University of Rhode Island

DigitalCommons@URI

Open Access Master's Theses

2020

EVALUATING THE EFFECTS OF MEDICATIONS FOR OPIOID USE DISORDER ON HIV RISK BEHAVIORS AMONG A NETWORK OF PEOPLE WHO INJECT DRUGS

Joseph Puleo

University of Rhode Island, jpuleo@uri.edu

Follow this and additional works at: <https://digitalcommons.uri.edu/theses>

Recommended Citation

Puleo, Joseph, "EVALUATING THE EFFECTS OF MEDICATIONS FOR OPIOID USE DISORDER ON HIV RISK BEHAVIORS AMONG A NETWORK OF PEOPLE WHO INJECT DRUGS" (2020). *Open Access Master's Theses*. Paper 1866.

<https://digitalcommons.uri.edu/theses/1866>

This Thesis is brought to you for free and open access by DigitalCommons@URI. It has been accepted for inclusion in Open Access Master's Theses by an authorized administrator of DigitalCommons@URI. For more information, please contact digitalcommons-group@uri.edu.

EVALUATING THE EFFECTS OF MEDICATIONS FOR OPIOID USE DISORDER
ON HIV RISK BEHAVIORS AMONG A NETWORK
OF PEOPLE WHO INJECT DRUGS

BY
JOSEPH PULEO

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
STATISTICS

UNIVERSITY OF RHODE ISLAND
2020

ABSTRACT

People who inject drugs (PWID) are a population with increased HIV risk due in part to sharing drug injection equipment. In networks of people connected through risk behavior, individuals can exert influence on each other. Evidence suggests that receiving medications for opioid use disorder (MOUD) can reduce HIV risk behaviors. However, limited studies have been conducted to determine if there are direct effects (among the participants treated) or disseminated effects (among participants who were not treated themselves, but shared connections with those treated) of MOUD among networks of PWID.

The main goal of this thesis is to evaluate if receiving MOUD causes reductions in HIV risk behavior and if being socially connected to people that receive MOUD also causes reductions in HIV risk behavior. To achieve this goal we analyzed a network of 246 PWID from the Transmission Reduction Intervention Project conducted in Athens, Greece 2013 to 2015. For our analysis we utilized methods for causal inference in the presence of dissemination in combination with methods for missing data imputation. To quantify the causal effects of MOUD on subsequent HIV risk behaviors, we employed a group-level inverse probability weighted approach to adjust for confounding. Specifically, we identified communities of participants in the network using modularity-based community detection. Two community detection methods, a "node-moving" algorithm and a "spectral" algorithm, were used to detect communities for a sensitivity analysis. We employed a finite sample correction for the standard errors of the estimators to account for the relatively small number of communities (20 to 21) in the network.

To impute missing data for covariates, we used a non-parametric random forest method. To impute missing outcome data we first assumed a missing at random (MAR) mechanism. We fit a longitudinal model with a random effect for the community to estimate the posterior distribution of the outcome. Based on this distribution, missing outcomes were imputed and a sensitivity analysis was conducted to account for the more realistic assump-

tion that the outcomes were missing *not* at random (MNAR).

The results of our analysis showed significant disseminated effects of MOUD on HIV risk behaviors. Although the magnitude of these effects was sensitive at times to the community detection method, the conclusion of the analysis remained unchanged. Analysis of the community structure suggests alternative approaches to identifying interference sets (i.e., communities or clusters) that consider the nearest-neighbors or the full network should be considered. A *nearest-neighbors* approach would define interference sets uniquely for each individual in the network. Interference sets would consist of the subjects to whom an individual is directly connected. Sensitivity analysis of the missing data mechanism shows strong evidence of significant causal effects under MAR and MNAR assumptions.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my advisors Drs. Ashley Buchanan and Natallia Katenka. I thank you both for all your help and support throughout this process of researching and writing my thesis. I appreciate your kindness, patience, and your encouragement.

Thank you to all my committee members and chair: Drs. Jing Wu, Stephen Kogut, and Ami Vyas for help and feedback. Thank you to Drs. Samuel R. Friedman, Georgios Nikolopoulos, and M. Elizabeth Halloran for your help and feedback and for allowing me to use the data.

Thank you to my parents, my sister and my brother-in-law for your encouragement and support throughout this experience.

Finally I would like to give a special thanks to Catherine Robinson, who initially encouraged me to pursue my degree in Statistics. Thank you as well to Drs. Prabhani Kurupmullage Don and Chad Estabrooks for your support at the start of my time at URI. Thank you all so much.

TABLE OF CONTENTS

ABSTRACT	ii
ACKNOWLEDGEMENTS	iv
TABLE OF CONTENTS	v
LIST OF TABLES	viii
LIST OF FIGURES	xiv
1 INTRODUCTION	1
1.1 People Who Inject Drugs	1
1.2 Dissemination	1
1.3 Transmission Reduction Intervention Project (TRIP)	2
1.4 Causal Inference in Presence of Dissemination	6
1.5 Community Detection	9
1.6 Finite Sample Correction	11
1.7 Missing Data	11
LIST OF REFERENCES	14
2 METHODOLOGY	19
2.1 Causal inference in the Presence of Dissemination	19
2.2 Community Detection	25

2.3	Finite Sample Correction	29
2.4	Missing Data	32
	LIST OF REFERENCES	38
3	RESULTS	41
3.1	Effect of MOUD on HIV Risk Behavior	41
3.2	Missing Data Results	46
3.3	Community Structure	50
3.4	Finite Sample Correction Confidence Intervals	53
3.5	Random Effects Normality	55
	LIST OF REFERENCES	56
4	DISCUSSION	57
4.1	Causal Estimation	57
4.2	Community Detection	58
4.3	Finite Sample Correction	59
	LIST OF REFERENCES	61
5	DIRECTIONS FOR FUTURE RESEARCH	62
6	CONCLUSION	65
7	APPENDIX A - CONCEPTS FROM INFORMATION THEORY	66
	LIST OF REFERENCES	70
8	APPENDIX B - PERMUTATION RANDOMIZATION	71
	LIST OF REFERENCES	73
9	APPENDIX C - FEATURES OF COMMUNITIES	74

LIST OF REFERENCES	104
BIBLIOGRAPHY	105

LIST OF TABLES

1	Illustration of LOCF for participants with missing outcome data. Estimated averages at baseline are carried forward to follow-up for those with missing outcome data at follow-up. We therefore assume that average outcomes at follow-up, $\hat{P}$, are equal to the estimated average outcome at baseline, $\hat{P}^{(0)}$. .	36
2	Contingency table showing the conditional distributions (shown as percentages) of the covariates among the overall (marginal) study sample, among those prescribed MOUD at baseline, and among those who were not prescribed MOUD at baseline.	41
3	Estimated risk differences (RDs) with 95% CIs of MOUD (Received vs. Not Received) on likelihood of engaging in HIV risk behavior in TRIP (coverage of received), not adjusted for baseline outcome	43
4	Estimated risk differences (RDs) with 95% CIs of MOUD (Received vs. Not Received) on likelihood of engaging in HIV risk behavior in TRIP (coverage of received), adjusted for baseline outcome	44
5	Proportion of Missing Covariate Data and the number missing in TRIP. Covariates data were only approximately 2.5% missing.	47
6	The finite sample corrected margin of error for the 5 causal estimands under both community detection methods. Results are non-informative, however, should be interpreted with caution as the method was only developed for an experimental setting.	54

- 7 Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 1 in TRIP from the largest connected component under the **node-moving** algorithm. The lower section of the table includes measures of cohesion in the community. 75

- 8 Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 2 in TRIP from the largest connected component under the **node-moving** algorithm. The lower section of the table includes network information about the community. 76

- 9 Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 3 in TRIP from the largest connected component under the **node-moving** algorithm. The lower section of the table includes network information about the community. 77

- 10 Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 4 in TRIP from the largest connected component under the **node-moving** algorithm. The lower section of the table includes network information about the community. 78

- 11 Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 5 in TRIP from the largest connected component under the **node-moving** algorithm. The lower section of the table includes network information about the community. 79

- 12 Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 6 in TRIP from the largest connected component under the **node-moving** algorithm. The lower section of the table includes network information about the community. 80

- 13 Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 7 in TRIP from the largest connected component under the **node-moving** algorithm. The lower section of the table includes network information about the community. 81
- 14 Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 8 in TRIP from the largest connected component under the **node-moving** algorithm. The lower section of the table includes network information about the community. 82
- 15 Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 9 in TRIP from the largest connected component under the **node-moving** algorithm. The lower section of the table includes network information about the community. 83
- 16 Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 10 in TRIP from the largest connected component under the **node-moving** algorithm. The lower section of the table includes network information about the community. 84
- 17 Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 11 in TRIP from the largest connected component under the **node-moving** algorithm. The lower section of the table includes network information about the community. 85
- 18 Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 12 in TRIP from the largest connected component under the **node-moving** algorithm. The lower section of the table includes network information about the community. 86

19	Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 13 in TRIP from the largest connected component under the node-moving algorithm. The lower section of the table includes network information about the community.	87
20	Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 14 in TRIP from the largest connected component under the node-moving algorithm. The lower section of the table includes network information about the community.	88
21	Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 15 in TRIP from the largest connected component under the node-moving algorithm. The lower section of the table includes network information about the community.	89
22	Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 1 in TRIP from the largest connected component under the spectral algorithm. The lower section of the table includes network information about the community.	90
23	Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 2 in TRIP from the largest connected component under the spectral algorithm. The lower section of the table includes network information about the community.	91
24	Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 3 in TRIP from the largest connected component under the spectral algorithm. The lower section of the table includes network information about the community.	92

25	Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 4 in TRIP from the largest connected component under the spectral algorithm. The lower section of the table includes network information about the community.	93
26	Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 5 in TRIP from the largest connected component under the spectral algorithm. The lower section of the table includes network information about the community.	94
27	Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 6 in TRIP from the largest connected component under the spectral algorithm. The lower section of the table includes network information about the community.	95
28	Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 7 in TRIP from the largest connected component under the spectral algorithm. The lower section of the table includes network information about the community.	96
29	Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 8 in TRIP from the largest connected component under the spectral algorithm. The lower section of the table includes network information about the community.	97
30	Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 9 in TRIP from the largest connected component under the spectral algorithm. The lower section of the table includes network information about the community.	98

31	Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 10 in TRIP from the largest connected component under the spectral algorithm. The lower section of the table includes network information about the community.	99
32	Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 11 in TRIP from the largest connected component under the spectral algorithm. The lower section of the table includes network information about the community.	100
33	Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 12 in TRIP from the largest connected component under the spectral algorithm. The lower section of the table includes network information about the community.	101
34	Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 13 in TRIP from the largest connected component under the spectral algorithm. The lower section of the table includes network information about the community.	102
35	Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 13 in TRIP from the largest connected component under the spectral algorithm. The lower section of the table includes network information about the community.	103

LIST OF FIGURES

1	Schematic showing the recruitment of participants in the Transmission Reduction Intervention Project. Limiting antigen avidity assay normalized optical density (LAg ODn) ≤ 1.5 implies a recency window period of 130 days [31]. Source: Nikolopoulos et. al (2016)	3
2	The TRIP network illustrated as a graph. Red nodes represent individuals who received MOUD (exposed) at baseline, and blue nodes represent individuals who did not receive MOUD (non-exposed) at baseline.	5
3	Illustration showing the effect of removing connected individuals in a network based on missing data. Removing connected individuals alters the observed coverage levels for those that are not removed. Black nodes represent exposed participants, white nodes represent non-exposed participants, and the numbers in the nodes indicate the coverage level for that participant.	12
4	Illustration of the causal estimands: <i>overall effect</i> , <i>composite effect</i> , <i>indirect effect</i> , and the <i>direct effect</i> . Arrows connecting regions represent the contrasts being made between the average outcomes. [6].	23
5	An illustration of the four parts of the dataset used by missForest, where $X_{(6)}$ has missing data. White cells indicate $x_{mis}^{(6)}$, the second darkest cells indicate $y_{obs}^{(6)}$, the third darkest cells indicate $y_{mis}^{(6)}$, and the dark cells indicate $x_{obs}^{(6)}$. Source: Misztal, Malforzata (2013)	33

6	A frequency histogram of the values of $ RD_{sp} - RD_{nm} $, the difference in the magnitudes of the estimated risk differences under the spectral method (sp) and the node-moving method (nm) for the models that are <i>not</i> adjusted for the outcome at baseline. Higher frequency where $ RD_{sp} - RD_{nm} $ is positive indicates the spectral method generally resulting larger magnitudes of RD's.	45
7	A histogram of the values of $ RD_{sp} - RD_{nm} $, the difference in the magnitudes of the estimated risk differences under the spectral method (sp) and the node-moving method (nm) for the models that <i>are</i> adjusted for the outcome at baseline.	45
8	A histogram of the values of $ RD_{sp} - RD_{nm} $, the difference in the magnitudes of the estimated risk differences under the spectral method (sp) and the node-moving method (nm) for both the adjusted and unadjusted models.	46
9	Sensitivity plots for the Direct Effect under exposure policy 0.25 across the full range of ψ . The vertical axis indicates the estimated RD. The horizontal axis indicates ψ ranging from 1 (where all missing outcomes are assigned the value 0) to 0 (where all missing outcomes are assigned the value 1).	48
10	Sensitivity plots for the Direct Effect under exposure policy 0.75 across the full range of ψ . The vertical axis indicates the estimated RD. The horizontal axis indicates ψ ranging from 1 (where all missing outcomes are assigned the value 0) to 0 (where all missing outcomes are assigned the value 1).	48
11	Sensitivity plots for the Indirect Effect, $IE(0.75, 0.25)$, across the full range of ψ . The vertical axis indicates the estimated RD. The horizontal axis indicates ψ ranging from 1 (where all missing outcomes are assigned the value 0) to 0 (where all missing outcomes are assigned the value 1).	49

12	Sensitivity plots for the Composite Effect, $CE(0.75, 0.25)$, across the full range of ψ . The vertical axis indicates the estimated RD. The horizontal axis indicates ψ ranging from 1 (where all missing outcomes are assigned the value 0) to 0 (where all missing outcomes are assigned the value 1). . . .	49
13	Sensitivity plots for the Overall Effect, $OE(0.75, 0.25)$, across the full range of ψ . The vertical axis indicates the estimated RD. The horizontal axis indicates ψ ranging from 1 (where all missing outcomes are assigned the value 0) to 0 (where all missing outcomes are assigned the value 1). . . .	50
14	Community structure found using the node-moving algorithm (left) and the spectral algorithm (right). Nodes represent communities and edges represent edges between communities.	51
15	Perturbation plots. Robustness of community structure to changes in proportion of randomly reassigning edges. Gray represents the VOI band for the null network, and red represents the VOI band for the TRIP network . . .	52
16	Observed coverage levels for both the communities found in the node-moving algorithm (left) and the spectral algorithm (right) in TRIP.	54
17	VOI Curves for 6 simulated networks. Each point represents the average VOI for 100 perturbed graphs. Source: Karrer et. al. (2008)	69

1 INTRODUCTION

1.1 People Who Inject Drugs

People who inject drugs (PWID) are a population with a high risk for HIV infection due to the high prevalence of risky behaviors such as condomless sex and the sharing of drug injection equipment. Despite advances in HIV/AIDS treatment and care, HIV/AIDS prevalence remains a crisis among high risk and vulnerable populations such as PWID [9]. Barriers to the continuum of HIV care remain significantly high for this population due to many factors including stigma associated with HIV and drug use, incarceration, lack of housing, and socioeconomic factors [6, 13]. Research has shown that participants who are treated for their opioid use disorder (OUD) may be less likely to engage in drug injection behavior as well as sexual risk behaviors [3, 7, 14, 24]. PWID are also embedded in a network and are connected by social interaction and shared risk behaviors [11]. Considering social influence within networks can allow for better quantification of the causal effect of not only a person's exposure status (being prescribed MOUD or not) on that person's HIV risk behaviors, but also the effect of *other* individuals' exposures on that person's HIV risk behaviors.

1.2 Dissemination

Many public health interventions provide benefits that extend beyond those that receive it and impact people in close physical or social proximity to those who did not receive the intervention themselves [2]. Treating highly-connected individuals may help to ensure dissemination to individuals who themselves were unwilling or unable to engage with the intervention delivered by the study investigators and/or medical staff. Those who are treated for OUD are less likely to engage in injection behavior with their contacts, protecting those contacts from HIV risk behavior and HIV acquisition if they are HIV-negative [4,5]. Due to

the possibility of dissemination, intervening on highly connected individuals may be a more effective means of reducing risk behaviors, and thus the spread of HIV and other infectious diseases such as Hepatitis C. Valid estimation of dissemination thus calls for methods that consider the community structure and the shared influence among participants.

1.3 Transmission Reduction Intervention Project (TRIP)

For the analysis in this study, we used data from the TRIP collected from participants in Athens, Greece between 2013 and 2015 [32]. The study recruited adults with a heightened risk of HIV infection and members of their social networks to prevent or treat HIV. PWID are a population with a considerably high risk for HIV infection due to the high prevalence of risky sexual behaviors and the sharing of drug injection equipment. The study used a treatment as prevention (TasP) intervention, identifying and treating recently-infected persons rapidly after infection and then linking them to continued care for treatment [12, 32]. The primary aim of the TRIP study was to find ways to recruit recently-infected people and refer them to antiretroviral treatment while they were still highly infectious. The study also aimed to warn their network members to use extreme care to avoid risk for several months because there were recently-infected people in their networks [11]. Initial participants, known as "seeds," were referred to TRIP by collaborating HIV testing sites and were HIV-positive. Based on HIV testing results and history, seeds were categorized into two arms of the study: Recent Seeds (RS), and Control Seeds with Long-term HIV infection (LCS). RS were the intervention arm and included those who were recently-infected (in the 6 months prior to baseline) and therefore could be acutely infectious. LCS, the seeds who were newly diagnosed with HIV but not recently-infected, were the control arm [32]. To construct the HIV risk network, recruitment of the seeds' contacts was conducted using two-wave contact tracing. Network connections were based on social contact in the 6 months prior to baseline including sexual contact and/or shared drug use, identifying someone as a friend

or acquaintance, or frequenting the same venues.

Figure 1: Schematic showing the recruitment of participants in the Transmission Reduction Intervention Project. Limiting antigen avidity assay normalized optical density (LAg ODn) ≤ 1.5 implies a recency window period of 130 days [31].

Source: Nikolopoulos et. al (2016)

These contacts were then tested for HIV and interviewed to collect data on demographics, past HIV testing and treatment, risk behaviors, social connections, substance abuse treatment, use of pre-exposure prophylaxis (PrEP), access to care, HIV knowledge, and injection norms. Every HIV-positive contact who was recently infected (or "borderline recent", i.e., those who marginally failed to meet the operational definition of "recent") was then considered a new seed, and two-wave contact tracing ensued for them. After the initial two waves of contact tracing were done for the seeds, one-wave contact tracing was conducted for those newly recruited contacts that were long-term infected or HIV-negative. The study also recruited 79 initial participants who were HIV-negative, serving as a control group, and their contacts were not traced, although they were connected to the TRIP participants who were identified as contacts. Follow-up interviews were conducted approximately 6 months after each participant's baseline interview to collect data on behavioral changes since baseline, utilization of care through the TRIP study, mental health, attitudes

toward healthcare providers, and HIV status for those who were HIV-negative at baseline. The full network contains 356 participants with 829 dyadic connections. After removing 27 (7.6%) participants who were not injecting drugs at baseline and the 83 (23.3%) who did not share any edges with anyone else in the study, the network includes 246 vertices and 496 dyadic connections.

Many statistical methods rely on the assumption of independence of participants in the data and no interference. Interference occurs when the individuals' outcomes are affected by the treatment (or exposure assignment) of other individuals. Failing to account for the dependency structure among participants may lead to invalid standard errors, while ignoring dissemination or spillover effects leads biased point estimates which may underestimate the full effect of the intervention.

To represent the PWID network in TRIP, we use a mathematical graph representation. The network graph $G = (V, E)$ is comprised of a set of nodes (or vertices), V , where each node represents a study participant; and a set of edges, E , a set of connected pairs of nodes, where two participants were connected (i.e., share an edge) if they were identified as having had sexual contact and/or shared drug use, identified as a friend or acquaintance, or frequented the same venues. Figure 2 below illustrates the network with blue nodes indicating participants who were not exposed to MOUD at baseline and red nodes indicating participants who were exposed to MOUD at baseline.

Figure 2: The TRIP network illustrated as a graph. Red nodes represent individuals who received MOUD (exposed) at baseline, and blue nodes represent individuals who did not receive MOUD (non-exposed) at baseline.

