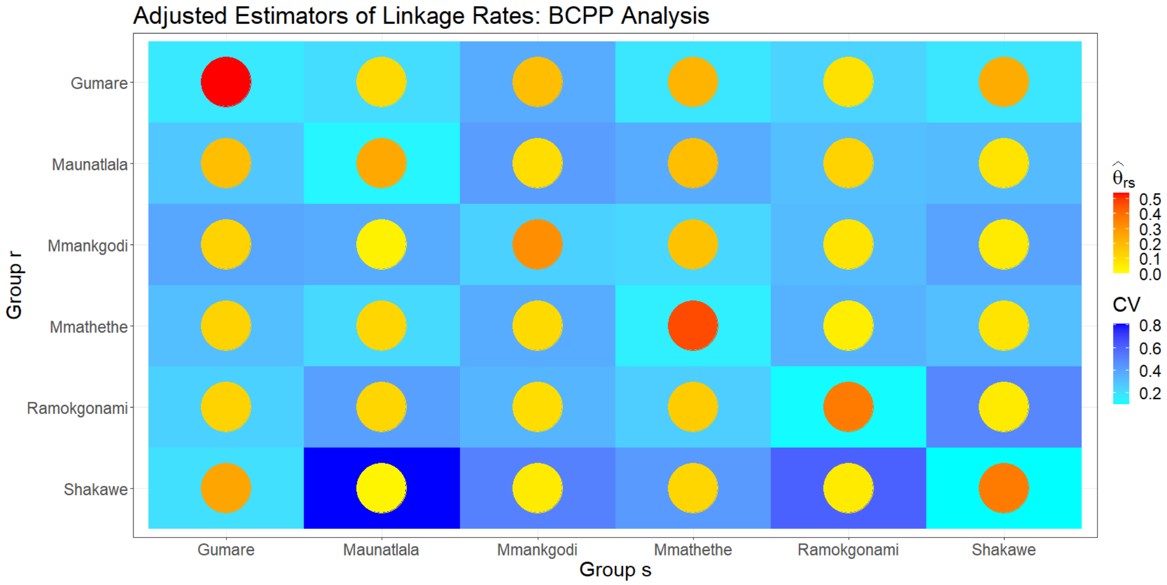


Figure 2.4. Adjusted estimators for linkage rates between the communities from the BCPP and coefficients of variation (CV) indicated by colors of circle and cell, respectively.

for 4 of the 6 communities, the proportions were over 40% but for 2, they were below 30%.

Table 2.2. The proportion (p) and number (n) of HIV+ individuals in each community that participated in the BCPP.

	Gumare	Maunatlala	Mmankgodi	Mmathethe	Ramokgonami	Shakawe
p	0.29	0.52	0.26	0.44	0.48	0.48
n	325	363	270	336	350	484

We applied our methods to adjust for the incompleteness of the same. Figure 2.4 provides a heat map of the intensity of linkage after adjustment for missing data, within and across the communities as well as the variability associated with these estimates of linkage. For within community analyses, Gumare and Mmathethe have the highest linkage rates. Across communities, the high linkage rates are between Shakawe and Gumare in both directions.

2.6 Discussion

Viral genetic linkage analysis can play an important role in molecular epidemiology in that it can reveal features of transmission patterns within and across communities. This knowledge can aid in control of outbreaks and by helping direct resources where they might be most effectively deployed. Because observed transmission networks are rarely complete, and because VGL analyses are particularly sensitive to the impact of missing data—even when observations are MCAR within communities under study—adjustment for incompleteness of samples is essential.

While methods have been proposed for such analyses, this paper is the first to ground such methods in statistical theory. Through the use of the U-statistics framework, we were able to show consistency of our estimator under assumptions about the nature of the VGL network that are consistent with available literature and also to prove asymptotic normality, thereby permitting development of confidence interval estimates. We demonstrate that the methods work well when sampling proportion is greater than 0.4; but even in a setting with lower sampling rates, the adjustment greatly improve performance of estimators compared to those that ignore missing data. How to make further improvements to our estimator when applied to data with low sampling rates is a topic for further research.

Our illustrative example made use of data from the HIV prevention study in Botswana — the BCPP. We demonstrated that VGL linkage across communities is common—which implies that a treatment-as-prevention intervention applied at the village level will likely have effects on HIV incidence that are attenuated compared to effects that would occur if all relationships took place within villages. Furthermore such estimates would also be attenuated compared to an estimand of interest—the counterfactual expected difference in incidence between a setting in which the intervention was implemented in all villages and a setting in which it in none. Hence these VGL analyses are useful in both design and interpretation of cluster randomized trials for control of endemic diseases or disease outbreaks.

We note that our methods would apply not only to VGL or transmission networks but to networks of all types, for which sampling of nodes is not complete. Our focus was on scale-free network; future work is required for extension to networks more generally.

2.7 Acknowledgements

Chapter 2, in full, is a reprint of the materials as it appears in Estimating Viral Genetic Linkage Rates in the Presence of Missing Data 2022. Victor De Gruttola, Tu, Xin, Vladimir Novitsky, Jingjing Zou, Tuo Lin. The dissertation author was the primary investigator and author of this paper.

Chapter 3

Estimating Probabilities of Linkage in the Presence of Missing Data

3.1 Introduction

Molecular epidemiology is increasingly used to investigate patterns of HIV transmission, epidemic dynamics; in addition, both the CDC and NIH have proposed that such analyses be used to guide resources intended to end the AIDS epidemic [8]. An important feature of such analyses is investigation of HIV genetic linkage; such linkage can be based on the genetic distance between genetic sequences taken from pairs of individuals from whom HIV transmission may have occurred. Such analyses can reveal which viral strains are propagating within and between communities, the characteristics of people infected with such strains, and the effects of interventions designed to control HIV on the rates at which viral genetic clusters grow. However, in the presence of potentially informatively missing data, observed viral genetic linkage networks do not represent the true underlying networks in populations under study, rendering inferences based on sampled networks unreliable [5]. Specifically, estimates of probabilities of linkage, specifically defined the probability of linkage between viral genetic sequences from two individuals selected at random from their respective groups, that ignore the impact of missing data (henceforth referred to as unadjusted estimators) will be biased.

Carnegie et.al. provided consistent estimates of probabilities of linkage under the assumption that viral genetic sequences were missing at random (MAR) given group membership.

However, they did not demonstrate asymptotic normality for this estimator. It follows in our previous work, under the assumption that viral sequences were missing completely at random, we developed an unbiased estimator through a subsampling approach and demonstrated consistency and asymptotic normality using a U-Statistics framework. However, in our previous work, demonstrating consistency required strict conditions. Specifically the network generating process of the complete network had to be known and the degree distribution for the complete network would be approximately the same as that of the sampled network (which seemed to be feasible when the sampling proportion was at least 0.40. In this paper, we propose a more flexible approach that allows data to be MAR given continuous covariates and yields a consistent and asymptotically normal estimator without the strict conditions required in our previous work.

We consider linkage to occur between two individuals if the pairwise genetic distance between their viral genetic sequences is less than some threshold. Obtaining asymptotic properties for estimators of probabilities of linkage — which are informative regarding linkage rates–is challenging, because indicators of linkage across pairs of individuals are between-, rather than, within-subject attributes in conventional statistical analyses, and as such standard asymptotic methods such as the central limit theorem and law of large numbers cannot be directly applied to these estimators [15, 16]. In this paper, we develop estimators for probabilities of linkage under the assumption that unobserved viral genetic sequences are missing at random (MAR) and derive asymptotic properties for these estimators.

The choice of the threshold indicating linkage is an important scientific question in the analysis of viral genetic data. In general it may be best to investigate the sensitivity of findings, but the methods developed here apply regardless of the threshold value.

We apply the proposed methods to analyze HIV sequences from the Botswana Combination Prevention Project (BCPP), which has motivated the development of the proposed approach, but we note that our methods apply in any setting wherein nodes are sampled from networks. We demonstrate that the methods can be applied to networks more generally.

3.2 Botswana Combination Prevention Project (BCPP)

The BCPP was a large cluster-randomized trial of a combination HIV prevention intervention compared to standard of care in 30 villages in Botswana. In this section we review the sampling design of the BCPP along with the layers of missingness in this study.

3.2.1 Study Introduction

At baseline, 20% of the households in each community in Botswana were targeted for participation in a baseline household survey, which collected demographic and household data among those household members willing to participate. For those unwilling to participate, such demographic and house data were generally provided by heads of households. All participants were tested for HIV infection and virus from blood samples were sequenced for all HIV+ participants; the remaining participants who were HIV- form the incidence cohort. For the next two years the HIV incidence cohort was annually tested for HIV; once again, all virus from those participants who became HIV+ was sequenced. At the end of the BCPP, six communities were selected to participate in a survey of all households, denoted the End of Survey Study (ESS). Because of the inclusiveness of this survey, we illustrate our methods using data from ESS villages.

As our research question focuses on viral linkage without regard to timing of infection—in other words on a static VGL network– we do not consider time as variable in our models. Dynamic VGL models have been described but require information about time of infection , which is generally not available in our study population.

3.2.2 Missing Data

In BCPP, we have two layers of missingness in the observed viral genetic linkage (VGL) network data. First, HIV status is unknown for non-participating household members. Note that unlike common survey studies, demographic data for non-participating household members are also observed (obtained through head of household), provided that the head of the household participated in the BCPP. Second, genetic sequences are unavailable for those who were not tested; hence they are available for only a subsample of those who tested positive; hence they are available for only a subsample of those who were HIV+. Third, we do not have any observed data on households that did not participate in the ESS. In this paper we will only be addressing the first two layers of missingness and hence, assume that our population of interest to consist of only individuals from participating households.

A common approach for addressing non-response in survey studies is to model this missingness probability, or propensity score, using all observed participants' information such as demographic and HIV status in the current study and then use the inverse of the propensity as propensity score weights, in addition to weights due to multi-stage sampling frames if applicable, to construct consistent population-level estimators under the missing at random (MAR) mechanism [26]. Because the second layer of missingness causes all genetic links to be missing for those who were HIV+ but never tested, this usual approach cannot be applied to address non-response for BCPP. We propose instead to prior estimates of HIV prevalence in Botswana to address this missing not at random (MNAR) mechanisms in the current BCPP. We consider this analysis in two steps: 1) to address the missing HIV status of non-participants, and 2) to address the missingness of links among HIV+ nonparticipants and between them and others who might have been linked to them.

3.3 Notation and Setting

Consider a population of individuals, Ω, of finite size N partitioned into w disjoint groups, $\Omega_1, \ldots, \Omega_w$, with N_r being the number of individuals in group r. Let $\mathbf{y}_{ri}$ be a vector denoting the viral sequence of the ith individual in group r. We consider two individuals to be linked if the pairwise distance between their viral genetic sequences is less than some given threshold, c. Note that all individuals in the VGL network are HIV+ (HIV- individuals cannot have a viral strain). Let the VGL network be represented by $G = (\Omega, E)$ where E is the set of links between individuals, $E \subset \Omega \times \Omega$. As the direction of transmission is unknown, G is an undirected network. We assume that $N_1, \dots, N_w$ are known.

Let D_{ij}^{rs} be the pairwise distance between individuals i and j from groups r and s, respectively. Note that we have $D_{ij}^{rs} = D_{ji}^{sr}$ under the assumptions. Our focus is on inference about the probability of linkage between viral sequences from two individuals selected at random from their respective groups. We denote the probability of linkage as the following:

$$\gamma_{rs} = \Pr(D_{ij}^{rs} \leq c)$$

where $\Pr()$ is defined with respect to the underlying superpopulation [26]. Note that $\gamma_{rs} = \gamma_{sr}$.

In practice, we estimate γ_{rs} based on a sample from Ω. Consider a sample of individuals from Ω of size n, such that the same household sampling is performed as in the BCPP. We denote this sample by S_n. We denote the sample from Ω_r as $S_{n(r)}$ and the size of $S_{n(r)}$ as n_r. Then $n = \sum_{r=1}^{w} n_r$.

3.4 Methodology

As described above, obtaining an estimator for γ_{rs} is challenging, because of 1) the missing HIV statuses of non-participants and 2) the informative missingness of the BCPP network data. To accommodate the first layer of missingness, we apply a multiple imputation approach. It follows that to address the second layer of missingness, we model and estimate the missingness probability using logistic regression [13].

As with any estimator, we hope to derive desirable asymptotic properties for it. Since indicators for linkage are dependent, to derive asymptotic properties for such an estimator, Central Limit Theorem and Law of Large Number cannot be directly applied. Hence, we use a U-Statistics based weighted generalized estimating equations to derive an estimator for γ_{rs} that is asymptotically normal and consistent for γ_{rs} [13].

