UNIVERSITY OF CALIFORNIA

Los Angeles

Agent-Based Modeling for HIV Prevention

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

by

David Scott Boren

2015

ABSTRACT OF THE DISSERTATION

Agent-Based Modeling for HIV Prevention

By

David Scott Boren
Doctor of Philosophy in Biostatistics
University of California, Los Angeles, 2015
Professor Ronald S. Brookmeyer, Chair

Progress in HIV (human immunodeficiency virus) prevention interventions has been made in recent years. These developments have raised questions concerning the impact and optimal combination of biomedical and behavioral interventions. Agent-based models offer a way to compare intervention combinations that could not be readily accomplished with large scale prevention trials because of cost and logistical considerations. An agent-based model is a community of individuals whose behavior and interactions are simulated according to well-defined rules.

We describe two agent-based social network models and methods of calibration to sexual contact parameters relevant to MSM (men who have sex with men) in South Africa: a model with static partnerships and a novel model featuring self-reinforcing partnership edges. We only have sexual contact summary statistics (i.e. coarse network statistics) and these models are calibrated to replicate these statistics. A calibration approach was developed using recent advances in multi-objective optimization and applied to the self-reinforcing model.

We use these models to evaluate the varying impact of HIV prevention combinations by simulating community-randomized trials across realistic values of several parameters: ART (anti-retroviral therapy) uptake, PREP (pre-exposure prophylaxis) uptake, HIV testing uptake, and CAI (condomless anal intercourse) reduction. For the static partnership agent-based model, we create a statistical model to describe quantitatively the effects of these parameters and variation. For the self-reinforcing model we model sensitivity across several key parameters. We develop sample size and power formula for community randomized trials that incorporate estimates of variation and effect sizes determined from the agent-based model. We find that traditional sample size approaches that rely on binomial (or Poisson) models are inadequate and can lead to underpowered studies. In our studies of variation and effect sizes, we identified important sensitivities to the network contact structure. We conclude that agent-based models offer a useful tool in the design of HIV prevention trials.

The dissertation of David Scott Boren is approved.

Catherine Crespi-Chun

Robert Erin Weiss

Pamina M. Gorbach

Ronald S. Brookmeyer, Committee Chair

University of California, Los Angeles

2015

To my wife, Heather, who married a man with negative assets.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

A version of the material present in chapters 4 and 7 has been published in *Statistics in Medicine* [47]. Additionally, chapter 6 contains a version of material published in *PloS one* [48]. Chapter 5 contains figures and sections from a paper currently submitted to *Network Science.* Many authors contributed to these papers, but the statistical work was accomplished by myself under the direction of my advisor, Dr. Ron Brookmeyer.

This work was supported by a grant from the National Institutes of Health (R01AI094575), the Biostatistics in AIDS Training Grant Award (T32AI007370), the Human Sciences Research Council, and was facilitated by the Emory Center for AIDS Research (P30 AI050409), the UCLA Center for AIDS Research (AI028697) and the Johns Hopkins University Center for AIDS Research, an NIH funded program (1P30AI094189).

My committee members Dr. Crespi, Dr. Gorbach, and Dr. Weiss were instrumental in the editing and review process. None of this work would have been possible without the aid of my mentor and advisor, Dr. Ron Brookmeyer, and the constant and unwavering support of my loving wife, Heather. Thank you all.

VITA

NAME OF AUTHOR:

David Scott Boren

DEGREES AWARDED:

B. S. in Bioengineering, University of California, Los Angeles, 2007

M. S. in Biomedical Engineering, University of California, Los Angeles, 2008

PROFESSIONAL EXPERIENCE:

Graduate Student Researcher in Biostatistics, University of California, Los Angeles, 2010-2015

Tutor and Teacher for MCAT, Kaplan, Los Angeles, 2008-2010

Teaching Assistant, University of California, Los Angeles, 2007-2008

PROFESSIONAL PUBLICATIONS:

Boren, David, Patrick S. Sullivan, Chris Beyrer, Stefan D. Baral, Linda, Gail Bekker, and Ron Brookmeyer. "Stochastic variation in network epidemic models: implications for the design of community level HIV prevention trials." Statistics in Medicine 33, no. 22 (2014): 3894-3904.

Brookmeyer, Ron, David Boren, Stefan D. Baral, Linda-Gail Bekker, Nancy Phaswana-Mafuya, Chris Beyrer, and Patrick S. Sullivan. "Combination HIV prevention among MSM in South Africa: results from agent-based modeling." PloS One 9, no. 11 (2014): e112668.

Ivask, Angela, Elizabeth Suarez, Trina Patel, David Boren, Zhaoxia Ji, Patricia Holden, Donatello Telesca, Robert Damoiseaux, Kenneth A. Bradley, and Hilary Godwin. "Genome-wide bacterial toxicity screening uncovers the mechanisms of toxicity of a cationic polystyrene nanomaterial." Environmental Science & Technology 46, no. 4 (2012): 2398-2405.

Kim, Sanggu, Namshin Kim, Beihua Dong, David Boren, Serena A. Lee, Jaydip Das Gupta, Christina Gaughan et al. "Integration site preference of xenotropic murine leukemia virus-related virus, a new human retrovirus associated with prostate cancer." Journal of Virology 82, no. 20 (2008): 9964-9977.

HONORS AND AWARDS:

Biostatistics in AIDS Training Grant Award · UCLA · 2012 - 2015

Departmental Award for Outstanding Master's Student in Biomedical Engineering , University of California, Los Angeles, 2008

PRESENTATIONS:

Joint Statistical Meeting, San Diego, August 2012

Joint Statistical Meetings, Montreal, August 2013

Biometric Society Meeting (WNAR), Honolulu, June 2014

Chapter 1: Introduction

1.1 Motivations for the Work

The subject of this dissertation was motivated originally by the Sibanye Health Project [1]. The project is part of the "Methods of Prevention Packages Program" (MP3) funded by the National Institute of Health. It is a 4-phase prevention intervention project with the specific aim of testing combination HIV prevention interventions and services for men who have sex with men (MSM) in Southern Africa. It will determine the acceptability as well as develop a rational and well-informed proposal for efficacy trials of a particular prevention package. The program has four phases, including: a comprehensive literature review to summarize the current body of knowledge regarding HIV prevention interventions for MSM, qualitative studies to obtain information regarding the acceptability and feasibility of the package, a mathematical model of HIV transmission in Southern African MSM, and finally a pilot study of 400 MSM at 2 locations in South Africa. Our contribution is the mathematical model of transmission.

The project began in 2011. Our collaborators include Johns Hopkins University, the Desmond Tutu HIV Foundation, and the Human Sciences Research Council. Currently HIV prevalence in South Africa among MSM stands at 25.5% [2], and the small number of clinical trials specific to this group means that the data at hand are minimal and incomplete. In addition, it is only recently that it has been properly established that anti-retroviral treatment reduces the probability of HIV transmission.

1.2 Combinations of interventions

One of the primary questions of interest in this work was the effect that multiple forms of HIV treatment might have in a South-African MSM scenario. This project dealt with the recent field of biomedical HIV intervention. There have been significant advances in HIV prevention interventions in recent years [3,4]. Trials have identified effective interventions to prevent acquisition of HIV infection including circumcision, antiretroviral therapy (ART) for HIV infected persons, and pre-exposure prophylaxis (PREP) for high risk uninfected persons [5,6]. These recent successes were preceded by a number of earlier HIV prevention trials that failed to detect benefits of various interventions [7]. In some instances, the failures of earlier trials to detect significant effects were attributed to underpowered trials with inadequate sample sizes [8]. It is with these issues in mind that simulation work was carried out and the results of the simulations were thoroughly analyzed to answer questions regarding power and sample size.

1.3 Overview of Research

In this dissertation we first examine the history of two forms of modeling that at first can seem quite disparate. The first is what might be called a bottom up approach, known as agent-based modeling, and focuses on the creation of an interactive system of agents (simulated individuals) with the aim of replicating the broad epidemic trends of the population of interest. The second, known as network regression, focuses on identifying the determinants of associations (sexual, social, or otherwise) between individuals. Following these background

sections we discuss the work that we've done on our own agent-based model, including both the mechanistic portions and the analysis of results. This portion describes our first attempt at developing a new agent-based model: the static partnership model, as well our recently developed self-reinforcing model. Following this, we discuss our approach to calibrating these models and the complications that arise in this process. For the calibration of our self-reinforcing model we applied an advanced algorithm for calibrating to noisy data. This method avoids the trade-offs that come from combining objectives for classical optimization [9, 10]. We discuss the results of the models and their significance in application to community-randomized trials. Finally, we review some of the current limitations to this approach to agent-based modeling, and the further research that would help address these limitations and extend the overall model relevance.

Chapter 2: Background and History of Agent-Based Models

2.1 A Quick History Lesson - The Reed-Frost Theory

One of the earliest attempts at agent-based modeling was the Reed Frost theory of epidemics, which was derived from earlier work by Soper [11], and makes the following assumptions:

1. Time of infectiousness is short relative to infection, meaning that the time that an infected person can infect others is minimal compared to the duration of the infection.
2. All individuals have equal susceptibility to the disease
3. All individuals have equal transmission capacity
4. All individuals pass out of observation when this transmission period is finished, meaning that, for the purposes of the formula, individuals that are no longer infectious do nothing, they cannot infect anyone and they cannot be reinfected.
5. It is a closed population, meaning no individuals leave or enter the population
6. There is uniform mixing, meaning that the probability of transmission between any two individuals is the same for every pair of individuals

Under these assumptions, one can produce the following formula for calculating the expected number of cases at a given time period given the state at the previous time period,

$$C_{t+1} = S_t * (1 - q^{C_t}) . \quad (1)$$

Here *t* represents a particular discrete interval of time, where all intervals in this model are equally wide. C_t stands for the expected number of contagious individuals at the end of interval t, S_t stands for the number of susceptible individuals at the end of interval *t*, and *q* stands for the probability that an individual does not have contact with another individual over one iteration of time. Consequently q^{C_t} represents the probability that an individual does not have contact with all of the contagious individuals at time t and $(1 - q^{C_t})$ is the probability that at least one of the C_t contagious individuals has contact with this individual. This is known as Soper's equation. Under this model a given time step represents the average length of the incubation period, and it is precisely the assumption of short period of infectiousness relative to incubation period that allows the entire contagious population to pass out of observation before the next set of individuals becomes infectious. In other words, the period of infectiousness is so short that by the time the newly infected individuals finish their incubation period, all previously infectious individuals are no longer infectious. Notably this model is deterministic, in that it ignores the random chance that may be associated with each step. Reed and Frost further developed this model through the addition of this missing variation by assuming that the cases drawn in a given time step follow a binomial probability distribution, where $(1 - q^{C_t})$ is the probability that one of S_t individuals becomes contagious and $1 - (1 - q^{C_t}) = q^{C_t}$ is the probability that one of S_t individuals does not become contagious. Given that there are S_t individuals from which to

choose the next pool of contagious individuals, we can write the probability that there are C_{t+1} contagious individuals in the next time interval as

$$P(C_{t+1}) = \frac{S_t!}{C_{t+1}!\,S_{t+1}!}(1-q^{C_t})^{C_{t+1}}(q^{C_t})^{S_{t+1}} \quad , \qquad (2)$$

where $S_{t+1} = S_t - C_{t-1}$. Thus an entire epidemic can be simulated through the successive draws from this binomial distribution, and the concept of an agent-based simulation was formalized, albeit in a restrictive, uniform sense, where all individuals are alike and all associations are equal.

In a critique of this theory in 1952, Helen Abbey [12] outlined the particular situations where the assumptions of the model might be left intact. The requirement of a closed population with uniform mixing might be found within families or institutions. The short period of infectiousness might be found in certain diseases of childhood, such as measles or chickenpox. She took data from 19 well documented outbreaks that fit this criteria and, through careful consideration of sampling error (among susceptible agents), was able to remove some of the immediate biases, but even then she found some very large overall differences between observed and expected numbers, as measured by a chi-square test. She postulates three potential sources of error: Miscounting of the at-risk population, changes in contact rate with time, and variation in contact rate among individuals.

Computational progress since then has allowed for much more varied and complex agent-based models, where each agent is unique and parameters are subject to change over time, and the assumptions of uniform mixing and even closed population are no longer troublesome. Dr.

Abbey's other sources of error, however, remain relevant. Correctly specifying both the at-risk population and the contact rate configurations are vital to an agent-based model's usefulness.

2.2 Agent-Based Modeling – Current

So then, in simple terms how can one define an agent-based model? Perhaps Dr. Joshua Epstein does this best by stating that agent-based models are in fact "artificial societies," where every individual in the society of interest is represented as a distinct agent [13]. An agent-based model is a community of individuals whose behavior and interactions are simulated according to well-defined rules. This means that the behavior and attributes of every agent can be unique and calibrated to the most accurate and knowable behavior of the real-life individual that he or she represents. They may change their behavior in response to other agents or changes in the environment [14]. This means we eliminate most of the assumptions that are present in our earlier models, such as homogeneity and perfect mixing. It also means, however, that the model is computation-limited, since each one of these distinct individuals requires computation, rather than just pools of individuals. Indeed, it is only since the 1990s that agent-based models have become feasible on a useful scale.

Recently, agent-based modeling has been applied to various fields including the social sciences [15], spatial patterns of health [16], and the spread and control of infectious diseases such as smallpox [17, 18] and pandemic influenza [19]. Agent-based models for the spread of infectious diseases depend on assumptions about the networks of contacts between persons [20], which we will discuss in detail later.

2.3 Features/Benefits of Agent-Based Models

Often the "statistical" approach can be thought of as top-down, where certain regular trends are observed with a high degree of confidence. Epstein [21], in his article "Agent-Based Computational Models and Generative Social Science," asserts that, in the context of agent-based models, the natural question that arises is: "How could the decentralized local interactions of heterogeneous autonomous agents generate the given regularity?" Or, in other words, how can the interactions of extremely different individuals in a complex environment produce the strangely regular trends observed in statistics.

He then argues that agent-based computational models are well suited to exploring this question, since they possess the following features: Heterogeneity- each agent can have completely unique behaviors and characteristics; Autonomy – each agent can make its own decisions independently of all other agents (though this not need be the case); Explicit Space – agents are defined spatially in relation to each other in some form: for example in the Reed-Frost Model they all occupy the same unit space, having equal probabilities of interacting with one another; Local Interactions – the uniform mixing of the Reed Frost model is not a common case in agent-based models, but rather individuals who are neighbors in the explicit space are more likely to interact; and Bounded Rationality- agent-based models represent finite numbers of actors that may be completely out of equilibrium, and thus offer insights into smaller systems where asymptotic behaviors may not be applicable.

Bonabeau [14] summarized the benefits of agent-based modeling in 3 ways: it allows for the natural description of a system, it is flexible, and it captures emergent phenomena. Now, by "natural description," Bonabeau refers to rules that make intuitive sense at the level of the agents or groups of agents, such as how often an individual interacts with another individual in his age group or how often he visits the doctor for a check-up. The flexibility of the model lies in both the way that often simple agent-level rules have the propensity to drastically alter the system wide results, but also in how these agents can be fine-tuned to whatever behaviors, groupings, and interactions are required. Emergent phenomena, on the other hand, is an interesting and somewhat vague concept. Bonabeau describes emergent phenomena as the phenomena that result from the natural, low-level (individual or groups of individual) descriptions of the system. Indeed, by definition, they cannot be reduced to the system's parts without losing some information. It is the interactions between these parts that allow the whole to be more than the sum of the parts. In other words, an emergent phenomenon can have properties that are not apparent even from a detailed analysis of the parts, which is why a simulation of agents must be carried out to determine the effect of the low-level design.

2.3 Agent-Based vs Differential Equation models

The moniker of "simulation'' often is applied to two ostensibly different approaches: agent-based and differential equation-based models. An agent-based model, as we have said, allows for heterogeneous agents to make individual decisions and have unique interactions with other agents, which allow for the propagation of the epidemic. A differential equation model, on

the other hand models a compartment of individuals using differential equations. One of the more notable examples is the Kermack McKendrick Model [22]. It is a SIR model, which is a model that keeps track of the numbers of susceptible individuals (S), infected individuals (I), and recovered individuals (R) as they change over time. It also makes the following assumptions:

1. Population size is fixed (i.e., no births, deaths due to disease, or deaths by natural causes)
2. Incubation period of the infectious agent is instantaneous
3. Duration of infectivity is same as length of the disease.
4. A completely homogeneous population with no age, spatial, or social structure.

The Kermack McKendrick Model assumes continuous time and consists of 3 nonlinear ordinary differential equations, each of which keep track of one the three components of SIR,

$$\frac{\partial S}{\partial t} = -\beta SI \tag{3}$$

$$\frac{\partial I}{\partial t} = \beta SI - \gamma I \tag{4}$$

$$\frac{\partial R}{\partial t} = \gamma I \ , \tag{5}$$

where t is time, S(t), I(t), and R(t) are as defined by a SIR model, β is the infection rate, and γ is the recovery rate.

Notice already that the homogenous population assumption already distinguishes the differential equation approach from that of an agent-based model in a very large way, but also every other assumption is not necessarily present in the implementation of an agent-based model.

Indeed, these assumptions are more akin to a Reed-Frost model. They can become more complex, but the point remains that there are some very basic homogeneity assumptions that a differential equation model will never be able to escape, simply because it monitors compartments rather than individuals.

So the question is: when should one use differential equations and when should one use agent-based models? Bonabeau [14] believes that the entire dichotomization into two approaches is itself incorrect, "as a set of differential equations, each describing the dynamics of one of the system's constituent units, is an agent-based model." Indeed, one can observe the differential equation model above and note that each of the three groups of individuals could be considered an "agent" that interacts with the other two agents, changing their counts.