The study population consists of 246 individuals that were injecting drugs and have either been referred to the TRIP study by another HIV treatment program and are socially connected to others in the TRIP study. In our study, the exposure is MOUD in the 6 months prior to the baseline interview, and the outcome is having reported engaging in any HIV risk behavior during the time between the baseline interview and the follow-up interview (where the follow-up interview took place approximately 6 months after the baseline interview). The exposure is represented as the binary variable A , where $A = 1$ if the participant was prescribed MOUD in the 6 months prior to the baseline interview and $A = 0$ if the participant was not. We considered the following HIV injection risk behaviors: 1.) shared a syringe that someone else had previously used to inject, 2.) gave someone a syringe to use that the interviewed participant already injected with, 3.) shared a cooker, filter or rinsed water that someone else had previously used to inject, 4.) gave someone a cooker, filter or rinsed water that the interviewed participant had previously used to inject, and 5.) backloaded (piggy-backed) to share injection drugs. The outcome in our study is a binary variable denoted Y , where $Y = 1$ if the participant reported having engaged in any of the

5 risk behaviors in the 6 months prior to follow-up, and $Y = 0$ if the participant reported having *not* engaged in any of the injection risk behaviors listed above in the 6 months prior to follow-up.

Covariates included in this study are HIV status, education level, employment status, marital status, homelessness, heroin use, being able to access medical care, and having been encouraged by their contacts to share a works, and the risk behavior at baseline. These were chosen on the basis of other studies on the causal effects of MOUD on HIV risk behaviors and were known or suspected risk factors for the outcome [27].

1.4 Causal Inference in Presence of Dissemination

One of the critical challenges in causal inference in networks is the violation of the "stable unit-treatment value" assumption, known as SUTVA. SUTVA implies that one participant's outcomes are unaffected by any other participant's treatment assignment (or exposure) [8]. This assumption also includes causal consistency, meaning there is intervention variation irrelevance, i.e., there are not multiple versions of treatment and the exposure maps onto unique potential outcomes [35]. In their discussion of the setting where SUTVA does not hold, Hudgens and Halloran point out that "when there is the dependency of events, interventions can have different types of effects." [17] They defined the *direct*, *indirect*, *composite*, and *overall* effects in the presence of partial interference [16, 19]. Partial interference implies that within communities (i.e., interference sets), a participant's outcome can be affected by their own exposure as well as the exposures of all participants in their community. We define a community (or cluster) to be a group of participants that are densely connected, with only sparser connections to other groups [23]. In a counterfactual scenario, the exposure policy α is the probability of any individual being independently assigned the exposure. Conditioning on covariates, one can assume a Bernoulli individual group counterfactual exposure policy, where exposure is assigned independently within

levels of the covariates according to, and possibly contrary to fact, exposure policy α [25]. The *direct effect* is a contrast between the average outcome when an individual is exposed and the average outcome when an individual is not exposed, both under the same counterfactual exposure policy. It measures the causal effect of receiving the intervention beyond being in a community with a certain proportion of participants who receive the intervention. The *indirect effect* (also called the "disseminated effect") measures the effect of the intervention on the outcomes of individuals who were not exposed themselves but shared risk connections with those who were, contrasting two different counterfactual exposure policies. The *composite effect* contrasts the outcomes of those who receive the intervention under one exposure policy, say α , to those who do not receive the intervention under a different exposure policy, say α' . The *composite effect* (or *total effect*) can be thought of as a maximal intervention impact contrasting the exposed and the unexposed under different exposure policies. The *overall effect* contrasts the population mean outcome under one exposure policy (α), to the population mean outcome under another exposure policy (α'), marginalizing over the individual-level exposures.

Two-stage randomization leads to these effects being readily identifiable from the observed data. In such an experiment, randomization takes place at the following two stages. The first is the community level, where exposure policies (i.e., treatment allocation strategies) are randomly assigned to groups or communities of highly connected individuals. The second stage is at the individual-level within communities, where exposure (MOUD) is assigned to individuals according to the exposure policy assigned to each group or community. Tchetgen Tchetgen and VanderWeele (TV) extended the methodology to allow for estimation of causal effects in a two-stage randomized design for observational studies [35]. In this thesis, the proposed causal inference methods rely on the notion of the potential outcome framework. Jerzy Neyman was the first to discuss this framework in his Master's Thesis in 1923 in the context of time-fixed treatments in a randomized experiment. This framework was further developed and extended to time-fixed observational studies by Don-

ald Rubin in 1974 [15, 28, 33, 34]. In the case of a binary treatment (or exposure) where SUTVA holds, each participant has two potential outcomes only, the outcome when exposed (possibly contrary to fact) and the outcome when not exposed (possibly contrary to fact). To measure the causal effect of the exposure for a single individual, we may consider subtracting the potential outcome under no exposure from the potential outcome under the exposure for that individual. If this quantity is non-zero, then there is a causal effect for that individual. If this quantity is zero, then there is no causal effect. However, the fundamental problem of causal inference prevents us from measuring this specific quantity. The fundamental problem of causal inference is that only one potential outcome, the one corresponding to the exposure actually received, for any individual is ever observed. Although we can never directly measure an individual's causal effect (except under a few limited circumstances), we can estimate the *average* causal effects provided that the following three identifiability conditions are satisfied, namely intervention variation irrelevance, positivity, and exchangeability [26]. The identifiability conditions, as the name suggests, allow for the average causal effect to be identifiable from the observed data and are defined as follows. For the first condition, intervention variation irrelevance, to be met for a binary exposure, the exposure must be well-defined such that only a single version of the exposure exists, allowing for no ambiguity in the exact meaning of having received the exposure (or not), and mapping onto unique potential outcomes.. The second condition, positivity, requires that all participants in the study are eligible for the intervention within levels of measured confounders. Thus, the probability of receiving the intervention is greater than zero for all participants within levels of the covariates. The third condition, exchangeability, requires that there are no systematic differences between the exposed and the unexposed that would affect the outcome within levels of confounders. Confounding is the bias that arises when the exposure assignment *and* the outcome share a common cause [26]. Randomized control experiments allow for exchangeability by randomly assigning the intervention independent of any variables that would otherwise affect both treatment assignment and the outcome by

a third variable [26].

In observational studies, the exposure assignment is not randomized, resulting in possible confounding of the exposure and the outcome. Causal inference methods for observational studies allow the investigators to replicate the conditions of randomization that take place in randomized control experiments under the following assumptions: all confounders are controlled for in the analysis, thus allowing for *conditional* exchangeability, there is no selection bias (i.e., the bias which arises from the procedure by which individuals are selected into the analysis such as by complete case analysis when there is loss to follow-up), and there is no measurement error (i.e., all variables are perfectly measured and defined).

In this study, we relax the assumption of no interference and allow for partial interference that can occur between individuals within communities but not between individuals in different communities. We assume stratified interference, which implies that dissemination depends on the proportion of people exposed in a community, rather than who was exposed [1, 17].

In an ideal scenario, all counterfactual outcomes would be observed in the study population and we could then calculate individual and average causal effects. However, we only observe one counterfactual outcome per individual in our study population. Using the method of inverse probability weighting, we estimated the *average* outcomes among all individuals in a setting where all participants received the intervention and the average outcomes of all individuals in the setting where all participants did not receive the intervention and then took the difference, thus consistently estimating the causal effect (measured as a risk-difference (RD)).

1.5 Community Detection

As our analysis aims to use causal inference methods for observational data to replicate the conditions of a two-stage randomized design, identifying the communities in the network

is of course necessary determine the interference sets. We define a community (or cluster) to be a group of vertices that are densely connected, with only sparser connections to other groups [23].

The primary goal of community detection is to find the natural divisions of a network into groups of participants such that there are many edges within communities and few between them. The most widely used approach to measuring the quality of a partition is known as *modularity*. The modularity of a partition (division of the network into some set of non-overlapping groups) is found by subtracting the proportion of *randomly* assigned edges that occur within groups from the proportion of edges within groups under the partition. Thus, the partition with the highest modularity score is identified as the community structure in the network [29]. Two widely used modularity maximizing algorithms are the iterative *node-moving* algorithm, and the *spectral modularity maximization* algorithm. The node-moving algorithm searches over the space of all possible partitions by iteratively modifying each partition one node at a time. A faster method is the spectral algorithm which uses partition vectors and the *modularity matrix* to optimize modularity. The partition vector which maximizes modularity happens to be the eigenvector which corresponds to the largest eigenvalue of the modularity matrix [29, 30].

Both community detection algorithms can ascertain good approximations to the maximum modularity partition of the network [29]. Although each algorithm takes a different approach to calculation, they both approximately maximize the same quality function. Therefore, we generally expect that differences between the two partitions should be minimal and ideally have little to no effect on the results of the estimation of the causal parameters.

Any network will have competing candidate partitions with similarly high modularity scores, but a network with a strong community structure will have competing partitions that contain negligible differences in information. If competing candidate partitions have highly different groupings, then there is evidence of a weak community structure [29]. Karrer et al (2008) argue that "the defining property of significant community structure is not a high

[or uniquely high] modularity, but robustness against small changes (or perturbations) of the edge assignments. They provide a method of testing the robustness, and therefore the significance or strength, of the detected community structure, employing concepts from *information theory* [10, 22].

1.6 Finite Sample Correction

The asymptotic properties of the TV estimator may not apply when the number of communities is less than 30 [36]. Many cluster randomized trials do not have a large number of communities, increasing the likelihood of Type I errors [20]. Tchetgen Tchetgen and Vanderweele propose finite-sample confidence intervals for the four causal estimands of interest [35]. These intervals are designed for a two-stage permutation randomized trial. We extended this approach to an observational setting to apply the finite sample correction. We divided the communities into two groups, a treatment arm and a control arm, according to the proportion of exposed participants in each community. We then did a sensitivity analysis for the effect of changing the criteria which assigns treatment or control to communities.

1.7 Missing Data

Missing Covariates

Under the assumption that random variables across participants are independently and identically distributed (iid), we can consider removing the participants with missing data not at the expense of unbiased estimation but a possible loss of efficiency. In a network setting however, we can only consider removing entire independent units (i.e., the communities) rather than individuals. Removal of any *individuals* on the basis of missing data may result in bias because the individuals are *not* the independent units in the data. Excluding

individuals with missing data could change the definition of "indirect" exposure, which is essentially defined by the level of coverage (the observed proportion of exposed participants in a cluster). Figure 3 shows a simple illustration of the effect that removing individuals can have on the coverage levels for connected participants. Black nodes represent exposed participants, white nodes represent non-exposed participants, and the numbers in the nodes indicate the coverage level for that participant. Consider a cluster of 4 individuals with no missing data on the exposure, but on some covariate. If 50% of individuals in that cluster are exposed, removal of one exposed individual with missing covariate information significantly changes how that cluster is defined with respect to coverage. By removing an exposed individual, the remaining exposed individual will have gone from coverage of 33.3% to 0%, likely a non-trivial change in coverage. For networks with only large communities, this may be of less concern. However, the TRIP network has under the node-moving algorithm has 8 to 11 communities with less than 5 individuals, which make up 38% to 50% of the communities in the network (under the node-moving method and spectral method respectively).

Figure 3: Illustration showing the effect of removing connected individuals in a network based on missing data. Removing connected individuals alters the observed coverage levels for those that are not removed. Black nodes represent exposed participants, white nodes represent non-exposed participants, and the numbers in the nodes indicate the coverage level for that participant.

Under the assumption of partial interference, removal of units for the purpose of complete case analysis can only be conducted by excluding *communities* based on missing information of its members, rather than excluding *individuals*. Rather than removing communities

from the network, we used a random forest method to impute the missing data for the covariates.

Missing Outcomes

When data are missing at random (MAR) the missing values can be predicted by the observed data. However, it seems entirely plausible that *missingness* of the outcome is associated with the *unobserved* values of the outcome. Such a missing data mechanism is called "missing not at random" (MNAR). Under the assumption that higher proportions of risky behavior are associated with a higher likelihood of loss to follow-up, we conduct a sensitivity analysis similar to that of Hedecker et. al. (2007) for the robustness of causal effect estimates under various assumptions regarding the MNAR mechanism. The methods used here employ an approach in Jackson et. al. (2014), where a similar but more in-depth analysis is conducted specifically on missing binary outcome data [21]. The Jackson et. al. method builds on the work of Hedecker et. al. (2007), which similarly considers a range of assumptions about the effect of the outcome on its missingness. Hedecker et. al. conduct sensitivity analysis on a plausible range of odds ratios of the outcome for missing participants (i.e., the odds of the outcome for the missing participants divided by the odds of the outcome for the non-missing participants) [18].

LIST OF REFERENCES

- [1] Basse, Guillaume and Feller, Avi. Analyzing two-stage experiments in the presence of interference. *Journal of the American Statistical Association*, 113(521):41–55, 2018.
- [2] Benjamin-Chung, Jade and Arnold, Benjamin F and Berger, David and Luby, Stephen P and Miguel, Edward and Colford Jr, John M and Hubbard, Alan E. Spillover effects in epidemiology: parameters, study designs and methodological considerations. *International Journal of Epidemiology*, 47(1):332–347, 2017.
- [3] Camacho, L Mabel and Bartholomew, Norma G and Joe, George W and Cloud, Melynda A and Simpson, D Dwayne. Gender, cocaine and during-treatment HIV risk reduction among injection opioid users in methadone maintenance. *Drug and Alcohol Dependence*, 41(1):1–7, 1996.
- [4] Carlson, Robert G and Daniulaityte, Raminta and Silverstein, Sydney M and Nahhas, Ramzi W and Martins, Silvia S. Unintentional drug overdose: Is more frequent use of non-prescribed buprenorphine associated with lower risk of overdose? *International Journal of Drug Policy*, 79:102722, 2020.
- [5] Carroll, Jennifer J and Rich, Josiah D and Green, Traci C. The more things change: buprenorphine/naloxone diversion continues while treatment remains inaccessible. *Journal of Addiction Medicine*, 12(6):459, 2018.
- [6] Degenhardt, Louisa and Peacock, Amy and Colledge, Samantha and Leung, Janni and Grebely, Jason and Vickerman, Peter and Stone, Jack and Cunningham, Evan

- B and Trickey, Adam and Dumchev, Kostyantyn and others. Global prevalence of injecting drug use and sociodemographic characteristics and prevalence of HIV, HBV, and HCV in people who inject drugs: a multistage systematic review. *The Lancet Global Health*, 5(12):e1192–e1207, 2017.
- [7] Edelman, E Jennifer and Chantararat, Tongtan and Caffrey, Sarah and Chaudhry, Amina and O'Connor, Patrick G and Weiss, Linda and Fiellin, David A and Fiellin, Lynn E. The impact of buprenorphine/naloxone treatment on HIV risk behaviors among HIV-infected, opioid-dependent patients. *Drug and Alcohol Dependence*, 139:79–85, 2014.
- [8] Fienberg, Stephen E. A brief history of statistical models for network analysis and open challenges. *Journal of Computational and Graphical Statistics*, 21(4):825–839, 2012.
- [9] Centers for Disease Control & Prevention. HIV in the United States and Dependent Areas, Sep 2019.
- [10] Fortunato, Santo. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.
- [11] Samuel R Friedman, Leslie Williams, April M Young, Jennifer Teubl, Dimitrios Paraskevis, Evangelia Kostaki, Carl Latkin, Danielle German, Pedro Mateu Gelabert, Honoria Guarino, et al. Network research experiences in New York and Eastern Europe: lessons for the southern US in understanding HIV transmission dynamics. *Current HIV/AIDS Reports*, 15(3):283–292, 2018.
- [12] Friedman, Samuel R and Downing, Martin J and Smyrnov, Pavlo and Nikolopoulos, Georgios and Schneider, John A and Livak, Britt and Magiorkinis, Gkikas and Slobodanyk, Liudmyla and Vasylyeva, Tetyana I and Paraskevis, Dimitrios and

- others. Socially-integrated transdisciplinary HIV prevention. *AIDS and Behavior*, 18(10):1821–1834, 2014.
- [13] Galea, Sandro and Vlahov, David. Social determinants and the health of drug users: socioeconomic status, homelessness, and incarceration. *Public Health Reports*, 117(Suppl 1):S135, 2002.
- [14] Gowing, Linda R and Farrell, Michael and Bornemann, Reinhard and Sullivan, Lynn E and Ali, Robert L. Brief report: Methadone treatment of injecting opioid users for prevention of HIV infection. *Journal of General Internal Medicine*, 21(2):193–195, 2006.
- [15] Greenland, Sander and Robins, James M. Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology*, 15(3):413–419, 1986.
- [16] M Elizabeth Halloran, Michael Haber, Ira M Longini Jr, and Claudio J Struchiner. Direct and indirect effects in vaccine efficacy and effectiveness. *American Journal of Epidemiology*, 133(4):323–331, 1991.
- [17] M Elizabeth Halloran and Michael G Hudgens. Dependent happenings: a recent methodological review. *Current Epidemiology Reports*, 3(4):297–305, 2016.
- [18] Hedeker, Donald and Mermelstein, Robin J and Demirtas, Hakan. Analysis of binary outcomes with missing data: missing= smoking, last observation carried forward, and a little multiple imputation. *Addiction*, 102(10):1564–1573, 2007.
- [19] Michael G Hudgens and M Elizabeth Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008.
- [20] NM Ivers, M Taljaard, Stephanie Dixon, C Bennett, A McRae, J Taleban, Z Skea, JC Brehaut, RF Boruch, MP Eccles, et al. Impact of CONSORT extension for cluster

- randomised trials on quality of reporting and study methodology: review of random sample of 300 trials, 2000-8. *Bmj*, 343:d5886, 2011.
- [21] Jackson, Dan and White, Ian R and Mason, Dan and Sutton, Stephen. A general method for handling missing binary outcome data in randomized controlled trials. *Addiction*, 109(12):1986–1993, 2014.
- [22] Karrer, Brian and Levina, Elizaveta and Newman, Mark EJ. Robustness of community structure in networks. *Physical Review E*, 77(4):046119, 2008.
- [23] Eric D Kolaczyk. *Statistical Analysis of Network Data*. Springer, 2010.
- [24] Larney, Sarah. Does opioid substitution treatment in prisons reduce injecting-related HIV risk behaviours? A systematic review. *Addiction*, 105(2):216–223, 2010.
- [25] Liu, Lan and Hudgens, Michael G and Becker-Dreps, Sylvia. On inverse probability-weighted estimators in the presence of interference. *Biometrika*, 103(4):829–842, 2016.
- [26] Hernán MA and Robins (2020) JM. Causal Inference. What If. Boca Raton: Chapman Hall/CRC.
- [27] MacArthur, Georgie J and Minozzi, Silvia and Martin, Natasha and Vickerman, Peter and Deren, Sherry and Bruneau, Julie and Degenhardt, Louisa and Hickman, Matthew. Opiate substitution treatment and HIV transmission in people who inject drugs: systematic review and meta-analysis. *Bmj*, 345:e5945, 2012.
- [28] Departments of Epidemiology Miguel Hernán and Biostatistics. Definition of Causal Effect.
- [29] Newman, Mark. *Networks*. Oxford University Press, 2018.
- [30] Newman, Mark EJ. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3):036104, 2006.

- [31] Nikolopoulos, Georgios K and Katsoulidou, Antigoni and Kantzanou, Maria and Rokka, C and Tsiara, C and Sypsa, V and Paraskevis, D and Psychogiou, M and Friedman, S and Hatzakis, A. Evaluation of the limiting antigen avidity EIA (LAg) in people who inject drugs in Greece. *Epidemiology & Infection*, 145(2):401–412, 2017.
- [32] Nikolopoulos, Georgios K and Pavlitina, Eirini and Muth, Stephen Q and Schneider, John and Psychogiou, Mina and Williams, Leslie D and Paraskevis, Dimitrios and Sypsa, Vana and Magiorkinis, Gkikas and Smyrnov, Pavlo and others. A network intervention that locates and intervenes with recently HIV-infected persons: the Transmission Reduction Intervention Project (TRIP). *Scientific Reports*, 6:38100, 2016.
- [33] Robins, James. A new approach to causal inference in mortality studies with a sustained exposure period? Application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512, 1986.
- [34] Rubin, Donald B. Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4):472–480, 1990.
- [35] Eric J Tchetgen Tchetgen and Tyler J VanderWeele. On causal inference in the presence of interference. *Statistical Methods in Medical Research*, 21(1):55–75, 2012.
- [36] van der Laan, Mark J. Causal inference for a population of causally connected units. *Journal of Causal Inference J. Causal Infer.*, 2(1):13–74, 2014.