3.4.1 Estimating Probability of Linkage

We denote the indicator linkage in the sample as

$$X_{ij}^{rs} = I(D_{ij}^{rs} \leq c)$$

so

$$\gamma_{rs} = E(X_{ij}) = \Pr(D_{ij}^{rs} \leq c)$$

where $E()$ is again defined with respect to the superpopulation. Let

$$z_{ri} = I(\mathbf{y}_{ri} \in S_{n(r)})$$
$$z_{sj} = I(\mathbf{y}_{sj} \in S_{n(s)}).$$

If non-response were MCAR, then by the Theory of U-Statistics [13] the following would be a consistent estimate for γ_{rs}:

$$\frac{1}{n_{rs}} \sum_{\mathbf{y}_{ri} \in \Omega_r} \sum_{\mathbf{y}_{sj} \in \Omega_s} X_{ij} z_{ri} z_{sj}$$

where $n_{rs} = n_r n_s$.

However, in BCPP, non-response may not arise from the missing-completely-at-random (MCAR) mechanism; the probability of non-participation is likely to depend on observed demographic variables and HIV status. (A low participation rate among young males was observed in the BCPP as a whole.) To accommodate for such selection bias, we model participation probability as:

$$w_{ij} = \Pr(\mathbf{y}_{ri}, \mathbf{y}_{sj} \in S_n \mid C_{ri}, C_{sj}, \text{ HIV+})$$

where C_{ri} and C_{sj} are a vector of covariates for $\mathbf{y}_{ri}$ and $\mathbf{y}_{sj}$, respectively. Note that we define w_{ij} by conditioning on the HIV+ subsample, as the linkage indicator can only ascertained among these participants. We assume that when conditioning on C_{ri} and C_{sj} , HIV+ individuals with non-response are MCAR so that the inverse probability weighted estimator:

$$\frac{1}{\sum_{\mathbf{y}_{ri}\in\Omega_r}\sum_{\mathbf{y}_{sj}\in\Omega_s} w_{ij}z_{ri}z_{sj}} \sum_{\mathbf{y}_{ri}\in\Omega_r}\sum_{\mathbf{y}_{sj}\in\Omega_s} w_{ij}X_{ij}z_{ri}z_{sj}$$

will be consistent for γ_{rs}. Below we consider estimation of the weights.

3.4.2 Estimating Participation Probability

Under the (previous) assumption that subjects are independently sampled, we have:

$$\begin{aligned} w_{ij} &= \Pr(\mathbf{y}_{ri}, \mathbf{y}_{sj} \in S_n \mid C_{ri}, C_{sj},\ \text{HIV+}) \\ &= \Pr(\mathbf{y}_{ri} \in S_n \mid C_{ri},\ \text{HIV+})\Pr(\mathbf{y}_{si} \in S_n \mid C_{si},\ \text{HIV+}) \end{aligned}$$

Here we assume C_{ri} and C_{sj} are only to be demographic variables. We note that it would be possible for z_{ri} and z_{sj} could be correlated, especially when two subjects are within the same household but we do not address this possibility.

Without loss of generality, we now focus only on modeling $\Pr(\mathbf{y}_{ri} \in S_n \mid C_{ri},\ \text{HIV+})$, for which we use logistic regression

$$\pi(C_{ri};\beta)) = \Pr(\mathbf{y}_{ri} \in S_n \mid C_{ri},\ \text{HIV+})$$

where

$$\text{logit}(\pi(C_{ri};\beta)) = \beta_0 + \beta_1 C_{ri}$$

for the vector of parameters $\beta = (\beta_0, \beta_1)$. In our setting, the variables for C_{ri} are community and age.

We condition on HIV+ individuals only, because only they can contribute to the VGL network. To fit this logistic regression model, we need to accommodate the uncertainty regarding which non-responders are HIV+. As noted earlier, this information is not available since HIV status can only be obtained for BCPP responders. Although for nonparticipants in the BCPP, gender and age are provided by other members within their household, it is still not possible to estimate the parameters in the logistic regression model. We discuss next how to address this MNAR mechanism using population level data.

3.4.3 Addressing Missing HIV+ Nonparticipants

Because HIV status is missing for all non-participants, HIV status is missing not at random (MNAR) [26]; hence we cannot fit the logistic regression model in the previous section directly using observed BCPP data. To address this MNAR mechanism, we will use estimates of HIV prevalence in each community obtained from a national household survey (BAIS). We assume that within HIV+ individuals, non-response in the BCPP is MAR, depending only demographic covariates. Under this assumption, the HIV+ individuals in the BCPP sample is a random subsample of the HIV+ subpopulation of the targeted population of interest. Thus, given the covariates, prevalence of HIV is assumed to be the same in the sampled and non-sampled HIV+.

We propose to use multiple imputation to address the missing HIV status of the nonparticipants in the BCPP. Each imputation is obtained by randomly selecting a proportion of the nonparticipants to be assigned status of HIV+. This proportion is given by the HIV prevalence in each community which was obtained in the BAIS study. Thus all village residents–including those not selected to be in the original household incidence cohort–were used to estimate w_{ij}.

Following Rubin et. al., we conduct this imputation 10 times resulting in 10 imputed samples [23] and use the methods in Section 4 to estimate γ_{rs} for each imputed sample. We use the following formulas to calculate the mean probability of linkage across imputed-complete samples, the variance, standard deviation and standard error of the mean.

Following the notation from Rubin, the m imputed samples have corresponding m estimated statistics $[Q_{.1}, Q_{.2}, \ldots, Q_{.m}]$ and variance-covariance matrices $[U_{.1}, U_{.2}, \ldots, U_{.m}]$ [23]. The repeated-imputation estimate is $Q_m = \sum_{l=1}^{m} Q_{.l}/m$, the associated variance-covariance of Q_m is $T_m = U_m + \frac{m+1}{m} B_m$, where $U_m = \sum_{l=1}^{m} U_{.l}/m$ is the within-imputation variability, which is calculated using the binomial variance, and $B_m = \frac{1}{m-1} \sum_{I=1}^{m} (Q_{.I} - Q_m)(Q_{.I} - Q_m)'$ is the between-imputation variability.

3.4.4 Inference on Probability of Linkage

First, we assume that $X_{ij} \perp z_{ri}, z_{sj} \mid \widetilde{\mathbf{C}}_{ij}$. We then note that if S_n were a completely random sample, then $X_{ij} \perp z_{ri}, z_{sj}$ and it follows that

$$\gamma_{rs} = E(X_{ij}) = \frac{1}{p_r p_s} E(z_{ri} z_{sj} X_{ij}),$$

where p_r and p_s are the sampling proportions of community r and s, respectively. A consistent estimate is then given by:

$$\widehat{\gamma}_{rs} = \frac{1}{p_r p_s} \frac{1}{n_{rs}} \sum_{\Omega_r} \sum_{\Omega_s} X_{ij} z_{ri} z_{sj}.$$

However, in our case, we do not have a completely random sample, so we cannot use $\frac{1}{p_r}$ and $\frac{1}{p_S}$ as sampling weights as demonstrated above. To adjust for selection bias, we use the following:

$$\widehat{\gamma}_{rs} = \sum_{\Omega_r} \sum_{\Omega_s} \frac{X_{ij} z_{ri} z_{sj}}{\pi_{ri} \pi_{sj}}.$$

where $\pi_{ri} = \pi(C_{ri};\beta))$ and $\pi_{sj} = \pi(C_{sj};\beta))$.

One approach is to estimate β for $\pi_r i$ and $\pi_s j$ and substitute such estimates in the estimator for $\gamma_r s$. However, this approach does not take into account sampling variability when estimating $\pi_r i$ and $\pi_s j$. By utilizing the functional response models (FRM), we can jointly estimate π_{ri}, π_{sj} and γ_{rs}. To this end, consider the following FRM.

$$\begin{aligned}
&E(\mathbf{f}_{ij} \mid \widetilde{\mathbf{C}}_{ij}) = \mathbf{h}_{ij}(\widetilde{\mathbf{C}}_{ij};\theta), \quad \mathbf{f}_{ij} = (f_{ij1}, f_{ij2})^\top, \quad \mathbf{h}_{ij} = (h_{ij1}, h_{ij2})^\top, \quad \widetilde{\mathbf{C}}_{ij} = \{C_{ri}, C_{sj}\}, \\
&\quad f_{ij1} = X_{ij}, \quad f_{ij2} = z_{ri} z_{sj}, \quad h_{ij1} = \gamma_{rs}, \quad h_{ij2} = \pi_{ri}\pi_{sj}, \\
&\quad \pi_{ri} = \pi(C_{ri};\beta) = \operatorname{logit}^{-1}(\beta_0 + \beta_1 C_{ri}), \quad \pi_{sj} = \pi\left(C_{sj};\beta\right) = \operatorname{logit}^{-1}\left(\beta_0 + \beta_1 C_{sj}\right).
\end{aligned}$$

In the model above, the response $\mathbf{f}_{ij}$ is indexed by a pair of subjects. It is a member of a class of functional response models (FRM). This class of models is useful, because of its ability to model relationships of interest that involve interactions between subjects [7, 30]. For inference, consider a class of U-Statistics based weighted generalized estimating equations (UWGEE):

$$\mathbf{U}_N(\theta) = \sum_{\mathbf{y}_{ri}\in\Omega_r} \sum_{\mathbf{y}_{sj}\in\Omega_s} \mathbf{U}_{N,ij} = \sum_{\mathbf{y}_{ri}\in\Omega_r} \sum_{\mathbf{y}_{sj}\in\Omega_s} D_{ij} V_{ij}^{-1} \boldsymbol{\Delta}_{ij} S_{ij} = \sum_{\mathbf{y}_{ri}\in\Omega_r} \sum_{\mathbf{y}_{sj}\in\Omega_s} D_{ij} V_{ij}^{-1} \boldsymbol{\Delta}_{ij} \left(\mathbf{f}_{ij} - \mathbf{h}_{ij}\right), \tag{3.1}$$

where

$$\begin{aligned}
&Var\left(f_{ij1} \mid \widetilde{\mathbf{C}}_{ij}\right) = \gamma_{rs}(1-\gamma_{rs}), \\
&Var\left(f_{ij2} \mid \widetilde{\mathbf{C}}_{ij}\right) = \pi_{ri}\pi_{sj}(1-\pi_{ri}\pi_{sj}), \\
&\mathbf{S}_{ij} = \mathbf{f}_{ij} - \mathbf{h}_{ij}, \quad V_{ij} = \begin{pmatrix} Var(f_{ij1}) & 0 \\ 0 & Var\left(f_{ij2}\right) \end{pmatrix}, \\
&D_{ij} = \frac{\partial}{\partial\theta}\mathbf{h}_{ij}, \quad \boldsymbol{\Delta}_{ij} = \begin{pmatrix} \frac{z_{ri} z_{sj}}{\pi(C_{ri};\beta)\pi(C_{sj};\beta)} & 0 \\ 0 & 1 \end{pmatrix}, \quad \theta = (\gamma_{rs}, \beta)^\top.
\end{aligned}$$

To show the UWGEE in (3.1) is unbiased, i.e.,

$$E\left[\mathbf{U}_N(\theta)\right] = \sum_{\mathbf{y}_{ri}\in\Omega_r}\sum_{\mathbf{y}_{sj}\in\Omega_s} E\left(\mathbf{U}_{N,ij}\right) = \sum_{\mathbf{y}_{ri}\in\Omega_r}\sum_{\mathbf{y}_{sj}\in\Omega_s} E\left(D_{ij}V_{ij}^{-1}\mathbf{\Delta}_{ij}\mathbf{S}_{ij}\right) = \mathbf{0}.$$

we only need to show that $E\left(D_{ij}V_{ij}^{-1}\mathbf{\Delta}_{ij}S_{ij}\right) = \mathbf{0}$. To this end, we have:

$$\begin{aligned}
E\left(D_{ij}V_{ij}^{-1}\mathbf{\Delta}_{ij}S_{ij}\right) &= E\left[E\left(D_{ij}V_{ij}^{-1}\mathbf{\Delta}_{ij}S_{ij}\right) \mid \widetilde{\mathbf{C}}_{ij}\right] = E\left[D_{ij}V_{ij}^{-1}\mathbf{\Delta}_{ij}E\left(S_{ij} \mid \widetilde{\mathbf{C}}_{ij}\right)\right] \\
&= E\left\{D_{ij}V_{ij}^{-1}\mathbf{\Delta}_{ij}E\left[\begin{pmatrix} f_{ij1}-h_{ij1} \\ f_{ij2}-h_{ij2} \end{pmatrix} \mid \widetilde{\mathbf{C}}_{ij}\right]\right\} \\
&= E\left\{D_{ij}V_{ij}^{-1}\mathbf{\Delta}_{ij}\begin{pmatrix} E\left[\left(f_{ij1}-h_{ij1}\right) \mid \widetilde{\mathbf{C}}_{ij}\right] \\ E\left[\left(f_{ij2}-h_{ij2}\right) \mid \widetilde{\mathbf{C}}_{ij}\right] \end{pmatrix}\right\} \\
&= E\left\{D_{ij}V_{ij}^{-1}\mathbf{\Delta}_{ij}\begin{pmatrix} E\left(f_{ij1}\right)-h_{ij1} \\ E\left(f_{ij2} \mid \widetilde{\mathbf{C}}_{ij}\right)-h_{ij2} \end{pmatrix}\right\} \\
&= 0.
\end{aligned}$$

Thus, the UWGEE is unbiased.