Less dismissive is Hazhir Rahmandad [23], whose work in Management Science delved specifically into the comparison of differential equation and agent-based models and gave the simple assessment that differential equation models are computationally efficient, but rely upon the assumptions of perfect mixing and homogeneity of individuals within compartments. Agent-based models, on the other hand, increase the need for computation, which in turn constrains the reach of sensitivity analyses, but also more easily establishes the relationships of individuals in a more realistic network form along with their attribute and contact heterogeneities. Rahmandad goes on to specifically address types of heterogeneities in the context of epidemics, stating that "Heterogeneity in individual contact rates causes slightly earlier mean peak times as high contact individuals rapidly seed the epidemic, followed by lower diffusion levels as the high-contact individuals are removed, leaving those with lower average transmission probability and a smaller

reproduction rate…Such dynamics were also observed in the HIV epidemic, where initial diffusion was rapid in sub-populations with high contact rates." Thus the decision of which simulation methodology to employ really does vary with the nature of the epidemic (heterogeneity of behavior, attributes, mixing) and the need for sensitivity analysis, which, computationally speaking, can be prohibitively expensive in a situation where your simulation produces wide distributions.

2.4 Scaling up - Model properties to statistical properties

So agent-based modeling does indeed appear to be the appropriate simulation tactic for HIV epidemiology, but the question then becomes to what extent does the model accurately, in practice, reflect the reality of the epidemic? In a short article to *The Bulletin of the Santa Fe Institute* in 1994, Robert L. Axtell and Joshua L. Epstein [24] defined four levels at which a model might accurately reflect the population it claims to represent:

Level 0: The model is a caricature of reality, as established through the use of simple graphical devices (e.g., allowing visualization of agent motion);

Level 1: The model is in qualitative agreement with empirical macrostructures, as established by plotting, say, distributional properties of the agent population;

Level 2: The model produces quantitative agreement with empirical macrostructures, as established through on-board statistical estimation routines; and finally,

Level 3: The model exhibits quantitative agreement with empirical microstructures, as determined from cross-sectional and longitudinal analysis of the agent population.

Now, in the case of our South-African data, we currently have access to some very fundamental statistics, which we will cover in detail later. This means that even with the most rigorous of calibration schemes we can hope at best to obtain a level 2 agreement with the empirical macrostructures, simply because many of the microstructures for our MSMs in South Africa are unknown. So the goal might very well be to produce a model that has multiple parameter fits, such that all the fits accurately reflect the macrostructures, while at the same time cycling through different microstructures that might create such macrostructures.

2.5 Agent-Based Models - Not Quantitatively Perfect

In a small opinion piece in Nature entitled "Modeling to contain pandemics" Epstein [13] outlines some of the applications for agent-based models in an epidemiological context, even if these models are not "crystal balls." He gives an example of the process of epidemiological agent-based simulation, starting with the primary simulation. This first simulation produced is not a prediction, but rather a base case, which by design is highly unrealistic, since it ignores "pharmaceuticals, quarantines, school closures and behavioral adaptations," all of which might be relevant to the progress of an epidemic. Nevertheless, this base case allows one to rerun the model, this time perturbing parameters that relate specifically to questions in which health

agencies might be interested. Such questions in the face of a flu pandemic might take the form "What is the best way to allocate limited supplies of vaccine or antiviral drugs?" or "How effective are school or work closures?" Thus, an agent-based model that accurately represents macrostructures might be able to accurately quantify comparative intervention plans. This was one of our principle goals in this work, to give quantitative comparisons of various HIV intervention plans relevant to HIV among MSMs in South Africa.

Chapter 3: Modeling Considerations in Networks

3.1 Early models

Looking at an agent-based model approach, one might question how exactly it is organized. Especially in a sexual contact modeling scenario, it would be improper to assume that agents simply make contact at random or haphazardly. Rather sexual contacts are the result of a pre-existing network structure. As a result, it would not be prudent to model HIV transmission as a series of random contacts. Instead, it might be best to examine work in graph (network) theory to best establish the relationships that define these contacts.

All the way back in 1981, Holland and Leinhardt [25] laid down much of the foundation for graph theory in the paper "An exponential family of probability distributions for directed graphs." Here they defined a graph as a specified set of nodes (agents or individuals) and a set of lines (edges or relationships) that connect certain pairs of these nodes. They were also the first to model probability as an exponential family and defined their model as a group of independent dyads, meaning that the probability of an edge existing between two individuals *i* and *j*, is independent of their other edge statuses with other individuals (e.g., between *i* and *k)* as well as those between two unrelated individuals (e.g., between *k* and *l)*.

Since a graph is just a representation of pairwise relationships, it can be expressed in matrix form, known as the adjacency matrix. If the graph is known as "undirected," then all relationships are reciprocated, such that every edge from individual *i* to *j* encoded in the

adjacency matrix is accompanied by an edge from individual *j* to *i* and the adjacency matrix is symmetric. If there are *g* nodes in an undirected graph Y, then a symmetric adjacency matrix A is coded as

$$\begin{pmatrix} 0 & \cdots & A_{1j} & \cdots & A_{1g} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ A_{j1} & \cdots & 0 & \cdots & A_{jg} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ A_{g1} & \cdots & A_{gj} & \cdots & 0 \end{pmatrix},$$

where

$$A_{ij} = \begin{cases} 1, & \text{if edge exists between nodes i and j} \\ 0, & \text{otherwise} \end{cases}.$$

Notice here there are no self-ties ($A_{ii} = 0 \forall i$) and, since the network is undirected, a tie from *i* to *j* is the same as that from j to i, or $A_{ij} = A_{ji}$. This means there are $\binom{g}{2}$ possible edges. In our case the nodes would represent MSMs and the edges would represent sexual partnerships.

In 1985 Fienberg and Meyer extended this p1 model to include block levels, where dyads were still independent, but the probabilities of edge existences varied with block-categorized relationships [26]. In 1986 Ove and Strauss introduced a dependence structure into graph relationships, such that that the probability of a tie A_{ij} existing was dependent not only on individual and block-level characteristics, but also dependent upon all other ties in the graph [27].

3.2 The ERGM Framework

More recently advances in MCMC have allowed the area of graph theory to move to the ERGM (exponential-family random graph models) framework. In this framework the distribution of the adjacency matrix Y can be parameterized in the form [28]

$$P_{\eta,\Psi}(A=a)=\frac{exp(\eta^T g(a))}{\kappa(\eta,\Psi)}, \qquad a\in\Psi$$
$$\kappa(\eta,\Psi)=\sum_{z\in\Psi} exp(\eta^T g(z)). \qquad (6)$$

Here *a* is a particular form of the adjacency matrix A, η is a vector of model coefficients, and *g(a)* is the vector of network statistics given the adjacency matrix *a*. These statistics could be the number of triangles in the network, the number of individuals of connected to k individuals, or simply the total number of edges in the network. Additional covariates *X* can easily be added to this network through inclusion in the function g, such that *g = g(a,X)*. The parameter ψ can be thought of as the set of all possible configurations of the adjacency matrix, while $\kappa(\theta,\Psi)$ is the normalizing factor that makes this a probability distribution. It might be apparent now that this normalizing factor is the sum of an exceedingly large space even for networks of a modest size. For example, a network of ten nodes has 2^(10*9/2)= 2^45 possible graphs. As a result, the model is more intelligible when put into a pairwise probabilistic form, or in other words, the probability of edge formation given the rest of the network structure. Under the ERGM framework takes the form [28]

$$\text{logit}[P_{\eta,\Psi}(A_{ij}=1 \mid A_{ij}^{c}=a_{ij}^{c})] = \log\left[\frac{P_{\eta,\Psi}(A_{ij}=1 \mid A_{ij}^{c}=a_{ij}^{c})}{1-P_{\eta,\Psi}(A_{ij}=1 \mid A_{ij}^{c}=a_{ij}^{c})}\right] = \eta^{T}\delta_{g}(a)_{ij} \quad (7)$$

$$\delta_{g}(a)_{ij} = g(a_{ij}^{+}) - g(a_{ij}^{-}).$$

Here $\delta_{g}(a)_{ij}$, referred to as a change statistic, is the difference between the network statistics with (+) and without (-) the addition of edge *ij*. The θ vector then can be clearly interpreted as the increase in the conditional log-odds of tie-formation in the network per unit increase in the corresponding component of g(y). Notice here that the probability of an edge between two individuals is conditional on the rest of the network a_{ij}^{c}. Also of note is that under this framework when an edge is independent of the rest of the network the probability becomes a simple logit model.

So it is this ERGM model, which takes into account dependency among all ties, that would truly offer the best relationship format for our agent-based modeling. Unfortunately, to utilize such a model, we would first need to have complete or nearly complete structural data of the South African network of MSM contacts. Such information is lacking, and as a result, our simulation could not incorporate a dependent network structure.

Now, one additional thing that might be apparent is that these network forms are really non-temporal. An ERGM defines the ties between nodes in a network, but does not define how these ties behave over time. In order to do this one might use what is known as a STERGM [29], or a "Separable Temporal ERGM." The basic formulation of this model requires the addition of a time series aspect, where time is measured in discrete steps, and instead of looking at the probabilities of edges existing between pairs, one examines the probability of edge formation (+) and dissolution (-), such that at any given time-step you have these two different adjacency

matrices of ‘’changes’’: formation and dissolution, which are incorporated into the adjacency matrix at time-step t: a_t . This turns equation 6 into

$$
\begin{aligned}
P_{\eta,\Psi}(A^+ = a^+ \mid \mathrm{a}_t) &= \frac{exp(\eta^{-T} g(a^+))}{\kappa(\eta^+, \Psi^+(a^t))}, \qquad a^+ \in \Psi^+(\mathrm{a}_t) \\
P_{\eta,\Psi}(A^- = a^- \mid \mathrm{a}_t) &= \frac{exp(\eta^{-T} g(a^-))}{\kappa(\eta^-, \Psi^-(a^t))}, \qquad a^- \in \Psi^-(\mathrm{a}_t) \ . \\
\kappa(\eta, \Psi) &= \sum_{z \in \Psi} exp(\eta^T g(z)),
\end{aligned}
\tag{8}
$$

Here we have the adjacency matrix at time t, a^t, the formation adjacency matrix a^+, the dissolution adjacency matrix a^-, the space of all possible edge formations at time t, $\psi^+(a^t)$, and the space of all possible edge dissolutions at time t, $\psi^-(a^t)$. In addition, η^+ and η^- are vectors of model coefficients for the formation and dissolution networks, respectively. All other parameters are unchanged from the regular ERGM model.

It is this structure that would be ideal for an agent-based model, if the data for the calibration of such a network existed to such an extent that a STERGM regression could be performed. Since the data we have are insufficient, however, we recognize that such a regression (and its subsequent use in calibrating our agent-based model) is unreachable at this time.

Chapter 4: Network Epidemic Models for HIV Prevention

4.1 Available Data

Our methodological work grew out of the Sibanye Health Project which is an HIV prevention project to develop and test combination HIV prevention interventions among men who have sex with men (MSM) in Southern Africa. To contribute to this project we had to rely upon broad descriptive statistics, what we call "coarse" network data: very basic partnership distribution statistics obtained through surveys that pertain to a particular period of time. These statistics include the mean unique partners in 6 months, % individuals with >5 unique partners in 6 months, mean CAI contacts in 2 months, and the ratio of the mean number of unique partners an individual belonging to a main partnership has in twelve months to that of an individual who does not belong to a main partnership. Many other parameters have been incorporated into the model and a full list of these parameters and their values is in Table 1.

Table 1 classifies inputs into 3 categories: attributes assigned to each person at start, daily updates, and prevention interventions. For example, the attributes assigned to each person at start section includes the probability that an individual is assigned a positive HIV serostatus at the start of the simulation, *P(HIV+)=0.255*. The daily updates section includes the decrease in CD4 count, which is linearly interpolated from the equation for the yearly decrease, *Yearly CD4 Decrease* = $4.26\sqrt{CD4}$. Finally the prevention interventions section describes the four interventions that are implemented in the agent-based models.

Table 1: Main characteristics of agent-based model for combination HIV prevention among MSM in peri-urban South Africa (additional information and specific parameter values are in the Supporting Information)

Attributes assigned to each person at start	Value (if applicable)
Frequency of sexual activity	
HIV status at start	P(HIV+)=0.255
CD4 count at start if HIV +	ln (CD4) ~ N(6.2, 0.81),
Knowledge of HIV status at start (yes, no)	P(Has knowledge)=0.25
Sexual role preference (insertive, receptive, versatile)	P(Role Preference)=1/3 for all
HIV testing frequency (3 levels: moderate, low, never)	P(Test in 6-mon period) =0.17, 0.085,0
Some assigned a main partner	P(Has Main Partner) = 0.46
Proportion of sexual contacts that are CAI (2 levels)	P(Accepts contact as CAI)=0.4,1
Sexual networks of regular partners (with allowance for sero-sorting)	Varied
Daily updates	
Daily sexual contacts depends on type of partnership	Likelihood of contact (in decreasing order): main, regular, casual, have other main partners
HIV testing possible	
CAI rate adjusted if learns knowledge of HIV status	Reduction of CAI = 1/3
CD4 levels updated for HIV positive	Yearly CD4 Decrease = $4.26\sqrt{CD4}$
Infection status updated	
Prevention Interventions	
ART for eligible HIV positives	
Eligible: HIV test within 6 months and CD4<350	
Considered varying levels of coverage	
PREP for eligible HIV negatives	
Eligible: in last 6 months had both HIV test and >12 CAIs or in sero-discordant main partnership.	
Considered varying levels of PREP acceptance with two levels of adherence (low and high)	
Reduction in CAI frequency (considered varying levels)	
Increase in HIV testing: convert 50% of the never testers to low frequency testers	

Each simulated run of the agent-based model consists of 1000 persons (agents) whose interactions and infection status are simulated over 5 years of whom an expected N=745 persons were initially uninfected. We randomly assign covariates based on distributions of the covariates from the South African setting [2]. Each person is assigned: a level of sexual activity based on the distribution of reported numbers of partners in 6 months among South African MSMs: predominant type of sexual activity(e.g., primarily the receptive or insertive partner in anal intercourse (the risk of transmission depends on the sexual role[30]); and frequency of HIV antibody test screening, all of which are assumed to be independent.

4.2 Static Partnership Model: Description

Our original work focused on the creation two networks of individuals: a main and regular partnership network and a daily sexual contact network. The partnership network is created at the time of agent initialization. Persons are assigned into networks of regular sexual partners; one of those regular partners may also be assigned to be the person's main sexual partner (46% of MSMs in South Africa are estimated to be in main partnerships [2]). Partners who are not in each others' network of regular partners are potentially "casual" partners. On every time-step of the simulation a sexual contact network is formed, where a sexual contact between two persons depends on whether the partnership is between main partners (most likely), regular partners (somewhat less likely) or casual partners (least likely).

Figure 1: A schematic demonstrating the two types of networks present in the model. The partnership network is created upon initialization of the agents. The daily sexual contact network is formed at every time-step during the simulation, and is influenced by the structure of the first network.

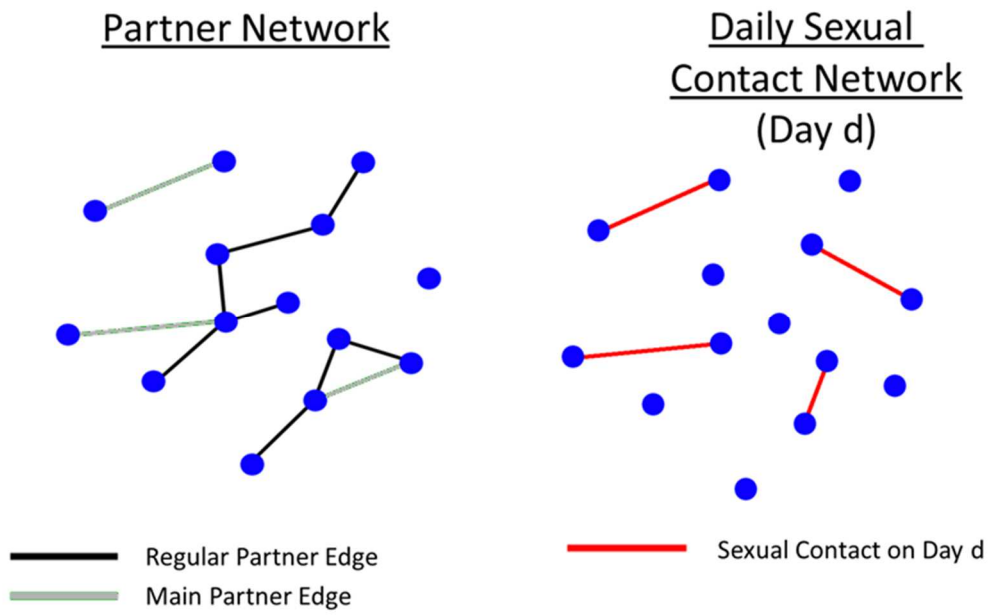


Figure 1 depicts a simple implementation of this static network agent-based model. It consists of a partnership network and a sexual contact network formed at every time-step (here a day). Main partnerships are more likely to experience sexual contact on a given day, regular partnerships slightly less so, and unconnected agents even less so.

4.3 Static Partnership Model: Model Specification and Notation

A main partnership not only illustrates a higher probability of contact on any given day, but also the exclusivity of this partnership against other regular partnerships to which the two individuals might belong. As previously stated, structural network data was not available for MSMs in south Africa, and as a result there is no way to produce an extensive ERGM regression for use in our agent-based model. Instead, a simplified network format was used, such that there is complete dyadic independence between pairs of individuals, or

$P(Y_{ij} = 1 \,\&\, Y_{jk} = 1) = P(Y_{ij} = 1) * P(Y_{jk} = 1)$, where i, j, k, and l are all unique individuals. So our original ERGM equation 7 reduces to a dyad-independent ERGM, which removes the conditioning upon the rest of the network, a_{ij}^{c} [28]. Prior to network formation we assign each individual an activity value, which represents the person's level of sexual activity, and is drawn from a distribution that possesses the mean, median, and percent with greater than five unique partners in six months. We then create the network using our dyad-independent network structure, where the probability that persons *i* and *j* are regular sexual partners, r_{ij}, is

$$\text{logit}(r_{ij}) = \alpha_0 + \alpha_1 X_{ij1} + \alpha_2 X_{ij2} \qquad (9)$$

where X_{ij1} is the sum of sexual activity levels for persons *i* and *j*, and X_{ij2} indicates whether the infection status of the two partners are the same or not at baseline. Due to independence of dyads, this network can be formed through a series of Benoulli trials. This partnership network is established before any sexual contact simulation occurs. Once it is established, main partnerships are randomly selected, while all other connections are considered regular partnerships. This model allows for overlapping networks of variable size and a degree of

assortative mixing because persons with the same infection status (sero-concordant) are assigned a higher probability of being regular partners than sero-discordant persons.