2 METHODOLOGY

To estimate causal effects with observational data in the presence of partial interference, we employed the inverse probability weighting methods and finite sample correction introduced by Tchetgen Tchetgen and VanderWeele [27]. The following four distinct causal effects were estimated: *direct effect* (DE), *indirect effect* (IE), *composite* (CE), and *overall effect* (OE) [6, 27].

2.1 Causal inference in the Presence of Dissemination

Notation & Assumptions

In our study, the exposure is receipt of MOUD prescription and the outcomes are self-reported HIV injection risk behaviors. The exposure and the outcomes are binary variables A and Y . A key assumption in our study is that of partial interference, meaning that the exposure of one individual affects the outcome of another individual *only within* defined interference sets. The interference sets are the communities identified by the community detection algorithms. To guarantee the validity of causal estimation we rely on the three identifiability conditions. Conditional on baseline confounders, we assume exchangeability between communities, and conditional on baseline confounders (of exposure status and individual outcomes) within communities, we assume exchangeability between exposed and unexposed individuals. Thus, within levels of baseline confounders at the community-level, we can estimate the average counterfactual outcomes for communities with different exposure policies. Within levels of individual-level baseline confounders, we can estimate the average counterfactual outcomes for individuals with different exposure statuses.

Assume we have N communities and each community has n_i individuals for $i = 1, 2, \dots, N$, then Y_{ij} , A_{ij} , x_{ij} represent the observed outcome, observed exposure status, and the baseline covariate vector for the j^{th} individual in community i . $\mathbf{A}_i$ is a vector of exposure

allocations, and $\mathbf{X}_i$ is a matrix of baseline covariates for members within community i . Let $\mathcal{A}(n)$ be the set of all vectors possible exposure allocations of length n . For example, under a binary intervention, A , takes the value $a \in \{a = 0, a = 1\}$ where $a = 0$ represents control and $a = 1$ represents being assigned the intervention, if a community consists of only three individuals, then all possible exposure allocations are represented as $\mathcal{A}(3) = \{(1,1,1), (1,1,0), (1,0,1), (1,0,0), (0,1,1), (0,1,0), (0,0,1), (0,0,0)\}$. Thus, there would be $2^3 = 8$ possible exposure allocations for that community. Generally, $|\mathcal{A}(n_i)| = 2^{n_i}$. We denote the potential outcome for individual j in community i as $Y_{ij}(a_{ij}, a_{i,-j}) = Y_{ij}(a_i)$ if the community received exposure vector a_i . Thus, we consider the exposure status of individual j (denoted a_{ij}), as well as the exposure vector that includes the exposures of everyone else in the community excluding person j (denoted $a_{i,-j}$) when indexing the potential outcomes. We used inverse probability weighting to estimate the effects described above. We assume a group-level generalization of randomized exposure policy within levels of covariates for individuals. We standardize to a population in which the counterfactual exposure assignment was randomly allocated according to a Bernoulli distribution conditional on covariates [15].

Thus, we assume that conditional on $\mathbf{X}_i$, the counterfactual outcomes are independent of the exposure allocation $\mathbf{A}_i$. The second assumption is conditional positivity at the community level, i.e., $P(\mathbf{A}_i = a_i | \mathbf{X}_i) > 0, \forall a_i \in \mathcal{A}(n_i)$. To define the potential outcomes, we assume a Bernoulli individual group exposure policy. In a counterfactual scenario, the exposure policy α is the probability of any individual being independently assigned the exposure. Conditioning on covariates, we assume independent exposure assignments within levels of the covariates. Thus, the distribution of the exposure under a counterfactual exposure policy α is used to define the counterfactual estimands [16, 23, 27]. The probability of

the exposure vector for community i under exposure policy α is

$$\pi_i(\mathbf{A}_i; \alpha) = \prod_{j=1}^{n_i} \alpha^{A_{ij}} (1 - \alpha)^{1-A_{ij}}$$

and the probability of the exposure vector for community i which excludes individual j is

$$\pi_i(\mathbf{A}_{i,-j}; \alpha) = \prod_{k=1, k \neq j}^{n_i} \alpha^{A_{ik}} (1 - \alpha)^{1-A_{ik}}$$

When individual j is assigned to the intervention $a \in \{a = 0, a = 1\}$ with probability α , an individual's average potential outcome is denoted by

$$\bar{Y}_{ij}(a; \alpha) = \sum_{a_{i,-j} \in \mathcal{A}(n_i-1)} Y_{ij}(a_{ij}, a_{i,-j}) \pi_i(a_{i,-j}; \alpha)$$

The marginal individual average potential outcome is defined by

$$\bar{Y}_{ij}(\alpha) = \sum_{a_i \in \mathcal{A}(n_i)} Y_{ij}(a_i) \pi_i(a_i; \alpha)$$

We express the community level potential outcome as follows, where we take each individual's average potential outcome and average all of those across all individuals in the community:

$$\bar{Y}_i(a; \alpha) = \frac{1}{n_i} \sum_{j=1}^{n_i} \bar{Y}_{ij}(a; \alpha)$$

The population-level average potential outcome is found by averaging the community level

potential outcomes across the communities:

$$\bar{Y}(a; \alpha) = \frac{1}{N} \sum_{i=1}^N \bar{Y}_i(a; \alpha)$$

Thus, the following represent the four causal effects proposed by Hudgens and Halloran [8]. The *direct effect* measures the causal effect of receiving the intervention beyond being in a community with a fixed counterfactual exposure allocation strategy,

$$\overline{DE}(\alpha) = \bar{Y}(a = 1; \alpha) - \bar{Y}(a = 0; \alpha)$$

The *indirect effect* (also called the "disseminated effect") measures the effect of the intervention on the outcomes of individuals who were not exposed themselves but shared risk connections with those who were, contrasting two different coverage levels, and where $\alpha > \alpha'$

$$\overline{IE}(\alpha, \alpha') = \bar{Y}(a = 0; \alpha) - \bar{Y}(a = 0; \alpha')$$

The *composite effect* (or *total effect*) contrasts the outcomes of those who receive the intervention under one exposure policy (the probability, say α , of individual-level exposure assignment within a group or community) to those who do not receive the intervention under a different exposure policy (say α'), where $\alpha > \alpha'$. The *composite effect* can be thought of as a maximal intervention impact contrasting the exposed with coverage level α , and the unexposed with coverage level α' .

$$\overline{CE}(\alpha, \alpha') = \bar{Y}(a = 1; \alpha) - \bar{Y}(a = 0; \alpha')$$

The *overall effect* contrasts the population mean outcome under one exposure policy (α), to

the population mean outcome under another exposure policy (α'), where $\alpha > \alpha'$, marginalizing over the individual level exposures.

$$\overline{OE}(\alpha, \alpha') = \bar{Y}(\alpha) - \bar{Y}(\alpha')$$

Figure 4 gives an illustration of the causal estimands of interest. Arrows connecting regions represent the contrasts being made between the average outcomes.

Figure 4: Illustration of the causal estimands: *overall effect*, *composite effect*, *indirect effect*, and the *direct effect*. Arrows connecting regions represent the contrasts being made between the average outcomes. [6].

Inverse Probability Weighted (IPW) Estimators

To define the weights for the IPW estimator, we use the community level propensity score, which is the probability of the observed exposure vector in a community. When the community level propensity scores are known and the identifiability assumptions hold, IPW estimators are unbiased. As we were analyzing data from an observational study, we do not know the true propensity scores. We first estimated the propensity scores at the individual level using a mixed effects logit model with a random effect to account for dependency within communities. We then estimate the community level propensity scores by integrating the estimated probability of observed intervention vectors across the distribution of random effects [27].

$$f_{A|X}(\mathbf{A}_i|X_i) = \int \prod_{j=1}^{n_i} h_{A|X}(A_{ij}|X_{ij}, \mathbf{b}_i)^{A_{ij}} \left\{ 1 - h_{A|X}(A_{ij}|X_{ij}, \mathbf{b}_i) \right\}^{1-A_{ij}} f_b(\mathbf{b}_i|\sigma_b^2) d\mathbf{b}_i$$

where $h_{A|X}(A_{ij}|X_{ij}, \mathbf{b}_i)$ is the inverse-logit of the longitudinal regression model which estimates the individual level propensity score for individual j in community i , X_{ij} denotes baseline confounders for individual j in community i , and $\mathbf{b}_i$ is the random effect associated with community i , which follows a normal distribution $f_b(\mathbf{b}_i|\sigma_b^2)$ with mean 0 and variance σ_b^2 . We tested the assumption of normality of random effects in mixed effects models using the "glmerGOF" package [2, 26]. The IPW community-level average potential outcome and the marginal average potential outcome are obtained by the following estimators respectively:

$$\hat{Y}_i^{IPW}(a, \alpha) = \frac{\sum_{j=1}^{n_i} \pi_i(\mathbf{A}_{i,-j}; \alpha) I(A_{ij} = a) Y_{ij}}{n_i f_{A|X}(\mathbf{A}_i|X_i)}$$

$$\hat{Y}_i^{IPW}(\alpha) = \frac{\sum_{j=1}^{n_i} \pi_i(\mathbf{A}_i; \alpha) I(\mathbf{A}_i; \alpha) Y_{ij}}{n_i f_{A|X}(\mathbf{A}_i|X_i)}$$

The IPW estimators for the four causal effects of interest are as follows, and were calculated in R Studio using the "inference" package [1]:

$$\begin{aligned}\widehat{DE}(\alpha) &= \widehat{Y}^{IPW}(a = 1; \alpha) - \widehat{Y}^{IPW}(a = 0; \alpha) \\ \widehat{IE}(\alpha, \alpha') &= \widehat{Y}^{IPW}(a = 0; \alpha) - \widehat{Y}^{IPW}(a = 0; \alpha') \\ \widehat{CE}(\alpha, \alpha') &= \widehat{Y}^{IPW}(a = 1; \alpha) - \widehat{Y}^{IPW}(a = 0; \alpha') \\ \widehat{OE}(\alpha, \alpha') &= \widehat{Y}^{IPW}(\alpha) - \widehat{Y}^{IPW}(\alpha')\end{aligned}$$

2.2 Community Detection

Modularity

We define a community (or community) is to be a group of participants that are densely connected, with only sparser connections to other groups [13]. Community detection is conducted on the largest connected component in the network and separately on the unconnected components. Let $\mathcal{C} = \{C_1 \dots C_N\}$ be a given candidate partition and define $f_{ij}(\mathcal{C})$ to be the fraction of edges in the original network that connect vertices in C_i with vertices in C_j . The modularity of $\mathcal{C}$ is defined by Newman and Girvan as the following [18]:

$$Q(\mathcal{C}) = mod(\mathcal{C}) = \sum_{k=1}^N [f_{kk}(\mathcal{C}) - f_{kk}^*]^2$$

where f_{kk}^* is the expected value of $f_{kk}(\mathcal{C})$ under random edge assignment. The calculation of the modularity score however generally requires a re-parameterization that allows us to more explicitly show the modularity as a function of the given network and some partition $\mathcal{C}$. We will start by defining two key quantities, the number of edges within communities (rather than between communities) and the expected number of edges within communities

if edges were assigned at random. Starting with the number of edge within communities, we have $\frac{1}{2} \sum_{ij} A_{ij} \delta_{ij}$, where A is the adjacency matrix of the network graph, and $\delta_{ij} = 1$ if vertices i and j are in the same group (community) and $\delta_{ij} = 0$ otherwise. To solve for the number of edges within communities under random edge assignment, we first let m be the number of edges in the network, and let k_i be the degree of vertex i . The expected number of edges connecting nodes i and j is therefore $\frac{k_i k_j}{2m}$. The expected number of edges within communities when edges are assigned at random is then $\frac{1}{2} \sum_{ij} \frac{k_i k_j}{2m} \delta_{ij}$. We can now calculate the difference between the number of edges within communities and the expected number of edges within communities under random edge assignment by subtracting the second quantity from the first.

$$\frac{1}{2} \sum_{ij} A_{ij} \delta_{ij} - \frac{1}{2} \sum_{ij} \frac{k_i k_j}{2m} \delta_{ij} = \frac{1}{2} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta_{ij}$$

Dividing by the number edges gives us the difference between the proportion of edges within communities and the expected proportion of edges within communities under random edge assignment. Thus, modularity Q can be written as [19]:

$$\begin{aligned} Q &= \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta_{ij} \\ &= \frac{1}{2m} \sum_{ij} B_{ij} \delta_{ij} \end{aligned}$$

Node-moving algorithm

One of the most common classes of methods for community detection is hierarchical communitying [14]. These methods search over the space of all possible partitions of the network by iteratively modifying successive candidate partitions until the highest modularity score is reached. The partition that maximizes the modularity is chosen under the commu-

nity detection algorithm [13]. The optimization of this quantity Q starts with each vertex being the sole member of a community of size 1. The next step is to join pairs of communities and choosing the set of pairings that result in the largest increase in Q . This process is repeated until the Q no longer increases [3, 20]. Community detection was done in R Studio using the "igraph" package [5].

Spectral modularity maximization

An alternative method of community detection is spectral modularity maximization, which utilizes results from eigen-analysis of certain matrices, particularly the *modularity matrix* of the graph. Let B be a $n \times n$ matrix where element $B_{ij} = A_{ij} - \frac{k_i k_j}{2m}$, as shown above. The expression, $\frac{1}{4m} s^T B s$, represents the modularity score where s is a n degree vector where s_i indicates the partition to which vertex i belongs. It can be shown that maximizing the expression, $\frac{1}{4m} s^T B s$ with respect to the partition vector s and solving for s results in the eigenvector associated with the largest eigenvalue of the matrix B [19]. The result is a bisection of graph G . The two communities are then subdivided further using the same process until Q no longer increases [21, 22].

Both algorithms result in a set of communities (or clusters) such that each individual is a member of only one community, and participants within a community interact highly with each other. participants in different communities that share edges are defined as having a social connection, but interference is assumed to only take place within communities. Community detection was done in R Studio using the "igraph" package [5].

Robustness of Community Structure

Karrer et al. point out that although networks with strong community structure have high modularity, not all networks with high modularity have strong community structure [12]. Thus, the question of significance of an identified community structure is essentially a

question about the robustness of that structure against random reassignment of edges (or perturbations of edges) [12]. Quantifying the robustness of a community structure utilizes concepts from information theory such as *conditional entropy* and *variation of information*. Let u_i be the label of the community to which vertex i belongs in partition C , and let z_i be the label of the community to which vertex i belongs in partition C' . The variation of information between partitions C and C' is then defined as follows [12].

$$\begin{aligned} V(C, C') &= H(U|Z) + H(Z|U) \\ &= - \sum_{uz} P(u, z) \log \frac{P(u, z)}{P(z)} - \sum_{uz} P(u, z) \log \frac{P(u, z)}{P(u)} \end{aligned}$$

where $H(U|Z)$ is the *conditional entropy* of U given Z , i.e., the additional information needed to describe U once Z is known. The *variation of information* is the sum of information needed to describe C from C' and the information needed to describe C' from C [12].

Karrer et al give the following method to test for the strength of the community structure in the network [12]. We find the partition, C , which maximizes modularity. We then perturb the network by randomly reassigning edges with probability γ . The partition of this new perturbed network is found, C' , which maximizes modularity and the variation of information, $V(C, C')$, is calculated. Many perturbed networks are created, each time with the same probability of edge reassignment, γ , and the average of $V(C, C')$ is calculated for parameter γ , i.e., $\bar{V}(\gamma)$. This process is repeated for values of γ ranging from 0, which corresponds to no perturbation of the network, to 1 which corresponds to changing the entire network. A "null network" (or configuration network) is created for comparison. The null network has the same degree sequence as the actual network, but is defined to have no community structure. Plotting $\bar{V}(\gamma)$ for both networks allows us to compare the sensitivity of the community structure of our network to that of a network which we know has no community structure. In this work, we make the additional consideration of the *range* of

$V(\gamma)$ by plotting $\min\{V(\gamma)\}$ and $\max\{V(\gamma)\}$ along with $\overline{V}(\gamma) \forall \gamma \in [0, 1]$.

2.3 Finite Sample Correction

The asymptotic properties of the TV estimator may not apply when the number of communities is less than 30 [28]. Many community randomized trials do not have a large number of communities, increasing the likelihood of Type I errors [9]. Tchetgen Tchetgen and Vanderweele propose finite-sample confidence intervals for the four causal estimands of interest [27]. Finite sample confidence intervals were calculated in R Studio using the "interferenceCI" package [11].

Finite-sample confidence interval for the *direct effect*:

$$\left(\overline{DE}(\alpha) - \epsilon_{DE}^*(q, \kappa, N), \overline{DE}(\alpha) + \epsilon_{DE}^*(q, \kappa, N) \right)$$

where

$$\epsilon_{DE}^*(q, \kappa, N) = \sqrt{\frac{4\left(\frac{1}{q} - 1\right)^2 + \frac{\sum_{i=1}^N \left(\frac{L_{DE,i}}{q}\right)^2}{N}}{2N}} \ln\left(\frac{2}{\kappa}\right),$$

$$L_{DE,i}(\alpha) = 2\left(1 - \frac{1}{\left(\frac{n_i}{K_{0,i}}\right)}\right)$$

and where q is defined to be the proportion of communities assigned to exposure policy α , κ is the significance level, N is the number of communities, n_i denotes the number of individuals in community i , and $K_{0,i}$ denotes the number of exposed individuals in community i . This interval is analogously extended for the other three causal estimands of interest, the *indirect effect*, *composite*, and *overall effect*.

Finite-sample confidence interval for the *indirect effect*:

$$\left(\overline{IE}(\alpha, \alpha') - \epsilon_{IE}^*(q, \kappa, N), \quad \overline{IE}(\alpha, \alpha') + \epsilon_{IE}^*(q, \kappa, N) \right)$$

where

$$\epsilon_{IE}^*(q, \kappa, N) = \sqrt{\frac{\left[\max\left\{ \frac{1}{q^2}, \frac{1}{(1-q)^2} \right\} + \sum_i \frac{L_{IE,i}^2(q)}{N} \right]}{2N}} \ln\left(\frac{2}{\kappa}\right)$$

and

$$L_{IE,i}(q) = \max\left\{ \frac{1}{q} \left(1 - \frac{1}{\binom{n_i}{K_{0,i}}} \right), \quad \frac{1}{1-q} \left(1 - \frac{1}{\binom{n_i}{K_{0,i}}} \right) \right\}$$

Finite-sample confidence interval for the *composite effect*:

$$\left(\overline{CE}(\alpha, \alpha') - \epsilon_{CE}^*(q, \kappa, N), \quad \overline{CE}(\alpha, \alpha') + \epsilon_{CE}^*(q, \kappa, N) \right)$$

where

$$\epsilon_{CE}^*(q, \kappa, N) = \sqrt{\frac{\left[\max\left\{ \frac{1}{q^2}, \frac{1}{(1-q)^2} \right\} + \sum_i \frac{L_{CE,i}^2(q)}{N} \right]}{2N}} \ln\left(\frac{2}{\kappa}\right)$$

and

$$L_{CE,i}(q) = \max\left\{ \frac{1}{q} \left(1 - \frac{1}{\binom{n_i}{K_{0,i}}} \right), \quad \frac{1}{1-q} \left(1 - \frac{1}{\binom{n_i}{K_{0,i}}} \right) \right\}$$

Finite-sample confidence interval for the *overall effect*:

$$\left(\overline{OE}(\alpha, \alpha') - \epsilon_{OE}^*(q, \kappa, N), \overline{OE}(\alpha, \alpha') + \epsilon_{OE}^*(q, \kappa, N) \right)$$

where

$$\epsilon_{OE}^*(q, \kappa, N) = \sqrt{\frac{\left[\max\left\{ \frac{1}{q^2}, \frac{1}{(1-q)^2} \right\} + \sum_i \frac{L_{OE,i}^2(q)}{N} \right] \ln\left(\frac{2}{\kappa}\right)}{2N}}$$

and

$$L_{OE,i}(q) = \max\left\{ \frac{1}{q} \left(1 - \frac{1}{\binom{n_i}{K_{0,i}}} \right), \frac{1}{1-q} \left(1 - \frac{1}{\binom{n_i}{K_{0,i}}} \right) \right\}$$

To the best of our knowledge, no method currently exists for finite sample confidence intervals in the observational setting. However, we make an adaptation to the intervals shown in section 1.6 that allows us to approximate the finite sample intervals in an observational setting. The approach we take here is to approximate the parameter q , which is the proportion of communities assigned to the exposure arm of the study in the first stage of randomization. The first stage of randomization is the assignment of exposure strategy α and control strategy α_o . Since there is no stage-1 randomization in the observational setting, we assume that communities with observed coverage greater than or equal to some threshold $\psi \in [0, 1]$ were randomized to the treatment arm (i.e., had been randomized to α) and communities with observed coverage less than ψ were randomized to the control arm (i.e., had been randomized to α_o). Thus the proportion of communities in the exposure arm, q , is the proportion of communities with observed coverage greater than or equal to ψ .