As in the case of GEE, a working correlation structure $C(\alpha)$ between f_{ij1} and f_{ij2} parameterized by some vector α may be assumed and incorporated into V_{ij} to improve efficiency of estimates of θ. In this more general case, $V_{ij}(\alpha)$ depends on α as well. Like GEE, the UWGEE estimate $\widehat{\theta}$ by solving the equations above, the estimated parameter of interest is asymptotically normal by Theorem 1. For notational brevity, we only consider working independence structure below unless otherwise stated.

Theorem 3 *Let*

$$\mathbf{v}_{ri} = E(\mathbf{U}_{n,ij} \mid y_{ri}, z_{ri}, C_{ri}), \quad \mathbf{v}_{sj} = E(\mathbf{U}_{n,ij} \mid y_{sj}, z_{sj}, C_{sj}), \quad B = E\left(D_{ij}\boldsymbol{\Delta}_{ij}V_{ij}^{-1}D_{ij}^{\top}\right), \quad (3.2)$$

$$\Sigma_r = Var(\mathbf{v}_{ri}), \quad \Sigma_s = Var\left(\mathbf{v}_{sj}\right), \quad \rho_r^2 = \lim_{n\to\infty}\frac{N_{rs}}{N_r} < \infty, \quad \rho_s^2 = \lim_{n\to\infty}\frac{N_{rs}}{N_s} < \infty,$$

$$N_{rs} = N_r N_s, \quad \Sigma_U = \rho_r^2\Sigma_r + \rho_s^2\Sigma_s, \quad \Sigma_\theta = B\Sigma_U B^{\top}.$$

Then, under mild regularity conditions, we have

1. $\widehat{\theta}$ is consistent.

2. If $\sqrt{N_{rs}}(\widehat{\theta} - \theta) \to_d N(\mathbf{0}, \Sigma_\theta)$.

To estimate Σ_θ, we first estimate B by:

$$\widehat{B} = \frac{1}{N_{rs}} \sum_{\mathbf{y}_{ri}\in\Omega_r} \sum_{\mathbf{y}_{sj}\in\Omega_s} \widehat{D}_{ij}\widehat{\boldsymbol{\Delta}}_{ij}\widehat{V}_{ij}^{-1}\widehat{D}_{ij}^{\top},$$

where $\widehat{B}$ denotes B with θ substituted by $\widehat{\theta}$. We then estimate Σ_r and Σ_s by:

$$\widehat{\Sigma}_r = \frac{1}{N_r}\sum_{i=1}^{N_r} \widehat{\mathbf{v}}_{ri}\widehat{\mathbf{v}}_{ri}^{\top}, \quad \widehat{\mathbf{v}}_{ri} = \frac{1}{N_s}\sum_{j=1}^{N_s} \widehat{\mathbf{U}}_{N,ij},$$

$$\widehat{\Sigma}_s = \frac{1}{N_s}\sum_{i=1}^{N_s} \widehat{\mathbf{v}}_{sj}\widehat{\mathbf{v}}_{sj}^{\top}, \quad \widehat{\mathbf{v}}_{sj} = \frac{1}{N_r}\sum_{i=1}^{N_r} \widehat{\mathbf{U}}_{N,ij},$$

where $\widehat{\mathbf{U}}_{N,ij}$ denotes $\mathbf{U}_{N,ij}$ with θ substituted by $\widehat{\theta}$. A consistent estimator of Σ_θ is given by:

$$\widehat{\Sigma}_\theta = \widehat{B}\widehat{\Sigma}_U\widehat{B}^{\top} = \widehat{B}\left(\frac{N_{rs}}{N_r}\widehat{\Sigma}_r + \frac{N_{rs}}{N_s}\widehat{\Sigma}_s\right)\widehat{B}^{\top}.$$

3.5 Application

3.5.1 Simulation Study

We apply the proposed approach to simulated data to examine its performance. Consider two groups: Group r and s. Let $N_r = 1000$ and $N_s = 1200$. Note HIV networks have shown to have been follow power law distributions; we assume such a network for our simulation study:

$$\Pr(D_{rs} = k) = \beta k^{-\alpha}$$

for some value of α with β being the normalizing factor. Stumpf et. al., showed that a larger results in a heavier loss of data when sampling; hence we perform simulations on scale-free(a subset of power law) networks with $\alpha = 3$ [25].

For sampling, let

$$\begin{aligned}
&t_{ri} \sim \eta_1 + \eta_2 w_{ri} + \eta_4 D_{ri} \\
&t_{sj} \sim \eta_1 + \eta_2 w_{sj} + \eta_3 + \eta_4 D_{sj} \\
&w_{ri} \sim N(\mu_{wr}, \sigma_{wr}), \quad w_{sj} \sim N(\mu_{ws}, \sigma_{ws}) \\
&h_{ri} \sim Bern(p_r), \quad h_{sj} \sim Bern(p_s)
\end{aligned}$$

As in the BCPP, we will define an individual from Group r to be sampled, so $z_{ri} = 1$, if and only if $t_{ri} = 1$ and $h_{ri} = 1$ and $z_{ri} = 0$ otherwise,and use the same procedure for z_{sj}. We apply the proposed approach to this setting for each of the scenarios shown in Table 3.1. Figure 3.1 demonstrates that the adjusted estimators are considerably less biased than those of the unadjusted estimators. Additionally, Table 3.2 shows that the coverage probabilities for each scenarios are considerably high with the lowest coverage probability being 0.94 for Scenario 5.

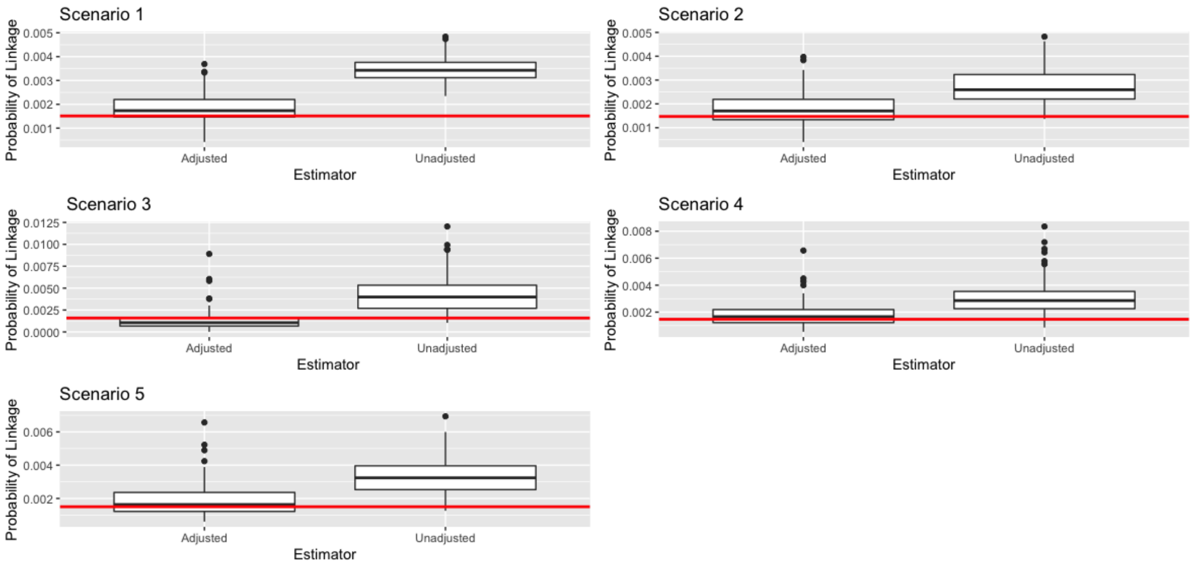


Figure 3.1. Distribution of adjusted and unadjusted estimators for Probabilities of Linkage from the simulation. The red horizontal line is the true value

Table 3.1. Each of the scenarios considered in the simulation study performed in Section 5.1.

Scenario	η	μ	σ	**p**
1	(0.50, -2, 2, 0.60)	(1,2)	(1.5, 1.5)	(0.50, 0.50)
2	(0.50, -2, 2, 0.60)	(1,1)	(1.5, 1.5)	(0.30, 0.25)
3	(1, -3, 1, 0.60)	(1,2)	(2, 1.5)	(0.15, 0.15)
4	(1, -2, 3, 0.6)	(2,2)	(1, 1)	(0.15, 0.15)
5	(1, -2, 3, .45)	(2, 2)	(1.5, 1.5)	(0.55, 0.20)

Table 3.2. The coverage probabilities for the adjusted estimators for probabilities of linkage from the simulation.

Scenario	1	2	3	4	5
Coverage	1	1	1	0.99	0.94

3.5.2 BCPP Application

At the end of BCPP study, all households were targeted for a survey in 6 of the 30 participating communities (Gumare, Mauntalala, Mmankgodi, Mmathethe, Ramokgonami and

Shakawe). For participants who chose to participate, the survey obtained demographic and household data as well as HIV status. For those who were HIV+, the viral genetic sequences are available. Table 3.3 provides the proportions of HIV+ in individuals that participated in the BCPP; for 4 of the 6 communities, the proportions were over 40% but for the other 2, they were below 30%.

Table 3.3. The proportion (p) and number (n) of HIV+ individuals in each community that participated in the BCPP.

	Gumare	Maunatlala	Mmankgodi	Mmathethe	Ramokgonami	Shakawe
p	0.29	0.52	0.26	0.44	0.48	0.48
n	325	363	270	336	350	484

We applied our methods to adjust for the sampling bias described above. As stated in Section 1, two individuals are considered linked if the pairwise genetic distance between their viral genetic sequences is less than some given threshold. Following Novistky et. al., we use a threshold of $c = 0.07$ to define genetic linkage. Figure 3.2 provides a heat map of the intensity of linkage rates after adjustment for missing data, within and across the ESS communities as well as the variability associated with these estimates of linkage. The analysis provides evidence of a larger within- than between-community linkage. Gumare, Mmathethe and Ramokgonami have the highest within-community linkage rates.

3.6 Discussion

Viral genetic linkage analysis play an important role in molecular epidemiology in it s ability to reveal features of transmission patterns within and across communities such analyses may prove useful in control of COVID-19 and other outbreaks. While methods have been proposed for viral genetic linkage analyses in the presence of sampling bias, this paper is the first to ground such methods in a statistical framework uniquely positioned to address between-subject, rather than within-subject attributes as as the primary focus of analyses. Through the

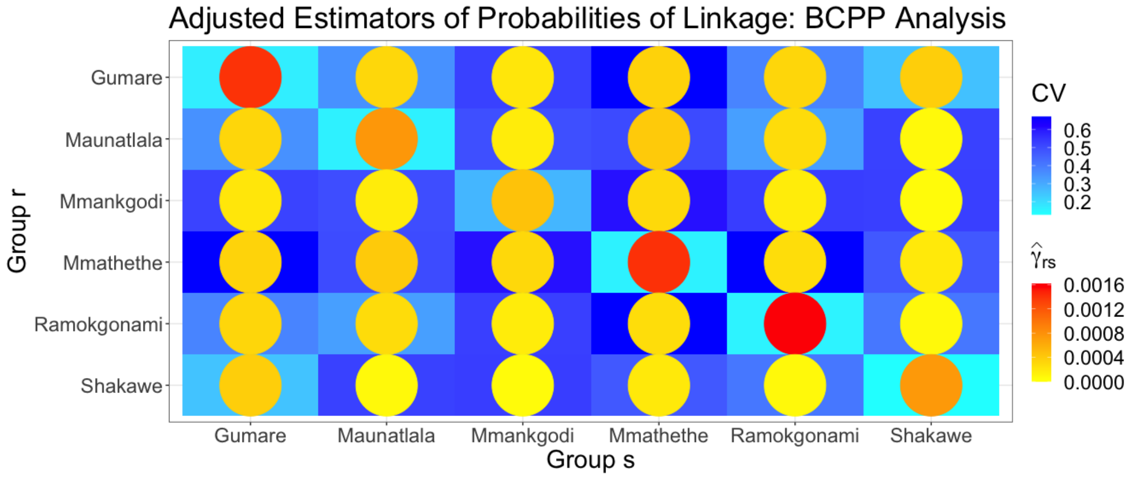


Figure 3.2. Adjusted estimators for probabilities of linkage between the communities from the BCPP and coefficients of variation (CV) indicated by colors of circle and cell, respectively.

use of FRM and UWGEE, we were able to show consistency and asymptotic normality of our estimators under the assumption that non-responses are MAR, thereby permitting unbiased point and interval estimates, as demonstrated by our simulation results.