A daily *dyad-independent* ERGM network for sexual contacts occurring is constructed as follows. The probability, c_{ij}, that persons *i* and *j* have sexual contact on a given day is time-invariant under this model and given by

$$\text{logit}(c_{ij}) = \beta_0 + \beta_1 M_{ij} + \beta_2 R_{ij} - \beta_3 (H_i + H_j) \qquad (10)$$

where $M_{ij} = 1$ if persons *i* and *j* form a main partnership, $R_{ij} = 1$ if persons *i* and *j* form a regular partnership, and $H_i = 1$ if person *i* is in a main partnership with anyone. Once again, due to dyadic independence this dynamic network can be established via simple Bernoulli trials. We now define our vector of unknown parameters, $\eta = (\alpha_0, \alpha_1, \alpha_2, \beta_0, \beta_1, \beta_2, \beta_3)$. These are the elements of the two networks that must be found through calibration, which will be detailed in chapter 5.

4.4 Self-Reinforcing Model: Description

Many previous agent-based models of HIV among MSM, and indeed our original model described in the previous sections, typically classify partnerships into main, regular, and casual. However, the distinction between these types of partnerships is not always clear and unambiguous. Some relationships may begin as casual partnerships and then evolve into main partnerships. Furthermore, previous models have assumed that the durations of partnerships are independent, and unrelated to ongoing frequency of sexual contact. In this section we describe

our recent work: an alternative model for sexual partnerships and contacts for use in HIV agent-based modeling. The proposed model allows for the strength of the relationships between partners to evolve over time based on their prior history of contacts and thereby incorporates feedback loops that allow relationships to reinforce, diminish or even spontaneously dissolve.

The key idea of our proposed self –reinforcing model for sexual networks is that the probability of sexual contact between two persons in a given time step depends on the prior history of contacts between the partners. A related idea has recently been used to model message sending behavior in corporate e-mail networks [31]. In what follows we describe specific model structure for application to sexual networks and contacts in the context of HIV transmission. We take the time step to be a single day.

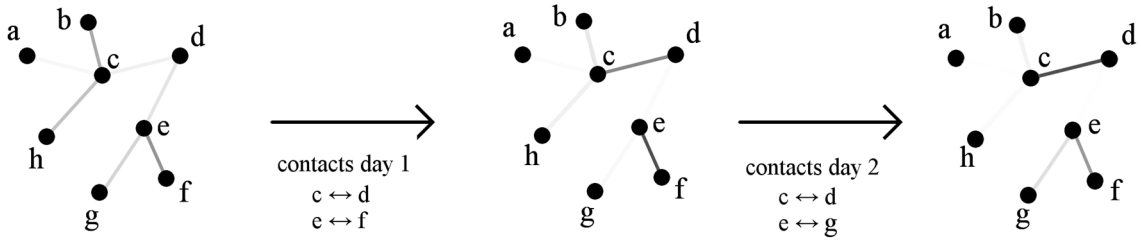
We assume that the probability of contact between two individuals on a given day increases by each prior contact but decreases with increasing duration of the absence of contacts between the partners. In this way, prior contacts reinforce or strengthen the relationship while lack of contact will ultimately diminish the strength of the relationship. Relationships may also spontaneously dissolve. The specific model assumptions are as follows:

1. The log odds of an initial sexual contact between persons *i* and *j* is γ_{0ij}, which may depend on characteristics of persons *i* and *j*
2. A first contact between two individuals increases the log-odds of a future contact by γ_1.
3. A subsequent contact after the initial contact increases the log-odds of a future contact by γ_2.
4. A day without contact decreases the log-odds of a future contact by γ_3.

5. The mean time from first contact until dissolution is λ years and is exponentially distributed.

With the above model assumptions, each relationship begins with a baseline log-odds of a contact intrinsic to the particular pair of individuals. If an initial contact does occur, then the log-odds of a subsequent contact increases. The log odds of each subsequent contact then increases and is reinforced by the prior number of contacts. If no contact occurs, then the log-odds decays. In addition, on any given day there is a chance that the relationship will spontaneously dissolve such that it will be reset back to baseline log-odds. It should be noted that this model assumes a maximum of one contact per pair of individuals per time-step, and thus the time-step should be scaled appropriately. Figure 2 is a schematic illustration of the self-reinforcing model.

Figure 2: A schematic demonstrating the self-reinforcing nature of the model. Each line represents the strength of the link between the two connected individuals (Darker = a stronger connection). The stronger the current connection the more likely a contact will occur on the subsequent day. Contacts that occur strengthen the connection, while contacts that do not occur weaken it. For example pair (e,f) starts out with a strong connection. On day 1 this pair experiences a contact, which strengthens the connection (darkens it). On day two, however, (e,f) does not experience a contact, and thus it's connection is subsequently weakened.



4.5 Self-Reinforcing Model: Model Specification and Notation

Our network is a representation of pairwise edges and thus can be expressed in matrix form, A_t, the adjacency matrix, where

$$A_{tij} = \begin{cases} 1, & \text{if edge exists between individuals } i \text{ and } j \text{ at timestep } t \\ 0, & \text{otherwise} \end{cases}.$$

This model can be expressed as an ERGM, with the specification that an edge represents a contact at a particular time-step rather than an entire relationship [28] Once again, we use a dyad-independent ERGM to represent the probability of a contact on a given day,

$$logit\left[P_\gamma(A_{tij} = 1) \right] = \eta^T \delta_{t_{ij}} . \qquad (11)$$

Here η is a vector of unknown parameters and δ_{tij}, referred to as a vector of change statistics, is the difference between the network statistics at time *t* with and without the addition of edge *ij*. This δ_{tij} includes historical contact data (all the prior contact matrices that have occurred up until time t). Notice that here we are not making an assumption of complete partnership independence, as we did in the case of the static network, but only the weaker assumption of independence conditional on the contact history.

From this point on we will not refer to the vectors γ and δ_{tij}, but to their individual components. To completely specify this model, we define the following terms:

N = number of individuals in the simulated population

λ = dissolution parameter

τ = total number of timesteps

$p_{ijt} = P_\gamma(A_{t_{ij}} = 1)$ = probability of a contact between i and j at timestep t

d_{ijt} = number of timesteps since first contact between i and j at timestep t

$$f_{ijt} = \begin{cases} 0 & \text{if i and j have had any prior contact} \\ 1 & \text{otherwise} \end{cases}$$

C = matrix of contacts of dimension $N \times N \times \tau$

$$c_{ijt} = \begin{cases} 1 & \text{if there is a contact between } i \text{ and } j \text{ at timestep } t \\ 0 & \text{otherwise} \end{cases}$$

C is simply an archive defined here separate from each daily network A_t merely for convenience and consistency of notation. Each slice $C_{..t}$ is equal to its respective realized daily network, which in classic ERGM notation would be written a_t.

The dissolution parameter, λ, is assumed to be the constant rate of spontaneous dissolution of a partnership per day. We will assume that spontaneous dissolution will reset the log odds of contact back to γ_{0ij}. We can thus write the probability of contact at time-step t for two individuals as

$$logit(p_{ijt}) = \gamma_{0ij} + f_{ijt} * (\gamma_1 + \sum_{k=0}^{t-1} c_{ijk} * \gamma_2 - (d_{ijt} - \sum_{k=0}^{t-1} c_{ijk}) * \gamma_3) \quad . \qquad (12)$$

The above equation also holds for a new relationship between a couple who had previously experienced a dissolution between them of past relationship provided c_{ijk} refers only to contacts

in the current ongoing relationship and not contacts in their previously dissolved relationship. The parameter γ_{0ij} performs the important role of incorporating predispositions to contact for pairs of individuals. It could be expressed as

$$\gamma_{0ij} = \gamma_0 + \beta^T Z_{ij} \quad , \tag{13}$$

where β is a vector of unknown parameters, and Z_{ij} is a vector of covariates that characterize the partnership. For example, Z_{ij} can include an ethnicity matching indicator variable ($\mathrm{I}_0(Ethnicity_i = Ethnicity_j)$), a difference in age variable ($abs(Age_i - Age_j)$), a measure of sexual activity, or a measure of degree separation. Degree separation in particular refers to the idea that "friends of friends" are more likely to have contact with one another.

4.6 Self-Reinforcing Model: Application to HIV transmission

Our first extension for our HIV-specific agent-based model lies in the specification of baseline parameter γ_{0ij} . For the entries of Z_{ij} we use a measure of overall sexual activity: $(a_i + a_j)$. Since there is no true measure of activity available to us, we use a surrogate measurement drawn from the distribution of unique six-month partners. Essentially, we assume that this activity level is directly proportional to the number of unique partners an individual

might have contact with over six months. We draw an activity parameter α_i for every individual i, resulting in the re-definition of $\gamma_{0ij} = \gamma_0 + \beta * (a_i + a_j)$.

We also extend the model to allow partnerships to evolve into main partnerships. Main partnership classification modifies the model in two ways: first, a partnership reaching a threshold value of log-odds can be classified as a "Main" partnership, and second, the concept of partial monogamy can now be included as in our static partner model, whereby the presence of a main-classified relationship decreases the likelihood of contacts outside of this relationship [2]. This introduces the following new terms into the model:

T = log-odds threshold at which a main partnership is formed

γ_4 = decrease in log odds of a contact due to main partnership exclusivity

$$M_{ijt} = \begin{cases} 1 & \text{if logit}(c_{ijt}) > T \\ 0 & \text{otherwise} \end{cases} .$$

$$e_{ijt} = \sum_{k \neq j}^{n} M_{ikt}$$

Exclusivity can be applied in various ways, one of which is a simple additive application, whereby the exclusivity term in the log-odds equation becomes

$$-\gamma_4 * (e_{ijt} + e_{jit}) .$$

Notice here that, because of the greater complexity of this model, we use a different notation for main partnership membership than in the static model (e_{ijt} and e_{jit} vs H_i and H_j). Our main partnerships here are time dependent and only look at partnerships outside of the pair i and j, while our static model included i and j main partnership membership in H_i and H_j, and simply compensated by increasing γ_1. As is the case in the static partnership model, $\gamma_4 \geq 0$ by definition.

An important case to consider, especially in the case of sexual contact networks, is the existence of an exclusivity-independent subpopulation, to which the concept of exclusivity might not apply (e.g. sex-workers), in which case every individual i in this subpopulation should have exclusivity term

$$e_{ijt} = 0 \; \forall \; j,t \; .$$

It is easy to assign this class based on a threshold value of the "activity" levels used in the re-parameterization of γ_{0ij}, but for our current work we have not implemented this subpopulation.

Relationship dissolution can also be made specific to each pair of individuals, such that instead of one λ that describes all pairs, the time until dissolution can then be drawn from the exponential distribution with a parameter λ. If one has access to a population-level estimate of the mean relationship duration, $\hat{\lambda}$, one option is to use the surrogate "activity" values used in the re-parameterization of γ_{0ij} to re-parameterize λ_{ij}. This is reasonable if one believes that a

higher activity level correlates (either positively or negatively) with relationship duration. If one assumes that higher activity inversely correlates with relationship duration, one can parameterize λ_{ij} as

$$E[\lambda_{ij}] = \theta \frac{1}{1+\alpha_i+\alpha_j} \ . \qquad (14)$$

This allows for a simple estimate of the proportionality constant θ:

$$\hat{\theta} = \frac{\hat{\lambda}}{\frac{1}{N(N-1)}\sum_{i \neq j}\frac{1}{1+\alpha_i+\alpha_j}} \ . \qquad (15)$$

Of course, much like the inclusion of the exclusivity-independent subpopulation, one could argue that a similar (or even the same) subpopulation of high-activity individuals might have their own dissolution rate, such that $\lambda_{ij} = \lambda_s \forall$ individual in this subpopulation . We do not go so far as to assume this population exists in our HIV model, but we do re-parameterize $\hat{\lambda}$ with $\hat{\theta}$.

Finally we decided for our model to utilize an alternative form of relationship decay, where the additive decay term

$$-(d_{ijt} - \sum_{k=0}^{t-1} c_{ijk}) * \gamma_3$$

is removed and replaced with a sequential decay term:

$$-[\sum_{k=r_{ij}}^{t-1}(1-c_{ijk})]*\gamma_3,$$

where r_{ij} represent the most recent day of contact. This term then states that the relationship decays with the most recent sequence of non-contacts. Or, in other words, every day that goes by without a contact contributes to the decay of the relationship, but if a contact occurs this slate is wiped clean and sequential decay begins anew. For simplicity we have kept the notation γ_3, since we include only one decay type in simulation in order to limit the number of required unknown parameters.

Our final equation used in our HIV simulation then becomes

$$\begin{aligned} logit(p_{ijt}) = \gamma_0 + \beta*(\alpha_i+\alpha_j) + f_{ijt}*[\gamma_1 + \sum_{k=0}^{t-1} c_{ijk}*\gamma_2 \\ -[\sum_{k=r_{ij}}^{t-1}(1-c_{ijk})]*\gamma_3 - \gamma_4*(e_{ijt}+e_{jit})] \end{aligned}, \qquad (16)$$

and for this model we define the vector of unknown parameters, $\eta = (\gamma_0, \beta, \gamma_1, \gamma_2, \gamma_3, \gamma_4, \mathrm{T})$.

Chapter 5: Calibration

5.1 Coarse Network Statistics

The lack of a detailed network structure of MSM (men who have sex with men) relationships, makes the simulation of a true ERGM difficult if not impossible. In places where slightly more detailed MSM network data is available, groups have taken to simulating an ERGM structure for "main partnerships" only, and a random contact structure for "casual contacts" between other individuals where sexual contacts [30]. Our approach differs in that we do not attempt to extrapolate an underlying partnership structure from our survey data, or coarse network statistics, but rather attempt to replicate survey data through our simulation while allowing for smooth growth and decay of relationships.

Since MSM in South Africa are comprised of a stigmatized and difficult to reach group of individuals and detailed survey data is lacking, we are often left with an incomplete idea of the network structure of this group [4]. We have access instead to what we refer to as "coarse network statistics." These take the form of published survey statistics, including the "mean number of unique partners in 6 months," "the percent with greater than 5 unique partners in 6 months," "the percent of individuals currently in a main partnership," and "the ratio of the mean number of unique partners an individual belonging to a main partnership has in twelve months to that of an individual who does not belong to a main partnership [2, 32]." The target values for these coarse network statistics are given in Table 2 of section 5.4.

This brings us to an interesting difference between other agent-based models and our own work. Often if an agent-based model is developed with complete network data, one can use snapshots of the network taken at different times to create formation and dissolution networks needed for an STERGM (see section 3.2). Our coarse network data are not snapshots of the network at different points in time, but rather summaries of contact behavior over time (e.g. unique partners contacted over six months). If we condense them into an instantaneous network we make the assumption that these partners are all concurrent (partners within a relationship at the same time). We do not make this assumption, but rather calibrate our model so that the simulated coarse network statistics match the data that we have. This requires a great deal of computation time because agent-based model simulations are required in the calibration process.

Thus, instead of attempting to determine an instantaneous network structure from survey statistics that are themselves representative of a history of sexual contact interactions, rather than a snapshot of a current network, we calibrate to the history itself. Here, calibration refers to the process of identifying values for the unknown parameter vectors, η, which are the inputs into our agent-based models and defined for the static partnership and self-reinforcing models on pages 25 and 34, respectively. For every point in parameter space (value of η), a series of simulations can be run and the coarse network statistics found for each. Objective functions measuring closeness to the goal coarse network statistics can then help determine how to progress further through the parameter space. This is a very high computational burden with only N=1000 agents modeled. Even with highly parallelized processing on a cluster, this difficult task requires an efficient search algorithm. In addition, because of the stochastic nature of these simulations,

there is noise accompanying these statistics. Thus it is a problem of multi-objective optimization with noisy data.

5.2 Static Partnership Model: Unified Objective Function

With the static partnership model we took a simplified approach to model calibration. Our vector of unknown parameters is defined as $\eta = (\alpha_0, \alpha_1, \alpha_2, \beta_0, \beta_1, \beta_2, \beta_3)$. We now constrain it so that we have the following operational definitions for partner types: a main partner is a partner with which an individual makes a sexual contact (not necessarily a CAI) on average once every two days; a regular partner is a partner with which an individual makes a sexual contact on average once a week, and the log-odds of two individuals being in each other's regular partner network increases by 3 when they are the same serostatus ($\alpha_2 = 3$). This reduces the dimensionality of η to a four-dimensional space and allows us to perform a grid search over the entire parameter space, minimizing the unified objective function

$$f(\eta) = \sum_{w=1}^{m} \sum_{v=1}^{p} \left[\frac{y_v - g(\eta)_{vw}}{y_v} \right]^2, \qquad (17)$$
$$\eta_{min} = \arg\min_{\eta} f(\eta)$$

where each y_v is a coarse network statistic, p is the total number of coarse network statistics, $g(\eta)_{vw}$ is the simulated coarse network statistic *v* for replicate *w*, *m* is the number of simulation replicates, and η_{min} is the value of the unknown parameter vector η that minimizes this objective function. The simulated coarse network statistic requires an agent-based model simulation. We

chose this objective function because of its simplicity. It was possible to find our minimizer η_{min} for this function using a simple grid search of the parameter space because of the constraints that we had placed on η. For our more recent, self-reinforcing model, we wished to eliminate these constraints. We desired a technique that was already well established and robust for calibrating a high-dimensional unknown parameter vector to multiple and noisy objectives. The noise here refers to the stochastic variability associated with each simulated coarse network statistic.