Tchetgen Tchetgen and VanderWeele note that the length of the confidence interval for the *direct effect*, $C_{DE}(\kappa, q, N)$, is proportional to $\frac{1}{\sqrt{N}}$ where N is the number of communities in the experiment [27]. For a fixed value of κ and q , the interval becomes more precise as the number of communities, N , increases. When N is small, $C_{DE}(\kappa, q, N)$ may not be

informative at all. They illustrate this point by considering a case where $1/\binom{n_i}{K_{0,i}}$ is approximately 0, and q is approximately $1/2$. The margin of error in this case, $\epsilon_{DE}^*(1/2, \kappa, N)$, is approximately equal to $6/\sqrt{N}$, guaranteeing that the interval will be non-informative if there are less than 9 communities [27]. A possible violation to the validity of the intervals comes from assuming a Bernoulli allocation strategy within communities, rather than permutation randomization. Under permutation randomization, communities are defined to have a coverage level that is actually equal to the value of α (or α_o for the control arm). Here we allow the assumed treatment arm to have observed coverage that differs from the assigned α as in the case of a Bernoulli allocation strategy. See Appendix C for a brief discussion on permutation randomization. We report the results of an analysis of the sensitivity of the confidence intervals to values of ψ which range from 0 to 1.

In addition, we restricted the analysis to communities with coverage of approximately 25% (i.e., 20% to 26%) and of approximately 50% (i.e., 45% to 55%). These coverage assignments could plausibly result from a two-stage permutation randomization with coverage levels of approximately 25% as the control arm and the communities with approximately 50% as the exposure arm.

2.4 Missing Data

Covariate Imputation

To impute values for missing covariate data, we use a non-parametric random forest method. Considering the network setting, we chose to *not* remove *individuals* with missing data, as this would have resulted in altering the level of coverage in for nearby individuals and alter the network structure. In addition, this method outperforms other methods such as mean imputation, hot deck imputation, and predictive mean matching [17]. This is true particularly when the data is missing completely at random (MCAR) or MAR. The random forest imputation method also outperforms these methods for MNAR, albeit by a narrow mar-

gin [17]. As pointed out by Tang & Ishwaran, missForest does not take outcome data into account is strictly a method of imputing "features" (or covariate data) [25].

Consider n participants with p variables. For an arbitrary variable X_s with missing values at entries $i_{mis}^{(s)} \in \{1, \dots, n\}$, the dataset can be divided into 4 parts as illustrated in Figure 5 [17,24].

- 1.) The observed values of X_s , denoted by $y_{obs}^{(s)}$;
- 2.) The missing values of X_s , denoted by $y_{mis}^{(s)}$;
- 3.) The variables other than X_s that are observed, denoted by $x_{obs}^{(s)}$;
- 4.) The variables other than X_s that are missing, denoted by $x_{mis}^{(s)}$.

x1	x2	x3	x4	x5	x6	class
36	702	7	NA	76,34	11	w
32	5000	7	1100	241,49	35	w
31	5000	13	1340,2	292,58	24	w
35	7710	NA	2446	356,53	35	w
32	7679	4	1750	NA	35	w
23	NA	3	560	205,5	18	w
28	1147	NA	676,9	114,91	12	w
20	1300	5	1950	77,36	23	w
36	1320	93	1011,3	142,24	11	w
31	1612	22	594	93,94	NA	w
46	2776	58	NA	272,4	12	s
41	NA	95	1752,2	154,7	18	s
23	4300	8	1450	195,31	35	s
40	3018	137	2416,2	137,08	NA	s
24	4950	20	1038,9	224,83	35	s
50	1000	34	1163,4	98,13	12	s
42	4742	231	1293	215,39	35	s
NA	3268	105	1940	375,15	10	s
41	3000	214	1539,1	350,55	NA	s
43	3000	71	1433,7	NA	11	s

Legend:

y_{obs}
y_{mis}
x_{obs}
x_{mis}

Figure 5: An illustration of the four parts of the dataset used by missForest, where $X_{(6)}$ has missing data. White cells indicate $x_{mis}^{(6)}$, the second darkest cells indicate $y_{obs}^{(6)}$, the third darkest cells indicate $y_{mis}^{(6)}$, and the dark cells indicate $x_{obs}^{(6)}$.
Source: Misztal, Malforzata (2013)

The procedure of missForest is described in the following steps [17,24]:

- 1.) Make initial guess for the missing values using any imputation method.

- 2.) Sort all the variables X_s , $s = 1, 2, \dots, p$, according to the amount of missing values, starting with the lowest amount.
- 3.) For each variable X_s fit a Random Forest with response $y_{obs}^{(s)}$ and predictors $s_{obs}^{(s)}$. Then, predict the missing values $y_{mis}^{(s)}$ by applying the trained random forest to $x_{mis}^{(s)}$.
- 4.) The imputation procedure is repeated until a stopping criterion is met, which takes place as soon as the difference between the newly imputed data matrix and the previous one increases for the first time [17].

Outcome Imputation

The outcome at follow-up was missing for 21% of the participants. Those with a missing outcome at the 6 months were the result of missing the 6-month follow-up visit. The methods used here are based on approaches proposed by Jackson et. al. (2014), where a similar but more in-depth analysis is conducted specifically on missing binary outcome data [10]. They fit the following longitudinal model.

$$\begin{aligned} \text{logit}\left(P[Y = 1|R, X, Z]\right) = & Z\left\{(1 - X)(\beta_1 + \beta_2(1 - R)) + X\left(\beta_3 + \beta_4(1 - R)\right)\right\} \\ & + (1 - Z)\left\{(1 - X)\left(\beta_5 + \beta_6(1 - R)\right) + X\left(\beta_7 + \beta_8(1 - R)\right)\right\} \end{aligned}$$

Where Y is the binary outcome at follow-up, R indicates if the subject is observed at baseline (i.e., $R = 1$ if the outcome is observed, $R = 0$ otherwise), X is binary outcome at baseline, and Z is the binary indicator of treatment group. The parameters β_2 , β_4 , β_6 , and β_8 denote the log-odds ratio of the outcome comparing those with missing outcomes to those with observed outcomes. A range of values which correspond to plausible assumptions about the MNAR mechanism are considered for these parameters. The Jackson et. al. method builds on the work of Hedecker et.

al. (2007), which similarly considers a range of assumptions about the effect of the outcome on its missingness [7]. One major difference between the method used in this thesis and the methods described above is that neither of the above methods are developed for the observational network setting. Further research in the area of missing data methods for longitudinal studies in the network setting is needed to address missing outcome data that is MNAR with a non-randomized intervention.

Missing outcome data were handled in two primary steps. The first step was to naively assume MAR mechanism and predict the outcome distributions using a longitudinal mixed effects model. The MAR assumption leads to biased results when the data are MNAR, likely underestimating the proportion of missing participants who had the risk behavior. The longitudinal model predicts the probability that participants have the risk behavior at baseline.

$$\text{logit}\left(P[Y_{ij}^{(t)} = 1|A_{ij}, x_{ij}, \mathbf{r}_i]\right) = \beta_0 + \beta_1 A_{ij} + \beta_2 x_{ij} + \beta_3 T + \mathbf{r}_i$$

where $Y_{ij}^{(t)}$ denotes the outcome at time t for individual j in community i , where $t = 0$ the outcome is at baseline and $t = 1$ the outcome is at follow-up. A_{ij} represents the observed exposure status, and x_{ij} represents the baseline covariate vector for the j^{th} individual in community i , and t indicates baseline ($t = 0$) or follow-up ($t = 1$), and $\mathbf{r}_i$ indicates the community random effect. Under this mixed effects model we assume $\mathbf{r}_i \sim N(0, \sigma_r^2)$ and is uncorrelated with covariates X_{ij} and exposure A_{ij} . The first step is to fit the model shown above to obtain the following:

$$\text{logit}\left(\hat{P}[Y_{ij}^{(t)} = 1|A_{ij}, x_{ij}, \mathbf{r}_i]\right) = \hat{\beta}_0 + \hat{\beta}_1 A_{ij} + \hat{\beta}_2 x_{ij} + \hat{\beta}_3 T + \mathbf{r}_i$$

The second step is to use the Last Observation Carried Forward (LOCF) method for the participants with missing follow-up data despite the bias that this method induces

which could be positive or negative. We carry forward the estimated *average* baseline outcome, i.e. $\hat{P}[Y_{ij}^{(0)} = 1 | A_{ij}, x_{ij}, \mathbf{r}_i]$ which we now denote simply as $\hat{P}_{ij}^{(0)}$, to impute the missing values. The *imputed* average outcome at follow-up for individual j in community i is denoted $\hat{P}_{ij}$. LOCF assumes that the outcomes that occur after dropout remain constant at the last observed value [4]. Table 1 illustrates this for participants, say 1 through 4, where participants 2 and 4 have missing follow-up data which are filled in with baseline data, and participants 1 and 3 have observed follow-up data.

LOCF Set-up		
Baseline		Follow-up
$\hat{P}_{i1}^{(0)}$		Y_{i1}
$\hat{P}_{i2}^{(0)}$	LOCF $\longrightarrow$	NA
$\hat{P}_{i3}^{(0)}$		Y_{i3}
$\hat{P}_{i4}^{(0)}$	LOCF $\longrightarrow$	NA

LOCF Result	
Baseline	Follow-up
$\hat{P}_{i1}^{(0)}$	Y_{i1}
$\hat{P}_{i2}^{(0)}$	$\hat{P}_{i2} \left(= \hat{P}_{i2}^{(0)} \right)$
$\hat{P}_{i3}^{(0)}$	Y_{i3}
$\hat{P}_{i4}^{(0)}$	$\hat{P}_{i4} \left(= \hat{P}_{i4}^{(0)} \right)$

Table 1: Illustration of LOCF for participants with missing outcome data. Estimated averages at baseline are carried forward to follow-up for those with missing outcome data at follow-up. We therefore assume that average outcomes at follow-up, $\hat{P}$, are equal to the estimated average outcome at baseline, $\hat{P}^{(0)}$.

In Table 1, the "NA" values on the right are replaced with $\hat{P}^{(0)}$, thus setting $\hat{P}_{ij}$ equal to $\hat{P}_{ij}^{(0)}$. For the third step we define the parameter $\psi \in [0, 1]$ to be a threshold such that anyone with missing outcome data and whose imputed $\hat{P}$, is greater than ψ is

assigned an outcome value of $Y = 1$. Anyone who has missing outcome data whose imputed $\hat{P}$, is less than or equal to ψ is then assigned an outcome value of $Y = 0$. This means that if $\psi = 1$, all the participants with missing outcome data would be assigned an outcome of 0. If $\psi = 0$, all the participants with missing outcome data would be assigned an outcome of 1. Here, we choose $\psi = 0.5$. The fourth step is to conduct sensitivity analysis for the robustness of the results against changes in ψ . Similar to the method of Jackson et. al. (2014), this method allows us to investigate a range of plausible possibilities with regard to the MNAR model [10]. We assume under a MAR mechanism that the estimated probabilities of the outcome are biased (which may be positive or negative). We therefore sequentially change the threshold ψ and analyze the effect this has on the results. We consider extreme cases ranging across the entire range of ψ , i.e., $[0, 1]$.

LIST OF REFERENCES

- [1] Bradley C. Saul and Michael G. Hudgens. A Recipe for inference: Start with Causal Inference. Add Interference. Mix Well with R. *Journal of Statistical Software*, 82(2):1–21, 2017.
- [2] Brian G. Barkley. *glmerGOF: Goodness of Fit of Random Effect Distribution in Logistic Mixed Model*, 2019. R package version 0.1.1.
- [3] Clauset, Aaron and Newman, Mark EJ and Moore, Cristopher. Finding community structure in very large networks. *Physical Review E*, 70(6):066111, 2004.
- [4] Garrett M Fitzmaurice, Nan M Laird, and James H Ware. *Applied Longitudinal Analysis*, volume 998. John Wiley & Sons, 2012.
- [5] Gabor Csardi and Tamas Nepusz. The igraph Software Package for Complex Network Research. *InterJournal*, Complex Systems:1695, 2006.
- [6] M Elizabeth Halloran and Michael G Hudgens. Dependent happenings: a recent methodological review. *Current Epidemiology Reports*, 3(4):297–305, 2016.
- [7] Hedeker, Donald and Mermelstein, Robin J and Demirtas, Hakan. Analysis of binary outcomes with missing data: missing= smoking, last observation carried forward, and a little multiple imputation. *Addiction*, 102(10):1564–1573, 2007.
- [8] Michael G Hudgens and M Elizabeth Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008.

- [9] NM Ivers, M Taljaard, Stephanie Dixon, C Bennett, A McRae, J Taleban, Z Skea, JC Brehaut, RF Boruch, MP Eccles, et al. Impact of CONSORT extension for cluster randomised trials on quality of reporting and study methodology: review of random sample of 300 trials, 2000-8. *Bmj*, 343:d5886, 2011.
- [10] Jackson, Dan and White, Ian R and Mason, Dan and Sutton, Stephen. A general method for handling missing binary outcome data in randomized controlled trials. *Addiction*, 109(12):1986–1993, 2014.
- [11] Joseph Rigdon. *interferenceCI: Exact Confidence Intervals in the Presence of Interference*, 2015. R Package Version 1.1.
- [12] Karrer, Brian and Levina, Elizaveta and Newman, Mark EJ. Robustness of community structure in networks. *Physical Review E*, 77(4):046119, 2008.
- [13] Eric D Kolaczyk. *Statistical Analysis of Network Data*. Springer, 2010.
- [14] Eric D Kolaczyk and Gábor Csárdi. *Statistical Analysis of Network Data With R*, volume 65. Springer, 2014.
- [15] Liu, Lan and Hudgens, Michael G. Large sample randomization inference of causal effects in the presence of interference. *Journal of the American Statistical Association*, 109(505):288–301, 2014.
- [16] Liu, Lan and Hudgens, Michael G and Becker-Dreps, Sylvia. On inverse probability-weighted estimators in the presence of interference. *Biometrika*, 103(4):829–842, 2016.
- [17] Misztal, Malgorzata. Some remarks on the data imputation using missforest method. 2013.
- [18] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113, 2004.
- [19] Newman, Mark. *Networks*. Oxford University Press, 2018.

- [20] Newman, Mark EJ. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69(6):066133, 2004.
- [21] Newman, Mark EJ. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3):036104, 2006.
- [22] Newman, Mark EJ. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.
- [23] Perez-Heydrich, Carolina and Hudgens, Michael G and Halloran, M Elizabeth and Clemens, John D and Ali, Mohammad and Emch, Michael E. Assessing effects of cholera vaccination in the presence of interference. *Biometrics*, 70(3):731–741, 2014.
- [24] Stekhoven, Daniel J and Buhlmann, Peter. Missforest non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.
- [25] Tang, Fei and Ishwaran, Hemant. Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 10(6):363–377, 2017.
- [26] Eric J Tchetgen and Brent A Coull. A diagnostic test for the mixing distribution in a generalised linear mixed model. *Biometrika*, 93(4):1003–1010, 2006.
- [27] Eric J Tchetgen Tchetgen and Tyler J VanderWeele. On causal inference in the presence of interference. *Statistical Methods in Medical Research*, 21(1):55–75, 2012.
- [28] van der Laan, Mark J. Causal inference for a population of causally connected units. *Journal of Causal Inference J. Causal Infer.*, 2(1):13–74, 2014.

3 RESULTS

3.1 Effect of MOUD on HIV Risk Behavior

Covariates included in this study are HIV status, education level, employment status, marital status, homelessness, heroin use, being able to access medical care, and having been encouraged by their contacts to share a works, and the risk behavior at baseline. These were chosen on the basis of other studies on the causal effects of MOUD on HIV risk behaviors and were known or suspected risk factors for the outcome [3]. Some covariates were not included on the basis of model convergence problems.

Table 2: Contingency table showing the conditional distributions (shown as percentages) of the covariates among the overall (marginal) study sample, among those prescribed MOUD at baseline, and among those who were not prescribed MOUD at baseline.

		Overall (n=246)	Exposed to MOUD (n=52)	Not Exposed to MOUD (n=194)
Baseline Outcome	No	16.26	19.23	15.46
	Yes	83.74	80.77	84.54
HIV Status	Negative	44.31	36.54	46.39
	Positive	55.69	63.46	53.61
Education	Primary or Less	32.93	21.15	36.08
	Some HS	55.28	69.23	51.55
	Vocational/College or More	11.79	9.62	12.37
Employed	Employed	28.05	21.15	29.90
	Unemployed or Unable	71.95	78.85	70.10
Partnered	Partnered	26.02	23.08	26.80
	Non-Partnered	73.98	76.92	73.20
Homeless	Not Homeless	71.14	76.92	69.59
	Homeless	28.86	23.08	30.41
Used Heroin	No	27.64	38.46	24.74
	Yes	72.36	61.54	75.26
Able to Get Medical Care	Agree	52.85	55.77	52.06
	Otherwise	47.15	44.23	47.94
Encouraged to Share Works	No	60.98	67.31	59.28
	Yes	39.02	32.69	40.72

Tables 3 and 4 below show the estimated causal effects under the various condi-

tions (community detection methods and adjusting for the baseline outcome). Table 3 shows the estimated RD's under both the node-moving community detection algorithm (left column) and the spectral method (right column) for the models which do *not* adjust for the baseline outcome. Table 4 shows the estimated RD's under both the node-moving community detection algorithm and the spectral method for the models which do adjust for the baseline outcome. The estimated causal effects in Tables 3 and 4 are measured on the difference scale. For example, Table 3 below shows among communities with 25% coverage of MOUD, participants who were treated with MOUD had a 19.8% reduction in the likelihood of reporting HIV injection risk behaviors at the 6-month visit when communities are identified using the node-moving method and we do not control for baseline outcomes. The reduction is 13.3% when communities are identified using the spectral method. Unexposed participants in communities with 25% coverage of MOUD had a 46.6% reduction in the likelihood of reporting HIV injection risk behaviors at the 6-month visit compared to unexposed participants in communities with 75% coverage when communities are identified using the node-moving method and we do not control for baseline outcomes. The reduction is 20.1% when communities are identified using the spectral method. Unexposed participants in communities with 25% coverage of MOUD had a 37.4% reduction in the likelihood of reporting HIV injection risk behaviors at the 6-month visit compared to exposed participants in communities with 75% coverage when communities are identified using the node-moving method and we do not control for baseline outcomes. The reduction is 38.1% when communities are identified using the spectral method. The overall population of participants in communities with with 25% coverage of MOUD had a 34.7% reduction in the likelihood of reporting HIV injection risk behaviors at the 6-month visit compared to the overall population of participants in communities with 75% coverage when communities are identified using the node-moving method and we do not control for baseline out-

comes. The reduction is 30.3% when communities are identified using the spectral method.

Table 3: Estimated risk differences (RDs) with 95% CIs of **MOUD** (Received vs. Not Received) on likelihood of **engaging in HIV risk behavior** in TRIP (coverage of received), **not adjusted for baseline outcome**

	Node-moving Method		Spectral Method	
Effect	RD	95% CI	RD	95% CI
Direct (0.25, 0.25)	-0.198	(-0.352 , -0.043)	-0.133	(-0.312 , 0.045)
Direct (0.5, 0.5)	-0.015	(-0.168 , 0.138)	-0.181	(-0.348 , -0.014)
Direct (0.75, 0.75)	0.092	(-0.019 , 0.202)	-0.180	(-0.373 , 0.012)
Indirect (0.75, 0.25)	-0.466	(-0.532 , -0.399)	-0.201	(-0.353 , -0.05)
Indirect (0.5, 0.25)	-0.261	(-0.326 , -0.197)	-0.146	(-0.231 , -0.061)
Indirect (0.75, 0.5)	-0.204	(-0.262 , -0.146)	-0.055	(-0.127 , 0.017)
Composite (0.75, 0.25)	-0.374	(-0.505 , -0.242)	-0.381	(-0.535 , -0.228)
Composite (0.25, 0.5)	-0.276	(-0.420 , -0.133)	-0.327	(-0.493 , -0.161)
Composite (0.75, 0.5)	-0.112	(-0.082 , -0.045)	-0.236	(-0.404 , -0.067)
Overall (0.75, 0.25)	-0.347	(-0.438 , -0.256)	-0.303	(-0.399 , -0.208)
Overall (0.5, 0.25)	-0.220	(-0.266 , -0.173)	-0.203	(-0.269 , -0.137)
Overall (0.75, 0.5)	-0.128	(-0.193 , -0.063)	-0.100	(-0.149 , -0.051)

Table 4 below shows that among communities with 25% coverage of MOUD, participants who were treated with MOUD had a 19.1% reduction in the likelihood of reporting HIV injection risk behaviors at the 6-month visit when communities are identified using the node-moving method and we do control for baseline outcomes. The reduction is 12.7% when communities are identified using the spectral method. Unexposed participants in communities with 25% coverage of MOUD had a 46.3% reduction in the likelihood of reporting HIV injection risk behaviors at the 6-month visit compared to unexposed participants in communities with 75% coverage when communities are identified using the node-moving method and we do control for baseline outcomes. The reduction is 20.1% when communities are identified using the spectral method. Unexposed participants in communities with 25% coverage of MOUD had a 37.9% reduction in the likelihood of reporting HIV injection risk behaviors at the 6-month visit compared to exposed participants in communities with

75% coverage when communities are identified using the node-moving method and we do control for baseline outcomes. The reduction is 38.9% when communities are identified using the spectral method. The overall population of participants in communities with with 25% coverage of MOUD had a 35.2% reduction in the likelihood of reporting HIV injection risk behaviors at the 6-month visit compared to the overall population of participants in communities with 75% coverage when communities are identified using the node-moving method and we do control for baseline outcomes. The reduction is 31.2% when communities are identified using the spectral method.