Our illustrative example made use of data from an HIV prevention study in Botswana — the BCPP. We demonstrated that VGL linkage across communities is common—which implies that a treatment-as-prevention intervention applied at the village level will likely have effects on HIV incidence that are attenuated compared to effects that would occur if all relationships took place within villages. Furthermore such estimates would also be attenuated compared to another estimand of interest—the counterfactual expected difference in incidence between a setting in which the intervention was implemented in all villages and a setting in which it was in none. Hence these VGL analyses are useful in both design and interpretation of cluster randomized trials for control of endemic diseases or disease outbreaks.

In many real studies, we can estimate missing response probabilities under the MAR assumption. In this case, the FRM with inference based on a class of UWGEE will provide valid inference about linkage among network nodes. In the BCPP study, data are missing on people within households who were enumerated but who did not provide blood samples (used to assess HIV status as well as to obtain sequences) —leading to data that are MNAR. the By utilizing population level estimates and multiple imputation, we addressed this statistical challenge. The idea is similar to raking–an approach used in survey research to improve estimation of sampling weights by utilizing aggregated population-level estimates. Our methods would apply networks of all types, for which sampling of nodes is not complete but for which there exist sufficient covariate information to help identify the MAR mechanism. This paper also illustrates how to address a type of MNAR mechanism in survey research by taking advantage of general information regarding the population survey.

3.7 Acknowledgements

Chapter 3, in full, is a reprint of the materials as it appears in Estimating Probabilities of Linkage in the Presence of Missing Data 2022. Victor De Gruttola, Tu, Xin, Jingjing Zou, Tuo Lin. The dissertation author was the primary investigator and author of this paper.

Chapter 4

Nonparametric Estimation of Network Properties in the Presence of Missing Data

4.1 Introduction

While networks have become widely used to analyze elements in a system and how these elements interconnect, the challenge of sampling complete network data remains a prevalent issue. In most instances, we only sample a portion of the nodes and hence don't observe the edges corresponding to the missing nodes. As a result, estimation of summary statistics related to the network and the network's generative process will be generally biased.

We addressed this issue for probabilities of linkage and linkage rates. However, the methodology shown in each of those works only applied for those specific summary statistics. Hence, in this chapter we focus on developing a method to generate consistent estimators for network properties while assuming nodes are missing completely at random. To do so, we developed a Monte Carlo Markov Chain (MCMC) edge toggling approach.

This brings a nonparametric approach to estimating various network properties such as degree distribution, average degree and totals. In doing so, the only assumptions required are the nodes are missing completely at random and properties pertaining to the network property of interest. Notably, this approach does not require any assumptions on structure of the complete network. Overall, this approach can be seen as a tool to derive consistent estimators for notable network properties under the case that we have a sampled network.

We apply these methods to analyses of HIV viral genetic linkage network in Botswana where the data is from a large cluster-randomized trial of a combination HIV prevention intervention - the Botswana Combination Prevention Project (BCPP) [21]. Specifically, we intend to estimate the power law parameters and degree distribution from the viral genetic linkage network within and between each of the communities in the BCPP. The interest in these analyses is to investigate the patterns of HIV transmission between communities in Botswana.

4.2 Notation and Setting

Consider a population of nodes, Ω, of finite size N. Let $G = (\Omega, E)$ denote a network where E is the set of edges between nodes, $E \subset \Omega \times \Omega$. Note that we assume that G is an undirected network. We note that we will assume that all nodes in G are known, but the edges are not assumed to be known.

Now let $\phi(\cdot)$ be some function that maps a given network to network property statistics. Examples of network property statistics are edge probability, degree distribution, clustering coefficient and values pertaining to the network generating process. We are interested in inference about $\phi(G)$ when nodes are MCAR.

Consider a random sample of nodes from Ω of size n, which we denote by S_n, such that the proportion of sampled nodes is defined as p (known). It is important to note how we denote the "sampled graph" - g_0. We then denote g_0 as the graph that includes all nodes in G, but only includes all observed edges from the random sample taken, so $g_0 = (\Omega, E_n)$ where E^n is the set of edges observed in S_n. This implies that although we sampled only a proportion of the nodes from G, we assume that the missing nodes are enumerated and are included in the g_0. However, only the edges that are observed in S_n are included in g_0.

Let E^* denote the set of all possible edges based off Ω. It follows that we define the "potential edges of G" as the following:

$$E_p = E^* - E^n$$

Therefore, E_p resembles the set of edges that may possibly exist if all the edges were observed while not including any edges that were observed in S_n.

4.3 Methodology

In estimating $\phi(G)$, Markov Chain Monte Carlo (MCMC) procedure is the basis for generating a collection of estimates for $\phi(G)$, which we call by $\{\phi_1, \dots, \phi_t\}$ where our final proposed estimator for $\phi(G)$ will be defined as $\widehat{\phi}$

$$\widehat{\phi} = \frac{1}{t} \sum_{m=1}^{t} \phi_m$$

with t denoting MCMC sample size. We first discuss the steps implemented to calculate ϕ_m for $m = 1, \dots, t$ and then we show consistency for $\widehat{\phi}$.

4.3.1 Calculating ϕ_m

We first calculate $\phi(g_0)$. We then randomly take M samples from g_0. Note that the choice of the number of subsamples M is arbitrary. Let g_{0k} be the graph that includes all nodes in G, but only includes all edges observed in the kth subsample from g_0. For each subsample taken, we calculate $\phi(g_{0k})$ and $d = ||\bar{\phi}_0 - \phi(g_0)||$ where $\bar{\phi}_0 = \frac{1}{M} \sum \phi(g_{0k})$ and $|| \cdot ||$ is the Euclidean distance. We randomly select $[d * b]$ potential edges from E_p where $[\cdot]$ represents the ceiling function and b represents is arbitrary such that if b is small then we get more accurate estimates for $\phi(G)$ at the cost of run-time and if b is large then we get less estimates for $\phi(G)$ with a faster run-time. Now we edge toggle all selected potential edges. Where an edge does exist, we remove it; for any potential edges where no edge exists, we add an edge. We define the graph for which edges have been toggled as g_o^p.

We obtain M completely random subsamples from g_0^p of size $[p*N]$. Let g_{0k}^p be the graph that includes all nodes in G, , but only includes edges that exist after edge toggling in the kth subsample from g_0^p. For each subsample taken, we calculate $\phi(g_{0k}^p)$ and $d_s = ||\bar{\phi} - \phi(g_0)||$ where $\bar{\phi} = \frac{1}{M}\sum\phi(g_{0k}^p)$. Let c be some pre-defined threshold. It follows that there are 3 potential cases to consider: 1) If $d_s < d$ and $d > c$, then let $d = d_s$ and $g_0 = g_0^p$, 2) if $d_s < d$ and $d \leq c$, then let $d = d_s$, $g_0 = g_0^p$ and $\phi_m = \phi(g_0)$ and 3) otherwise, restart back at the first step and repeat. After performing the described approach t times, we then calculate $\widehat{\phi}$, our proposed estimator for ϕ. Note that we outline the proposed approach Algorithm 1.

4.3.2 Consistency of $\widehat{\phi}$

Let $\Phi(g_0)$ be a random variable representing the network property statistics of interest when sampling from G. Thus, $\phi(g_0)$ is an instance of $\Phi(g_0)$. In Theorem 4, we show that the proposed approach leads to a consistent estimator for $\phi(G)$ under the following set of assumptions: 1) $||\ \bar{\phi} - E(\Phi(g_0))\ || \rightarrow 0$, as $N \rightarrow \infty$ and 2) there exists some continuous function f such that $f(E(\Phi(g_0))) = \phi(G)$ and 3) $f(\bar{\phi}) \rightarrow \phi(g)$ as $N \rightarrow \infty$. Theorem 4 can be shown using Continuous Mapping Theorem. Theorem 4 implies that if the expected network summary statistic for a sample from g_0 converges to the expected network summary statistic for a sample from G, then $\phi(g) \rightarrow \phi(G)$ as $N \rightarrow \infty$.

The first assumption is equivalent to performing a sufficient number of edges toggles on g_0 for this condition to hold: $||\ \bar{\phi} - E(\Phi(g_0))\ || \rightarrow 0$, as $N \rightarrow \infty$. To assess whether this assumption holds and given that $E(\Phi(g_0))$ is not observed, we estimate $d_s =||\ \bar{\phi} - E(\Phi(g_0))\ ||$ by replacing $E(\Phi(g_0))$ with $\phi(g_0)$ and assess whether $d_s \leq c$ to identify potential violations of this first assumption.

The second assumption implies that for our network property of interest, there exists some function that can map the network property of the sampled network to that of population network.To provide additional context, we show an example of f that satisfies the second assumption. $\phi(\cdot)$ be edge probability for a given network. Let X be the number of edges in G and T be the number of possible edges in G, so $T = \binom{N(N-1)}{2}$. Let X_s be number of observed edges in g_0 and T_s be the number of possible edges in g_0, so $T_s = \binom{n(n-1)}{2}$. It can be shown that

$$\frac{E(X_s)}{T_s} = \frac{X}{T}$$

We define f as the following:

$$f(X) = \frac{TX}{T_s}.$$

We consider another example of f that satisfies the second assumption, where interest lies in degree distribution. Suppose that we are interested in degree distribution. Let f_d and f_d^* be the true and observed frequency for degree d nodes in G and g_0, respectively. From [22], it can be shown that

$$E(f_d^*) = \sum_{d'=0}^{N-1} \Pr(d,d') f_{d'}$$

where

$$\Pr(d,d') = \frac{\binom{d'}{d}\binom{N-1-d'}{n-1-d}}{\binom{N-1}{n-1}}$$

It follows that we can use the above equation to create a system of equations to solve for $(f_0, f_1, \ldots, f_{N-1})$ which we would define as f. Additionally, it can be shown that such an f exists for other relevant network properties such as average degree, group size, clustering coefficient and totals [12].

4.3.3 Diagnostic

Our proposed diagnostic for assessing if the assumptions of Theorem 4 hold focuses on violation of the first assumption. Let $D_s = d_{s1}, \ldots, d_{st}$ be the collection containing each corresponding d_s from $\phi_1, \ldots, \phi_t$. Using D_s, we then estimate the standard error for d_s by calculating its asymptotic variance or through bootstrapping. We then calculate the 95% confidence interval for d_s.

After obtaining the confidence interval, we then focus on its upper bound which we denote as d_u. To assess if the first assumption holds, we check to see if $d_u \leq c$. If so, we infer that the first assumption is not violated. If $d_u \geq c$, then further edge toggling may be necessary to safely ensure that the first assumption holds.

Theorem 4 *Let $\Phi(g_0)$ be a random variable representing the network property statistics of interest when sampling from G (so $\phi(g_0)$ would be considered an instance of $\Phi(g_0)$). Suppose that $||\ \bar{\phi} - E(\Phi(g_0))\ || \rightarrow 0$, as $N \rightarrow \infty$. If there exists some continuous function f such that $f(E(\phi(G_0))) = \phi(G)$ and $f(\bar{\phi}) \rightarrow \phi(g)$ as $N \rightarrow \infty$, then $\phi(g_0) \rightarrow \phi(G)$ as $N \rightarrow \infty$.*

Proof. We note that from the first assumption, $\bar{\Phi}$ is a consistent estimator for $\phi(G_0)$, so

$$\bar{\phi} \rightarrow E(\Phi(G_0))$$

Hence, by Continuous Mapping Theorem, we have

$$\phi(g_0) = f(\bar{\phi}) \rightarrow f(E(\Phi(G_0))) = \phi(G)$$

■

Algorithm 1. Estimating ϕ_m

```
procedure ET($g_0$)
    System Initialization
    for $m = 1; m++;$ while $m <= t$ do
        while $d \geq c$ do
            Calculate $\phi(g_0)$
            Perform completely random subsamples from $g_0$ of size $[p * N]$
            Calculate $g_{0k}$, the graph that includes all nodes in $G$, but only includes all edges observed in the kth subsample from $g_0$
            Calculate $\phi(g_{0k})$ for all $k$
            Calculate $d = ||\bar{\phi}_0 - \phi(g_0)||$ where $\bar{\phi}_0 = \frac{1}{M} \sum \phi(g_{0k})$
            Randomly select $[d * b]$ potential edges in from $E_p$
            Edge toggle all $[d * b]$ selected potential edges and call this new graph $g_0^p$
            Perform completely random subsamples from $g_0^p$ of size $[p * n]$
            Calculate $g_{0k}^p$, the graph that includes all nodes in $G$, but only includes all edges observed in the kth subsample from $g_0^p$
            Calculate $\phi(g_{0k}^p)$ for all $k$
            Calculate $d_s = ||\bar{\phi} - \phi(g_0)||$ where $\bar{\phi} = \frac{1}{M} \sum \phi(g_{0k}^p)$
            if $d_s \leq d$ then
                Let $d = d_s$
                Let $g_0 = g_0^p$
            end if
        end while
        Let $\phi_m = \phi(g_0)$
    end for
    Calculate $\widehat{\phi} = \frac{1}{t} \sum \phi_m$
end procedure
```

4.4 Application

4.4.1 Simulation

In this section, we apply the proposed approach in a simulated setting. Stumpf et. al. performed a simulation with sampling from networks that follow a power law distribution [25] and follow the following probability distribution:

$$\Pr(D_i = k) = \begin{cases} \beta k^{-\alpha} & k \geq 1 \\ p & k = 0 \end{cases}$$

where α is considered as a scaling factor and β is a normalizing factor. Generally α is the parameter of interest and that will be our parameter of interest for this simulation study. Stumpf et. al. showed that the larger the value of α, the greater the sample degree distribution deviates from the true distribution. Since for most settings, $2 \leq \alpha \leq 3$ [25], we apply the proposed approach to a power law network with $\alpha = 3$ for various sampling proportions.