5.3 Self-Reinforcing Model: Multi-Objective Evolutionary Algorithms

Multi-objective evolutionary algorithms are a subset of stochastic methods of optimization that are ideal for the scenario of calibrating to noisy simulated objectives for two reasons. First, they are derivative-free, which is necessary since our output is a stochastic simulation (each simulated coarse network statistic $g(\eta)_{vw}$ is not differentiable with respect to η). Second, they are not as prone to getting trapped in local minima as other methods [33], which is a necessary requirement because the number of local minima in $f(\eta)$ is unknown. Recently, multi-objective optimizers have been developed to address the challenge of noisy optimization [34]. These multi-objective evolutionary algorithms are highly varied in their methodology but almost always make use of the following concepts: the Pareto frontier, crossing over, and mutation, all of which is explained in detail later in this section. Noisy optimization here simply refers to the act of minimizing objective functions when each evaluation of these functions is accompanied by noise, which in our application is synonymous with the stochastic variance of simulated coarse network statistics.

For the self-reinforcing model we define the unknown parameter vector $\eta = (\gamma_0, \beta, \gamma_1, \gamma_2, \gamma_3, \gamma_4, \mathrm{T})$. We elected to use the "Rolling Tide Evolutionary Algorithm" for our parameter search. This algorithm has been shown to have good convergence properties in the face of noise that changes as a function of the parameter space [35]. Figure 3 demonstrates how the variance of something as simple as the mean number of daily contacts can be highly dependent upon the position in the parameter space (the current value of unknown parameter vector η). We describe this algorithm in detail later in this section.

Figure 3: Mean number of daily contacts across individuals versus the simulation time-step, which here is one day. Each line represents a simulation from an agent-based model initialized with no prior contacts, with $\gamma_0 = -13.5$, $\beta = 0.2$, $\gamma_2 = 0.9$, $\gamma_3 = 0.5$, $\gamma_4 = 14.3$, and $T = -0.1$, corresponding to the best calibration of the HIV model. The solid lines correspond to $\gamma_1 = 5$ and show quick convergence and small final variance relative to the dotted lines, which correspond to $\gamma_1 = 25$. Both the number of time-steps to convergence as well as the variance of converged coarse network statistics are shown to vary within the parameter space.

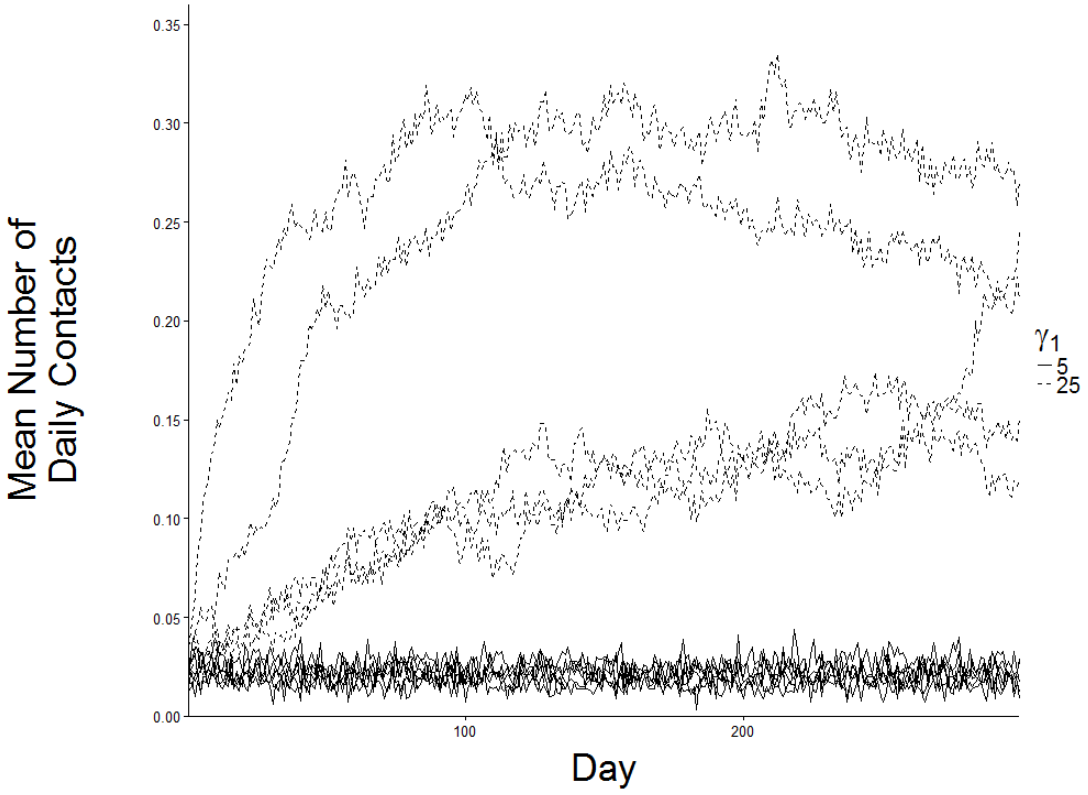


Figure 3 shows the behavior of a simple example coarse network statistic, the mean number of contacts across individuals, over time and for two different values of γ_1 . The model is initialized with zero previous contacts and each line represents a distinct starting network with randomized agents. Burn in here is considered to be the initial vertical climbs of the coarse network statistics. In Figure 3, for example, the burn in for the simulations corresponding to

$\gamma_1 = 25$ is accomplished by day 200, while for $\gamma_1 = 5$ it is accomplished nearly immediately and is not visible to the naked eye. This does not mean, however, that there will be no drift in coarse network statistic values over simulation time. Looking at the two highest lines in Figure 3, we can see that they experience a slight downward trend over time after day 200, while one of the lower dashed lines experiences a sharp incline after day 250. Because this is a stochastic simulation where we allow partnership strengths to form, dissolve, strengthen, and diminish over time, we expect a certain amount of drift to occur. No matter how long the simulations are run, they will always differ in their final coarse network statistic values by some amount. As we can see by comparing the $\gamma_1 = 25$ simulations to those of $\gamma_1 = 5$, this variance is dependent upon the value taken by the vector of unknown parameters, η (here $\eta = [-13.5, 0.2, 5.0, 0.9, 0.5, 14.3, -0.1]$ for $\gamma_1 = 5$ and $\eta = [-13.5, 0.2, 25.0, 0.9, 0.5, 14.3, -0.1]$ for $\gamma_1 = 25$).

This dependence of the variance of the simulated coarse network upon the value of η is addressed by the Rolling Tide Algorithm, which was specifically designed to work in the context of variance changing as a function of the parameter space [35]. To ensure convergence of the coarse network statistics within individual simulations, we found that a 1000 time-step burn-in was more than sufficient for all simulated coarse network statistics used in our calibration.

Here we briefly review the main ideas of the Rolling Tide algorithm. To use this algorithm we first create an objective function, $f_v(\eta)$, for each coarse network statistic defined as

$$f_v(\eta) = \frac{1}{m}\sum_{w=1}^{m} | y_v - g(\eta)_{vw} | \quad , \qquad (18)$$
$$v = 1 \ldots p$$

where $g(\eta)_{vw}$ is once again the *w*th replicate of the *v*th coarse network statistic simulated at parameter vector value η, p is the total number of coarse statistics, y_v is the desired value for the *v*th coarse network statistic as determined from literature and survey data, and m is the number of simulation replicates performed at η. Each objective function then corresponds to the distance from "true" coarse network statistics and our goal is thus to minimize each element of the vector

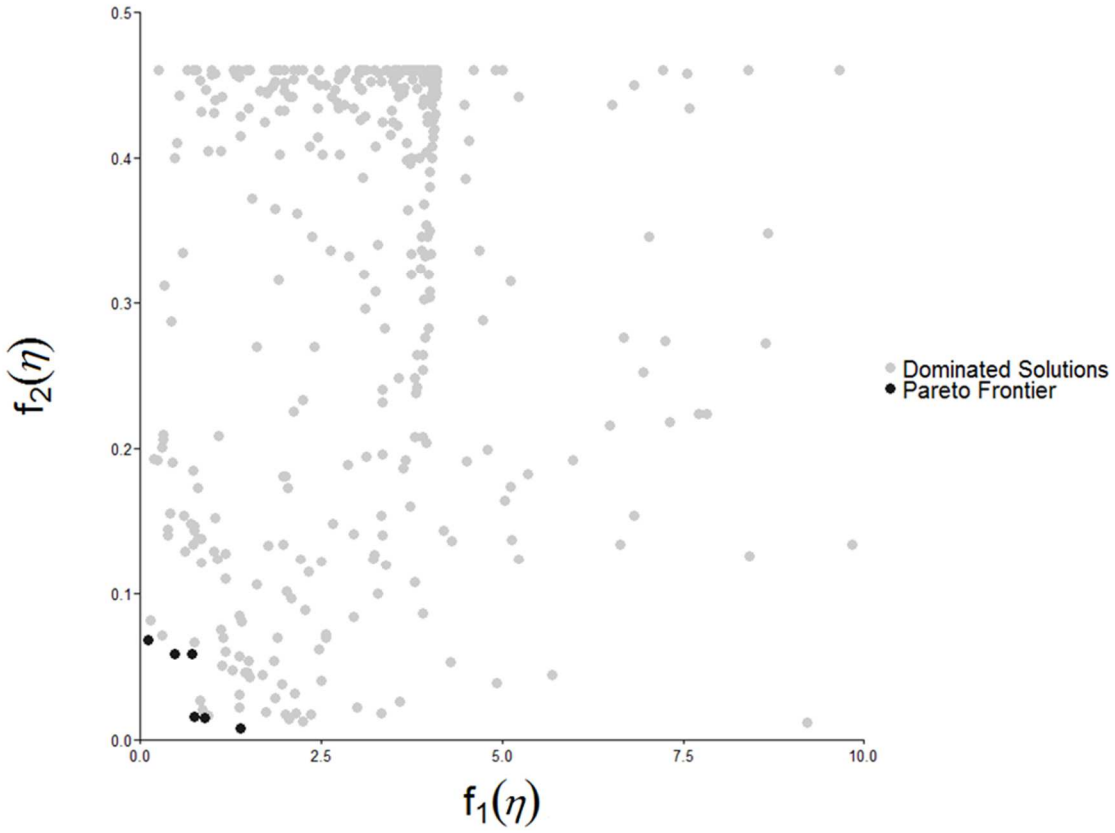
$$f(\eta) = [f_1(\eta), \ldots, f_p(\eta)] \ ,$$

and we define the result of this minimization as

$$\eta_{min} = \arg\min_{\eta} f(\eta) \ .$$

Additional constraints can and should be placed on the η vector to lessen the computational burden. For our purposes we refer to these constraints as the parameter window sizes.

Figure 4: Value of $f_1(\eta)$ = Objective function 1 = Distance to "Mean Unique Partners over 6 months" vs $f_2(\eta)$ = Objective function 2 = Distance to objective "Proportion in Main Partnerships." This graph shows a simplified Pareto frontier achieved between these two objectives at the end of the HIV model calibration. The true calculated Pareto frontier exists across all objectives and is much more difficult to visualize.



The algorithm makes use of the concept of the Pareto frontier, which was first introduced in the 1800s [36]. A solution is said to be a member of the Pareto frontier if it is not dominated by any other solution, which means that no solution exists that is superior for at least one objective function and equal or superior for all other objective functions. Another way to think of the Pareto frontier is as the set of solutions for which performance on one objective cannot be improved without reducing performance on at least one other objective. Thus, all solutions

within the Pareto frontier can be considered globally optimal solutions. It is an alternative to the option of aggregating solutions into a single objective function for minimization, which requires that one combine objectives that exist on different scales, with different boundaries, into a single measurement, as we did with our static partnership calibration. Figure 4 is a visualization of a simple two-objective Pareto frontier extracted from the simulated dataset after calibration to our HIV model. In practice it is difficult to visualize this frontier in more than two to three dimensions.

Figure 5: Schematic of the Rolling Tide Algorithm

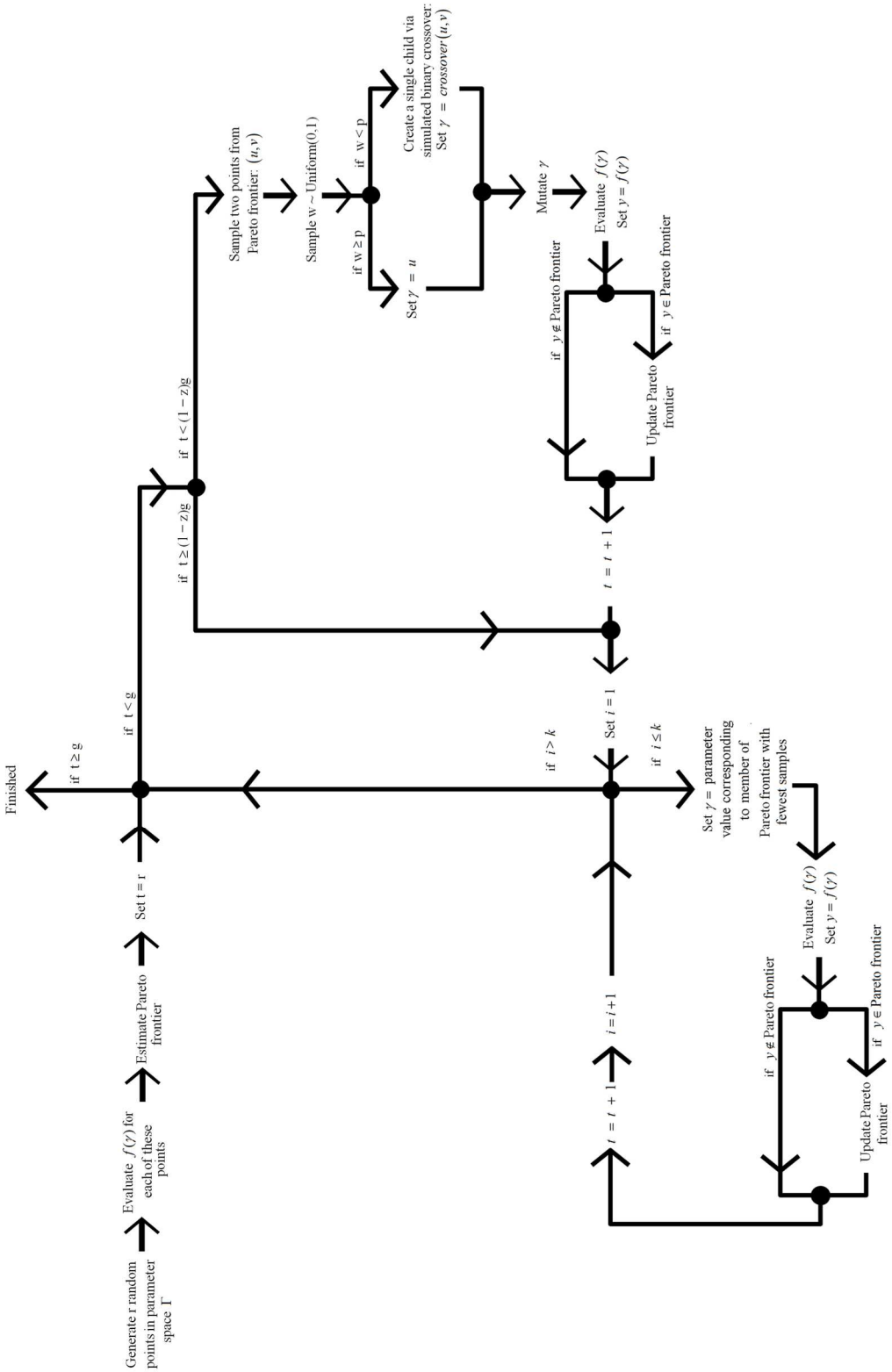


Figure 5 is a schematic illustration of the Rolling Tide algorithms. This algorithm requires the following parameters for execution:

r = Number of random locations to initially sample
p = Probability of a crossover occuring
s = mutation window size
g = Total number of function evaluations
k = Number of archive resamples per iteration
z = Proportion of run to be spent solely refining the archive

It begins by evaluating r points chosen randomly over the parameter space, Γ, and then calculating the Pareto frontier. It continuously both re-evaluates points (values of the vector γ) at this frontier in addition to generating new potential points by mixing these members of the Pareto frontier. It is in this mixing portion that the "evolutionary algorithm" aspect comes into play. In order to mix two points from the frontier it employs both "crossing-over" and "mutation." When generating new points, crossing-over occurs with probability p and in our setting is simply the act of taking a random proportion of elements from the parameter vector of one member of the frontier and combining it with the remaining elements of another member of the frontier. For example, γ_2 and γ_3 might be taken at random from one member and the remainder from the other member of the frontier.

Mutation in our setting is taking each element of the parameter vector and altering it within a fixed mutation window size s. We run our algorithm with $s = 1/10^{th}$ parameter window size, a value suggested by previous multi-objective optimization literature [37]. For example, if

our search uses an acceptable parameter window of $(-1,-21)$ for element γ_0 and we wish to mutate it from a value of -6.25, we would add to -6.25 a random variable drawn from a uniform distribution over interval $(-1,1)$, since our mutation window has width $0.1*20=2.$ It is in this manner that members of the Pareto frontier are mixed and modified over the course of the algorithm.

Next, the total number of function evaluations, *g*, refers to the total number of simulations that we choose to run over the course of the calibration. The archive resamples, *k*, refers to the number of times in a calibration loop that a member of the Pareto frontier will be re-evaluated with an additional simulation replicate. The refining proportion, *z*, refers to the proportion of calibration loop devoted only to these re-evaluations. It is there essentially to ensure that all points on the final Pareto frontier have enough replicates to compensate for their variances.

5.4 Self-Reinforcing Model: Multi-Objective Optimization Calibration Results

The Rolling Tide Algorithm was implemented with parameter windows designed to be as wide as possible while still reflecting the structure of the model, which predetermines sign. The values of these windows are given in the description of Table 5 in section 5.5. We performed 5000 iterations of the Rolling Tide Algorithm to calibrate this HIV model. In order to quantify the improvement of the Pareto frontier over the course of the calibration, we combine objective

functions into the metric $\sum_{j=1}^{p} \frac{f_j(\eta)}{y_j}$, normalizing each objective function by its target coarse network statistic.