Table 4: Estimated risk differences (RDs) with 95% CIs of **MOUD** (Received vs. Not Received) on likelihood of **engaging in HIV risk behavior** in TRIP (coverage of received), **adjusted for baseline outcome**

	Node-moving Method		Spectral Method	
Effect	RD	95% CI	RD	95% CI
Direct (0.25, 0.25)	-0.191	(-0.342 , -0.04)	-0.127	(-0.309 , 0.055)
Direct (0.5, 0.5)	-0.022	(-0.161 , 0.116)	-0.182	(-0.344 , -0.020)
Direct (0.75, 0.75)	0.084	(0.003 , 0.164)	-0.180	(-0.353 , -0.006)
Indirect (0.75, 0.25)	-0.463	(-0.52 , -0.405)	-0.209	(-0.336 , -0.082)
Indirect (0.5, 0.25)	-0.254	(-0.293 , -0.215)	-0.150	(-0.226 , -0.074)
Indirect (0.75, 0.5)	-0.209	(-0.266 , -0.151)	-0.059	(-0.116 , -0.002)
Composite (0.75, 0.25)	-0.379	(-0.506 , -0.252)	-0.389	(-0.539 , -0.239)
Composite (0.25, 0.5)	-0.276	(-0.421 , -0.131)	-0.332	(-0.494 , -0.169)
Composite (0.75, 0.5)	-0.125	(-0.091 , -0.003)	-0.239	(-0.401 , -0.077)
Overall (0.75, 0.25)	-0.352	(-0.439 , -0.265)	-0.312	(-0.402 , -0.223)
Overall (0.5, 0.25)	-0.217	(-0.264 , -0.17)	-0.209	(-0.269 , -0.149)
Overall (0.75, 0.5)	-0.135	(-0.188 , -0.082)	-0.103	(-0.151 , -0.056)

With the exception of one direct effect in Tables 3 and 4 we generally see a protective causal effect of MOUD, typically leading to a statistically significant reduction in HIV injection risk behavior.

Figures 6, 7, and 8 below show there are some differences in the *magnitudes* of estimated RD's based on the community detection method used. The quantity $|RD_{sp}| - |RD_{nm}|$ measures the difference in magnitudes of the estimated RD's under the spectral method (sp) and the node-moving method (nm).

Figure 6: A frequency histogram of the values of $|RD_{sp}| - |RD_{nm}|$, the difference in the magnitudes of the estimated risk differences under the spectral method (sp) and the node-moving method (nm) for the models that are *not* adjusted for the outcome at baseline. Higher frequency where $|RD_{sp}| - |RD_{nm}|$ is positive indicates the spectral method generally resulting larger magnitudes of RD's.

Figure 7: A histogram of the values of $|RD_{sp}| - |RD_{nm}|$, the difference in the magnitudes of the estimated risk differences under the spectral method (sp) and the node-moving method (nm) for the models that *are* adjusted for the outcome at baseline.

Figure 8: A histogram of the values of $|RD_{sp}| - |RD_{nm}|$, the difference in the magnitudes of the estimated risk differences under the spectral method (sp) and the node-moving method (nm) for both the adjusted and unadjusted models.

3.2 Missing Data Results

Table 5 shows the proportions of missing data for the covariates. Only 2.5% of the covariate data are missing which typically implies that complete case analysis can be conducted with respect to the covariates to obtain valid estimates with a possible loss of efficiency. Due to the dependency between individuals, it is important to address missing information and we chose impute the missing data using a non-parametric random forest.

Table 5: Proportion of Missing Covariate Data and the number missing in TRIP. Covariates data were only approximately 2.5% missing.

Baseline Outcome	0.01 (2)
HIV Status	0.00 (0)
Education	0.00 (0)
Employment	0.00 (0)
Marital Status	0.01 (2)
Homelessness	0.00 (1)
Heroin Use	0.00 (0)
Able to Get Medical Care	0.00 (1)
Encouraged to Share Works	0.01 (2)
Total	0.025 (8)

The results of the sensitivity analysis for varying assumptions of the MNAR mechanism for the outcome are shown in Figures 9 to 13. The gray band represents a 95% confidence interval and the points represent the estimated RD's across the range of ψ . A horizontal line is drawn at RD= 0. The horizontal line overlapping the gray band indicates non-significant causal effects (i.e., 0 is included in the confidence interval) at the corresponding value of ψ . A vertical line is drawn at $\psi = 0.5$ indicating the threshold chosen for our main analysis as shown in Tables 3 and 4. The estimated RD's and 95% confidence intervals in Tables 3 and 4 correspond to $\psi = 0.5$. Figures 9 through 13 show the gray band changing very little with respect to ψ . This shows that the causal effect estimates in these cases are robust against varying assumptions about the MNAR mechanism for the outcome. As we decrease the threshold parameter ψ , more of the missing outcomes are imputed with 1 rather than 0. The *indirect*, *composite*, and *overall* effects remain statistically significant across the entire range of ψ .

Figure 9: Sensitivity plots for the Direct Effect under exposure policy 0.25 across the full range of ψ . The vertical axis indicates the estimated RD. The horizontal axis indicates ψ ranging from 1 (where all missing outcomes are assigned the value 0) to 0 (where all missing outcomes are assigned the value 1).

Node-moving Method

Spectral Method

Figure 10: Sensitivity plots for the Direct Effect under exposure policy 0.75 across the full range of ψ . The vertical axis indicates the estimated RD. The horizontal axis indicates ψ ranging from 1 (where all missing outcomes are assigned the value 0) to 0 (where all missing outcomes are assigned the value 1).

Node-moving Method

Spectral Method

Figure 11: Sensitivity plots for the Indirect Effect, $IE(0.75, 0.25)$, across the full range of ψ . The vertical axis indicates the estimated RD. The horizontal axis indicates ψ ranging from 1 (where all missing outcomes are assigned the value 0) to 0 (where all missing outcomes are assigned the value 1).

Node-moving Method

Spectral Method

Figure 12: Sensitivity plots for the Composite Effect, $CE(0.75, 0.25)$, across the full range of ψ . The vertical axis indicates the estimated RD. The horizontal axis indicates ψ ranging from 1 (where all missing outcomes are assigned the value 0) to 0 (where all missing outcomes are assigned the value 1).

Node-moving Method

Spectral Method

Figure 13: Sensitivity plots for the Overall Effect, $OE(0.75, 0.25)$, across the full range of ψ . The vertical axis indicates the estimated RD. The horizontal axis indicates ψ ranging from 1 (where all missing outcomes are assigned the value 0) to 0 (where all missing outcomes are assigned the value 1).

Node-moving Method

Spectral Method

3.3 Community Structure

Figure 14 below shows the identified community structure for both the node-moving and spectral algorithms. Node labels indicate the number of nodes (individuals) in the community, and the size of the edges are scaled to express the number of edges between communities. The node-moving algorithm yields a community structure with 382 edges within communities and 114 edges between communities, and the spectral method results in a community structure with 388 edges within communities and 108 edges between communities.

Figure 14: Community structure found using the node-moving algorithm (left) and the spectral algorithm (right). Nodes represent communities and edges represent edges between communities.

Figure 15 shows the $V(\gamma)$ curves (or VOI curves) where community detection is done using both the node-moving and spectral algorithms for both the largest connected component (giant component) and for the whole network. Each point represents the average VOI , $\bar{V}(\gamma)$, of 30 perturbed graphs. The upper and lower bounds of the bands around the points represent the maximum VOI , $\max\{V(\gamma)\}$, and minimum VOI , $\min\{V(\gamma)\}$, of the 30 perturbed graphs. The closer the gray and red points are to each other, the closer the network structure is to a random graph (null graph) which has *no* community structure [2]. We include the additional maximum and minimum to show variation of $V(\gamma)$ for both the null and real networks. Gray represents the VOI band for the null network, and red represents the VOI band for the TRIP network. The gray line, which has a slope of 1 indicates where the proportion of perturbed edges matches the proportion changed information of the detected communities. Thus, any points above the line represent sensitivity of the community structure to the corresponding change in edge assignments.

Node-moving
(Whole Network)

Spectral
(Whole Network)

Node-moving
(Giant Component)

Spectral
(Giant Component)

Figure 15: Perturbation plots. Robustness of community structure to changes in proportion of randomly reassigning edges. Gray represents the VOI band for the null network, and red represents the VOI band for the TRIP network

For both the node-moving and the spectral methods, we observe what appears to be a significant departure of the network from the null graph. When compared to the null graph the network appears to have a strong community structure because the gray and red bands rarely overlap for values of γ less than 0.5. When γ exceeds 0.5 we see that the null network and the observed network become very close, as larger values of γ cause the network to be closer and closer to a null graph (by definition). However, if we compare the red band to the 45° line, we can see that the community

structure is still highly sensitive to small perturbations. For both the node-moving and the spectral methods, perturbing 10% of the edges in the whole network results in 30% and 34% change in the community structure on average respectively. The VOI ranges between approximately 21% and 31% for the node-moving method and 25% to 35% for the spectral method. When 10% of giant component is perturbed we also see an average of 30% change in the community structure, but ranging from approximately 23% to 33%.

3.4 Finite Sample Correction Confidence Intervals

Table 6 below shows the finite corrected margins of error for all the causal estimators for when $\psi = 0.25$. This value of ψ corresponds to approximately one half of the communities being considered the exposure arm, and the other half as the control arm. The non-informative result is not surprising considering the discussion in Section 2.3 regarding non-informative intervals as well as not having permutation randomization. The margins of error for the finite sample confidence intervals are increasing functions of q and $\binom{n_i}{K_{0,i}}$, and are decreasing functions of N . Large community sizes, like the community sizes resulting from the both community detection methods, are likely to result in a larger value of $\binom{n_i}{K_{0,i}}$, therefore increasing the margins of error.

Effect	Node-Moving	Spectral
DE (0.25)	0.810	0.778
DE (0.75)	1.918	1.595
IE (0.25, 0.75)	2.431	1.871
CE (0.25, 0.75)	2.431	1.871
OE (0.25, 0.75)	2.431	1.871

Table 6: The finite sample corrected margin of error for the 5 causal estimands under both community detection methods. Results are non-informative, however, should be interpreted with caution as the method was only developed for an experimental setting.

To replicate a two-stage permutation randomization, we considered only the communities which had coverages of the two highest frequencies. We restricted the analysis to communities with coverage of approximately 25% (i.e., 20% to 26%) and of approximately 50% (i.e., 45% to 55%). These coverage assignments could plausibly result from a two-stage permutation randomization with coverage levels of approximately 25% as the control arm and the communities with approximately 50% as the exposure arm. The histograms below show the observed coverage levels for the detected communities in the TRIP network.

Figure 16: Observed coverage levels for both the communities found in the node-moving algorithm (left) and the spectral algorithm (right) in TRIP.

3.5 Random Effects Normality

As stated in Section 2.1.2 the random effect $\mathbf{b}_i$ associated with community i is assumed to follow a normal distribution $f_b(\mathbf{b}_i|\sigma_b^2)$ with mean 0 and variance σ_b^2 . Below we show the results of a test for normality [1, 4]. The null hypothesis is a normal distribution for the random effect. We fail to reject the null hypothesis under both community detection methods. Under the node-moving algorithm, the test statistic was 3.96 with a p-value of 0.949. Under the spectral method, the test statistic was 3.27 with a p-value of 0.974.

LIST OF REFERENCES

- [1] Brian G. Barkley. *glmerGOF: Goodness of Fit of Random Effect Distribution in Logistic Mixed Model*, 2019. R package version 0.1.1.
- [2] Karrer, Brian and Levina, Elizaveta and Newman, Mark EJ. Robustness of community structure in networks. *Physical Review E*, 77(4):046119, 2008.
- [3] MacArthur, Georgie J and Minozzi, Silvia and Martin, Natasha and Vickerman, Peter and Deren, Sherry and Bruneau, Julie and Degenhardt, Louisa and Hickman, Matthew. Opiate substitution treatment and HIV transmission in people who inject drugs: systematic review and meta-analysis. *Bmj*, 345:e5945, 2012.
- [4] Eric J Tchetgen and Brent A Coull. A diagnostic test for the mixing distribution in a generalised linear mixed model. *Biometrika*, 93(4):1003–1010, 2006.

4 DISCUSSION

4.1 Causal Estimation

The overall results show that the estimated causal effects under the various conditions (both community detection algorithms as well as controlling and *not* controlling for the baseline outcome) are protective and do achieve statistical significance at the 5% significance level. Considering the results of both tables, it appears the estimated *indirect effect* is statistically significant the most consistently. The other three estimated effects appear to be more sensitive to whether or not we adjust for the baseline outcomes.

The *magnitudes* of the estimated RD's appear to be sensitive to the community detection method used, as shown in Figures 6, 7, and 8. The quantity $|RD_{sp}| - |RD_{nm}|$ measures the difference in magnitudes of the estimated RD's under the spectral method (sp) and the node-moving method (nm). Differences between the RD's are as large as -0.35 , implying that for some causal effect estimate the node-moving algorithm resulted in an estimated RD that 35% larger in magnitude (on the difference scale) than the RD under the spectral method. All three histograms are approximately centered at -0.5 , which implies that the estimates from the node-moving method tend to result in larger magnitudes than the spectral method.

Another limitation of the estimation methods for this network is the implicit assumption of equally sized communities. Under the node-moving and spectral community detection methods there are communities in which the number of participants ranges from 2 to 35, and 1 to 48 respectively. Basse & Feller (2017) allow for varying community sizes using household-weighted (or "community-weighted" for our purposes) estimands in their study on student absenteeism [1]. Extending their methodology to the observational setting for analysis of the TRIP network will allow us to investigate the impact of considering smaller communities to be equally informative as larger

communities. The method used in this paper assumes that individuals, rather than communities, are of equal weight, and thus possibly places too much reliance on information from the larger communities.

4.2 Community Detection

We return to Figure 14 to show the identified community structure for both the node-moving and spectral algorithms. The node-moving algorithm yields a community structure with 382 edges within communities and 114 edges between communities, and the spectral method results in 388 edges within communities and 108 edges between communities. The variation between the two methods comes from the largest connected component, as the communities that are not in the largest connected component are the same in both methods. One feature of Figure 14 is the noticeably large number of edges between communities. This may imply possible violation of the *partial* interference assumption depending on the strength and number of connections between communities. Partial interference is certainly expected to hold for the communities that have few connections to other communities, but it would be difficult to argue for a lack of interference between say, the communities in Figure 14 (left) labeled "23" (blue) and "26" (orange) considering the large number of edges connecting them. The same can be said for the communities in Figure 14 (right) labeled "19" (green) and "33" (blue). Violation of the partial interference assumption implies that the interference sets are incorrectly identified, and thus the potential outcomes are incorrectly indexed. This is not to say that there are no distinct interference sets that allow for partial interference, but community detection may not be the proper method to identify them in such a highly connected network. Because we have seen some sensitivity of the causal estimates to the community detection algorithm and high connectivity between communities as shown in Figure 14, it should

be of little surprise we also found evidence of a weak community structure in the network.

Although the weak community structure may be the cause of varying point estimates between the two community detection algorithms, there does appear to be some consistency with respect to statistical significance and the direction of causation. Regardless of the magnitude of our estimates, there do appear to be significant causal effects of MOUD on reductions in risk behaviors for both community detection methods. The magnitude of these reductions in risk behavior is difficult to ascertain due to doubts about the partial interference assumption. More reliable conclusions are likely to be drawn from a study that defines the interference sets in a way that is more realistic in the setting of a network with weak community structure, such as using *nearest-neighbors* rather than non-overlapping communities that maximize modularity [2]. An additional approach for analyzing the TRIP network could be the auto g-computation approach which allows for arbitrary forms of interference [3].

4.3 Finite Sample Correction

The effect of allowing the parameter ψ to range from 0 to 1 shows almost no effect on how informative the confidence intervals are. This result further supports the necessity to develop methods of identifying interference sets in a network that allow for more communities which are smaller and defined more realistically. Although the modularity maximizing methods we have discussed may identify a community with as many as 48 individuals, it may be invalid to assume that any one person in this community is actually influenced by all 47 of the other people in the community. Identifying such large communities may also result in a value for N that is actually

smaller than the number of interference sets (depending on the dissemination mechanism). This suggests that methods which define interference sets according to each individual's neighbors (i.e., the people only individual j is connected to), such that each individual has their own uniquely defined interference set, may be more valid than community detection in certain cases. Caution should however be taken when making strong conclusions regarding the finite sample confidence intervals as one of the required conditions for their validity is violated by applying it to an observational setting. We essentially treated the second stage of randomization (i.e., the individual level within communities) as having been a Bernoulli allocation strategy when it should actually be permutation randomization. A permutation randomization would result in two arms of the study such that within each arm every community has an observed coverage level that actually corresponds to exposure policy (α or α_o it was assigned).

In addition to sensitivity analysis using the threshold ψ as described earlier, the observed coverage levels were also used to try to replicate the resulting conditions of a two-stage permutation randomization. Unfortunately, this method of restricting the analysis to communities that could be separated into two study arms from a possible two-stage permutation randomization, results in an even smaller number of communities where N becomes equal to 9 rather than 20 or 21. The finite sample confidence intervals in this setting were also non-informative.

LIST OF REFERENCES

- [1] Basse, Guillaume and Feller, Avi. Analyzing two-stage experiments in the presence of interference. *Journal of the American Statistical Association*, 113(521):41–55, 2018.
- [2] Forastiere, Laura and Airoidi, Edoardo M and Mealli, Fabrizia. Identification and estimation of treatment and interference effects in observational studies on networks. *arXiv preprint arXiv:1609.06245*, 2016.
- [3] Fulcher, Isabel Rose. *Statistical Inference for Causal Mechanisms: Mediation and Interference*. PhD thesis, 2019.

5 DIRECTIONS FOR FUTURE RESEARCH

Although many of the estimated causal effects of MOUD on HIV risk behaviors are statistically significant and protective, even under various assumptions about the missing data mechanism, there are some important limitations in this study. One limitation of this study is the non-generalizability of the results beyond Athens, Greece. The subjects in the TRIP dataset were not sampled randomly, and thus do not represent the population of all PWIDs in Athens, Greece, but rather represent PWIDs engaged in a contact tracing effort among those at risk for HIV. Another limitation is that the community detection methods used in this analysis result in large communities, some having over 35 individuals. The potential outcomes of individuals in these large communities, say a community of 48 individuals, are indexed by the remaining 47 individuals. It may be unlikely that any one individual's outcome is affected by the exposure status of all 47 people in their community; however, this depends on the contact structure within the cluster, as a highly connected cluster may have dissemination that reaches further. It may be more plausible that each individual has their own interference set uniquely defined by the individuals with which they share an observed direct connection. Defining interference sets in such a way would shrink the size of the interference sets and allow for more of them to be identified, therefore increasing number of independent units. Increasing the number of interference sets may allow for finite sample corrected confidence intervals that are actually informative. A possible direction for further research is further the relationship between the finite sample margins of error and the quantities $\binom{n_i}{K_{0,i}}$ and N . Further research should also focus on applying finite sample corrections to standard errors in an observational setting.

In our analysis the individuals are assumed to be of equal weight rather than weighting by community size. As a result, there is likely too much reliance on information

from the larger communities. Methodology that allows for communities to be of equal weight will place more reliance on the smaller communities where potential outcomes are more likely to be correctly defined.