Consider a network, $G = (V, E)$, where $\mid V \mid = 500$ with the following degree distribution:

$$\Pr(D_i = k) = \begin{cases} \beta k^{-3} & k \geq 1 \\ 0.50 & k = 0 \end{cases}$$

where $\Pr(D_i = k)$ represents the probability that a randomly selected node has degree k. It follows that we take a completely random sample from G and apply the proposed approach. We will take 500 completely random samples from G for each sampling proportion considered where $p = 0.20, 0.30, 0.40, 0.50, 0.60, 0.70$. We will then apply the approach.

Figure 4.1 shows the results from the simulation. For $p \geq 0.50$ and degree value between 0 and 3, we find that the estimated probabilities are accurate. Otherwise, the estimated probabilities show to be inaccuarate.

Figure 4.1 shows the results from the simulation. The estimated probabilities appear to be accurate only for values of $p \geq 0.50$ and degree value between 0 and 3,but not otherwise.

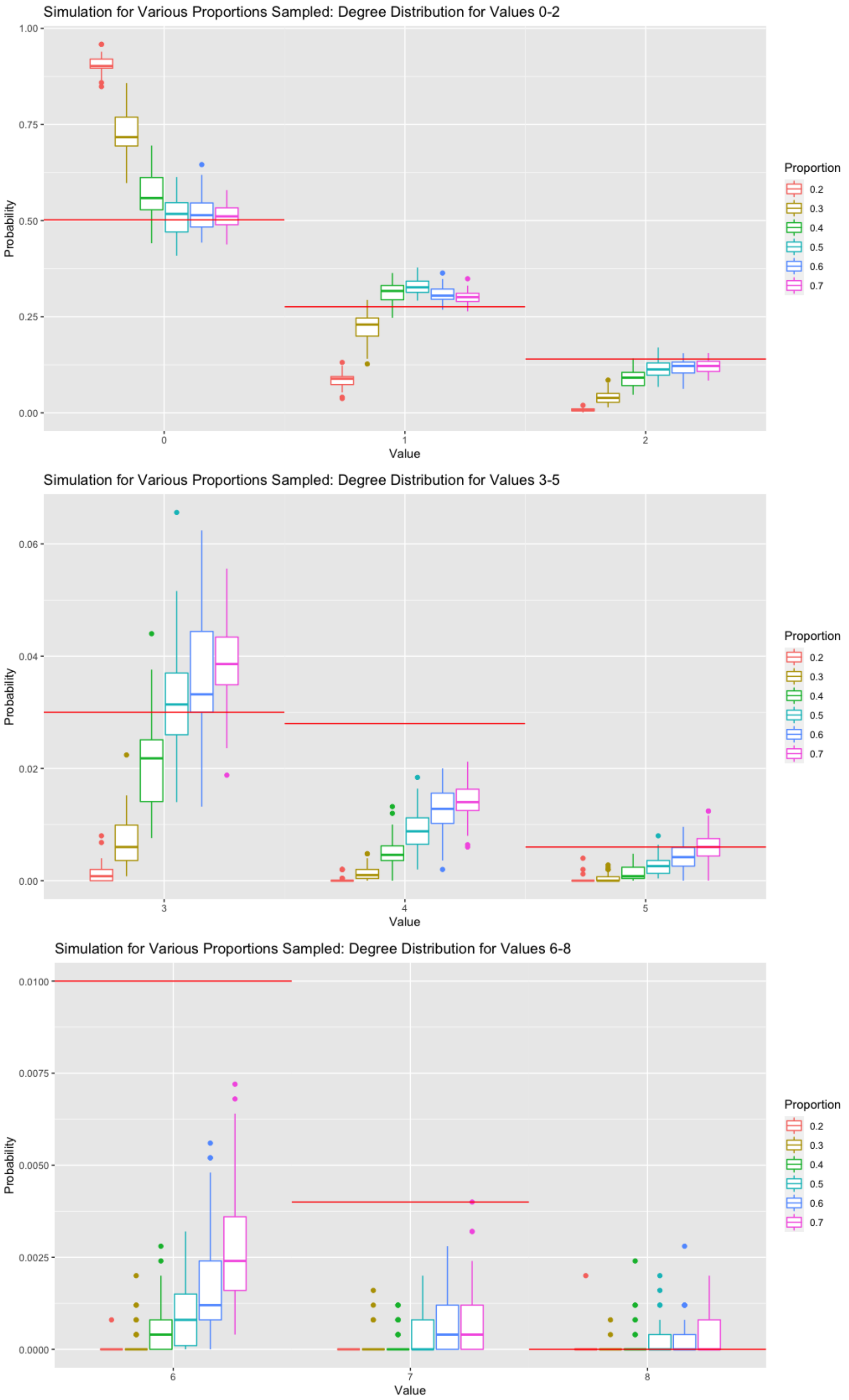
As stated in previous sections, the intent of the proposed method was not to estimate the degree distribution value itself, but instead to estimate the parameter, α pertaining to the network's generating distribution. Figure 4.2 shows that for all sampling proportions the estimated α is considerably closer to the true value, $\alpha = 3$, when fitting the proposed network's degree distribution to a sampled data from a power law network.

4.4.2 BCPP

As discussed above, we use this approach to estimate the network generating process for the HIV viral genetic linkage networks in Botswana. The data come from a large cluster-randomized trial of a combination HIV prevention intervention - the Botswana Combination Prevention Project (BCPP).

In the BCPP, all households were targeted for a survey in 6 of the 30 participating communities in Botswana. The communities that were selected are Gumare, Mauntalala, Mmankgodi, Mmathethe, Ramokgonami and Shakawe. Due to low sampling proportions for Gumare and Mmankgodi, we will only focus the following communities for the applied approach: Mauntalala, Mmathethe, Ramokgonami and Shakawe. For those that choose to participate in the survey, demographic and household data along with HIV status was ascertained. For those who were HIV+, the viral genetic sequences were obtained. Hence, missing data arise because some individuals chose not to participate in the BCPP. Nonetheless, in our analyses, individuals are assumed to be MCAR. We acknowledge the fact that this assumption may not entirely hold. Lastly, we note that an edge exists between two individuals when the pairwise distance between their viral genetic sequences are below the threshold 0.07.

We note that several investigators have shown that HIV genetic linkage networks appear to follow a power law distribution [14, 27, 28]. Obtaining an accurate for the power law parame-


Simulation for Various Proportions Sampled: Degree Distribution for Values 0-2

Simulation for Various Proportions Sampled: Degree Distribution for Values 3-5

Simulation for Various Proportions Sampled: Degree Distribution for Values 6-8

Figure 4.1. The proposed approach was applied to a power law network with $\alpha = 3$ and $N = 500$ for the following proportions: $0.20, 0.30, 0.40, 0.50, 0.60, 0.70$. The x axis represents the degree value and the y axis represents the proportion of nodes with the specified degree value. For values between 0 and 3, the degree distribution of the proposed approach diverges from the true degree distribution for proportions of 0.40 and less. For values between 4 and 7, the proposed degree distributions deviate considerably from the true values for all proportions.

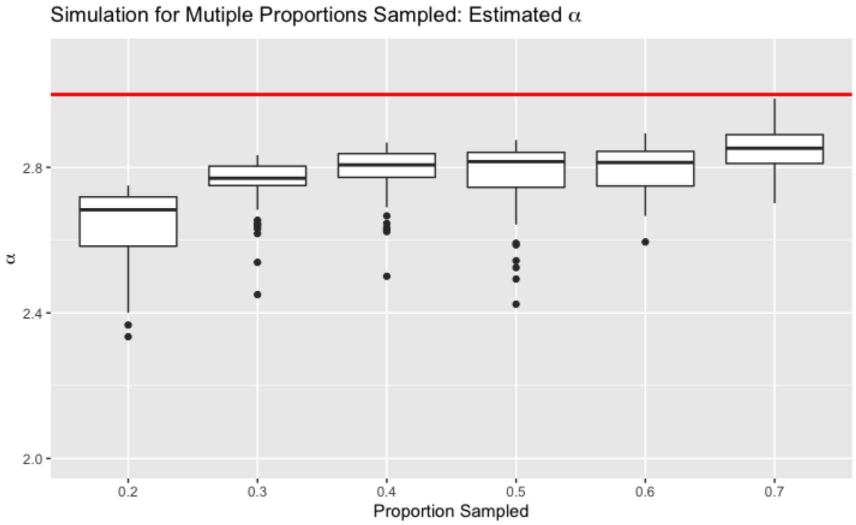


Figure 4.2. Estimated α for each sample performed in the simulation study.The red horizontal line represents the true value for α.

ter α is challenging. We apply the proposed approach to estimate the degree distributions for the communities of interest and use the estimated degree distributions to estimate α.

From Figure 3.3, we have the proposed degree distributions for Mauntalala, Mmathethe, Ramokgonami and Shakawe. For Mauntalala, Ramokgonami and Shakawe, the proportion of individuals that have zero edges are approximate to 0.60 and for Mmathethe the proportion of individuals that have zero edges is approximately 0.50 indicating that Mmathethe has the largest proportion of nodes that have an edge. Additionally Mauntalala and Ramokgonami appear to have a fairly large number of nodes with have six to eights edges–implying the existence of super spreaders in these communities. Figure 3.4 shows that the values of α for Mmathethe and Shakawe are both close to 2.28, whereas the values of alpha for Mauntalala and Ramokgonami are 2.42 and 2.38, respectively. Since Mauntalala and Ramokgonami have larger values for α compared to Mmathethe and Shakawe, then we can expect a higher probability for larger degree values in these two communities, this feature is displayed in Figure 3.3.

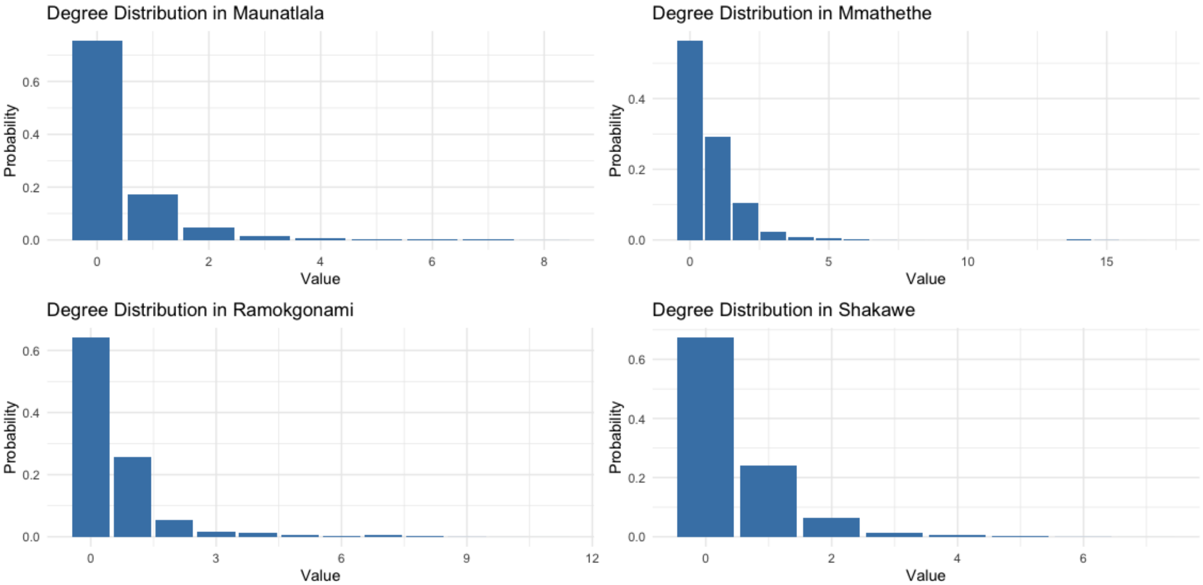


Figure 4.3. Proposed degree distributions for the communities Maunatlala, Mmathethe, Ramokgonami, Shakawe.