Figure 6: Iterations of the Rolling Tide Algorithm: Mean sum of normalized distance (= $\sum_{j=1}^{p} \frac{f_j(\eta)}{y_j}$) for closest ten members of the Pareto frontier vs. iteration number. As the calibration progresses, it is more difficult to improve this total distance, representing an expected diminishing returns aspect of the calibration process.

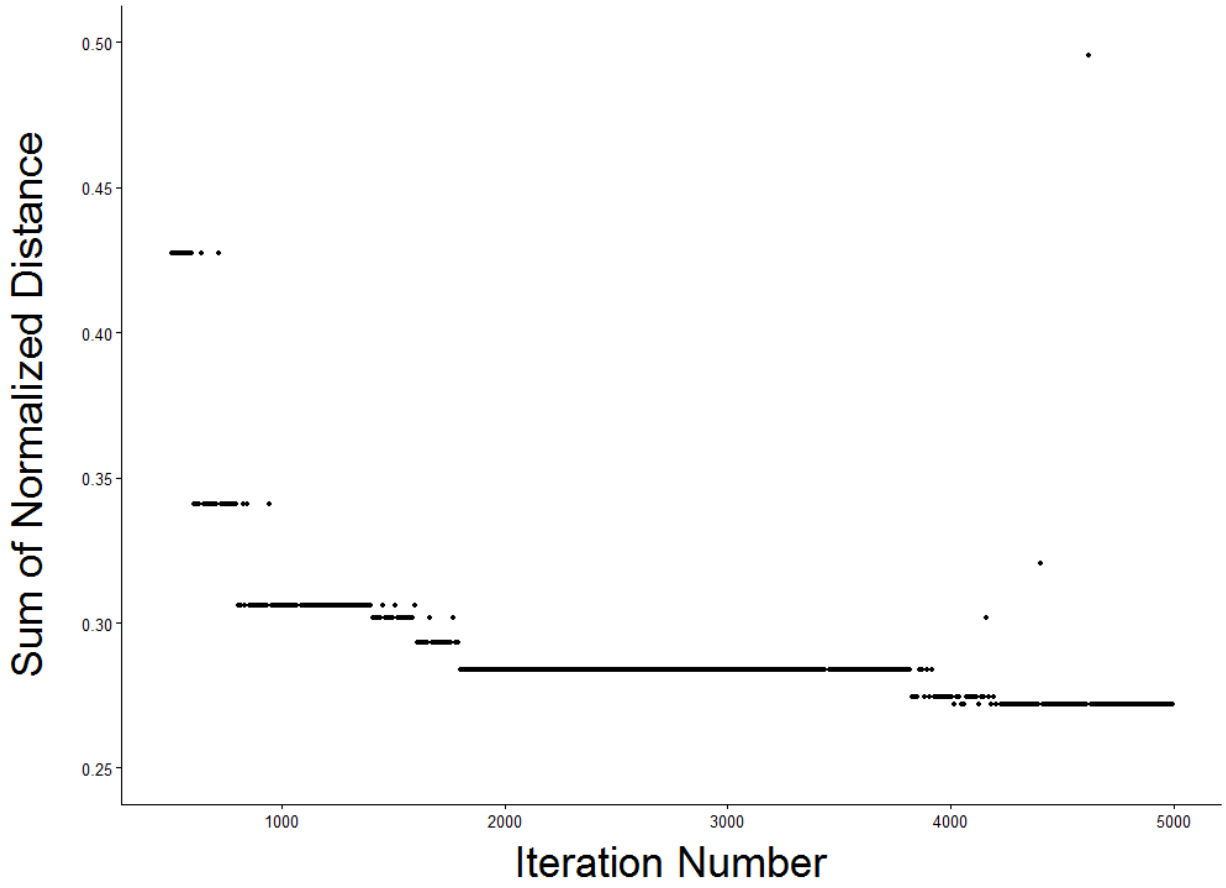


Figure 6 looks at the mean of the ten smallest values of this metric found within the Pareto frontier at each iteration. It shows that there are clear diminishing returns as more iterations are added to the Rolling Tide process.

Table 2: Coarse Network Statistics: The coarse network statistics available to the South-African MSM HIV agent-based model along with the value from the final best fit parameter set (Seen in Table 3). Notably the statistic that is most difficult to calibrate is the Ratio of Unique Partners.

Coarse Network Statistic	**Target Value**	**Calibrated Value**
Mean Unique Partners in 6 mo.	4.10	4.46
Percent in Main Partnerships	46	59.20
Median Unique Partners in 6 mo.	2	2.11
Percent w/ > 5 Partners in 6 mo.	17.7	14.50
Ratio of # of Unique Partners for someone in a Main Partnership vs someone not in a Main Partnership over 12 mo.	2.71	1.48
Yearly HIV Incidence %	7	5.60

Table 2 demonstrates the coarse network output for our best-case calibration, by which we mean the member of the final Pareto frontier with the minimum value of $\sum_{j=1}^{p} \frac{f_j(\gamma)}{y_j}$, performed on a network of N=1000 agents. This table gives the coarse network statistics used in

our HIV model, the post-calibration average values obtained from the simulation, and the target value obtained from literature. In this "best-case" calibration, the two largest trade-offs are given by the percent of individuals in main partnerships, which here is 59% rather than 46%, and the ratio of unique 12-month partners for someone not in a main partnership vs someone in a main partnership, which is under-shot by the model as 1.48 instead of 2.71, making it the largest model discrepancy.

Table 3: Descriptive statistics for the number of Unique Partners for N=1000 agents over various time-frames given in months, (time-step equals 1 day). Given within the parentheses next to each statistic is the standard deviation of said statistic. These results were obtained over 100 replicates at the best calibrated value for the parameter set.

	Months			
	2	**6**	**12**	**24**
Mean	1.7 (0.4)	4.4 (0.83)	7 .9(1.3)	14 (2)
%tile				
25th	0.01 (0.1)	0.93 (0.33)	2.2 (0.54)	5.1 (0.86)
50th	0.73 (0.44)	2 (0.47)	4.1 (0.69)	8 (1.1)
75th	1.5 (0.5)	3.8 (0.56)	6.8 (0.1)	12 (1.4)

In Table 3 we expand upon the concept of unique partners over additional time windows. We characterize the distribution of unique partners over 2, 6, 12, and 24 month periods, looking at mean as well as 25th, 50th, and 75th percentiles, accompanied by their respective standard deviations. Each of these could potentially serve as a coarse network statistic, should the

pertinent survey data be made available, but here it simply serves to describe the distributional shape of the relationships formed over time. By calibrating to 6 and 12 month coarse network statistics, the model also influences 2 and 24 month coarse statistics, and it is important to at least be aware of these effects when calibrating an agent-based model.

5.5 Self-Reinforcing Model: Sensitivity Analyses

In order to examine the sensitivity of both our Self-Reinforcing model as well as our approach to calibration, we performed independent Rolling Tide Calibrations across several key parameters: N, λ, and δ. The term δ has been mentioned but not introduced formally, and represents the log-odds increase due to two individuals having a degree separation of 1, which means that both individuals have had sexual contact with the same individual in the past. It seemed prudent to produce a variation of our model which included this aspect of human interaction. In Table 4 we present the six combinations of these parameter values at which we produced our simulations.

Table 4: The six combinations of parameter values at calibration of the remaining parameters was performed over 20000 iterations of the Rolling Tide Algorithm.

Parameter Combinations

N	**λ**	***δ***
1000	3	0
2000	3	0
5000	3	0
1000	1	0
1000	5	0
1000	3	1

Table 5: Parameter set resulting in best fit of coarse network statistics after calibration via the Rolling Tide Algorithm for N=1000, N=2000, and N=5000 agents. Each was calculated independently over 3000 iterations, and was selected by being the member of the Pareto frontier with the smallest value of $\sum_{j=1}^{p} \frac{f_j(\eta)}{y_j}$. These parameters were calibrated using the following bounds: $\gamma_0 \in (-20,-3)$, $\beta \in (0.01,2)$, $\gamma_1 \in (0.5,20)$, $\gamma_2 \in (0.001,1)$, $\gamma_3 \in (0.001,1)$, $\gamma_4 \in (0,20)$, $\lambda \in (1,3650)$, and $T \in (-6,3)$

	Calibrated Value		
Parameter	**N=1000**	**N=2000**	**N=5000**
γ_0	-13.458	-14.007	-14.000
β	0.183	0.169	0.121
γ_1	13.995	14.517	12.96
γ_2	0.880	0.538	0.431
γ_3	0.498	0.311	0.116
γ_4	14.275	6.135	4.88
T	-0.100	0.222	-2.369

In Table 5 we show the actual parameter values obtained for different size networks, once again using the smallest value of $\sum_{j=1}^{p} \frac{f_j(\eta)}{y_j}$ as the metric for selection. Here we show that all three are structurally very similar, with the N=5000 network showing an interesting downward shift in the threshold requirement for main partnership status *T*. We should, however, note that these are simply representative of the "best" parameter sets as selected by our simple normalized sum metric. In practice, every member of the Pareto frontier is a potential minimizer of our

group of objective functions, and thus an examination of the three "best" solutions according to our metric does not yield much more than a superficially interesting comparison. Instead, one might be better off choosing those parameter sets that produce similar patterns in the coarse network set, so as to ensure like behavior among networks of different sizes when running sensitivity analyses across network size. This is indeed what we have done, choosing the 5 "best" Pareto-optimal solutions for each parameter combination. Our simulation results are then an average over these 5 solutions.

Table 6: Mean coarse network statistics (over 200 total replicates) by parameter designation along with target coarse network statistics.

			N	
Coarse Network Statistic	**Target**	**1000**	**2000**	**5000**
Mean 6 mo. Partners	4.10	3.37 (0.85)	3.25 (0.67)	3.88 (0.73)
Percent in Main	46	71.4 (3.0)	56.3 (15)	71.5 (2.4)
Median 6 mo. Partners	2	2.15 (0.63)	1.76 (0.49)	1.65 (0.55)
Percent w/ > 5 Partners in 6 mo.	17.7	15.1 (5.9)	13.5 (4.4)	13.7 (3.0)
Ratio	2.71	1.82(0.40)	2.51 (0.53)	2.76 (1.09)
Incidence	7	5.56 (2.5)	4.95 (1.8)	6.76 (1.97)

Table 7: Mean coarse network statistics (over 200 total replicates) by parameter designation along with target coarse network statistics.

		δ	
Coarse Network Statistic	**Target**	**0**	**1**
Mean 6 mo. Partners	4.10	3.37 (0.85)	4.01 (0.72)
Percent in Main	46	71.4 (3.0)	51.9 (10)
Median 6 mo. Partners	2	2.15 (0.63)	2.14 (0.56)
Percent w/ > 5 Partners in 6 mo.	17.7	15.1 (5.9)	18.0 (4.5)
Ratio	2.71	1.82 (0.40)	2.14 (0.55)
Incidence	7	5.56 (2.5)	4.74 (2.8)

Table 8: Mean coarse network statistics (over 200 total replicates) by parameter designation along with target coarse network statistics.

		λ		
Coarse Network Statistic	**Target**	**1**	**3**	**5**
Mean 6 mo. Partners	4.10	3.67 (0.92)	3.37 (0.85)	3.46 (1.2)
Percent in Main	46	59.6 (5.1)	71.4 (3.0)	55.1 (13)
Median 6 mo. Partners	2	1.83 (0.50)	2.15 (0.63)	1.98 (0.75)
Percent w/ > 5 Partners in 6 mo.	17.7	15.8 (6.0)	15.1 (5.9)	14.7 (7.2)
Ratio	2.71	2.46 (0.51)	1.82 (0.40)	1.86 (0.59)
Incidence	7	5.60 (3.3)	5.56 (2.5)	5.45 (3.1)

Tables 6, 7, and 8 display the mean coarse network statistics across N, δ, and λ, respectively. Notice that even at 20,000 iterations there is no guarantee that the calibration will produce identical coarse network results across these parameters. It is unknown whether there even exists a solution within the parameter space that outputs coarse network statistics that perfectly matches target values, so it is logical that some of these structural changes that we make to the model involving N, δ, and λ might result in models that are more or less difficult to calibrate. In particular we notice that obtaining a stable main partnership percent of 46% appears to be difficult with this model, which we will address a bit further in our discussion.

Chapter 6: Simulation Results

After calibration, both agent-based simulations proceed day by day. On each day, an uninfected person who has sexual contact with an infected person has a transmission probability of becoming infected, and Bernoulli trials with the transmission probability simulate whether or not infection occurs. The transmission probability is determined by the type of sexual contact and the presence of any prevention interventions, such as antiretrovirals treatments, which would modify the transmission probabilities. We considered four prevention interventions and combinations of those interventions. The first intervention was treatment of HIV infected persons with ART. HIV infected persons with a CD4<350 who had an HIV test within the preceding 6 months were eligible to receive ART.

For the stationary network we considered an array of values for our four interventions. The first intervention indicated the proportion (λ_1) of eligible persons who actually receive ART (λ_1=0.05, 0.25, 0.5, 0.75, and 0.95). The second intervention was prophylactic treatment of high risk HIV uninfected persons to reduce risk of acquisition of HIV infection (PREP). HIV uninfected persons who had an HIV test within the preceding 6 months and were at high risk (defined as either >12 acts of condomless anal intercourse (CAI) in the preceding 6 months or having a main partner who is HIV infected) were eligible to receive PREP. We considered various values for the proportion (λ_2) of eligible persons who are offered and accepted PREP (λ_2=0.05, 0.25, 0.5, 0.75, and 0.95). Persons on PREP were classified as either a low or high adherer (the effectiveness of PREP depends on level of adherence). The third intervention was a counseling and condom promotion program to reduce condomless sexual contacts. We considered the impact of an intervention that could reduce the proportion of CAI contacts by λ

$_3$x 100% (λ_3=.05, .25, .50, .75 and .95). The fourth intervention was a program to increase HIV antibody testing. We considered an intervention that decreases by one half the proportions of persons who never receive an HIV antibody test, from 1/3 to 1/6. We will indicate that this intervention by the indicator λ_4=1.

We ran simulations of the agent-based model for most combinations of these four interventions over a 5 year period, including all combinations of interventions with increase in ART, PREP, and CAI reduction, yielding 162 distinct combinations. We performed multiple replications for each combination. The mean number of replicates performed for each combination was 13 with a minimum of 5 replicates always performed. We performed 60 replicates for the control setting of no intervention. These simulations produced a data set of 2157 runs of the agent-based models corresponding to the 162 distinct combinations of the prevention interventions.

For our self-reinforcing network these interventions were reduced simply to the baseline intervention (where all intervention parameters equal zero) and a single intervention of interest where λ_1, λ_2, λ_3, and λ_4 take on values 0.5, 0.5, 0.075, and 1, respectively.

6.1 Static Partnership Model: Analysis of Agent-Based Simulation Results

We analyzed the dataset of the results from 2157 simulation runs of our static partnership driven agent-based model. The goal was to determine a model for $\text{var}\left(\hat{P} \mid \theta\right)$, the variance of the

proportion who became infected over 5 years where the vector $\theta = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ defines the prevention interventions that are in place. We fit a generalized linear model for the mean of structure $E[\hat{P} \mid \theta] = P(\theta)$ and ultimately decided, after model fitting and regression diagnostics, on a logistic link of the form

$$logit(P(\theta)) = \beta_0 + \beta_1\lambda_1 + \beta_2\lambda_2 + \beta_3\lambda_3 + \beta_4\lambda_3^2 + \beta_5\lambda_3^3 + \beta_6\lambda_2\lambda_4 \qquad (19)$$

We modeled the variance $\text{var}\left(\hat{P} \mid \theta\right)$ using the empirical sample variances of $\hat{P}$ as the observed dependent variable. After model fitting and regression diagnostics, we ultimately decided that it was adequate to model the variance as a function only of $P(\theta)$ using a cubic polynomial model,

$$var(\hat{P} \mid \theta) = \beta_1 P(\theta) + \beta_2 P(\theta)^2 + \beta_3 P(\theta)^3 \qquad (20)$$

To estimate the parameters in equations 19 and 20, we used iteratively reweighted least squares whereby updated estimates of the parameters were obtained from fitting equation 19 by weighting by the inverse variances obtained from equation 20 at the previous step [38]. The parameter estimates from equation 20 were determined by least squares weighted by the inverse of the current estimate of $P(\theta)$.

Table 9: Regression coefficients for model of combination HIV prevention packages among MSM in peri-urban South Africa. The modeled percentages infected over five years is given by $P = 100e^{Z} / \left(1 + e^{Z}\right)$ where $Z = \beta_0 + \beta_1\lambda_1 + \beta_2\lambda_2 + \beta_3\lambda_3 + \beta_4\lambda_3^2 + \beta_5\lambda_3^3 + \beta_6\lambda_2\lambda_4$ and the λ's are the covariates that define the components of the combination HIV prevention package (λ_1 is percentages of eligible persons receiving ART[1]; λ_2 is percentages of eligible persons accepting PREP[2]; λ_3 is percentage reductions in CAIs[3]; λ_4 is set to 1 if the percentage of persons who never received an HIV test is reduced in half[4].

Model component	coefficients (β)	Standard error (β)
Intercept	-1.086	0.01263
λ_1	-0.000936	0.00018
λ_2	-0.00266	0.000185
λ_3	-0.04137	0.00155
λ_3^2	0.000642	0.0000461
λ_3^3	-0.0000087	0.00000034
$\lambda_2\lambda_4$	-0.00119	0.000469

[1] λ_1 is the percentage eligible person receiving ART. Persons eligible to receive ART are those persons with a CD4<350 and HIV test within the preceding 6 months.

[2] λ_2 is the percentage of eligible persons who are offered and accept PREP. Persons eligible for PREP are HIV uninfected persons who had an HIV test within the preceding 6 months and are at high risk(high risk is defined here as either >12 acts of condomless anal intercourse (CAI) in the preceding 6 months, or having a main partner who is HIV infected.)

[3] λ_3 is the percentage reduction in the probability a sexual contact is a CAI.