Another limitation of this study is that community detection implicitly assumes a strong underlying community structure. As we have seen, different community detection methods which optimize the same quality function (modularity in this case) result in different magnitudes of the estimated RD's. The implication is that there are competing partitions of the network with similarly high modularity scores, but are different enough to change the results of the causal estimation. This suggests a weak community structure in the network as evidenced by the *VOI* curves in Section 4.3. We also see many connections between communities, which is not accounted for by our methods. In fact, the proportion of edges between communities (22 to 23%), as shown in Figure 14, adds doubt to the validity of the partial interference assumption. It is possible that community detection based on covariates could help to alleviate the problem of oversized communities. However, this approach may also fail to identify interference sets in such a way that actually represents the nature of the dissemination.

An alternative approach to community detection should result in the identification of interference sets that not only validate the partial interference assumption, but allow for smaller sizes of interference sets and therefore a larger number of them. A *nearest-neighbor* approach which identifies unique interference sets for each individual will more likely allow for valid assumptions of partial interference and for correctly defined potential outcomes. Another approach which may be of interest for future research would be to index potential outcomes by edge information such as the number of edges within a community. For example, consider community i that has a large number of edges within it and another community i' of equal size and coverage but with many fewer edges within it. It seems plausible that because of the

higher connectivity in community i , dissemination may happen more in community i than in community i' . Some method of indexing potential outcomes in such a way that accounts for such network features may be of interest for future research.

6 CONCLUSION

Besides the causal estimates we have obtained, there are three main conclusions from this thesis. The first is that an analysis of the community structure must precede and guide any causal analysis in the presence of dissemination. A decision should be made based on this analysis as to the most valid method for identifying interference sets. If a network has a strong community structure and reasonably small community sizes, then community detection may be a valid choice. If a network has a weak community structure, or unreasonably large community sizes, then an alternative approach, such as nearest-neighbors, should be considered. The second conclusion is that the method of identifying interference sets should be highly informed by the mechanism of dissemination for the particular experiment being analyzed. The third is that valid finite sample corrections for the standard errors of the causal estimators used in this thesis become non-informative in the presence of large community sizes in conjunction with a small number of communities. Therefore, experiments should be setup in such a way that allow for smaller values of $\binom{n_i}{K_{0,i}}$, and of course larger values of N .

7 APPENDIX A - CONCEPTS FROM INFORMATION THEORY

Self-information is a basic quantity derived from the probability of a particular event occurring from a random variable. It can be interpreted as quantifying the level of "surprise" of a particular outcome, say ω_n . Information, $I(\omega_n)$, is transferred from an originating entity possessing the information to a receiving entity only when the receiver had not known the information a priori [2, 3].

$$P(\omega_n) = 1 \implies I(\omega_n) = 0$$

$$P(\omega_n) < 1 \implies I(\omega_n) > 0$$

Accordingly, the amount of self-information contained in a *message* conveying content informing an occurrence of event, ω_n , depends only on the probability of that event.

$$I(\omega_n) = f(P(\omega_n))$$

If a message informing us about event C is the intersection of independent events A and B , then the information and probability of the event C are

$$I(C) = I(A \cap B) = I(A) + I(B)$$

$$P(C) = P(A \cap B) = P(A)P(B)$$

Applying function f from two slides ago gives the following:

$$\begin{aligned} f(P(C)) &= f(P(A)) + f(P(B)) \\ &= f(P(A)P(B)) \end{aligned}$$

Thus, f is the logarithm function having the property $f(uz) = f(u) + f(z)$. Logarithms of different bases differ only by some scaling constant, k , so $f(u) = k\log(u)$. Since $0 \geq P(\bullet) \leq 1$, then $k < 0$. Therefore, self-information can be written as

$$I(\omega_n) = f(P(\omega_n)) = -\log(P(\omega_n)) = \log\left(\frac{1}{P(\omega_n)}\right)$$

The **entropy** of discrete random variable U , is defined to be the expected information content of measurement of variable U . It can be interpreted as the average level of "surprise", or "uncertainty" inherent in the variable's possible outcomes [2, 3].

$$\begin{aligned} H(U) &= E[I(U)] \\ &= E_u[f(P(u))] \\ &= E_u[-\log(P(u))] \\ &= -\sum_{i=1}^n P(u_i)\log(P(u_i)) \end{aligned}$$

Conditional entropy quantifies the amount of information needed to describe the outcome of a random variable U given that the value of another random variable Z

is known [2, 3].

$$\begin{aligned}
H(Z|X) &= \sum_i P(u_i) H(Z|U = u_i) && \text{(definition)} \\
&= - \sum_i P(u_i) \sum_j P(z_j|u_i) \log P(z_j|u_i) \\
&= - \sum_i \sum_j P(z_j, u_i) \log P(z_j|u_i) \\
&= - \sum_{i,j} P(z_j, u_i) \log P(z_j|u_i) \\
&= - \sum_{i,j} P(z_j, u_i) \log \frac{P(z_j, u_i)}{P(z_j)} \\
&= \sum_{i,j} P(z_j, u_i) \log \frac{P(z_j)}{P(z_j, u_i)}
\end{aligned}$$

The **variation of information** (*VOI*) is a measure of how different a perturbed network's community structure is from the community structure in the unperturbed network. Let u_i be the label of the community to which vertex i belongs in partition C , and let z_i be the label of the community to which vertex i belongs in partition C' . The variation of information between partitions C and C' is then defined as follows [2, 3].

$$\begin{aligned}
V(C, C') &= V(U, Z) \\
&= H(U|Z) + H(Z|U) \\
&= - \sum_{ij} P(u_j, z_i) \log \frac{P(u_j, z_i)}{P(z_i)} - \sum_{ij} P(u_j, z_i) \log \frac{P(u_j, z_i)}{P(u_j)}
\end{aligned}$$

The *variation of information* is the sum of information needed to describe C from C' and the information needed to describe C' from C .

Figure 17 below shows results from a simulation study by Karrer et. al., where *VOI* curves are plotted for varying strengths of community structure. For the 6

simulated networks "b" represents the number of between-group connections, and α denotes the proportion of changed edge assignments in the graph (i.e., γ in our study). The strongest community structure, where $b=7$, results in an "s"-like curve where small proportions of perturbation result in even smaller changes in the community structure [1].

Figure 17: *VOI* Curves for 6 simulated networks. Each point represents the average *VOI* for 100 perturbed graphs.
Source: Karrer et. al. (2008)

LIST OF REFERENCES

- [1] Karrer, Brian and Levina, Elizaveta and Newman, Mark EJ. Robustness of community structure in networks. *Physical Review E*, 77(4):046119, 2008.
- [2] Newman, Mark. *Networks*. Oxford University Press, 2018.
- [3] Pathria, RK and Beale, Paul D. Statistical Mechanics. Elsevier. *New York*, 2011.

8 APPENDIX B - PERMUTATION RANDOMIZATION

Suppose a cluster randomized trial where randomization takes place at the cluster level and the individual level within clusters. Also suppose 2-stage *permutation* randomization (sometimes called a completely randomized design). In an experimental setting where there is 3 clusters, 4 individuals per cluster, 2 treatment allocation policies (α and α_o), and a single binary treatment, a 2-stage permutation randomization would proceed as follows [1, 2].

Randomization 1

Let q be the proportion of clusters to be assigned the treatment policy α where q equals say, $1/3$. Let $S_i = 1$ if cluster i is assigned to α , and let $S_i = 0$ if cluster i is assigned to α_o . Let $\mathcal{S}(3)$ be the vector of length 3 (since our example has 3 clusters) such that $\mathcal{S}(3) = (S_1, S_2, S_3)$. Let $\mathcal{S}(3, 1)$ be the set of all possible vectors $\mathcal{S}(3)$ such that $\sum_{i=1}^3 S_i = 1$. This means that $\mathcal{S}(3, 1)$ represents all the possible treatment allocation strategies where the proportion of clusters assigned to policy α is equal to q . Thus,

$$\mathcal{S}(3, 1) = \left\{ (1, 0, 0), (0, 1, 0), (0, 0, 1) \right\}$$

To complete the first stage of randomization, one of the above vectors from the set $\mathcal{S}(3, 1)$ is randomly selected to be the allocation strategy at the cluster level.

Randomization 2

Let's say for example that the second element of $\mathcal{S}(3, 1)$, $(0, 1, 0)$ is selected at the first randomization. The first cluster is then assigned α_o , the second cluster is assigned α , and the third cluster is assigned α_o . Suppose $\alpha = 1/2$ and $\alpha_o = 0$. Let

$A_{ij} = 0, 1$ for individual j in cluster i . Let $\mathcal{A}_i(4)$ be the vector $(A_{i1}, A_{i2}, A_{i3}, A_{i4})$ which represents the allocation strategy for cluster i . Let $\mathcal{A}(4, 2)$ be the set of all possible vectors $\mathcal{A}_i(4)$ such that $\sum_{j=1}^4 A_{ij} = 2$. The set of all possible treatment allocations for the second cluster, which was assigned $\alpha = 1/2$ is thus:

$$\mathcal{A}(4, 2) = \left\{ (1, 1, 0, 0), (1, 0, 1, 0), (1, 0, 0, 1), (0, 0, 1, 1), (0, 1, 0, 1), (0, 1, 1, 0) \right\}$$

One of the elements of $\mathcal{A}(4, 2)$, i.e., one of the above vectors, is randomly chosen to be the individual allocation strategy within the second cluster. As the first and third clusters are assigned treatment policy α_o , they are each assigned allocation strategy $(0, 0, 0, 0, 0)$.

LIST OF REFERENCES

- [1] Onghena, Patrick. Randomization tests or permutation tests? A historical and terminological clarification. *Randomization, masking, and allocation concealment*, pages 209–227, 2018.
- [2] Yu, Chong Ho. Resampling methods: concepts, applications, and justification. *Practical Assessment, Research, and Evaluation*, 8(1):19, 2002.

9 APPENDIX C - FEATURES OF COMMUNITIES

Tables 7 through 35 are contingency tables showing the conditional distributions of covariates (shown as percentages) for individuals in each community of the largest connected component. The following measures of cohesion are reported for each community: *density*, *global transitivity*, *average transitivity*, and *assortativity*. Density characterizes to extent to which nodes in a graph are connected. It is the ratio of the number of edges in community to the number of possible edges in the community. Let H be a subgraph, a set of nodes (V_H) and edges (E_H) in a community. The value of the density, $\text{den}(H)$, will lie between 0 and 1 [1,2].

$$\text{den}(H) = \frac{|E_H|}{|V_H|(|V_H| - 1)/2}$$

Global transitivity (also called the clustering coefficient of the graph) measures the ratio of the number of triangles (where three nodes are all connected to each other by three edges) to connected triples (where three nodes are connected with two edges) in the graph. Local transitivity of a node is the ratio of the triangles connected to the node, to the triples centered on the node. Taking the average of these gives the average transitivity [1,2]. Assortativity measures the level of homophily of the graph. Homophily is the tendency of individuals to associate with other individuals who exhibit similar attributes [3]. The attribute used in the tables below is the baseline outcome. If the assortativity is positive, connected nodes tend to have similar baseline outcomes. If the assortativity is negative, connected nodes tend to have different baseline outcomes. Assortativity ranges between -1 and 1 , although never actually reaches -1 [1,2].

Table 7: Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 1 in TRIP from the largest connected component under the **node-moving** algorithm. The lower section of the table includes measures of cohesion in the community.

		Overall	MOUD	Non-MOUD
Baseline Outcome	No	18.18	11.11	23.08
	Yes	81.82	88.89	76.92
HIV Status	Negative	54.55	55.56	53.85
	Positive	45.45	44.44	46.15
Education	Primary or Less	18.18	22.22	15.38
	Some HS	72.73	66.67	76.92
	Vocational/College or More	9.09	11.11	7.69
Employed	Employed	22.73	11.11	30.77
	Unemployed or Unable	77.27	88.89	69.23
Partnered	Partnered	22.73	22.22	23.08
	Non-Partnered	77.27	77.78	76.92
Homeless	Not Homeless	72.73	66.67	76.92
	Homeless	27.27	33.33	23.08
Used Heroin	No	27.27	44.44	15.38
	Yes	72.73	55.56	84.62
Able to Get Medical Care	Disagree	27.27	0	46.15
	Agree	72.73	100	53.85
Encouraged to Share Works	No	22.73	22.22	23.08
	Yes	77.27	77.78	76.92

Order	Size (in)	Size (out)	Density	Global Transitivity	Average Transitivity	Assortativity
22	28	5	0.12	0.3	0.46	-0.1

Order is the number of individuals (nodes) in the community. Size (in) is the number of edges connecting nodes inside the community. Size (out) is the number of edges connecting nodes in the community to nodes outside of the community. Assortativity is with respect to the baseline outcome.

Table 8: Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 2 in TRIP from the largest connected component under the **node-moving** algorithm. The lower section of the table includes network information about the community.

		Overall	MOUD	Non-MOUD
Baseline Outcome	No	0	0	0
	Yes	100	100	100
HIV Status	Negative	52.63	42.86	58.33
	Positive	47.37	57.14	41.67
Education	Primary or Less	5.26	14.29	0
	Some HS	68.42	57.14	75
	Vocational/College or More	26.32	28.57	25
Employed	Employed	5.26	0	8.33
	Unemployed or Unable	94.74	100	91.67
Partnered	Partnered	26.32	42.86	16.67
	Non-Partnered	73.68	57.14	83.33
Homeless	Not Homeless	78.95	71.43	83.33
	Homeless	21.05	28.57	16.67
Used Heroin	No	42.11	57.14	33.33
	Yes	57.89	42.86	66.67
Able to Get Medical Care	Disagree	21.05	42.86	8.33
	Agree	78.95	57.14	91.67
Encouraged to Share Works	No	52.63	71.43	41.67
	Yes	47.37	28.57	58.33

Order	Size (in)	Size (out)	Density	Global Transitivity	Average Transitivity	Assortativity
19	29	15	0.17	0.29	0.43	NA

Order is the number of individuals (nodes) in the community. Size (in) is the number of edges connecting nodes inside the community. Size (out) is the number of edges connecting nodes in the community to nodes outside of the community. Assortativity is with respect to the baseline outcome.

Table 9: Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 3 in TRIP from the largest connected component under the **node-moving** algorithm. The lower section of the table includes network information about the community.

		Overall	MOUD	Non-MOUD
Baseline Outcome	No	26.92	25	27.78
	Yes	73.08	75	72.22
HIV Status	Negative	30.77	50	22.22
	Positive	69.23	50	77.78
Education	Primary or Less	30.77	0	44.44
	Some HS	57.69	87.5	44.44
	Vocational/College or More	11.54	12.5	11.11
Employed	Employed	34.62	37.5	33.33
	Unemployed or Unable	65.38	62.5	66.67
Partnered	Partnered	19.23	12.5	22.22
	Non-Partnered	80.77	87.5	77.78
Homeless	Not Homeless	76.92	87.5	72.22
	Homeless	23.08	12.5	27.78
Used Heroin	No	23.08	37.5	16.67
	Yes	76.92	62.5	83.33
Able to Get Medical Care	Disagree	38.46	25	44.44
	Agree	61.54	75	55.56
Encouraged to Share Works	No	50	62.5	44.44
	Yes	50	37.5	55.56

Order	Size (in)	Size (out)	Density	Global Transitivity	Average Transitivity	Assortativity
26	30	22	0.09	0.13	0.27	0.07

Order is the number of individuals (nodes) in the community. Size (in) is the number of edges connecting nodes inside the community. Size (out) is the number of edges connecting nodes in the community to nodes outside of the community. Assortativity is with respect to the baseline outcome.

Table 10: Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 4 in TRIP from the largest connected component under the **node-moving** algorithm. The lower section of the table includes network information about the community.

		Overall	MOUD	Non-MOUD
Baseline Outcome	No	23.08	0	27.27
	Yes	76.92	100	72.73
HIV Status	Negative	26.92	0	31.82
	Positive	73.08	100	68.18
Education	Primary or Less	42.31	0	50
	Some HS	50	100	40.91
	Vocational/College or More	7.69	0	9.09
Employed	Employed	7.69	0	9.09
	Unemployed or Unable	92.31	100	90.91
Partnered	Partnered	38.46	50	36.36
	Non-Partnered	61.54	50	63.64
Homeless	Not Homeless	65.38	100	59.09
	Homeless	34.62	0	40.91
Used Heroin	No	38.46	50	36.36
	Yes	61.54	50	63.64
Able to Get Medical Care	Disagree	38.46	50	36.36
	Agree	61.54	50	63.64
Encouraged to Share Works	No	46.15	25	50
	Yes	53.85	75	50

Order	Size (in)	Size (out)	Density	Global Transitivity	Average Transitivity	Assortativity
26	60	44	0.18	0.36	0.46	0.09

Order is the number of individuals (nodes) in the community. Size (in) is the number of edges connecting nodes inside the community. Size (out) is the number of edges connecting nodes in the community to nodes outside of the community. Assortativity is with respect to the baseline outcome.

Table 11: Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 5 in TRIP from the largest connected component under the **node-moving** algorithm. The lower section of the table includes network information about the community.

		Overall	MOUD	Non-MOUD
Baseline Outcome	No	7.69	0	10
	Yes	92.31	100	90
HIV Status	Negative	15.38	33.33	10
	Positive	84.62	66.67	90
Education	Primary or Less	46.15	33.33	50
	Some HS	53.85	66.67	50
	Vocational/College or More	0	0	0
Employed	Employed	15.38	33.33	10
	Unemployed or Unable	84.62	66.67	90
Partnered	Partnered	7.69	0	10
	Non-Partnered	92.31	100	90
Homeless	Not Homeless	61.54	100	50
	Homeless	38.46	0	50
Used Heroin	No	23.08	33.33	20
	Yes	76.92	66.67	80
Able to Get Medical Care	Disagree	46.15	66.67	40
	Agree	53.85	33.33	60
Encouraged to Share Works	No	69.23	0	90
	Yes	30.77	100	10

Order	Size (in)	Size (out)	Density	Global Transitivity	Average Transitivity	Assortativity
13	12	7	0.15	0	0	-0.09

Order is the number of individuals (nodes) in the community. Size (in) is the number of edges connecting nodes inside the community. Size (out) is the number of edges connecting nodes in the community to nodes outside of the community. Assortativity is with respect to the baseline outcome.

Table 12: Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 6 in TRIP from the largest connected component under the **node-moving** algorithm. The lower section of the table includes network information about the community.

		Overall	MOUD	Non-MOUD
Baseline Outcome	No	8.57	33.33	3.45
	Yes	91.43	66.67	96.55
HIV Status	Negative	42.86	33.33	44.83
	Positive	57.14	66.67	55.17
Education	Primary or Less	40	66.67	34.48
	Some HS	42.86	33.33	44.83
	Vocational/College or More	17.14	0	20.69
Employed	Employed	42.86	50	41.38
	Unemployed or Unable	57.14	50	58.62
Partnered	Partnered	22.86	16.67	24.14
	Non-Partnered	77.14	83.33	75.86
Homeless	Not Homeless	40	33.33	41.38
	Homeless	60	66.67	58.62
Used Heroin	No	22.86	16.67	24.14
	Yes	77.14	83.33	75.86
Able to Get Medical Care	Disagree	37.14	16.67	41.38
	Agree	62.86	83.33	58.62
Encouraged to Share Works	No	34.29	66.67	27.59
	Yes	65.71	33.33	72.41

Order	Size (in)	Size (out)	Density	Global Transitivity	Average Transitivity	Assortativity
35	63	41	0.11	0.22	0.45	-0.04

Order is the number of individuals (nodes) in the community. Size (in) is the number of edges connecting nodes inside the community. Size (out) is the number of edges connecting nodes in the community to nodes outside of the community. Assortativity is with respect to the baseline outcome.

Table 13: Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 7 in TRIP from the largest connected component under the **node-moving** algorithm. The lower section of the table includes network information about the community.

		Overall	MOUD	Non-MOUD
Baseline Outcome	No	14.29	NA	14.29
	Yes	85.71	NA	85.71
HIV Status	Negative	64.29	NA	64.29
	Positive	35.71	NA	35.71
Education	Primary or Less	35.71	NA	35.71
	Some HS	64.29	NA	64.29
	Vocational/College or More	0	NA	0
Employed	Employed	28.57	NA	28.57
	Unemployed or Unable	71.43	NA	71.43
Partnered	Partnered	21.43	NA	21.43
	Non-Partnered	78.57	NA	78.57
Homeless	Not Homeless	92.86	NA	92.86
	Homeless	7.14	NA	7.14
Used Heroin	No	7.14	NA	7.14
	Yes	92.86	NA	92.86
Able to Get Medical Care	Disagree	14.29	NA	14.29
	Agree	85.71	NA	85.71
Encouraged to Share Works	No	35.71	NA	35.71
	Yes	64.29	NA	64.29

Order	Size (in)	Size (out)	Density	Global Transitivity	Average Transitivity	Assortativity
14	19	2	0.21	0.24	0.38	-0.12

Order is the number of individuals (nodes) in the community. Size (in) is the number of edges connecting nodes inside the community. Size (out) is the number of edges connecting nodes in the community to nodes outside of the community. Assortativity is with respect to the baseline outcome.