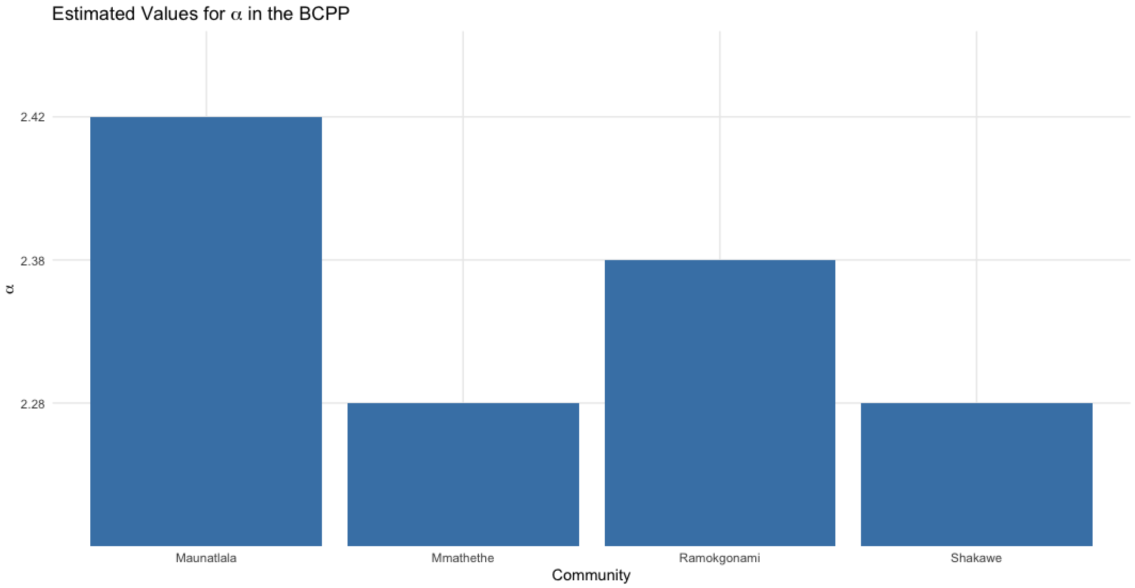


Figure 4.4. Estimated values for α for the communities Maunatlala, Mmathethe, Ramokgonami, Shakawe.

4.5 Discussion

This paper presents novel methods to estimate network properties in the presence of missing data. While methods have been proposed for such analyses, this paper is, to our knowledge, the first to ground a non-parametric method in statistical theory. Through an MCMC edge-toggling approach, we are able to show consistency of our estimator. We demonstrate that the methods work well when sampling proportion is greater than 0.4. How to make further improvements to our estimator when applied to data with low sampling rates is a topic for further research.

Our illustrative example made use of data from the HIV prevention study in Botswana — the BCPP. For the viral genetic linkage networks, we estimated the degree distribution within the communities and values pertaining to the communities' network generating process. These results can be then leveraged to understand transmission patterns for HIV in Botswana.

4.6 Acknowledgements

Chapter 4, in full, is a reprint of the materials as it appears in Nonparametric Estimation of Network Properties in the Presence of Missing Data 2022. Victor De Gruttola, Ravi Goyal, Tu, Xin, Jingjing Zou, Tuo Lin. The dissertation author was the primary investigator and author of this paper.

Chapter 5

Conclusions and Future Work

This dissertation illustrates novel methods to develop consistent and asymptotically normal estimators for various network properties of interest that adjust for the bias that results from missing data. The motivation behind developing such methods was to analyze the viral genetic linkage data from the BCPP. Specifically, my goal was to identify HIV transmission patterns within and between-communities in Botswana; analyses in this dissertation focuses on the 3 pairs of communities in the BCPP ESS.

The analyses from Chapters 2 and 3 provide evidence of a larger within- than between-community linkage and identify pairs of communities with high levels of linkage. The significance of this work lies in making accurate estimates of linkage between communities in the presence of missing data. Further in Chapter 4, the analyses provide evidence about whether some communities may have a fairly large number of superspreaders, individuals who are linked to a disproportionately high number of people, by identification of the network generating distribution that would lead to within-community linkage. These advancements allowed us to identify which pairs of communities that are in different treatment groups in the BCPP that have high levels of linkage between one another indicating evidence of mixing. Therefore, in order to accurately estimate the causal effect of the BCPP intervention, mixing between these identified pairs of communities will need to be addressed.

As stated in previous chapters, the main limitation of the methods presented is their relatively poor performance—including a considerable amount of bias–when the sampling proportion is below 40%. Thus, future work is required to extend the methods in ways that improve performance when sampling proportions are small.

Additionally, the methods in this dissertation expand the theory of network analysis to a missing data setting, and is the first, to our knowledge, that shows showing consistency and asymptotic normality of estimators that accommodate missing data, and thereby provide reliable inference on networks. Further, I expand upon this by developing a non-parametric approach to generate consistent estimators for network properties that generalize over a range of network properties under specified assumptions. These advancements will be useful in developing a method to investigate the effect of an intervention for HIV, in the presence of mixing between randomized cluster-level interventions.

A majority of analyses regarding HIV transmission networks have not adjusted for the bias that arises from missing data and as shown in this dissertation can have dramatic effects on inference and conclusions on HIV transmission patterns [1, 4]. Thus, the results from this dissertation bridge the gap between HIV networks and missing data ultimately allowing for more accurate and correct inference and conclusions regarding HIV transmission patterns.

Appendix A

Estimating Viral Genetic Linkage Rates in the Presence of Missing Data

A.1 Proof of Theorem 1

Proof. Without loss of generality suppose $p_s n_s \in \mathbb{N}$. If $p_s n_s \notin \mathbb{N}$, then we take $p_s n_s = \lceil p_s n_s \rceil$. We denote $\widetilde{D}^s_{ri}$ to be the number of individuals in $S_{n(s)}$ linked to subject $\mathbf{y}_{ri} \in S_{n(r)}$, i.e., a sample version of D^s_{ri}. Note that

$$P\left(\widetilde{D}^s_{ri} \geq 1 \mid D^s_{ri} \geq 1\right) = P(v^s_{ri} = 1 \mid u^s_{ri} = 1) = \pi_{rs}$$

and

$$\widetilde{D}^s_{ri} \mid D^s_{ri} \geq 1, D^s_{ri} = d \sim \text{HyperGeometric}\left(N_s, d, n_s\right)$$

where N_s is the size of the population of group s, d is the number of individuals in Ω_s that are linked to $\mathbf{y}_{ri} \in S_{n(r)}$ and n_s is the size of the random sample taken from Ω_s. We have

$$\begin{aligned}
P\left(\widetilde{D}^s_{ri} \geq 1 \mid D^s_{ri} \geq 1, D^s_{ri} = d\right) &= 1 - \frac{\binom{N_s-d}{n_s}}{\binom{N_s}{n_s}} \\
&= 1 - \frac{(N_s-d)!}{n_s!(N_s-d-n_s)!}\frac{n_s!(N_s-n_s)!}{N_s!} \\
&= 1 - \frac{(N_s-d)!}{(N_s-d-n_s)!}\frac{(N_s-n_s)!}{N_s!} \\
&= 1 - \frac{\prod_{k=0}^{n_s-1}(N_s-k-d)}{\prod_{k=0}^{n_s-1}(N_s-k)} \\
&= 1 - \prod_{k=0}^{n_s-1}(1 - \frac{d}{N_s-k})
\end{aligned}$$

Take $A_N(d) = \prod_{k=0}^{n_s-1}(1 - \frac{d}{N_s-k})$, and let

$$\begin{aligned}
A^l_N(d) &= \log(A_N(d)) \\
&= \sum_{k=0}^{n_s-1} \log(1 - \frac{d}{N_s-k})
\end{aligned}$$

We note that when $x \to 0$, we have $\log(1+x) \to x + O(x^2)$. For $N_s \to \infty$, We have

$$\begin{aligned}
A_N^l(d) &= -\sum_{k=0}^{n_s} \frac{d}{N_s - k} \\
&= -d \sum_{k=0}^{n_s} \frac{1}{N_s - k} \\
&= -d \sum_{j=(1-p_s)N_s}^{N_s} \frac{1}{j} \\
&= -d \sum_{j=(1-p_s)N_s}^{N_s} \frac{1}{N_s} \frac{N_s}{j} \\
&= -d \int_{(1-p_s)N_s}^{N_s} \frac{1}{N_s} \frac{N_s}{j} dj \\
&= -d \int_{1-p_s}^{1} \frac{1}{x} dx \\
&= -d(-\log(1-ps)) \\
&= d\log(1-p_s).
\end{aligned}$$

Thus, $A_N(d) \to (1-p_s)^d$, and

$$P\left(\widetilde{D}_{ri}^s \geq 1 \mid D_{ri}^s \geq 1, D_{ri}^s = d\right) \to 1 - (1-p_s)^d.$$

Let $D_{N(rs)}^{max} = \max\{d : \Pr(D_{ri}^s = d)\}$. It follows that we have

$$\begin{aligned}
\pi_{rs} &= P\left(\widetilde{D}_{ri}^s \geq 1 \mid D_{ri}^s \geq 1\right) \\
&= \sum_{d=1}^{D_{N(rs)}^{max}} P\left(\widetilde{D}_{ri}^s \geq 1 \mid D_{ri}^s = d\right) P(D_{ri}^s = d \mid D_{ri}^s \geq 1) \\
&\to \sum_{d=1}^{D_{N(rs)}^{max}} (1 - (1-p_s)^d) P(D_{ri}^s = d \mid D_{ri}^s \geq 1)
\end{aligned}$$

Let $D_{n(rs)}^{max} = \max\{d : \Pr(\widetilde{D}_{ri}^s = d)\}$. Similarly, if we treat S_n as the population and S_m as the sample from S_n of size pn, we can do the same as above and get the following:

$$\begin{aligned}
\widetilde{\pi}_{rs} &= \Pr(\widetilde{\widetilde{\widetilde{D}^s_{rs}}} \geq 1 \mid \widetilde{D}^s_{ri} \geq 1) \\
&= \sum_{d=1}^{D^{max}_{n(rs)}} P(\widetilde{\widetilde{\widetilde{D}^s_{rs}}} \geq 1 \mid \widetilde{D}^s_{ri} = d) P(\widetilde{D}^s_{ri} = d \mid \widetilde{D}^s_{ri} \geq 1) \\
&\to \sum_{d=1}^{D^{max}_{N(rs)}} (1-(1-p_s)^d) P(D^s_{ri} = d \mid D^s_{ri} \geq 1)
\end{aligned}$$

Thus, $\widetilde{\pi}_{rs}$ is a consistent estimator for π_{rs}.