[4] λ_4 is set to 1 if an intervention reduces the percentage of persons who never receive an HIV antibody test by one half, from 33.3% to 16.7%

Now, the quantity we are most interested in is not the proportion infected, but rather the proportion of infections prevented by the intervention, $\grave{o} = 1 - \frac{P(\theta_2)}{P(\theta_1)}$, where $P(\theta_1)$ and $P(\theta_2)$ are the proportion infected for the control group and the treatment groups, respectively. Let Σ be the covariance matrix for the vector of parameters $\hat{\beta}$. Define $\Lambda = (\theta_1, \theta_2)$ as a matrix combining our intervention parameters. Also define $\hat{b}_1 = \hat{\beta}\theta_1$ and $\hat{b}_2 = \hat{\beta}\theta_2$, where θ_1 is the control intervention and θ_2 is the treatment intervention. Also define a vector composite of the two $B = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$ Then B would be distributed according to $\hat{\beta}\Lambda = B \sim N(\beta\Lambda, \Lambda^T \Sigma \Lambda)$. Call this covariance between estimates $C = \Lambda^T \Sigma \Lambda$. The quantity we are interested in is a function of both b_1 and b_2. Call this function h and define it as

$$h(b_1, b_2) = 1 - \frac{\frac{exp(b_1)}{1 + exp(b_1)}}{\frac{exp(b_2)}{1 + exp(b_2)}} \quad . \tag{21}$$

For the delta method we first establish the gradient of this function and find that it can be expressed in terms of probabilities rather than the pre-transformed estimates (see appendix for details),

$$\nabla h(b_1, b_2) = \begin{pmatrix} \frac{\partial h}{\partial b_1} \\ \frac{\partial h}{\partial b_2} \end{pmatrix} = \begin{pmatrix} \frac{-p_1(1-p_1)}{p_2} \\ \frac{p_1(1-p_2)}{p_2} \end{pmatrix}, \tag{22}$$

and utilize the delta method to find our new variance of proportion prevented v^*,

$$v^* = \text{"}\ h(b_1, b_2)^T C \text{"}\ h(b_1, b_2)\ . \qquad (23)$$

We then used this variance to create confidence intervals for the proportion prevented using standard methods so that the interval is represented by: $\grave{o} \pm 1.96 * \sqrt{v^*}$. This is then multiplied by 100 to produce intervals for the percent prevented.

Table 10: Contribution of four components of an HIV prevention package to percent infections prevented. Components include ART (50% ART coverage of eligible persons); PREP (50% acceptance of PREP among eligible persons); CAI reduction (15% reduction), and HIV testing increase (50% reduction of persons who never have an HIV test)

Prevention package component	**percent infections prevented due to addition of component (95% CI)** [1]
ART	3.812 (2.65, 4.98)
PREP	14.6 (11.87, 17.33)
CAI reduction	20.86 (19.85, 21.87)
HIV testing increase	4.828 (1.81, 7.84)
Total[2]	34.16 (31.79, 36.53)
Prediction Interval for Total	(-12.60, 80.93)

Table 10 illustrates the estimated percentage of infections prevented by the various components of the chosen combination intervention package, along with the confidence interval for each of this estimates. These intervals were derived from the variance obtained via the delta method shown above. The effects seen here are far from additive, and show the difficulty to be found in eliminating larger proportions of the epidemic.

The last interval, shown below the confidence interval for the total, is the prediction interval for the total percent prevented. Prediction intervals were derived for the total through the use of our variance equation created for our iterative glm. Using this variance equation we can get the variance $var(\hat{P} \mid \theta)$ estimated for our prevention package component in Table 8. This variance is then used in a simple normal approximation to obtain a confidence interval: $\grave{o} \pm 1.96 * \sqrt{\text{var}(\hat{P} \mid \theta)}$. The variance of the estimate is ignored here due to the fact that the predictive variance is much larger. It is important to note here that the prediction interval covers zero. Even though there is, on average, a 34% decrease in the size of the epidemic, it is very possible that in the comparison of two communities this difference would not be present due to the high variability of the individual trials.

Figure 7: HIV infections prevented over 5 years from combination prevention interventions with four components: early ART, PREP with 50% acceptance (dotted lines), 15% CAI reduction (blue lines; no CAI change are in red) and increase in HIV testing (black triangles). See Table 1 for further details about the components of the prevention interventions.

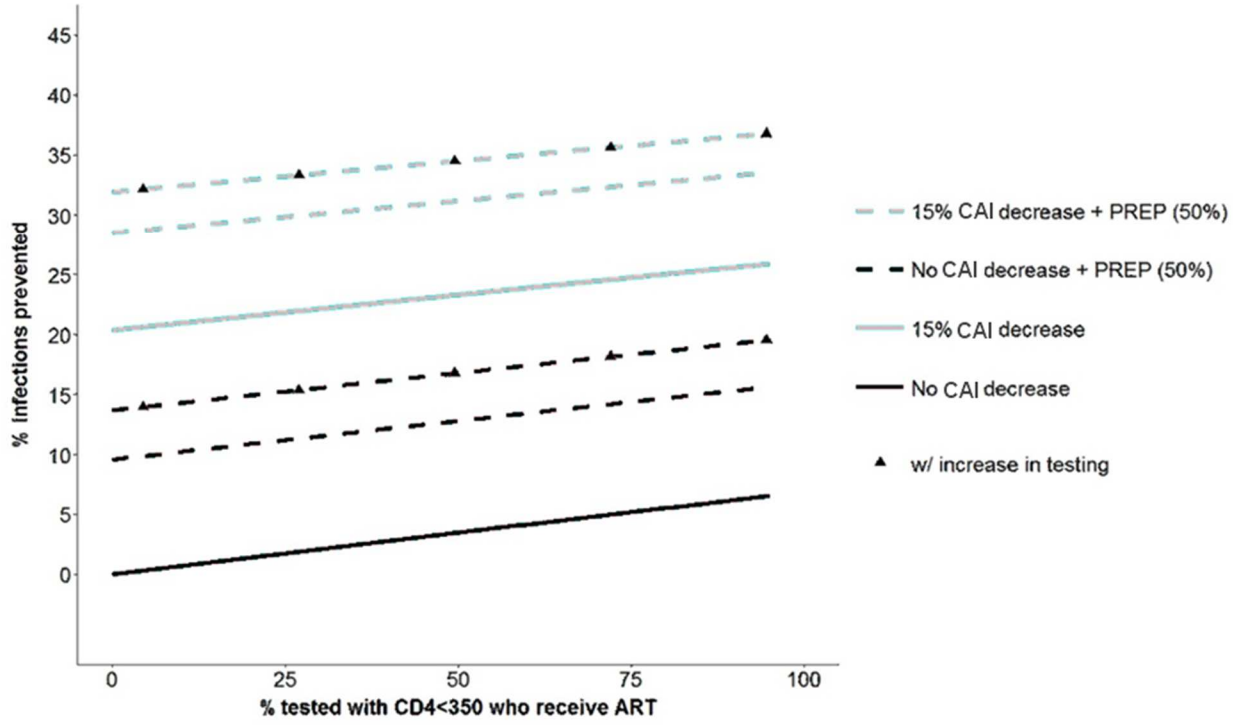


Figure 7 illustrates the iteratively reweighted model estimates for the % infections prevented under various plausible conditions. The most striking effect seen is that of an across-the-board CAI decrease, which, although realistically minimal in its intervention value of a 15% reduction in CAIs across the community, produces the greatest overall effect on prevention. It can be seen here, however, that a modest combination intervention of 50% PREP acceptance, 75% ART acceptance, 15% CAI, and greater testing penetration into the non-testing group can have an impactful effect on the epidemic rates within a community.

Figure 8: Empirical variances of proportions infected in 5 years ($\hat{P}$) for a given θ, versus fitted proportions (from equation 19). Also shown is the fitted variance function from equation 20 [$var(\hat{P} \mid \theta) = 0.05205P(\theta) - 0.1127P(\theta)^3$] and the naïve binomial variance. The fitted proportions are based on equation 8 with β_0=-1.086, β_1=-0.000936, β_2=-0.00266, β_3= -0.04137, β_4 =0.000642, β_5=-0.0000087, β_6=-0.00119

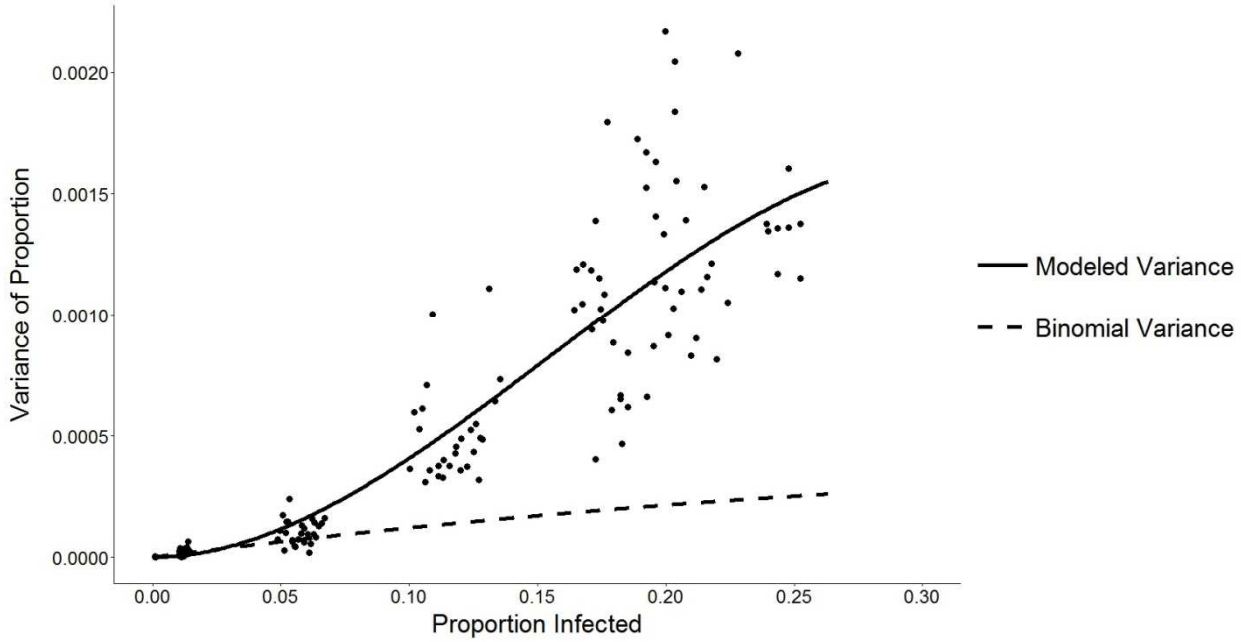


Figure 8 shows the empirical sample variances of $\hat{P}$. Each data point is the result of simulated replications of the agent-based model for a specific combination of interventions. We have plotted the empirical sample variance versus the fitted values of $P(\theta)$ obtained from fitting of equation 12. We found a small but significant decreasing (with increasing proportion infected) trend in the coefficient of variation, which ranged from 0.196 to 0.155. Figure 8 also shows the fitted curve for $var(\hat{P} \mid \theta)$ obtained from fitting equation 13 along with the naïve

binomial variance, $P(\theta)\left(1-P(\theta)\right)/N$. The figure illustrates that the naive binomial variance significantly underestimates the variance induced by the agent-based model by at least 50%.

6.2 Self-Reinforcing Model: Results

For our Self-Reinforcing Model we took a more direct approach to Prevention Package analysis, choosing to focus on one feasible combination intervention. For each sensitivity designation we performed 200 simulations of baseline (no prevention package) and 200 simulations at the prevention package combination: 50% ART acceptance, PREP with 50% acceptance, 15% CAI reduction, and increase in HIV testing. This is the same package featured in Table 10.

Table 11: Yearly mean percent incidence of HIV (standard deviation) for baseline and intervention prevention packages (based on 500 replicates each) as well as the corresponding percent prevented for each pair of packages, given across levels of N.

N	Baseline Percent Incidence	Package Percent Incidence	Percent Prevented
1000	5.56 (2.54)	5.31 (2.80)	4.51
2000	4.95 (1.76)	4.51 (1.58)	8.92
5000	6.76 (1.97)	6.05 (1.75)	10.5

Table 11 gives both the raw percent incidence for each sensitivity designation across baseline and prevention packages as well as the resulting percent prevented across values of N. Notice that, although the package implementation is identical to that implemented with the linear model of the static partnership simulation shown in Table 10, the percent prevented effect is much lower here. Additionally, there appears to be a small upward trend in percent prevented with larger N. However, keeping in mind that each level of N was calibrated independently, it is probably beneficial not to read too much into this trend.

In fact over the next two tables it is important to emphasize not the direction of the effect across the sensitivity-analysis parameters, but rather the consistency with which this self-reinforcing model diverges from the static partnership model in estimating a much lower percent prevented, even though each calibration of the model is entirely independent. This suggests that lower package efficacy is indeed a property of networks that exhibit self-reinforcing behavior.

Table 12: Yearly mean percent Incidence of HIV (standard deviation) for baseline and intervention prevention packages as well as the corresponding percent prevented for each pair of packages, given across levels of δ.

δ	Baseline Percent Incidence	Package Percent Incidence	Percent Prevented
0	5.56 (2.54)	5.31 (2.80)	4.51
1	4.74 (2.77)	4.38 (2.70)	7.43

Table 12 demonstrates once again a consistently lower percent prevented across independent calibrations. As in the case of the sensitivity analysis across N, we do not purport to make any assumptions about the meaning of increased percent prevented with the addition of the parameter δ (a log-odds increase of 1 for individuals with a first contact degree-separation of 2).

Table 13: Yearly Incidence Across λ : Yearly mean percent Incidence of HIV (standard deviation) for baseline and intervention prevention packages as well as the corresponding percent prevented for each pair of packages, given across levels of λ.

λ	Baseline Percent Incidence	Package Percent Incidence	Percent Prevented
1	5.60(3.25)	5.25(3.30)	6.23
3	5.56 (2.54)	5.31 (2.80)	4.51
5	5.44(3.06)	5.06(2.99)	7.02

Finally, Table 13 displays the results across λ (mean time to random dissolution). Illustrated here is a case where there is no obvious trend across the sensitivity parameter. The percent prevented, however, are consistently at the same rough level as they were across other sensitivity parameters, indicating that our calibration efforts yielded similar-behaving simulations independently.

By far the most surprising result is that our flagship intervention of 50% ART acceptance, 50% PREP acceptance, 15% reduction in CAI, and 50% reduction in non-testers is not nearly as effective under this model. Where previously our Percent Prevented under the static partnership model lay at 34%, we see here a range of roughly 4.5% to 10.5%. It is also worth noting that the standard deviation of our percent infected is roughly twice that of the static partnership model.

Because this deviation from our original static partnership model was so great, we went to additional lengths to verify that this much lowered percent prevented was a model-dependent

feature rather result than of a single static partnership calibration or worse yet, a programming error. To this end we created a simple random-contact model calibrated only by incidence and constructed a table showing the raw intervention results across model types. This random contact model only features parameter γ_0, such that contacts between all pairs of individuals have exactly the same probability of occurring $\left(\frac{\exp \gamma_0}{1+\exp \gamma_0} \right)$.

Table 14: Yearly mean percent Incidence of HIV (standard deviation) for baseline and intervention prevention packages as well as the corresponding percent prevented for each pair of packages, shown across models.

Model	Baseline Percent Incidence	Package Percent Incidence	Percent Prevented
Static Partnership	5.27(0.75)	3.70(0.64)	29.8
Self-Reinforcing	5.56 (2.54)	5.31 (2.80)	4.51
Random-Contact	5.82(0.64)	4.35(0.62)	25.2

Table 14 shows incidences and percent-prevented across our three models. The first model shown is the static-partnership model. Note here that, because these are the raw values, they differ from the values produced by our linear model (shown in Table 10). The second model is our self-reinforcing model and the third is the random-contact, identical log-odds model. We can clearly see that the random-contact model is much more similar in percent-prevented to the static-partnership model than the self-reinforcing model. These models are so complicated, however, that we don't make any assumptions about why the percent-prevented of the random-contact model lies between that of the self-reinforcing model and the static partnership model.

So what we have seen is that the self-reinforcing model assumptions dramatically change the expected efficacy of interventions in an MSM population. As you know, our static-partnership model featured unchanging partnerships over the 5-year period and did not assume that a contact between individuals predicted future contacts. In this way it was similar to the random-contact network, even though it had many more complex layers to each relationship. Of course, it is largely unknown to what degree sexual contact solidifies a relationship in this MSM population, and thus it should not be taken for granted that our self-reinforcing model lies

quantitatively closer to the truth in its estimates of incidence and percent-prevented. However, our self-reinforcing model serves as precautionary warning that the assumption of "one-off" contacts, where individuals do not increase the likelihood of future sexual contacts through their first sexual contact, might result in an overestimation of intervention efficacy.

Chapter 7: Implications for the Design of Community Level HIV Prevention

7.1 Static Partnership Model: Power and Sample Size Considerations

Sample size and power calculations for HIV prevention trials rely on critical assumptions about HIV incidence, effect sizes of interventions, and participant attrition rates. They also rely on assumptions about the stochastic variation in numbers of incident infections. While binomial or Poisson models for the variance are often used, the assumptions that justify those models do not automatically apply in epidemic settings for several reasons. First, there is variation in both behavioral (e.g., numbers and types of sexual contacts) and biological (e.g., circumcision) risk factors that are not accounted for by these models. Second, infections are not independent events. A person is more likely to become infected if he/she is in the same sexual network as another infected person. Epidemics may burn through a community rapidly if infections are introduced into large inter-connected sexual networks, or alternatively, slowly if infections are introduced into small more isolated networks. The objective of this work is to understand and quantify sources of variation in the spread of HIV in communities induced by the complexities of overlapping sexual networks, and biological or behavioral heterogeneities in populations. Our approach utilizes agent-based models. We show how the approach can help design of community (or cluster) randomized HIV prevention trials [39, 40].

Sample size and design considerations of community randomized trials have received considerable attention in the literature [41-44]. Hayes and Bennet derive sample size formula for

the numbers of clusters and individuals per cluster in two arm trials [45]. Those formulas are expressed in terms of the between-cluster coefficient of variation, (i.e., the standard deviation of the incidence rates between clusters divided by the mean incidence rate averaged over communities). However, as noted by Hayes and Bennet, a critical problem is that adequate information on between community variations is seldom available at the design stage of trials. The lack of information on between community variation is an especially acute problem in HIV prevention because of the challenges in obtaining reliable estimates of HIV incidence rates. While data on HIV prevalence rates are more readily available, variation in current prevalence rates between communities is not a reliable surrogate for the variation in future HIV incidence rates between communities.