Table 14: Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 8 in TRIP from the largest connected component under the **node-moving** algorithm. The lower section of the table includes network information about the community.

		Overall	MOUD	Non-MOUD
Baseline Outcome	No	0	0	0
	Yes	100	100	100
HIV Status	Negative	30.77	25	33.33
	Positive	69.23	75	66.67
Education	Primary or Less	30.77	50	22.22
	Some HS	53.85	50	55.56
	Vocational/College or More	15.38	0	22.22
Employed	Employed	46.15	25	55.56
	Unemployed or Unable	53.85	75	44.44
Partnered	Partnered	15.38	0	22.22
	Non-Partnered	84.62	100	77.78
Homeless	Not Homeless	84.62	75	88.89
	Homeless	15.38	25	11.11
Used Heroin	No	38.46	50	33.33
	Yes	61.54	50	66.67
Able to Get Medical Care	Disagree	23.08	0	33.33
	Agree	76.92	100	66.67
Encouraged to Share Works	No	30.77	25	33.33
	Yes	69.23	75	66.67

Order	Size (in)	Size (out)	Density	Global Transitivity	Average Transitivity	Assortativity
13	17	9	0.22	0.28	0.31	NA

Order is the number of individuals (nodes) in the community. Size (in) is the number of edges connecting nodes inside the community. Size (out) is the number of edges connecting nodes in the community to nodes outside of the community. Assortativity is with respect to the baseline outcome.

Table 15: Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 9 in TRIP from the largest connected component under the **node-moving** algorithm. The lower section of the table includes network information about the community.

		Overall	MOUD	Non-MOUD
Baseline Outcome	No	0	NA	0
	Yes	100	NA	100
HIV Status	Negative	75	NA	75
	Positive	25	NA	25
Education	Primary or Less	25	NA	25
	Some HS	75	NA	75
	Vocational/College or More	0	NA	0
Employed	Employed	75	NA	75
	Unemployed or Unable	25	NA	25
Partnered	Partnered	0	NA	0
	Non-Partnered	100	NA	100
Homeless	Not Homeless	100	NA	100
	Homeless	0	NA	0
Used Heroin	No	0	NA	0
	Yes	100	NA	100
Able to Get Medical Care	Disagree	100	NA	100
	Agree	0	NA	0
Encouraged to Share Works	No	50	NA	50
	Yes	50	NA	50

Order	Size (in)	Size (out)	Density	Global Transitivity	Average Transitivity	Assortativity
4	4	2	0.67	0.6	0.78	NA

Order is the number of individuals (nodes) in the community. Size (in) is the number of edges connecting nodes inside the community. Size (out) is the number of edges connecting nodes in the community to nodes outside of the community. Assortativity is with respect to the baseline outcome.

Table 16: Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 10 in TRIP from the largest connected component under the **node-moving** algorithm. The lower section of the table includes network information about the community.

		Overall	MOUD	Non-MOUD
Baseline Outcome	No	21.74	100	14.29
	Yes	78.26	0	85.71
HIV Status	Negative	39.13	0	42.86
	Positive	60.87	100	57.14
Education	Primary or Less	30.43	0	33.33
	Some HS	56.52	100	52.38
	Vocational/College or More	13.04	0	14.29
Employed	Employed	47.83	100	42.86
	Unemployed or Unable	52.17	0	57.14
Partnered	Partnered	21.74	0	23.81
	Non-Partnered	78.26	100	76.19
Homeless	Not Homeless	60.87	100	57.14
	Homeless	39.13	0	42.86
Used Heroin	No	60.87	50	61.9
	Yes	39.13	50	38.1
Able to Get Medical Care	Disagree	47.83	100	42.86
	Agree	52.17	0	57.14
Encouraged to Share Works	No	34.78	50	33.33
	Yes	65.22	50	66.67

Order	Size (in)	Size (out)	Density	Global Transitivity	Average Transitivity	Assortativity
23	54	53	0.21	0.45	0.63	-0.07

Order is the number of individuals (nodes) in the community. Size (in) is the number of edges connecting nodes inside the community. Size (out) is the number of edges connecting nodes in the community to nodes outside of the community. Assortativity is with respect to the baseline outcome.

Table 17: Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 11 in TRIP from the largest connected component under the **node-moving** algorithm. The lower section of the table includes network information about the community.

		Overall	MOUD	Non-MOUD
Baseline Outcome	No	27.27	0	33.33
	Yes	72.73	100	66.67
HIV Status	Negative	45.45	0	55.56
	Positive	54.55	100	44.44
Education	Primary or Less	45.45	50	44.44
	Some HS	36.36	50	33.33
	Vocational/College or More	18.18	0	22.22
Employed	Employed	9.09	0	11.11
	Unemployed or Unable	90.91	100	88.89
Partnered	Partnered	45.45	100	33.33
	Non-Partnered	54.55	0	66.67
Homeless	Not Homeless	81.82	100	77.78
	Homeless	18.18	0	22.22
Used Heroin	No	9.09	0	11.11
	Yes	90.91	100	88.89
Able to Get Medical Care	Disagree	45.45	100	33.33
	Agree	54.55	0	66.67
Encouraged to Share Works	No	72.73	100	66.67
	Yes	27.27	0	33.33

Order	Size (in)	Size (out)	Density	Global Transitivity	Average Transitivity	Assortativity
11	23	23	0.42	0.61	0.75	0.59

Order is the number of individuals (nodes) in the community. Size (in) is the number of edges connecting nodes inside the community. Size (out) is the number of edges connecting nodes in the community to nodes outside of the community. Assortativity is with respect to the baseline outcome.

Table 18: Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 12 in TRIP from the largest connected component under the **node-moving** algorithm. The lower section of the table includes network information about the community.

		Overall	MOUD	Non-MOUD
Baseline Outcome	No	25	0	50
	Yes	75	100	50
HIV Status	Negative	75	50	100
	Positive	25	50	0
Education	Primary or Less	0	0	0
	Some HS	100	100	100
	Vocational/College or More	0	0	0
Employed	Employed	0	0	0
	Unemployed or Unable	100	100	100
Partnered	Partnered	25	50	0
	Non-Partnered	75	50	100
Homeless	Not Homeless	100	100	100
	Homeless	0	0	0
Used Heroin	No	25	0	50
	Yes	75	100	50
Able to Get Medical Care	Disagree	25	50	0
	Agree	75	50	100
Encouraged to Share Works	No	75	50	100
	Yes	25	50	0

Order	Size (in)	Size (out)	Density	Global Transitivity	Average Transitivity	Assortativity
4	3	1	0.5	0	0	-0.2

Order is the number of individuals (nodes) in the community. Size (in) is the number of edges connecting nodes inside the community. Size (out) is the number of edges connecting nodes in the community to nodes outside of the community. Assortativity is with respect to the baseline outcome.

Table 19: Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 13 in TRIP from the largest connected component under the **node-moving** algorithm. The lower section of the table includes network information about the community.

		Overall	MOUD	Non-MOUD
Baseline Outcome	No	0	NA	0
	Yes	100	NA	100
HIV Status	Negative	100	NA	100
	Positive	0	NA	0
Education	Primary or Less	20	NA	20
	Some HS	60	NA	60
	Vocational/College or More	20	NA	20
Employed	Employed	60	NA	60
	Unemployed or Unable	40	NA	40
Partnered	Partnered	40	NA	40
	Non-Partnered	60	NA	60
Homeless	Not Homeless	20	NA	20
	Homeless	80	NA	80
Used Heroin	No	0	NA	0
	Yes	100	NA	100
Able to Get Medical Care	Disagree	40	NA	40
	Agree	60	NA	60
Encouraged to Share Works	No	40	NA	40
	Yes	60	NA	60

Order	Size (in)	Size (out)	Density	Global Transitivity	Average Transitivity	Assortativity
5	6	2	0.6	0.55	0.75	NA

Order is the number of individuals (nodes) in the community. Size (in) is the number of edges connecting nodes inside the community. Size (out) is the number of edges connecting nodes in the community to nodes outside of the community. Assortativity is with respect to the baseline outcome.

Table 20: Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 14 in TRIP from the largest connected component under the **node-moving** algorithm. The lower section of the table includes network information about the community.

		Overall	MOUD	Non-MOUD
Baseline Outcome	No	75	66.67	100
	Yes	25	33.33	0
HIV Status	Negative	75	66.67	100
	Positive	25	33.33	0
Education	Primary or Less	0	0	0
	Some HS	75	66.67	100
	Vocational/College or More	25	33.33	0
Employed	Employed	25	0	100
	Unemployed or Unable	75	100	0
Partnered	Partnered	25	0	100
	Non-Partnered	75	100	0
Homeless	Not Homeless	100	100	100
	Homeless	0	0	0
Used Heroin	No	25	33.33	0
	Yes	75	66.67	100
Able to Get Medical Care	Disagree	50	33.33	100
	Agree	50	66.67	0
Encouraged to Share Works	No	75	66.67	100
	Yes	25	33.33	0

Order	Size (in)	Size (out)	Density	Global Transitivity	Average Transitivity	Assortativity
4	3	1	0.5	0	0	-0.5

Order is the number of individuals (nodes) in the community. Size (in) is the number of edges connecting nodes inside the community. Size (out) is the number of edges connecting nodes in the community to nodes outside of the community. Assortativity is with respect to the baseline outcome.

Table 21: Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 15 in TRIP from the largest connected component under the **node-moving** algorithm. The lower section of the table includes network information about the community.

		Overall	MOUD	Non-MOUD
Baseline Outcome	No	25	NA	25
	Yes	75	NA	75
HIV Status	Negative	75	NA	75
	Positive	25	NA	25
Education	Primary or Less	50	NA	50
	Some HS	50	NA	50
	Vocational/College or More	0	NA	0
Employed	Employed	50	NA	50
	Unemployed or Unable	50	NA	50
Partnered	Partnered	75	NA	75
	Non-Partnered	25	NA	25
Homeless	Not Homeless	75	NA	75
	Homeless	25	NA	25
Used Heroin	No	0	NA	0
	Yes	100	NA	100
Able to Get Medical Care	Disagree	50	NA	50
	Agree	50	NA	50
Encouraged to Share Works	No	50	NA	50
	Yes	50	NA	50

Order	Size (in)	Size (out)	Density	Global Transitivity	Average Transitivity	Assortativity
4	4	1	0.67	0.6	0.78	-0.33

Order is the number of individuals (nodes) in the community. Size (in) is the number of edges connecting nodes inside the community. Size (out) is the number of edges connecting nodes in the community to nodes outside of the community. Assortativity is with respect to the baseline outcome.

Table 22: Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 1 in TRIP from the largest connected component under the **spectral** algorithm. The lower section of the table includes network information about the community.

		Overall	MOUD	Non-MOUD
Baseline Outcome	No	24.24	0	29.63
	Yes	75.76	100	70.37
HIV Status	Negative	36.36	16.67	40.74
	Positive	63.64	83.33	59.26
Education	Primary or Less	36.36	0	44.44
	Some HS	54.55	100	44.44
	Vocational/College or More	9.09	0	11.11
Employed	Employed	6.06	0	7.41
	Unemployed or Unable	93.94	100	92.59
Partnered	Partnered	36.36	50	33.33
	Non-Partnered	63.64	50	66.67
Homeless	Not Homeless	69.7	100	62.96
	Homeless	30.3	0	37.04
Used Heroin	No	36.36	33.33	37.04
	Yes	63.64	66.67	62.96
Able to Get Medical Care	Disagree	33.33	50	29.63
	Agree	66.67	50	70.37
Encouraged to Share Works	No	51.52	33.33	55.56
	Yes	48.48	66.67	44.44

Order	Size (in)	Size (out)	Density	Global Transitivity	Average Transitivity	Assortativity
33	82	40	0.16	0.33	0.43	0.04

Order is the number of individuals (nodes) in the community. Size (in) is the number of edges connecting nodes inside the community. Size (out) is the number of edges connecting nodes in the community to nodes outside of the community. Assortativity is with respect to the baseline outcome.

Table 23: Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 2 in TRIP from the largest connected component under the **spectral** algorithm. The lower section of the table includes network information about the community.

		Overall	MOUD	Non-MOUD
Baseline Outcome	No	16.67	0	28.57
	Yes	83.33	100	71.43
HIV Status	Negative	50	40	57.14
	Positive	50	60	42.86
Education	Primary or Less	25	40	14.29
	Some HS	58.33	40	71.43
	Vocational/College or More	16.67	20	14.29
Employed	Employed	41.67	20	57.14
	Unemployed or Unable	58.33	80	42.86
Partnered	Partnered	16.67	0	28.57
	Non-Partnered	83.33	100	71.43
Homeless	Not Homeless	75	60	85.71
	Homeless	25	40	14.29
Used Heroin	No	25	60	0
	Yes	75	40	100
Able to Get Medical Care	Disagree	16.67	0	28.57
	Agree	83.33	100	71.43
Encouraged to Share Works	No	8.33	0	14.29
	Yes	91.67	100	85.71

Order	Size (in)	Size (out)	Density	Global Transitivity	Average Transitivity	Assortativity
12	12	4	0.18	0.19	0.29	-0.26

Order is the number of individuals (nodes) in the community. Size (in) is the number of edges connecting nodes inside the community. Size (out) is the number of edges connecting nodes in the community to nodes outside of the community. Assortativity is with respect to the baseline outcome.

Table 24: Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 3 in TRIP from the largest connected component under the **spectral** algorithm. The lower section of the table includes network information about the community.

		Overall	MOUD	Non-MOUD
Baseline Outcome	No	21.05	100	16.67
	Yes	78.95	0	83.33
HIV Status	Negative	42.11	0	44.44
	Positive	57.89	100	55.56
Education	Primary or Less	36.84	0	38.89
	Some HS	47.37	100	44.44
	Vocational/College or More	15.79	0	16.67
Employed	Employed	47.37	100	44.44
	Unemployed or Unable	52.63	0	55.56
Partnered	Partnered	21.05	0	22.22
	Non-Partnered	78.95	100	77.78
Homeless	Not Homeless	57.89	100	55.56
	Homeless	42.11	0	44.44
Used Heroin	No	63.16	100	61.11
	Yes	36.84	0	38.89
Able to Get Medical Care	Disagree	52.63	100	50
	Agree	47.37	0	50
Encouraged to Share Works	No	36.84	0	38.89
	Yes	63.16	100	61.11

Order	Size (in)	Size (out)	Density	Global Transitivity	Average Transitivity	Assortativity
19	43	54	0.25	0.46	0.66	-0.1

Order is the number of individuals (nodes) in the community. Size (in) is the number of edges connecting nodes inside the community. Size (out) is the number of edges connecting nodes in the community to nodes outside of the community. Assortativity is with respect to the baseline outcome.

Table 25: Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 4 in TRIP from the largest connected component under the **spectral** algorithm. The lower section of the table includes network information about the community.

		Overall	MOUD	Non-MOUD
Baseline Outcome	No	8.33	0	11.11
	Yes	91.67	100	88.89
HIV Status	Negative	41.67	0	55.56
	Positive	58.33	100	44.44
Education	Primary or Less	41.67	33.33	44.44
	Some HS	50	66.67	44.44
	Vocational/College or More	8.33	0	11.11
Employed	Employed	8.33	0	11.11
	Unemployed or Unable	91.67	100	88.89
Partnered	Partnered	50	66.67	44.44
	Non-Partnered	50	33.33	55.56
Homeless	Not Homeless	83.33	66.67	88.89
	Homeless	16.67	33.33	11.11
Used Heroin	No	8.33	0	11.11
	Yes	91.67	100	88.89
Able to Get Medical Care	Disagree	41.67	66.67	33.33
	Agree	58.33	33.33	66.67
Encouraged to Share Works	No	58.33	66.67	55.56
	Yes	41.67	33.33	44.44

Order	Size (in)	Size (out)	Density	Global Transitivity	Average Transitivity	Assortativity
12	22	18	0.33	0.59	0.62	-0.07

Order is the number of individuals (nodes) in the community. Size (in) is the number of edges connecting nodes inside the community. Size (out) is the number of edges connecting nodes in the community to nodes outside of the community. Assortativity is with respect to the baseline outcome.

Table 26: Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 5 in TRIP from the largest connected component under the **spectral** algorithm. The lower section of the table includes network information about the community.

		Overall	MOUD	Non-MOUD
Baseline Outcome	No	6.25	6.67	6.06
	Yes	93.75	93.33	93.94
HIV Status	Negative	37.5	40	36.36
	Positive	62.5	60	63.64
Education	Primary or Less	20.83	26.67	18.18
	Some HS	62.5	53.33	66.67
	Vocational/College or More	16.67	20	15.15
Employed	Employed	22.92	26.67	21.21
	Unemployed or Unable	77.08	73.33	78.79
Partnered	Partnered	12.5	20	9.09
	Non-Partnered	87.5	80	90.91
Homeless	Not Homeless	75	73.33	75.76
	Homeless	25	26.67	24.24
Used Heroin	No	43.75	60	36.36
	Yes	56.25	40	63.64
Able to Get Medical Care	Disagree	31.25	33.33	30.3
	Agree	68.75	66.67	69.7
Encouraged to Share Works	No	56.25	60	54.55
	Yes	43.75	40	45.45

Order	Size (in)	Size (out)	Density	Global Transitivity	Average Transitivity	Assortativity
48	70	32	0.06	0.2	0.29	-0.08

Order is the number of individuals (nodes) in the community. Size (in) is the number of edges connecting nodes inside the community. Size (out) is the number of edges connecting nodes in the community to nodes outside of the community. Assortativity is with respect to the baseline outcome.

Table 27: Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 6 in TRIP from the largest connected component under the **spectral** algorithm. The lower section of the table includes network information about the community.

		Overall	MOUD	Non-MOUD
Baseline Outcome	No	37.5	0	50
	Yes	62.5	100	50
HIV Status	Negative	12.5	50	0
	Positive	87.5	50	100
Education	Primary or Less	37.5	0	50
	Some HS	50	100	33.33
	Vocational/College or More	12.5	0	16.67
Employed	Employed	25	50	16.67
	Unemployed or Unable	75	50	83.33
Partnered	Partnered	50	50	50
	Non-Partnered	50	50	50
Homeless	Not Homeless	62.5	100	50
	Homeless	37.5	0	50
Used Heroin	No	0	0	0
	Yes	100	100	100
Able to Get Medical Care	Disagree	50	50	50
	Agree	50	50	50
Encouraged to Share Works	No	62.5	100	50
	Yes	37.5	0	50

Order	Size (in)	Size (out)	Density	Global Transitivity	Average Transitivity	Assortativity
8	8	6	0.29	0.2	0.36	-0.27

Order is the number of individuals (nodes) in the community. Size (in) is the number of edges connecting nodes inside the community. Size (out) is the number of edges connecting nodes in the community to nodes outside of the community. Assortativity is with respect to the baseline outcome.

Table 28: Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 7 in TRIP from the largest connected component under the **spectral** algorithm. The lower section of the table includes network information about the community.

		Overall	MOUD	Non-MOUD
Baseline Outcome	No	15	33.33	11.76
	Yes	85	66.67	88.24
HIV Status	Negative	75	100	70.59
	Positive	25	0	29.41
Education	Primary or Less	30	0	35.29
	Some HS	70	100	64.71
	Vocational/College or More	0	0	0
Employed	Employed	20	0	23.53
	Unemployed or Unable	80	100	76.47
Partnered	Partnered	30	66.67	23.53
	Non-Partnered	70	33.33	76.47
Homeless	Not Homeless	85	66.67	88.24
	Homeless	15	33.33	11.76
Used Heroin	No	5	0	5.88
	Yes	95	100	94.12
Able to Get Medical Care	Disagree	25	0	29.41
	Agree	75	100	70.59
Encouraged to Share Works	No	35	33.33	35.29
	Yes	65	66.67	64.71

Order	Size (in)	Size (out)	Density	Global Transitivity	Average Transitivity	Assortativity
20	29	3	0.15	0.32	0.57	-0.12

Order is the number of individuals (nodes) in the community. Size (in) is the number of edges connecting nodes inside the community. Size (out) is the number of edges connecting nodes in the community to nodes outside of the community. Assortativity is with respect to the baseline outcome.