■

A.2 Derivation of Consistent Estimators for $\mathbf{h}_{rs1}(\mathbf{y}_{ki})$

A.2.1 A consistent estimator for $h^1_{rs1}(\mathbf{y}_{ki})$

To find $\widehat{h}^1_{rs1}(\mathbf{y}_{ki})$, a consistent estimator for $h^1_{rs1}(\mathbf{y}_{ki})$, we first note that

$$\begin{aligned}
h^1_{rs1}(\mathbf{y}_{ki}) &= E\left[\widehat{\gamma}_{rs1}(\mathbf{y}_{r1}, \mathbf{y}_{r2}, \ldots, \mathbf{y}_{rn_r}; \mathbf{y}_{s1}, \mathbf{y}_{s2}, \ldots, \mathbf{y}_{sn_s}) \mid \mathbf{y}_{ki}\right] \\
&= \binom{n_r}{m_r}^{-1} \binom{n_s}{m_s}^{-1} \sum_{S_{m(r)} \in C^{n(r)}_{m(r)}} \sum_{S_{m(s)} \in C^{n(s)}_{m(s)}} \left\{ \frac{1}{m_r} \sum_{j=1}^{m_r} E\left[\widetilde{v}^s_{ri} \mid \mathbf{y}_{ki}\right] \right\} \\
&= \binom{n_r}{m_r}^{-1} \binom{n_s}{m_s}^{-1} \sum_{S_{m(r)} \in C^{n(r)}_{m(r)}} \sum_{S_{m(s)} \in C^{n(s)}_{m(s)}} G^1_{ki},
\end{aligned}$$

where

$$G^1_{ki} = G^1_{ki}\left(S_{m(r)}, S_{m(s)}\right) = \frac{1}{m_r} \sum_{j=1}^{m_r} E\left[\widetilde{v}^s_{ri} \mid \mathbf{y}_{ki}\right].$$

If $k = r$ and $\mathbf{y}_{ri} \in S_{m(r)}$, then

$$G^1_{ki} = \frac{1}{m_r}\left[\widetilde{v}^s_{ri} + (m_r - 1)\gamma_{rs1}\right].$$

Otherwise,

$$G^{1}_{ki} = \gamma_{rs1}.$$

Therefore, we have $h^{1}_{rs1}(\mathbf{y}_{si}) = \gamma_{rs1}$, which implies

$$\widehat{h}^{1}_{rs1}(\mathbf{y}_{si}) = \widehat{\gamma}_{rs1}.$$

For $h^{1}_{rs1}(\mathbf{y}_{ri})$, we have there are $\binom{n_r-1}{m_r-1}$ subsets of size m_r from $S_{n(r)}$ that contain $\mathbf{y}_{ri}$ and the remaining $\binom{n_r-1}{m_r}$ subsets of $S_{n(r)}$ that do not contain $\mathbf{y}_{ri}$. It follows that we derive the explicit form of $h^{1}_{rs1}(\mathbf{y}_{ri})$:

$$\begin{aligned}
h^{1}_{rs1}(\mathbf{y}_{ri}) &= \binom{n_r}{m_r}^{-1}\binom{n_s}{m_s}^{-1}\left[\sum_{S_{m(r)}\in C^{n(r)}_{m(r)}:\mathbf{y}_{ri}\in S_{m(r)}}\ \sum_{S_{m(s)}\in C^{n(s)}_{m(s)}} G^{1}_{ri} + \sum_{S_{m(r)}\in C^{n(r)}_{m(r)}:\mathbf{y}_{ri}\notin S_{m(r)}}\ \sum_{S_{m(s)}\in C^{n(s)}_{m(s)}} G^{1}_{ri}\right]\\
&= \binom{n_r}{m_r}^{-1}\binom{n_s}{m_s}^{-1}\sum_{S_{m(r)}\in C^{n(r)}_{m(r)}:\mathbf{y}_{ri}\in S_{m(r)}}\ \sum_{S_{m(s)}\in C^{n(s)}_{m(s)}} G^{1}_{ri} + \binom{n_r}{m_r}^{-1}\binom{n_r-1}{m_r}\gamma_1\\
&= \binom{n_r}{m_r}^{-1}\binom{n_s}{m_s}^{-1}\sum_{S_{m(r)}\in C^{n(r)}_{m(r)}:\mathbf{y}_{ri}\in S_{m(r)}}\ \sum_{S_{m(s)}\in C^{n(s)}_{m(s)}} \frac{1}{m_r}\left[\widetilde{v}^{s}_{ri} + (m_r-1)\gamma_1\right] + \frac{n_r-m_r}{n_r}\gamma_1\\
&= \binom{n_r}{m_r}^{-1}\binom{n_s}{m_s}^{-1}\sum_{S_{m(r)}\in C^{n(r)}_{m(r)}:\mathbf{y}_{ri}\in S_{m(r)}}\ \sum_{S_{m(s)}\in C^{n(s)}_{m(s)}} \frac{1}{m_r}\widetilde{v}^{s}_{ri} + \binom{n_r}{m_r}^{-1}\binom{n_r-1}{m_r-1}\frac{m_r-1}{m_r}\gamma_1 +\\
&\quad \frac{n_r-m_r}{n_r}\gamma_1\\
&= \binom{n_r}{m_r}^{-1}\binom{n_s}{m_s}^{-1}\sum_{S_{m(r)}\in C^{n(r)}_{m(r)}:\mathbf{y}_{ri}\in S_{m(r)}}\ \sum_{S_{m(s)}\in C^{n(s)}_{m(s)}} \frac{1}{m_r}\widetilde{v}^{s}_{ri} + \frac{n_r-1}{n_r}\gamma_1\\
&= \left(\frac{1}{m_r}\binom{n_r}{m_r}^{-1}\binom{n_r-1}{m_r-1}\right)\binom{n_r-1}{m_r-1}^{-1}\sum_{S_{m(r)}\in C^{n(r)}_{m(r)}:\mathbf{y}_{ri}\in S_{m(r)}}\binom{n_s}{m_s}^{-1}\sum_{S_{m(s)}\in C^{n(s)}_{m(s)}}\widetilde{v}^{s}_{ri} +\\
&\quad \frac{n_r-1}{n_r}\gamma_1\\
&= \frac{1}{n_r}\binom{n_r-1}{m_r-1}^{-1}\sum_{S_{m(r)}\in C^{n(r)}_{m(r)}:\mathbf{y}_{ri}\in S_{m(r)}}\binom{n_s}{m_s}^{-1}\sum_{S_{m(s)}\in C^{n(s)}_{m(s)}}\widetilde{v}^{s}_{ri} + \frac{n_r-1}{n_r}\gamma_1
\end{aligned}$$

Therefore,

$$\widehat{h}^{1}_{rs1}(\mathbf{y}_{ri}) = \frac{1}{n_r}\binom{n_r-1}{m_r-1}^{-1}\sum_{S_{m(r)}\in C^{n(r)}_{m(r)}:\mathbf{y}_{ri}\in S_{m(r)}}\binom{n_s}{m_s}^{-1}\sum_{S_{m(s)}\in C^{n(s)}_{m(s)}}\widetilde{v}^{s}_{ri}+\frac{n_r-1}{n_r}\widehat{\gamma}_1$$

A.2.2 A consistent estimator for $h^2_{rs1}(\mathbf{y}_{ki})$

To find $\widehat{h}^2_{rs1}(\mathbf{y}_{ki})$, a consistent estimator for $h^2_{rs1}(\mathbf{y}_{ki})$, we first note that

$$\begin{aligned}
h^2_{rs1}(\mathbf{y}_{ki}) &= E\left[\widehat{\gamma}_{rs2}(\mathbf{y}_{r1},\ldots,\mathbf{y}_{rn_r};\mathbf{y}_{s1},\ldots,\mathbf{y}_{sn_s}) \mid \mathbf{y}_{ki}\right] \\
&= \binom{n_r}{m_r}^{-1}\binom{n_s}{m_s}^{-1}\sum_{S_{m(r)}\in C^{n(r)}_{m(r)}}\sum_{S_{m(s)}\in C^{n(s)}_{m(s)}}\left\{\frac{1}{m_r}\sum_{j=1}^{m_r}E[\widetilde{u}^{s}_{ri}\mid \mathbf{y}_{ki}]\right\} \\
&= \binom{n_r}{m_r}^{-1}\binom{n_s}{m_s}^{-1}\sum_{S_{m(r)}\in C^{n(r)}_{m(r)}}\sum_{S_{m(s)}\in C^{n(s)}_{m(s)}}G^2_{ki},
\end{aligned}$$

where

$$G^2_{ki} = G^2_{ki}(S_{m(r)},S_{n(s)}) = \frac{1}{m_r}\sum_{j=1}^{m_r}E[\widetilde{u}^{s}_{ri}\mid \mathbf{y}_{ki}].$$

If $k=r$ and $\mathbf{y}_{ri}\in S_{m_r}$, then

$$G^2_{ki} = \frac{1}{m_r}\left[\widetilde{u}^{s}_{ri}+(m_r-1)\,\gamma_{rs2}\right].$$

Otherwise,

$$G^2_{ki} = \gamma_{rs2}.$$

Therefore, we have $h^2_{rs1}(\mathbf{y}_{si}) = \gamma_{rs2}$, which implies a consistent estimator $\widehat{h}^2_{rs1}(\mathbf{y}_{si})$ for $h^2_{rs1}(\mathbf{y}_{si})$ is defined as follows:

$$\widehat{h}^2_{rs1}(\mathbf{y}_{si}) = \widehat{\gamma}_{rs2}.$$

Further, as $\widetilde{u}^{s}_{ri}$ is a connection indicator with respect to $S_{n(s)}$, as long as we know $\mathbf{y}_{ri}\in S_{m(r)}$ we have $\widetilde{u}^{s}_{ri}$ does not depend on S_m, i.e., if $S_{m(r)}$ and $S'_{m(r)}$ are both subsamples of size m_r from $S_{n(r)}$

that contains $\mathbf{y}_{ri} \in S_{m(r)}$, then

$$G_{ki}^{2}(S_{m(r)}, S_{n(s)}) = G_{ki}^{2}(S_{m(r)}, S_{n(s)}).$$

For $h_{rs1}^{2}(\mathbf{y}_{ri})$, we have $\binom{n_r-1}{m_r-1}$ subsets of size m_r from S_{n_r} that contain $\mathbf{y}_{ri}$ and there are the remaining $\binom{n_r-1}{m_r}$ subsets of size m_r from S_{n_r} that do not contain $\mathbf{y}_{ri}$. Fix S'_{m_r} to be any subset of size m_r from S_{n_r} that contain $\mathbf{y}_{ri}$. Thus,

$$\begin{aligned}
h_{rs1}^{2}(\mathbf{y}_{ri}) &= \binom{n_r}{m_r}^{-1}\binom{n_s}{m_s}^{-1}\left[\sum_{S_{m(r)} \in C_{m(r)}^{n(r)} : \mathbf{y}_{ri} \in S_{m(r)}} \; \sum_{S_{m(s)} \in C_{m(s)}^{n(s)}} G_{ri}^{2} + \sum_{S_{m(r)} \in C_{m(r)}^{n(r)} : \mathbf{y}_{ri} \notin S_{m(r)}} \; \sum_{S_{m(s)} \in C_{m(s)}^{n(s)}} G_{ri}^{2}\right] \\
&= \binom{n_r}{m_r}^{-1}\binom{n_r-1}{m_r-1} G_{ri}^{2}(S_{m'_r}, S_{n(s)}) + \binom{n_r}{m_r}^{-1}\binom{n_r-1}{m_r}\gamma_{rs2} \\
&= \frac{m_r}{n_r} G_{ri}^{2}(S_{m'_r}, S_{n(s)}) + \frac{n_r-m_r}{n_r}\gamma_{rs2} \\
&= \frac{1}{n_r}\left[\widetilde{u}_{ri}^{s} + (m_r-1)\gamma_{rs2}\right] + \frac{n_r-m_r}{n_r}\gamma_{rs2} \\
&= \frac{\widetilde{u}_{ri}^{s}}{n_r} + \frac{n_r-1}{n_r}\gamma_{rs2}
\end{aligned}$$

Therefore,

$$\widehat{h}_{rs1}^{2}(\mathbf{y}_{ri}) = \frac{\widetilde{u}_{ri}^{s}}{n_r} + \frac{n_r-1}{n_r}\widehat{\gamma}_{rs2}$$

A.2.3 A consistent estimator for $h_{rs1}^{3}(\mathbf{y}_{ki})$

To find $\widehat{h}_{rs1}^{3}(\mathbf{y}_{ki})$, a consistent estimator for $h_{1}^{3}(\mathbf{y}_{i})$, we first note that

$$\begin{aligned} h_1^3(\mathbf{y}_{ki}) &= E[\widehat{\gamma}_{rs3}(\mathbf{y}_{r1},\ldots,\mathbf{y}_{rn_r};\mathbf{y}_{s1},\ldots,\mathbf{y}_{sn_s}) \mid \mathbf{y}_{ki}] \\ &= \frac{1}{n_r}\sum_{j=1}^{n_r} E[v_{rj}^s = 1 \mid \mathbf{y}_{ki}] \\ &= \begin{cases} \frac{1}{n_r}[v_{ni}^{rs} + (n_r-1)\gamma_{rs3}], k = r \\ \gamma_{rs3}, \qquad k = s \end{cases} \end{aligned}$$

If $k = r$ and $j = i$, then

$$E[v_{rj}^s \mid \mathbf{y}_{ki}] = v_{ri}^s.$$

Otherwise,

$$E[v_{rj}^s] = \gamma_{rs3}.$$

Therefore,

$$\widehat{h}_1^3(\mathbf{y}_{ki}) = \begin{cases} \frac{v_{ni}^{rs}}{n_r} + \frac{n_r-1}{n_r}\gamma_{rs3}, k = r \\ \gamma_{rs3}, \qquad k = s \end{cases}$$

Thus,

$$\widehat{h}_1^3(\mathbf{y}_{ki}) = \begin{cases} \frac{v_{ni}^{rs}}{n_r} + \frac{n_r-1}{n_r}\widehat{\gamma}_{rs3}, k = r \\ \widehat{\gamma}_{rs3}, \qquad k = s \end{cases}$$

A.3 A Consistent Estimator for the Variance of $\widetilde{\theta}_{rs}$

Note that

$$\widetilde{\theta}_{rs} = \widetilde{\theta}_{rs}(\mathbf{y}_{r1},\ldots,\mathbf{y}_{rn_r};\mathbf{y}_{s1},\ldots,\mathbf{y}_{sn_s}) = \frac{1}{n_r}\sum_{i=1}^{n_r} v_{ri}^{s}.$$