7.2 Static Partnership Model: Variation in Community-Randomized Trials

In this section we describe a framework for assessing the sources of variation in incidence of infection between communities in randomized community prevention trials, and then show its application to the simulation dataset for our static partnership model. Suppose a prevention trial consists of two arms. Each arm includes k communities, and each community consists of N uninfected persons and M infected persons. Random samples of n persons from the N uninfected persons in each community are enrolled in the study and followed for a fixed duration. We observe the number of incident infections that occur over the follow-up period, x_i, and the proportion who become infected, $\hat{p}_i = x_i / n$ among the enrolled samples of n uninfected persons in the i^{th} community. The number and proportion that become infected in the entire i^{th}

community of N uninfected persons are X_i and $\hat{P}_i = X_i / N$, respectively. While x_i and p_i are observed, X_i and P_i are not observed. Each of our simulations provides a single X_i.

We decompose the variance of $\hat{p}$ into three sources. To simplify notation we will drop the subscript i indexing the community in the following development. The first source arises from differences in community attributes that are associated with HIV incidence rates. These attributes may include distributions of numbers of sexual partners, circumcision rates, condom usage rates, availability of HIV counseling and frequencies of HIV testing in the community. Once again, we call the vector of these community attributes that affect HIV incidence, θ.

The second source arises from the *stochasticity of epidemics.* By this term we are referring to the notion that X and $\hat{P}$ in the community (and not just the study sample of n enrolled persons) will vary between communities even if all the attributes (θ) are the same for each community. The conditional variance of $\hat{P}$ given θ, $\text{var}\left(\hat{P} \mid \theta\right)$, quantifies this source of variation. A challenge is how to determine this variance. Naive models, such as the binomial (i.e., $\text{var}\left(\hat{P} \mid \theta\right) = E\left(\hat{P} \mid \theta\right)\left(1 - E\left(\hat{P} \mid \theta\right)\right)\Big/ N$) or the Poisson (i.e., $\text{var}\left(\hat{P} \mid \theta\right) = E\left(\hat{P} \mid \theta\right)$), do not automatically apply because underlying assumptions required to justify these models do not hold in complex epidemic settings where the virus is spread through sexual networks of heterogeneous populations. For example, some epidemics may be more explosive than others, if by chance, the virus is introduced into a large, highly inter-connected sexual network as opposed to an isolated network. Further, the individuals in the community are not identical but rather are heterogeneous with respect to risks for acquisition of HIV infection. As such, the conditional variance $\text{var}\left(\hat{P} \mid \theta\right)$ depends on a multitude of factors such as the size and overlap of

sexual networks and variation among individuals in risks for HIV acquisition. We use our agent-based models to aid in assessing $\text{var}\left(\hat{P}\,|\,\theta\right)$.

The third source of variation of $\hat{p}$ results from the random sampling of n study participants from among N persons in the community. We do not determine the infection status on all persons in the community but only a randomly selected sample and that introduces additional variation into $\hat{p}$. We formalize the three sources of variation discussed above as follows. First, we consider the variance of $\hat{p}$ conditional on the community attributes θ. We designate the expected proportion that becomes infected in a community with attributes θ as $E\left(\hat{P}\,|\,\theta\right) = P\left(\theta\right)$. Then,

$$\text{var}\left(\hat{p}|\,\theta\right) = E\,\text{var}\left(\hat{p}|\hat{P},\theta\right) + \text{var}\,E\left(\hat{p}|\hat{P},\theta\right)\ . \qquad (24)$$

If the n study participants are a random sample of the N persons in the community then it follows from results in survey sampling [46] that

$$\text{var}\left(\hat{p}|\hat{P},\theta\right) = f_1\frac{\hat{P}(1-\hat{P})}{n}\ , \qquad (25)$$

where $f_1 = \left(\frac{N-n}{N-1}\right)$ is a finite population correction factor. From equations 17 and 18 and $E\left(\hat{p}|\hat{P},\theta\right) = \hat{P}$ we have

$$\text{var}\left(\hat{p}\,|\,\theta\right) = f_1 \frac{P(\theta)\left(1-P(\theta)\right)}{n} + f_2 \,\text{var}\left(\hat{P}\,|\,\theta\right), \qquad (26)$$

where $f_2 = \frac{N(n-1)}{n(N-1)}$. Equation 19 decomposes the variance of $\hat{p}$ conditional on θ into two components. The first component on the right side of equation 19 accounts for variation from random sampling and the second component accounts for variation from the stochasticity of epidemics. If $n=N$ then $f_1=1$ and equation 19 reduces to $\text{var}\left(\hat{p}\,|\,\theta\right) = \text{var}\left(\hat{P}\,|\,\theta\right)$. If N is large and n is small ($n<<N$), then $f_1 \approx 1$ and $f_2 \approx 1$ and $\text{var}\left(\hat{p}|\,\theta\right)$ is approximately the sum of the usual binomial variance of a proportion and $\text{var}\left(\hat{P}\,|\,\theta\right)$ .We use our agent-based model simulations and subsequent weighted iterative model to obtain both the P(theta) and $\text{var}\left(\hat{P}\,|\,\theta\right)$ for this equation.

Figure 9: The variance of the proportion in the study sample that become infected, $var(\hat{p} \mid \theta)$, plotted versus fitted proportions (from equation 12). The variance is shown decomposed into the random sampling and stochastic epidemic components with sampling sizes *n*=50, 100 and 200.

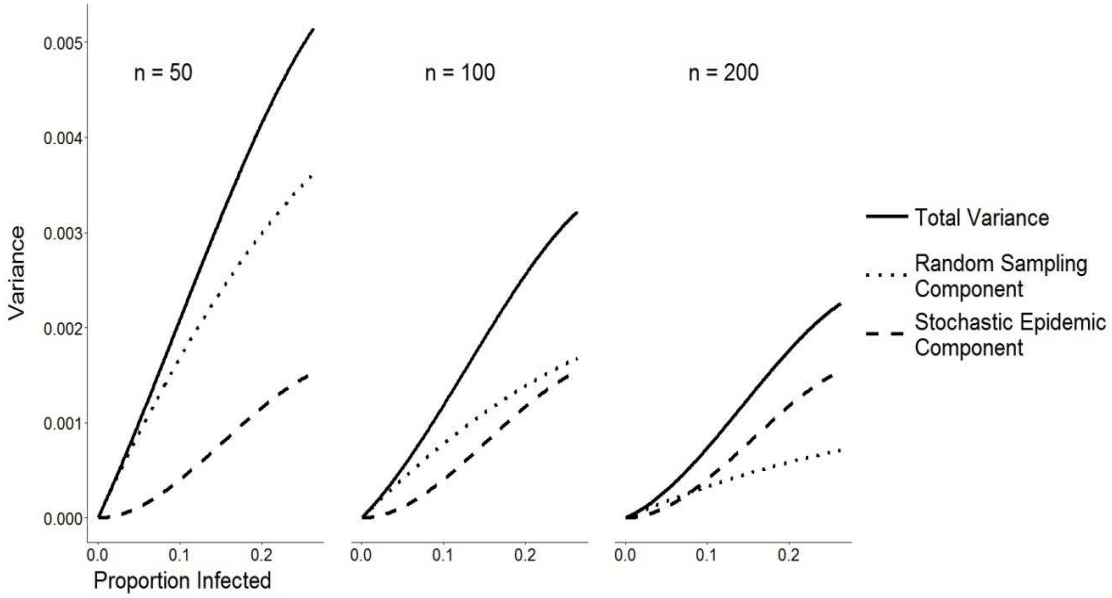


Figure 9 shows the decomposition of the variance $var(\hat{p} \mid \theta)$ (from equation 19) into random sampling component and the stochastic epidemic component with sample sizes *n*=50, 100 and 200. The figure illustrates that the stochastic epidemic component ($var(\hat{P} \mid \theta)$) can be an important source of the total variance of $var(\hat{p} \mid \theta)$.

7.3 Static Partnership Model: Power and Clusters

In this section, we consider the implications of our results for the design of community randomized trials. We consider testing the null hypothesis that the expected proportions infected

over the course of the trial in the control and intervention arms, called P_1 and P_2 respectively, are equal. The test statistic is based on the mean proportions infected among the k communities in each arm. We calculated the power under the alternative hypothesis that $(P_1 - P_2)/P_1 = \varepsilon$, where ε can be interpreted as the proportion of infections prevented by the intervention. We find (for a two sided test with type 1 error= α)

$$Power = P\left(Z > \frac{Z_{1-\frac{\alpha}{2}}\sqrt{\frac{2}{k}V_1} - (P_1\grave{o})}{\sqrt{\frac{1}{k}(V_1 + V_2)}} \right), \qquad (27)$$

where $V_1 = var(\hat{p}|\theta_1)$ for the control arm and $V_2 = var(\hat{p}|\theta_2)$ for the intervention arm; these variances are obtained by substituting var($\hat{P} \mid \theta$) from equation 19 into equation 27. When we solve for the number of clusters per arm necessary to obtain a power of 1-β we obtain

$$k = 1 + \left(\frac{Z_{1-\frac{\alpha}{2}}\sqrt{2V_1} + Z_{1-\beta}\sqrt{V_1 + V_2}}{P_1\grave{o}} \right)^2 . \qquad (28)$$

Figure 10: Power by Percent Prevented: Power versus percent of infections prevented (ε x 100) with α=.05, sample size *n*=50, 100 and 200 and *k*= 5 and 10.

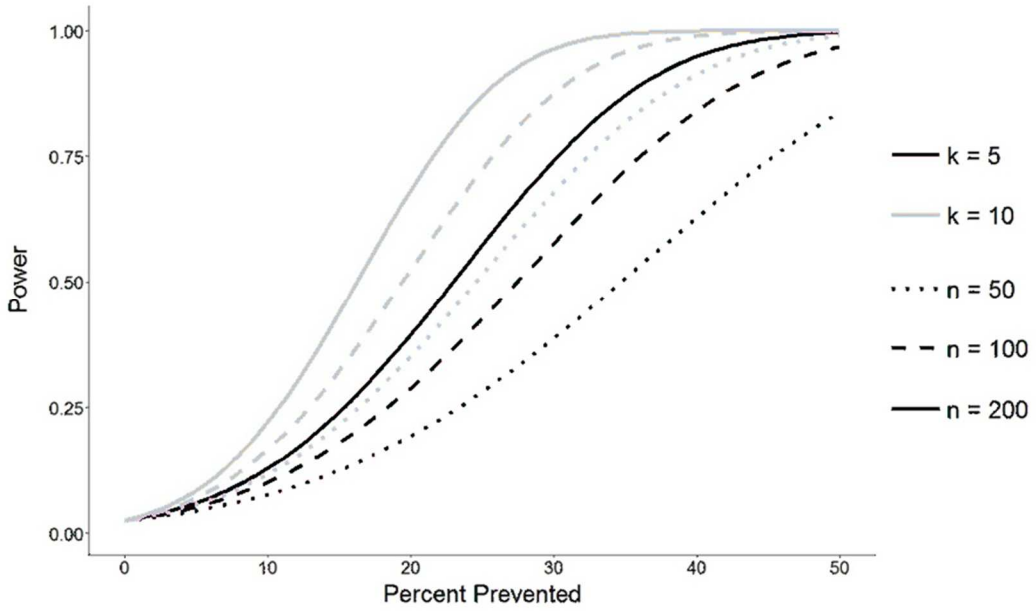


Figure 10 illustrates the relationship of power to *ε, k,* and sample size *n*. For example, the power to detect a significant effect with a baseline incidence of P_1=0.264 in the control arm, a true effect size ε =0.35, sample size *n*=200 and α=0.05 for *k*=5 and 10 are .87 and .99, respectively.

Figure 11: Clusters by Number Sampled: Number of clusters per arm (*k*) needed to obtain 90% power versus sample size *n* with α=.05 for percent infections prevented (ε x 100) = 25%, 35% and 50%

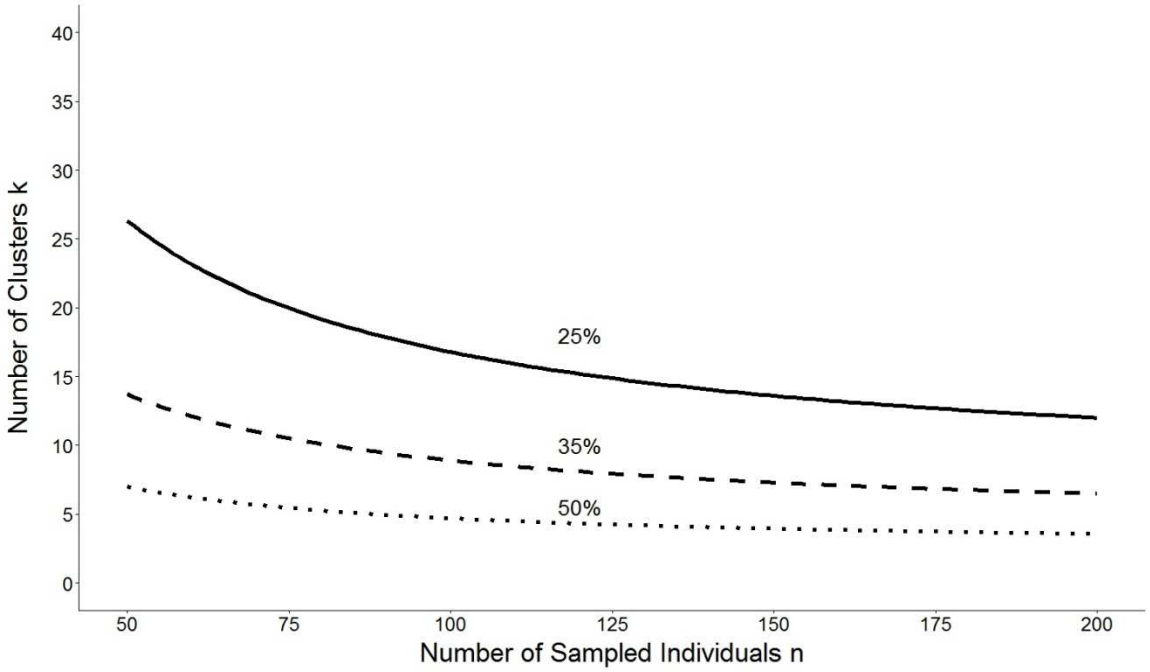


Figure 11 shows the numbers of communities per arm (*k*) versus sample size (*n*) to detect various effect sizes with power of .90; for example, the number of clusters per arm needed to detect effect sizes of ε =0.35 and ε =0.50 are 9 and 5, respectively.

We should note that we do not tabulate the expected power and required k-clusters for our self-reinforcing model. This is due to the fact that our intervention of interest (50% ART acceptance, 50% PREP acceptance, 15% reduction in CAI, and 50% decrease in nontesters) surprised us with a much lower effect size than we saw in our static-partnership model. This, combined with the much higher standard deviations predicted by the self-reinforcing model,

ensures that power will never be adequately achieved with that particular intervention and reasonable cluster sizes (less than 256).

Chapter 8: Discussion

8.1 Review

An objective of this work was to assess the stochastic variation of epidemics induced by sexual networks and heterogeneities in populations. Our approach was based on simulations of two unique agent-based models. We created a database of simulation results and used the simulated data to jointly model the mean and variance of the incidence of infection. We show how those results can be used to inform sample size and power calculations for community randomized HIV prevention trials. Failure to account for variation induced by risk factor heterogeneities and sexual networks in populations can lead to underpowered trials.

We implemented a simple static-edge agent-based model for our original work, and then augmented this transmission model with our self-reinforcing model of relationship development. This work has been published in both Statistics in Medicine [47] and PLOS One [48]. In comparing the intervention efficacy of the two models, we noted that there can exist a large disparity in percent-prevented estimates depending upon the model for sexual netowrks and contacts. We saw that our self-reinforcing model predicted an appreciably lower percent prevented.

The computation requirements for running large scale agent-based models can be enormous. In our models, because every individual had the potential for contact with every other

individual, N^2 =1,000,000 Bernoulli trials were performed each day , and over 5 years 1.825 x 10^9 Bernoulli trials were simulated for each simulation in both the static and self-reinforcing model.

This computational demand makes the calibration process difficult. For our static model we employed simplistic grid search over a limited parameter space. For our self-reinforcing model we implemented a recent algorithm from multi-objective optimization literature and calibrated over a series of 20000 iterations. Due to the huge number of iterations required, we implemented our agent-based models in R with the aid of the Hoffman2 cluster for both calibration procedures.

There are a number of limitations to our results and approaches that we hope to address in the future. The first limitation is that our numerical results on power in community randomized trails did not account for additional variation in baseline community attributes which could potentially be important source of variability. While matched designs where communities are matched on key attributes could help minimize that source of variation, it is very unlikely that perfect community matches could be achieved across all key baseline community attributes.

The second obvious limitation is the homogeneous behavior of the main partnerships in the self-reinforcing model. Main partnerships are reached at identical log-odds thresholds across individuals, while this threshold might be different for every pair of individuals. In addition choosing the log-odds as a threshold for main partnership formation was a subjective decision. As an alternative, the number of recent contacts could be used as a threshold, and this threshold itself could vary with other covariates (Age, ethnicity, or SES matching).

Another limitation lies in the calibration procedure. Each calibration across different values of fixed parameters N, λ, and δ requires an independent calibration using the Rolling Tide algorithm. It would be preferable to incorporate information from prior calibrations into calibrations of future models with additional parameters, or new values of these same parameters. However, there is no obvious way to incorporate these previous calibrations, as each parameter addition (or new parameter value of a fixed parameter) has an unknown effect on coarse network statistic output. It may be possible, however, to narrow the parameter calibration windows using previous calibration data. In our experience, we found that multiple, independent calibrations gave consistent and similar results, and thus the need for this improved calibration stems merely from its high computational demands.