Table 29: Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 8 in TRIP from the largest connected component under the **spectral** algorithm. The lower section of the table includes network information about the community.

		Overall	MOUD	Non-MOUD
Baseline Outcome	No	0	NA	0
	Yes	100	NA	100
HIV Status	Negative	66.67	NA	66.67
	Positive	33.33	NA	33.33
Education	Primary or Less	0	NA	0
	Some HS	100	NA	100
	Vocational/College or More	0	NA	0
Employed	Employed	100	NA	100
	Unemployed or Unable	0	NA	0
Partnered	Partnered	0	NA	0
	Non-Partnered	100	NA	100
Homeless	Not Homeless	100	NA	100
	Homeless	0	NA	0
Used Heroin	No	0	NA	0
	Yes	100	NA	100
Able to Get Medical Care	Disagree	100	NA	100
	Agree	0	NA	0
Encouraged to Share Works	No	33.33	NA	33.33
	Yes	66.67	NA	66.67

Order	Size (in)	Size (out)	Density	Global Transitivity	Average Transitivity	Assortativity
3	3	1	1	1	1	NA

Order is the number of individuals (nodes) in the community. Size (in) is the number of edges connecting nodes inside the community. Size (out) is the number of edges connecting nodes in the community to nodes outside of the community. Assortativity is with respect to the baseline outcome.

Table 30: Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 9 in TRIP from the largest connected component under the **spectral** algorithm. The lower section of the table includes network information about the community.

		Overall	MOUD	Non-MOUD
Baseline Outcome	No	0	0	0
	Yes	100	100	100
HIV Status	Negative	0	0	0
	Positive	100	100	100
Education	Primary or Less	25	50	0
	Some HS	75	50	100
	Vocational/College or More	0	0	0
Employed	Employed	25	0	50
	Unemployed or Unable	75	100	50
Partnered	Partnered	25	0	50
	Non-Partnered	75	100	50
Homeless	Not Homeless	75	100	50
	Homeless	25	0	50
Used Heroin	No	50	50	50
	Yes	50	50	50
Able to Get Medical Care	Disagree	0	0	0
	Agree	100	100	100
Encouraged to Share Works	No	0	0	0
	Yes	100	100	100

Order	Size (in)	Size (out)	Density	Global Transitivity	Average Transitivity	Assortativity
4	3	4	0.5	0	0	NA

Order is the number of individuals (nodes) in the community. Size (in) is the number of edges connecting nodes inside the community. Size (out) is the number of edges connecting nodes in the community to nodes outside of the community. Assortativity is with respect to the baseline outcome.

Table 31: Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 10 in TRIP from the largest connected component under the **spectral** algorithm. The lower section of the table includes network information about the community.

		Overall	MOUD	Non-MOUD
Baseline Outcome	No	0	0	NA
	Yes	100	100	NA
HIV Status	Negative	0	0	NA
	Positive	100	100	NA
Education	Primary or Less	0	0	NA
	Some HS	100	100	NA
	Vocational/College or More	0	0	NA
Employed	Employed	0	0	NA
	Unemployed or Unable	100	100	NA
Partnered	Partnered	0	0	NA
	Non-Partnered	100	100	NA
Homeless	Not Homeless	100	100	NA
	Homeless	0	0	NA
Used Heroin	No	50	50	NA
	Yes	50	50	NA
Able to Get Medical Care	Disagree	100	100	NA
	Agree	0	0	NA
Encouraged to Share Works	No	0	0	NA
	Yes	100	100	NA

Order	Size (in)	Size (out)	Density	Global Transitivity	Average Transitivity	Assortativity
2	1	1	1	NA	NA	NA

Order is the number of individuals (nodes) in the community. Size (in) is the number of edges connecting nodes inside the community. Size (out) is the number of edges connecting nodes in the community to nodes outside of the community. Assortativity is with respect to the baseline outcome.

Table 32: Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 11 in TRIP from the largest connected component under the **spectral** algorithm. The lower section of the table includes network information about the community.

		Overall	MOUD	Non-MOUD
Baseline Outcome	No	25	NA	25
	Yes	75	NA	75
HIV Status	Negative	75	NA	75
	Positive	25	NA	25
Education	Primary or Less	50	NA	50
	Some HS	50	NA	50
	Vocational/College or More	0	NA	0
Employed	Employed	50	NA	50
	Unemployed or Unable	50	NA	50
Partnered	Partnered	75	NA	75
	Non-Partnered	25	NA	25
Homeless	Not Homeless	75	NA	75
	Homeless	25	NA	25
Used Heroin	No	0	NA	0
	Yes	100	NA	100
Able to Get Medical Care	Disagree	50	NA	50
	Agree	50	NA	50
Encouraged to Share Works	No	50	NA	50
	Yes	50	NA	50

Order	Size (in)	Size (out)	Density	Global Transitivity	Average Transitivity	Assortativity
4	4	1	0.67	0.6	0.78	-0.33

Order is the number of individuals (nodes) in the community. Size (in) is the number of edges connecting nodes inside the community. Size (out) is the number of edges connecting nodes in the community to nodes outside of the community. Assortativity is with respect to the baseline outcome.

Table 33: Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 12 in TRIP from the largest connected component under the **spectral** algorithm. The lower section of the table includes network information about the community.

		Overall	MOUD	Non-MOUD
Baseline Outcome	No	50	66.67	33.33
	Yes	50	33.33	66.67
HIV Status	Negative	66.67	66.67	66.67
	Positive	33.33	33.33	33.33
Education	Primary or Less	16.67	0	33.33
	Some HS	83.33	100	66.67
	Vocational/College or More	0	0	0
Employed	Employed	66.67	33.33	100
	Unemployed or Unable	33.33	66.67	0
Partnered	Partnered	16.67	0	33.33
	Non-Partnered	83.33	100	66.67
Homeless	Not Homeless	100	100	100
	Homeless	0	0	0
Used Heroin	No	16.67	33.33	0
	Yes	83.33	66.67	100
Able to Get Medical Care	Disagree	0	0	0
	Agree	100	100	100
Encouraged to Share Works	No	83.33	100	66.67
	Yes	16.67	0	33.33

Order	Size (in)	Size (out)	Density	Global Transitivity	Average Transitivity	Assortativity
6	5	1	0.33	0	0	-0.43

Order is the number of individuals (nodes) in the community. Size (in) is the number of edges connecting nodes inside the community. Size (out) is the number of edges connecting nodes in the community to nodes outside of the community. Assortativity is with respect to the baseline outcome.

Table 34: Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 13 in TRIP from the largest connected component under the **spectral** algorithm. The lower section of the table includes network information about the community.

		Overall	MOUD	Non-MOUD
Baseline Outcome	No	75	66.67	100
	Yes	25	33.33	0
HIV Status	Negative	75	66.67	100
	Positive	25	33.33	0
Education	Primary or Less	0	0	0
	Some HS	75	66.67	100
	Vocational/College or More	25	33.33	0
Employed	Employed	25	0	100
	Unemployed or Unable	75	100	0
Partnered	Partnered	25	0	100
	Non-Partnered	75	100	0
Homeless	Not Homeless	100	100	100
	Homeless	0	0	0
Used Heroin	No	25	33.33	0
	Yes	75	66.67	100
Able to Get Medical Care	Disagree	50	33.33	100
	Agree	50	66.67	0
Encouraged to Share Works	No	75	66.67	100
	Yes	25	33.33	0

Order	Size (in)	Size (out)	Density	Global Transitivity	Average Transitivity	Assortativity
4	3	1	0.5	0	0	-0.5

Order is the number of individuals (nodes) in the community. Size (in) is the number of edges connecting nodes inside the community. Size (out) is the number of edges connecting nodes in the community to nodes outside of the community. Assortativity is with respect to the baseline outcome.

Table 35: Contingency table showing the conditional distributions of covariates (shown as percentages) for individuals in Community 13 in TRIP from the largest connected component under the **spectral** algorithm. The lower section of the table includes network information about the community.

		Overall	MOUD	Non-MOUD
Baseline Outcome	No	8.89	25	7.32
	Yes	91.11	75	92.68
HIV Status	Negative	44.44	50	43.9
	Positive	55.56	50	56.1
Education	Primary or Less	40	75	36.59
	Some HS	42.22	25	43.9
	Vocational/College or More	17.78	0	19.51
Employed	Employed	40	50	39.02
	Unemployed or Unable	60	50	60.98
Partnered	Partnered	22.22	25	21.95
	Non-Partnered	77.78	75	78.05
Homeless	Not Homeless	42.22	25	43.9
	Homeless	57.78	75	56.1
Used Heroin	No	20	0	21.95
	Yes	80	100	78.05
Able to Get Medical Care	Disagree	37.78	0	41.46
	Agree	62.22	100	58.54
Encouraged to Share Works	No	33.33	50	31.71
	Yes	66.67	50	68.29

Order	Size (in)	Size (out)	Density	Global Transitivity	Average Transitivity	Assortativity
45	76	45	0.08	0.19	0.36	-0.04

Order is the number of individuals (nodes) in the community. Size (in) is the number of edges connecting nodes inside the community. Size (out) is the number of edges connecting nodes in the community to nodes outside of the community. Assortativity is with respect to the baseline outcome.

LIST OF REFERENCES

- [1] Eric D Kolaczyk and Gábor Csárdi. *Statistical Analysis of Network Data With R*, volume 65. Springer, 2014.
- [2] Newman, Mark. *Networks*. Oxford University Press, 2018.
- [3] Nicholas A Christakis and James H. Fowler. Social contagion theory: examining dynamic social networks and human behavior. *Statistics in Medicine*, 32.4:556–577, 2013.

BIBLIOGRAPHY

- [1] Basse, Guillaume and Feller, Avi. Analyzing two-stage experiments in the presence of interference. *Journal of the American Statistical Association*, 113(521):41–55, 2018.
- [2] Benjamin-Chung, Jade and Arnold, Benjamin F and Berger, David and Luby, Stephen P and Miguel, Edward and Colford Jr, John M and Hubbard, Alan E. Spillover effects in epidemiology: parameters, study designs and methodological considerations. *International Journal of Epidemiology*, 47(1):332–347, 2017.
- [3] Bradley C. Saul and Michael G. Hudgens. A Recipe for inference: Start with Causal Inference. Add Interference. Mix Well with R. *Journal of Statistical Software*, 82(2):1–21, 2017.
- [4] Brian G. Barkley. *glmerGOF: Goodness of Fit of Random Effect Distribution in Logistic Mixed Model*, 2019. R package version 0.1.1.
- [5] Camacho, L Mabel and Bartholomew, Norma G and Joe, George W and Cloud, Melynda A and Simpson, D Dwayne. Gender, cocaine and during-treatment HIV risk reduction among injection opioid users in methadone maintenance. *Drug and Alcohol Dependence*, 41(1):1–7, 1996.
- [6] Carlson, Robert G and Daniulaityte, Raminta and Silverstein, Sydney M and Nahhas, Ramzi W and Martins, Silvia S. Unintentional drug overdose: Is

more frequent use of non-prescribed buprenorphine associated with lower risk of overdose? *International Journal of Drug Policy*, 79:102722, 2020.

- [7] Carroll, Jennifer J and Rich, Josiah D and Green, Traci C. The more things change: buprenorphine/naloxone diversion continues while treatment remains inaccessible. *Journal of Addiction Medicine*, 12(6):459, 2018.
- [8] Clauset, Aaron and Newman, Mark EJ and Moore, Cristopher. Finding community structure in very large networks. *Physical Review E*, 70(6):066111, 2004.
- [9] Degenhardt, Louisa and Peacock, Amy and Colledge, Samantha and Leung, Janni and Grebely, Jason and Vickerman, Peter and Stone, Jack and Cunningham, Evan B and Trickey, Adam and Dumchev, Kostyantyn and others. Global prevalence of injecting drug use and sociodemographic characteristics and prevalence of HIV, HBV, and HCV in people who inject drugs: a multistage systematic review. *The Lancet Global Health*, 5(12):e1192–e1207, 2017.
- [10] Edelman, E Jennifer and Chantarat, Tongtan and Caffrey, Sarah and Chaudhry, Amina and O’Connor, Patrick G and Weiss, Linda and Fiellin, David A and Fiellin, Lynn E. The impact of buprenorphine/naloxone treatment on HIV risk behaviors among HIV-infected, opioid-dependent patients. *Drug and Alcohol Dependence*, 139:79–85, 2014.
- [11] Fienberg, Stephen E. A brief history of statistical models for network analysis and open challenges. *Journal of Computational and Graphical Statistics*, 21(4):825–839, 2012.
- [12] Garrett M Fitzmaurice, Nan M Laird, and James H Ware. *Applied Longitudinal Analysis*, volume 998. John Wiley & Sons, 2012.
- [13] Centers for Disease Control & Prevention. HIV in the United States and Dependent Areas, Sep 2019.

- [14] Forastiere, Laura and Airoidi, Edoardo M and Mealli, Fabrizia. Identification and estimation of treatment and interference effects in observational studies on networks. *arXiv preprint arXiv:1609.06245*, 2016.
- [15] Fortunato, Santo. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.
- [16] Samuel R Friedman, Leslie Williams, April M Young, Jennifer Teubl, Dimitrios Paraskevis, Evangelia Kostaki, Carl Latkin, Danielle German, Pedro Mateu Gelabert, Honoria Guarino, et al. Network research experiences in New York and Eastern Europe: lessons for the southern US in understanding HIV transmission dynamics. *Current HIV/AIDS Reports*, 15(3):283–292, 2018.
- [17] Friedman, Samuel R and Downing, Martin J and Smyrnov, Pavlo and Nikolopoulos, Georgios and Schneider, John A and Livak, Britt and Magiorkinis, Gkikas and Slobodanyk, Liudmyla and Vasylyeva, Tetyana I and Paraskevis, Dimitrios and others. Socially-integrated transdisciplinary HIV prevention. *AIDS and Behavior*, 18(10):1821–1834, 2014.
- [18] Fulcher, Isabel Rose. *Statistical Inference for Causal Mechanisms: Mediation and Interference*. PhD thesis, 2019.
- [19] Gabor Csardi and Tamas Nepusz. The igraph Software Package for Complex Network Research. *InterJournal*, Complex Systems:1695, 2006.
- [20] Galea, Sandro and Vlahov, David. Social determinants and the health of drug users: socioeconomic status, homelessness, and incarceration. *Public Health Reports*, 117(Suppl 1):S135, 2002.
- [21] Gowing, Linda R and Farrell, Michael and Bornemann, Reinhard and Sullivan, Lynn E and Ali, Robert L. Brief report: Methadone treatment of injecting opioid users for prevention of HIV infection. *Journal of General Internal Medicine*, 21(2):193–195, 2006.

- [22] Greenland, Sander and Robins, James M. Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology*, 15(3):413–419, 1986.
- [23] M Elizabeth Halloran, Michael Haber, Ira M Longini Jr, and Claudio J Struchiner. Direct and indirect effects in vaccine efficacy and effectiveness. *American Journal of Epidemiology*, 133(4):323–331, 1991.
- [24] M Elizabeth Halloran and Michael G Hudgens. Dependent happenings: a recent methodological review. *Current Epidemiology Reports*, 3(4):297–305, 2016.
- [25] Hedeker, Donald and Mermelstein, Robin J and Demirtas, Hakan. Analysis of binary outcomes with missing data: missing= smoking, last observation carried forward, and a little multiple imputation. *Addiction*, 102(10):1564–1573, 2007.
- [26] Michael G Hudgens and M Elizabeth Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008.
- [27] NM Ivers, M Taljaard, Stephanie Dixon, C Bennett, A McRae, J Taleban, Z Skea, JC Brehaut, RF Boruch, MP Eccles, et al. Impact of CONSORT extension for cluster randomised trials on quality of reporting and study methodology: review of random sample of 300 trials, 2000-8. *Bmj*, 343:d5886, 2011.
- [28] Jackson, Dan and White, Ian R and Mason, Dan and Sutton, Stephen. A general method for handling missing binary outcome data in randomized controlled trials. *Addiction*, 109(12):1986–1993, 2014.
- [29] Joseph Rigdon. *interferenceCI: Exact Confidence Intervals in the Presence of Interference*, 2015. R Package Version 1.1.
- [30] Karrer, Brian and Levina, Elizaveta and Newman, Mark EJ. Robustness of community structure in networks. *Physical Review E*, 77(4):046119, 2008.
- [31] Eric D Kolaczyk. *Statistical Analysis of Network Data*. Springer, 2010.

- [32] Eric D Kolaczyk and Gábor Csárdi. *Statistical Analysis of Network Data With R*, volume 65. Springer, 2014.
- [33] Larney, Sarah. Does opioid substitution treatment in prisons reduce injecting-related HIV risk behaviours? A systematic review. *Addiction*, 105(2):216–223, 2010.
- [34] Liu, Lan and Hudgens, Michael G. Large sample randomization inference of causal effects in the presence of interference. *Journal of the American Statistical Association*, 109(505):288–301, 2014.
- [35] Liu, Lan and Hudgens, Michael G and Becker-Dreps, Sylvia. On inverse probability-weighted estimators in the presence of interference. *Biometrika*, 103(4):829–842, 2016.
- [36] Hernán MA and Robins (2020) JM. Causal Inference. What If. Boca Raton: Chapman Hall/CRC.
- [37] MacArthur, Georgie J and Minozzi, Silvia and Martin, Natasha and Vickerman, Peter and Deren, Sherry and Bruneau, Julie and Degenhardt, Louisa and Hickman, Matthew. Opiate substitution treatment and HIV transmission in people who inject drugs: systematic review and meta-analysis. *Bmj*, 345:e5945, 2012.
- [38] Departments of Epidemiology Miguel Hernán and Biostatistics. Definition of Causal Effect.
- [39] Misztal, Malgorzata. Some remarks on the data imputation using missforest method. 2013.
- [40] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113, 2004.
- [41] Newman, Mark. *Networks*. Oxford University Press, 2018.
- [42] Newman, Mark EJ. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69(6):066133, 2004.

- [43] Newman, Mark EJ. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3):036104, 2006.
- [44] Newman, Mark EJ. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.
- [45] Nicholas A Christakis and James H. Fowler. Social contagion theory: examining dynamic social networks and human behavior. *Statistics in Medicine*, 32.4:556–577, 2013.
- [46] Nikolopoulos, Georgios K and Katsoulidou, Antigoni and Kantzanou, Maria and Rokka, C and Tsiara, C and Sypsa, V and Paraskevis, D and Psychogiou, M and Friedman, S and Hatzakis, A. Evaluation of the limiting antigen avidity EIA (LA_g) in people who inject drugs in Greece. *Epidemiology & Infection*, 145(2):401–412, 2017.
- [47] Nikolopoulos, Georgios K and Pavlitina, Eirini and Muth, Stephen Q and Schneider, John and Psychogiou, Mina and Williams, Leslie D and Paraskevis, Dimitrios and Sypsa, Vana and Magiorkinis, Gkikas and Smyrnov, Pavlo and others. A network intervention that locates and intervenes with recently HIV-infected persons: the Transmission Reduction Intervention Project (TRIP). *Scientific Reports*, 6:38100, 2016.
- [48] Onghena, Patrick. Randomization tests or permutation tests? A historical and terminological clarification. *Randomization, masking, and allocation concealment*, pages 209–227, 2018.
- [49] Pathria, RK and Beale, Paul D. Statistical Mechanics. Elsevier. *New York*, 2011.
- [50] Perez-Heydrich, Carolina and Hudgens, Michael G and Halloran, M Elizabeth and Clemens, John D and Ali, Mohammad and Emch, Michael E. Assess-

- ing effects of cholera vaccination in the presence of interference. *Biometrics*, 70(3):731–741, 2014.
- [51] Robins, James. A new approach to causal inference in mortality studies with a sustained exposure period? Application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512, 1986.
- [52] Rubin, Donald B. Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4):472–480, 1990.
- [53] Stekhoven, Daniel J and Bühlmann, Peter. Missforest non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.
- [54] Tang, Fei and Ishwaran, Hemant. Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 10(6):363–377, 2017.
- [55] Eric J Tchetgen and Brent A Coull. A diagnostic test for the mixing distribution in a generalised linear mixed model. *Biometrika*, 93(4):1003–1010, 2006.
- [56] Eric J Tchetgen Tchetgen and Tyler J VanderWeele. On causal inference in the presence of interference. *Statistical Methods in Medical Research*, 21(1):55–75, 2012.
- [57] van der Laan, Mark J. Causal inference for a population of causally connected units. *Journal of Causal Inference J. Causal Infer.*, 2(1):13–74, 2014.
- [58] Yu, Chong Ho. Resampling methods: concepts, applications, and justification. *Practical Assessment, Research, and Evaluation*, 8(1):19, 2002.