We have that the arguments of $\widetilde{\theta}_{rs}$ are symmetric when when permuted with respect to each group and that

$$E(\widetilde{\theta}_{rs}) = \pi_{rs}\theta.$$

Thus, $\widetilde{\theta}_{rs}$ is a U-Statistic for $\pi_{rs}\theta$. Note that we denote $\widetilde{\theta}_{rs}$ and $\pi_{rs}\theta$ as $\widehat{\gamma}_3$ and γ_3, respectively. Let

$$\begin{aligned}
h_{rs1}^{3} &= E(h_{rs}^{3}(\mathbf{y}_{r1},\ldots,\mathbf{y}_{rn_r};\mathbf{y}_{s1},\ldots,\mathbf{y}_{sn_s})\,|\mathbf{y}_{ki}),\\
\widetilde{h}_{rs1}^{3}(\mathbf{y}_i) &= h_{rs1}^{3}(\mathbf{y}_i) - \gamma_{rs3}\\
\sigma_{h(3)}^{2} &= Var(\widetilde{h}_{rs1}^{3}(\mathbf{y}_{ki}))
\end{aligned}$$

By [13], it follows that

$$\sqrt{n_{rs}}\,(\widehat{\gamma}_{rs3} - \gamma_{rs3}) \rightarrow_d N\left(0, \sigma_{\gamma(3)}^{2} = \rho_r^2 n_r^2 \sigma_{r3}^2 + \rho_s^2 n_s^2 \sigma_{s3}^2\right).$$

where $\rho_k^2 = \lim_{n_{rs}\rightarrow\infty}\frac{n_{rs}}{n_k}$ and

$$n_{rs} = \begin{cases} n_r & r = s \\ n_r + n_s & r \neq s \end{cases}.$$

A consistent estimate of $\sigma^2_{\gamma(3)}$ is given by:

$$\widehat{\sigma}^2_{\gamma(3)} = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} \left(\widehat{h}^3_{rs1}\left(\mathbf{y}_{ki}\right) - \widehat{\gamma}_{rs3} \right)^2 ,$$

where $\widehat{h}^3_{rs1}\left(\mathbf{y}_{ki}\right)$ denotes a consistent estimator for $h^3_{rs1}\left(\mathbf{y}_{ki}\right)$.

A.4 Comparison of Adjusted and Unadjusted Probabilities of Linkage

A.4.1 Estimates for Probabilities of Linkage

Table A.1. Unadjusted (U) and adjusted (A) linkage rates for the communities investigated from the BCPP.

	Gumare		Maunatlala		Mmankgodi		Mmathethe		Ramokgonami		Shakawe	
	U	A	U	A	U	A	U	A	U	A	U	A
Gumare	0.26	0.54	0.08	0.11	0.06	0.19	0.12	0.22	0.07	0.09	0.131	0.24
Maunatlala	0.09	0.19	0.17	0.25	0.04	0.10	0.10	0.19	0.08	0.13	0.04	0.08
Mmankgodi	0.06	0.13	0.03	0.04	0.10	0.32	0.09	0.18	0.04	0.08	0.03	0.06
Mmathethe	0.05	0.13	0.07	0.12	0.04	0.11	0.26	0.47	0.03	0.05	0.05	0.08
Ramokgonami	0.06	0.13	0.07	0.12	0.03	0.10	0.08	0.15	0.23	0.37	0.03	0.06
Shakawe	0.10	0.26	0.02	0.03	0.02	0.06	0.06	0.12	0.03	0.06	0.22	0.37

A.4.2 Standard Error of Probabilities of Linkage

Table A.2. Standard errors of unadjusted (U) and adjusted (A) linkage rates for the communities investigated from the BCPP.

	Gumare		Maunatlala		Mmankgodi		Mmathethe		Ramokgonami		Shakawe	
	U	A	U	A	U	A	U	A	U	A	U	A
Gumare	0.13	0.09	0.07	0.02	0.02	0.07	0.02	0.04	0.02	0.02	0.20	0.04
Maunatlala	0.02	0.05	0.03	0.03	0.03	0.04	0.08	0.07	0.01	0.04	0.02	0.03
Mmankgodi	0.17	0.05	0.01	0.02	0.03	0.08	0.03	0.04	0.02	0.03	0.01	0.02
Mmathethe	0.16	0.04	0.04	0.03	0.01	0.04	0.09	0.07	0.04	0.02	0.06	0.02
Ramokgonami	0.02	0.03	0.04	0.05	0.03	0.03	0.08	0.04	0.09	0.04	0.04	0.03
Shakawe	0.03	0.05	0.02	0.03	0.02	0.03	0.07	0.05	0.04	0.03	0.34	0.04

Bibliography

[1] Jeannette L Aldous, Sergei Kosakovsky Pond, Art Poon, Sonia Jain, Huifang Qin, James S Kahn, Mari Kitahata, Benigno Rodriguez, Ann M Dennis, Stephen L Boswell, et al. Characterizing hiv transmission networks across the united states. *Clinical Infectious Diseases*, 55(8):1135–1143, 2012.

[2] Bluma G Brenner, Michel Roger, Daniela D Moisi, Maureen Oliveira, Isabelle Hardy, Reuven Turgel, Hugues Charest, Jean-Pierre Routy, Mark A Wainberg, et al. Transmission networks of drug resistance acquired in primary/early stage hiv infection. *AIDS (London, England)*, 22(18):2509, 2008.

[3] Nicole Bohme Carnegie, Rui Wang, Vladimir Novitsky, and Victor De Gruttola. Linkage of viral sequences among hiv-infected village residents in botswana: estimation of linkage rates in the presence of missing data. *PLoS computational biology*, 1 2014.

[4] Kristen Chalmet, Delfien Staelens, Stijn Blot, Sylvie Dinakis, Jolanda Pelgrom, Jean Plum, Dirk Vogelaers, Linos Vandekerckhove, and Chris Verhofstede. Epidemiological study of phylogenetic transmission clusters in a local hiv-1 epidemic reveals distinct differences between subtype b and non-b infections. *BMC infectious diseases*, 10(1):1–9, 2010.

[5] Sharoda Dasgupta, Anne Marie France, Mary-Grace Brandt, Jennifer Reuer, Tianchi Zhang, Nivedha Panneer, Angela L Hernandez, and Alexandra M Oster. Estimating effects of hiv sequencing data completeness on transmission network patterns and detection of growing hiv transmission clusters, 4 2019.

[6] Malia Duffy, Caitlin Madevu-Matson, Jessica E Posner, Hana Zwick, Melissa Sharer, and Antonia M Powell. Systematic review: Development of a person-centered care framework within the context of hiv treatment settings in sub-saharan africa. *Tropical Medicine & International Health*, 27(5):479–493, 2022.

[7] Abdulrahman M El-Sayed, Peter Scarborough, Lars Seemann, and Sandro Galea. Social network analysis and agent-based modeling in social epidemiology. *Epidemiologic Perspectives & Innovations*, 9(1):1, 2012.

[8] Anthony S. Fauci, Robert R. Redfield, George Sigounas, Michael D. Weahkee, and Brett P. Giroir. Ending the HIV Epidemic: A Plan for the United States. *JAMA*, 321(9):844–845, 03 2019.

[9] Tendani Gaolathe, Kathleen E Wirth, Molly Pretorius Holme, Joseph Makhema, Sikhulile Moyo, Unoda Chakalisa, Etienne Kadima Yankinda, Quanhong Lei, Mompati Mmalane, Vlad Novitsky, et al. Botswana's progress toward achieving the 2020 unaids 90-90-90 antiretroviral therapy and virological suppression goals: a population-based survey. *The lancet HIV*, 3(5):e221–e230, 2016.

[10] Gareth J Hughes, Esther Fearnhill, David Dunn, Samantha J Lycett, Andrew Rambaut, Andrew J Leigh Brown, and UK HIV Drug Resistance Collaboration. Molecular phylodynamics of the heterosexual hiv epidemic in the united kingdom. *PLoS pathogens*, 5(9):e1000590, 2009.

[11] Olivia Keiser, Patrick Taffé, Marcel Zwahlen, Manuel Battegay, Enos Bernasconi, Rainer Weber, Martin Rickenbach, Swiss HIV Cohort Study, et al. All cause mortality in the swiss hiv cohort study from 1990 to 2001 in comparison with the swiss population. *Aids*, 18(13):1835–1843, 2004.

[12] Eric D Kolaczyk and Gábor Csárdi. *Statistical analysis of network data with R*, volume 65. Springer, 2014.

[13] Jeanne Kowalski and Xin M Tu. Modern applied u-statistics, Apr 2007.

[14] Andrew J Leigh Brown, Samantha J Lycett, Lucy Weinert, Gareth J Hughes, Esther Fearnhill, and David T Dunn. Transmission network parameters estimated from hiv sequences for a nationwide epidemic. *Journal of Infectious Diseases*, 204(9):1463–1469, 2011.

[15] Tuo Lin, Tian Chen, Jinyuan Liu, and Xin M Tu. Extending the mann-whitney-wilcoxon rank sum test to survey data for comparing mean ranks. *Statistics in Medicine*, 40(7):1705–1717, 2021.

[16] J Liu, Xinlian Zhang, T Chen, T Wu, T Lin, L Jiang, S Lang, L Liu, L Natarajan, JX Tu, et al. A semiparametric model for between-subject attributes: Applications to beta-diversity of microbiome data. *Biometrics*, 2021.

[17] Shelley H Liu, Gabriel Erion, Vladimir Novitsky, and Victor De Gruttola. Viral genetic linkage analysis in the presence of missing data, 8 2015.

[18] Mark N Lurie and Samantha Rosenthal. Concurrent partnerships as a driver of the hiv epidemic in sub-saharan africa? the evidence is limited. *AIDS and Behavior*, 14(1):17–24, 2010.

[19] Lerato E Magosi, Yinfeng Zhang, Tanya Golubchik, Victor DeGruttola, Eric Tchetgen Tchetgen, Vladimir Novitsky, Janet Moore, Pam Bachanas, Tebogo Segolodi, Refeletswe Lebelonyane, et al. Deep-sequence phylogenetics to quantify patterns of hiv transmission in the context of a universal testing and treatment trial–bcpp/ya tsie trial. *Elife*, 11:e72657, 2022.

[20] Timothy L Mah and Daniel T Halperin. Concurrent sexual partnerships and the hiv epidemics in africa: evidence to move forward. *AIDS and Behavior*, 14(1):11–16, 2010.

[21] Vlad Novitsky, Melissa Zahralban-Steele, Sikhulile Moyo, Tapiwa Nkhisang, Dorcas Maruapula, Mary Fran McLane, Jean Leidner, Kara Bennett, Kathleen E Wirth, et al. Mapping of hiv-1c transmission networks reveals extensive spread of viral lineages across villages in botswana treatment-as-prevention trial. *The Journal of infectious diseases*, 222(10):1670–1680, 2020.

[22] Frank Ove. Statistical inference in graphs. 1971.

[23] Donald B Rubin. Multiple imputation after 18+ years. *Journal of the American statistical Association*, 91(434):473–489, 1996.

[24] Marko Sarstedt, Paul Bengart, Abdel Monim Shaltoni, and Sebastian Lehmann. The use of sampling methods in advertising research: A gap between theory and practice. *International Journal of Advertising*, 37(4):650–663, 2018.

[25] Michael PH Stumpf, Carsten Wiuf, and Robert M May. Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proceedings of the National Academy of Sciences*, 102(12):4221–4224, 2005.

[26] Wan Tang, Hua He, and Xin M Tu. *Applied categorical and count data analysis*. CRC Press, 2012.

[27] Joel O Wertheim, Sergei L Kosakovsky Pond, Lisa A Forgione, Sanjay R Mehta, Ben Murrell, Sharmila Shah, Davey M Smith, Konrad Scheffler, and Lucia V Torian. Social and genetic networks of hiv-1 transmission in new york city. *PLoS pathogens*, 13(1):e1006000, 2017.

[28] Joel O Wertheim, Andrew J Leigh Brown, N Lance Hepler, Sanjay R Mehta, Douglas D Richman, Davey M Smith, and Sergei L Kosakovsky Pond. The global transmission network of hiv-1. *The Journal of infectious diseases*, 209(2):304–313, 2014.

[29] S Yerly, S Vora, P Rizzardi, J-P Chave, PL Vernazza, Markus Flepp, A Telenti, M Battegay, A-L Veuthey, J-P Bru, et al. Acute hiv infection: impact on the spread of hiv and transmission of drug resistance. *Aids*, 15(17):2287–2292, 2001.

[30] Q Yu, W Tang, J Kowalski, and XM Tu. Multivariate u-statistics: a tutorial with applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(5):457–471, 2011.