A key assumption of the self-reinforcing model is that prior contacts influence the strength of the current relationship. While we applied the idea to sexual relationships in the context of HIV transmission, the self-reinforcing idea may have applicability to other social relationships. Social relationships may strengthen or weaken based on previous patterns of contacts. We considered one way of modeling the strength of relationships based on the past history of contacts. However, other models forms could be considered including adding linear, quadratic and higher order terms.

In addition, advances in HIV biomedical interventions might indicate that some of our assumptions regarding adherence should soon change. Given prophylactic measures that require less daily commitment, such as monthly injections, adherence levels might increase significantly. This would result in higher efficacy for PREP and is a variation worth exploring in future simulations.

8.2 Future Work

Uncertainty in Agent and Community Attributes

One issue in the work presented here is that we have taken population parameters that very well might vary among the different communities that we propose to simulate, and have used them as fixed parameters in the distributions from which we draw community properties. This most likely is not an accurate reflection of reality. For example, one could imagine that six-month partner distributions might vary considerable from one community to the next. A simple solution would be to introduce a "prior" distribution on this parameter, such that each community might have a different underlying distribution.

Unfortunately, such a method might overlook the correlation that would accompany such variation. A simple example of such a correlation might be the link between the sexual contact frequency parameters and the initial HIV prevalence, where higher frequency of contact should, in theory, lead to higher prevalence rates. If one simply drew the HIV prevalence and sexual contact frequency from two disparate distributions, one would actually be introducing more variability into the system than actually exists. This is an interesting conundrum, since often in agent-based modeling one imagines that there is always more variability than accounted for by the simple system of agent parameters.

Main Partnership Variation

In order to incorporate heterogeneity in main partnership formation as well as exclusivity, one of two things are required. Either, first, additional more detailed information relevant to South-African MSM regarding main partnership behaviors (such as heterogeneity of frequency of contact and exclusivity) must be made known. Or, second, this heterogeneity must be re-parameterized according to some variable we currently simulate (such as our activity attribute). One could assume that the more active someone is the weaker their tendency toward exclusivity or the higher their log-odds requirement for main partnership formation We have not included this re-parameterization up to this point specifically because we do not wish to create an over-dependency on this activity attribute, which is already incorporated into our baseline log-odds and dissolution parameter.

Refining Calibration

As we have stated, the calibration process would benefit from incorporation of prior calibration knowledge. If, however, one started with narrower calibration windows around previously calibrated members of the Pareto frontier, it would be necessary to incorporate a window-widening calibration procedure. A simple way to do this is to widen the parameter search window by a fixed amount whenever a member of the Pareto frontier appears within a predetermined width of the window-edge. Our primary efforts in this work, however, relied

upon the incorporation of a ready-made and simulation-tested algorithm, so that we would not have to perform these algorithm-quality tested ourselves.

Sampling Methodologies

One aspect of our power and cluster analyses that warrants further consideration is the assumption of random sampling. Such a thing will not occur in practice. Respondent-Driven Sampling is one option that can be used. RDS is similar to snowball sampling in that patient referrals are used to progressively sample deeper into the hidden population [49]. Our simulation already offers a solid framework for performing this sampling methodology. Those in recent testing or enrolled in an intervention of ART or PREP can be considered seeds and refer recent partners according to some probability distribution, and the variance can be empirically sampled from our simulation results. Of course, the variability assessed from respondent-driven sampling can be much greater than that from random sampling [50], so it will be interesting to see the effect on power and cluster estimation.

Another option is venue-based sampling. This would require more effort to simulate, but offers perhaps a more realistic approach to obtaining an unbiased estimate than respondent-driven sampling [51]. We recognize the important issues and caveats associated with these alternative sampling methods. Agent-based modeling offers an approach to help further quantify these issues.

8.3 Closing Thoughts

An important issues with agent-based models is the question of verification. Will we ever have enough data to verify completely all MSM distributions? Sexual network data is exceedingly difficult to obtain, and that which we do obtain is not complete network data. We do, however, have macroscopic trends and statistics, and can at least corroborate these large effects. So we know that our models produce these macroscopic epidemic trends because we've calibrated them to do so. But the work we have done produces a cautionary warning: that the variability we produce from community to community under both models far exceeds that estimated with a binomial variance assumption. The fact that both models support a higher variance suggests that the true variance is higher than a simplistic binomial model. It is difficult to say, however, whether we have overestimated the true variance. Usually we build all the variability into a system that we can account for at the time, but there are always things to miss, and consequently we often tend to underestimate the total variability.

We are never going to run enough clinical trials among MSM in South Africa to obtain a true variance between communities, so it is important that we prod our models further. This means exploring more of the parameter and model spaces. One important thing to note is that both of our models are somewhat unique in their approaches and other agent-based models have taken other routes in HIV epidemic modeling. We would stress that this is not a weakness, but rather a strength in that agent-based models, especially in the context of the HIV epidemic among hard-to-reach MSM, which must make many assumptions and simplifications in order to reach a useable simulation. It is precisely the lack of verification that necessitates that a multitude of approaches be examined. Agent-based modeling addresses exploratory questions

that cannot be answered through any other method, short of actually carrying out large-scale interventions. We conclude that agent-based models offer useful estimates and bounds on novel scenarios that can inform policy and influence the design of future HIV prevention trials.

APPENDIX

Derivation of " $h(b_1,b_2)$ for delta-method:

$$h(b_1,b_2) == 1-\frac{p_1}{p_2} = 1-\frac{\frac{exp(b_1)}{1+exp(b_1)}}{\frac{exp(b_2)}{1+exp(b_2)}}$$

$$" \ h(b_1,b_2) = \begin{pmatrix} \frac{\partial h}{\partial b_1} \\ \frac{\partial h}{\partial b_2} \end{pmatrix}$$

$$\frac{\partial h}{\partial b_1} = -\frac{1}{\frac{exp(b_2)}{1+exp(b_2)}} * \left(\frac{exp(b_1)}{1+exp(b_1)} - \frac{exp(b_1)}{(1+exp(b_1))^2} * exp(b_1) \right)$$

$$= -\frac{1}{\frac{exp(b_2)}{1+exp(b_2)}} * \left(\frac{exp(b_1)}{1+exp(b_1)} - \left(\frac{exp(b_1)}{1+exp(b_1)} \right)^2 \right) = -\frac{1}{p_2}\left(p_1 - p_1^2 \right) = -\frac{p_1(1-p_1)}{p_2}$$

$$\frac{\partial h}{\partial b_2} = -\frac{exp(b_1)}{1+exp(b_1)} * \frac{-1}{\left(\frac{exp(b_2)}{1+\exp(b_2)} \right)^2} * \left(\frac{exp(b_2)}{1+exp(b_2)} - \frac{exp(b_2)}{(1+exp(b_2))^2} * exp(b_2) \right)$$

$$= \frac{\frac{exp(b_1)}{1+exp(b_1)}}{\left(\frac{exp(b_2)}{1+exp(b_2)} \right)^2} * \left(\frac{exp(b_2)}{1+exp(b_2)} - \left(\frac{exp(b_2)}{1+exp(b_2)} \right)^2 \right) = \frac{p_1}{p_2^2}\left(p_2 - p_2^2 \right) = \frac{p_1(1-p_2)}{p_2}$$

REFERENCES

1. Sullivan , Patrick. Sibanye Health Project (MP3). Retrieved October 18, 2013 from http://prismhealth.us/research/active/sibanye-health-project-mp3/

2. Baral, Stefan, Earl Burrell, Andrew Scheibe, Ben Brown, Chris Beyrer, and Linda-Gail Bekker. "HIV risk and associations of HIV infection among men who have sex with men in peri-urban Cape Town, South Africa." *BMC Public health* 11, no. 1 (2011): 766.

3. Bekker, Linda-Gail, Chris Beyrer, and Thomas C. Quinn. "Behavioral and biomedical combination strategies for HIV prevention." *Cold Spring Harbor Perspectives in Medicine* 2, no. 8 (2012): a007435.

4. Sullivan, Patrick S., Alex Carballo-Diéguez, Thomas Coates, Steven M. Goodreau, Ian McGowan, Eduard J. Sanders, Adrian Smith, Prabuddhagopal Goswami, and Jorge Sanchez. "Successes and challenges of HIV prevention in men who have sex with men." *The Lancet* 380, no. 9839 (2012): 388-399.

5. Cohen, Myron S., Ying Q. Chen, Marybeth McCauley, Theresa Gamble, Mina C. Hosseinipour, Nagalingeswaran Kumarasamy, James G. Hakim et al. "Prevention of HIV-1 infection with early antiretroviral therapy." *New England Journal of Medicine* 365, no. 6 (2011): 493-505.

6. Grant, Robert M., Javier R. Lama, Peter L. Anderson, Vanessa McMahan, Albert Y. Liu, Lorena Vargas, Pedro Goicochea et al. "Preexposure chemoprophylaxis for HIV prevention in men who have sex with men." *New England Journal of Medicine* 363, no. 27 (2010): 2587-2599.

7. Lagakos, Stephen W., and Alicia R. Gable. "Challenges to HIV prevention—seeking effective measures in the absence of a vaccine." *New England Journal of Medicine* 358, no. 15 (2008): 1543-1545.

8. Gable, Alicia R., and Stephen W. Lagakos, eds. *Methodological Challenges in Biomedical HIV Prevention Trials.* National Academies Press, 2008.

9. Deb, Kalyanmoy. *Multi-objective optimization using evolutionary algorithms.* Vol. 16. John Wiley & Sons, 2001.

10. Jin, Yaochu, ed. Multi-objective machine learning. Vol. 16. *Springer Science & Business Media*, 2006.

11. Soper, H. E. "The interpretation of periodicity in disease prevalence." *Journal of the Royal Statistical Society* (1929): 34-73.

12. Abbey, Helen. "An examination of the Reed-Frost theory of epidemics." *Human Biology* 24, no. 3 (1952): 201.

13. Epstein, Joshua M. "Modelling to contain pandemics." *Nature* 460, no. 7256 (2009): 687-687.

14. Bonabeau, Eric. "Agent-based modeling: Methods and techniques for simulating human systems." *Proceedings of the National Academy of Sciences* 99, no. suppl 3 (2002): 7280-7287.

15. Epstein, Joshua M. "Agent-based computational models and generative social science." *Generative Social Science: Studies in Agent-Based Computational Modeling* 4, no. 5 (1999): 4-46.

16. Auchincloss, Amy H., and Ana V. Diez Roux. "A new tool for epidemiology: the usefulness of dynamic-agent models in understanding place effects on health." *American Journal of Epidemiology* 168, no. 1 (2008): 1-8.

17. Halloran, M. Elizabeth, Ira M. Longini, Azhar Nizam, and Yang Yang. "Containing bioterrorist smallpox." *Science* 298, no. 5597 (2002): 1428-1432.

18. Longini, Ira M., M. Elizabeth Halloran, Azhar Nizam, Yang Yang, Shufu Xu, Donald S. Burke, Derek AT Cummings, and Joshua M. Epstein. "Containing a large bioterrorist smallpox attack: a computer simulation approach." *International Journal of Infectious Diseases* 11, no. 2 (2007): 98-108.

19. Yang, Yang, Jonathan D. Sugimoto, M. Elizabeth Halloran, Nicole E. Basta, Dennis L. Chao, Laura Matrajt, Gail Potter, Eben Kenah, and Ira M. Longini. "The transmissibility and control of pandemic influenza A (H1N1) virus." *Science* 326, no. 5953 (2009): 729-733.

20. Goyal, Ravi, Rui Wang, and Victor DeGruttola. "Network epidemic models: assumptions and interpretations." *Clinical Infectious Diseases* (2012): cis388.

21. Epstein, Joshua M. "Agent-based computational models and generative social science." *Generative Social Science: Studies in Agent-Based Computational Modeling* 4, no. 5 (1999): 4-46.

22. Kermack, William O., and Anderson G. McKendrick. "A contribution to the mathematical theory of epidemics." *In Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 115, no. 772, pp. 700-721. The Royal Society, 1927.

23. Rahmandad, Hazhir, and John Sterman. "Heterogeneity and network structure in the dynamics of diffusion: Comparing agent-based and differential equation models." *Management Science* 54, no. 5 (2008): 998-1014.

24. Axtell, Robert L., and Joshua M. Epstein. "Agent-based modeling: understanding our creations." *The Bulletin of the Santa Fe Institute* 9, no. 2 (1994): 28-32.

25. Holland, Paul W., and Samuel Leinhardt. "An exponential family of probability distributions for directed graphs." *Journal of the American Statistical Association* 76, no. 373 (1981): 33-50.

26. Fienberg, Stephen E., Michael M. Meyer, and Stanley S. Wasserman. "Statistical analysis of multiple sociometric relations." *Journal of the American Statistical Association* 80, no. 389 (1985): 51-67.

27. Frank, Ove, and David Strauss. "Markov graphs." *Journal of the American Statistical Association* 81, no. 395 (1986): 832-842.

28. Hunter, David R., Mark S. Handcock, Carter T. Butts, Steven M. Goodreau, and Martina Morris. "ergm: A package to fit, simulate and diagnose exponential-family models for networks." *Journal of Statistical Software* 24, no. 3 (2008): nihpa54860.

29. Krivitsky, Pavel N., and Steven M. Goodreau. "STERGM-Separable Temporal ERGMs for modeling discrete relational dynamics with statnet." (2014).

30. Goodreau, Steven M., L. Pedro Goicochea, and Jorge Sanchez. "Sexual role and transmission of HIV Type 1 among men who have sex with men, in Peru." *Journal of Infectious Diseases* 191, no. Supplement 1 (2005): S147-S158.

31. Perry, Patrick O., and Patrick J. Wolfe. "Point process modelling for directed interaction networks." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75, no. 5 (2013): 821-849.

32. Rosenberg, Eli S., Patrick S. Sullivan, Elizabeth A. DiNenno, Laura F. Salazar, and Travis H. Sanchez. "Number of casual male sexual partners and associated factors among men who have sex with men: Results from the National HIV Behavioral Surveillance system." *BMC Public Health* 11, no. 1 (2011): 189.

33. Das, Sanjoy, and Bijaya K. Panigrahi. "Multi-Objective Evolutionary Algorithms." *Encyclopedia of Artificial Intelligence* 3 (2009): 1145-1151.

34. Tan, Kay Chen, Tong Heng Lee, and Eik Fun Khor. "Evolutionary algorithms for multi-objective optimization: performance assessments and comparisons." *Artificial Intelligence Review* 17, no. 4 (2002): 251-290.

35. Fieldsend, J., and R. Everson. "The Rolling Tide Evolutionary Algorithm: A Multi-Objective Optimiser for Noisy Optimisation Problems." (2013): 1-1.

36. Pareto, Vilfredo. *Manual of Political Economy: a Critical and Variorum Edition.* Oxford University Press, 2014.

37. Villa, Céline, Eric Lozinguez, and Raphaël Labayrade. "Multi-objective Optimization under Uncertain Objectives: Application to Engineering Design Problem." In Evolutionary Multi-Criterion Optimization, pp. 796-810. *Springer Berlin Heidelberg*, 2013.

38. Marx, Brian D. "Iteratively reweighted partial least squares estimation for generalized linear regression." *Technometrics* 38, no. 4 (1996): 374-381.

39. Hayes, Richard and L. Moulton. *Cluster Randomized Trials.* Chapman and Hall/CRC press: London, 2009

40. Davis, William W. "Design and Analysis of Cluster Randomization Trials in Health Research." *Journal of the American Statistical Association* 96, no. 456 (2001): 1529-1529.

41. Gail, Mitchell H., David P. Byar, Terry F. Pechacek, Donald K. Corle, and COMMIT Study Group. "Aspects of statistical design for the Community Intervention Trial for Smoking Cessation (COMMIT)." *Controlled clinical trials* 13, no. 1 (1992): 6-21.

42. Donner, Allan, Nicholas Birkett, and Carol Buck. "Randomization by cluster sample size requirements and analysis." *American Journal of Epidemiology* 114, no. 6 (1981): 906-914.

43. Koepsell, Thomas D., Donald C. Martin, Paula H. Diehr, Bruce M. Psaty, Edward H. Wagner, Edward B. Perrin, and Allen Cheadle. "Data analysis and sample size issues in evaluations of community-based health promotion and disease prevention programs: a mixed-model analysis of variance approach." *Journal of Clinical epidemiology* 44, no. 7 (1991): 701-713.

44. Hsieh, F. Y. "Sample size formulae for intervention studies with the cluster as unit of randomization." *Statistics in Medicine* 7, no. 11 (1988): 1195-1201.

45. Hayes, R. J., and S. Bennett. "Simple sample size calculation for cluster-randomized trials." *International Journal of Epidemiology* 28, no. 2 (1999): 319-326.

46. Cochran, William G. Sampling techniques. John Wiley & Sons, 2007.

47. Boren, David, Patrick S. Sullivan, Chris Beyrer, Stefan D. Baral, Linda, Gail Bekker, and Ron Brookmeyer. "Stochastic variation in network epidemic models: implications for the design of community level HIV prevention trials." *Statistics in Medicine* 33, no. 22 (2014): 3894-3904.

48. Brookmeyer, Ron, David Boren, Stefan D. Baral, Linda-Gail Bekker, Nancy Phaswana-Mafuya, Chris Beyrer, and Patrick S. Sullivan. "Combination HIV prevention among MSM in South Africa: results from agent-based modeling." *PloS one* 9, no. 11 (2014): e112668.

49. Magnani, Robert, Keith Sabin, Tobi Saidel, and Douglas Heckathorn. "Review of sampling hard-to-reach and hidden populations for HIV surveillance." *Aids* 19 (2005): S67-S72.

50. Salganik, Matthew J. "Variance estimation, design effects, and sample size calculations for respondent-driven sampling." *Journal of Urban Health* 83, no. 1 (2006): 98-112.

51. Muhib, Farzana B., Lillian S. Lin, Ann Stueve, Robin L. Miller, Wesley L. Ford, Wayne D. Johnson, Philip J. Smith, and Community Intervention Trial for Youth Study Team. "A venue-based method for sampling hard-to-reach populations." *Public Health Reports* 116, no. Suppl 1 (2001): 216.