Seeding a Simple Contagion*

Evan Sadler†

October 29, 2024

Abstract

I propose a method for selecting seeds to maximize contagion. First, fit a random graph model using a coarse categorization of individuals. Next, compute a *seed multiplier* for each category—this is the average number of new infections a seed generates. Finally, seed the category with the highest multiplier. Relative to the most common methods, my approach requires far less granular data, and it consumes less computing power—the problem scales with the number of categories, not the number of individuals. I validate the methodology through simulations using real network data.

1 Introduction

Imagine that you aim to impart information to a group of people using a small budget. While you can only inform a few individuals directly, these seeds may share what they learn

*I dedicate this paper to my third son, Norbert Cnut Sadler, who was born just as I started work on the revision—while this did occasion some delay, Bertie has been as accommodating in his sleep habits as one could ever hope. I am especially grateful to Mohammad Akbarpour and participants at the 2021 Retreat on Information, Networks, and Social Economics (RINSE) for comments and suggestions. I also thank Yeon-Koo Che, Navin Kartik, and seminar participants at Bocconi, Northwestern Kellogg, and Columbia for helpful feedback. The National Science Foundation (Grant SES-2018068) has generously supported the research in this paper. Finally, I thank Palaash Bhargava for excellent research assistance.

†Columbia University – es3668@columbia.edu

with others, so it pays to be choosy. How should you choose? Although targeted seeding has a long history in fields ranging from marketing to development economics, practitioners face persistent challenges. Collecting detailed network data is often infeasible or impractical, making it hard to implement many targeting strategies. What can we do with less granular data? Perhaps we have basic demographics and can estimate summary statistics to describe connections across different groups. For instance, you might track down a few random men and women and ask them: How many friends do you have, how many are men, and how many are women? This paper studies how to effectively use such limited information.

I propose a three-step method. First, fit a random graph model using network data—this involves sorting individuals into *types* based on observed attributes and estimating the average number of connections between individuals of different types. Next, compute a *seed multiplier* for each type, giving the number of new infections we should expect a seed to generate. Finally, seed the types that generate more infections. This method adapts to the data at hand—we can fit finer or coarser network models—and requires only low dimensional summary statistics to describe the network. Simulations suggest that we can obtain meaningful improvements over random seeding at low cost.

Our seed multipliers derive from a simple contagion model. We initially expose a set of seeds, each of whom gets infected with some probability. Infected individuals expose all of their neighbors in turn, spreading the contagion. I generate the network using a multi-type configuration model, which essentially takes a uniform random draw among graphs that match particular summary statistics. Each individual has one of finitely many types, which determine both susceptibility to the contagion and the distribution of connections. Our problem is to choose who to seed, based on observed types, to maximize the spread. The seed multiplier for type t measures how many additional infections we can expect on average if we expose one more type t individual, assuming a small seed budget. In the absence of more detailed network information, the optimal strategy targets the type with the highest

multiplier, and my results furnish explicit formulas for these seed multipliers.

To give a bit more detail, the extent of contagion hinges on the distribution of connected components among susceptible individuals—exposing one person in such a component eventually infects everyone in that component. My analysis distinguishes the non-viral case, in which the percolated graph is relatively sparse and all components are small, from the viral case, in which a single seed could infect a large fraction of the population. I focus on the former because it is more relevant empirically, and there is more to gain from targeting.[1] In a sparse network, each seed likely falls in its own component of susceptible individuals, and the seed multiplier is simply the average component size in the percolated graph. A key insight is that, although the network structure depends on the entire degree distribution for each type, the average component size depends only on the first two moments.

In order to apply the model, one needs data to define and identify types, to estimate network parameters, and to estimate susceptibility to the contagion. In Section 5, I discuss two approaches that use aggregated relational data (Breza et al., 2020), highlighting existing statistical methods that allow joint estimation of types and network parameters. Following this discussion, I briefly address the viral case—though less relevant in practice, a subtle point emerges from the analysis. Earlier work notes that even random targeting is very likely to hit the giant component, leaving little room for gains on this margin (Akbarpour et al., 2020). However, targeted seeding can still improve our performance among more peripheral individuals, and here the best targets differ from those in the non-viral case. Well-connected individuals are likely part of the giant component and are almost sure to get exposed, so seeding them is redundant. Consequently, the best targets often have fewer connections because these individuals are more likely to lie in the periphery.

I next turn to simulations, validating the model using real network data from Banerjee

[1]Applications in development economics often confront little to no diffusion of beneficial technologies (Banerjee et al., 2013, 2019; Beaman et al., 2021). Even in online settings, viral contagion is the exception and not the rule, as most cascades are quite small (Leskovec et al., 2006; Goel et al., 2012, 2016).

et al. (2013). The data include network connections and demographic information for a collection of 75 rural Indian villages. I conduct two exercises using this data. First, for each village, I fit a simple network model with two types, men and women, and compare calculated multipliers to estimates based on 2000 simulations of the diffusion process—note that in every village, men on average have more connections than women. In order to cover both the non-viral case and the viral case, I adapt the formal framework to allow transmission that occurs with a probability less than one.[2] If infected individuals have a 10% chance of exposing each neighbor, the network model falls into the non-viral case for most villages, while increasing to 25% puts most villages in the viral case. In both cases, the model correctly predicts which group has a higher multiplier in nearly all village networks. Although point estimates differ from the computed multipliers—in particular, the model systematically predicts a higher gain from targeting than what the simulations suggest—these differences are in line with what we should expect from using large network limits to approximate small graphs.

In a second exercise, I randomly add links between individuals in the largest 10 villages to create a single network, and I fit a model with 8 types.[3] Using a transmission rate that puts us firmly in the non-viral case, I again compare computed multipliers with multipliers estimated through simulations. In addition to closely matching the predicted ranking of types, our point estimates are much closer to computed multipliers in this larger graph. Moreover, with a finer partition of types, the gains from targeting become more meaningful—seeding the optimal type generates 23% more infections per seed than seeding at random.

After the simulations, I discuss possible consequences of model misspecification and a comparison with alternative seeding strategies. If we model the network using too few types,

[2]This is equivalent to thinning out the degree distribution, deleting each link independently with some probability. I additionally introduce an adjustment to account for clustering. In real networks, individuals typically have shared neighbors, but this does not occur in the configuration model. Consequently, two individuals might both expose a common neighbor, leading to just one new infection rather than two. To adjust for this, I further thin the degree distribution based on the village clustering coefficients.

[3]Here, a type describes an individual's sex and caste.

this can produce biased estimates of the seed multipliers. Nevertheless, I show that as long as the network exhibits homophily with respect to the unobserved traits, we will underestimate the true extent of contagion. The primary advantage over alternative methods from computer science lies in the ease with which we can gather the relevant data. With detailed network information, we should expect alternatives to perform better, but type-based seeding offers a way to achieve targeting gains when detailed data are unavailable or too expensive. With a relatively small set of types, we should also realize significant computational savings.

Whether or not my exact methodology gains traction, this paper makes a conceptual contribution, shifting our focus to targeting at the level of groups rather than individuals. While seeding problems are a major research topic across many fields,[4] essentially all of the proposed strategies require identifying specific opinion leaders or central individuals, entailing similar challenges with data and computational complexity. Finding optimal targets requires detailed information on the underlying network, but in practice such data are prohibitively expensive and extremely noisy.[5] Targeting groups, rather than individuals, offers a way to obtain smaller but still meaningful gains in a much larger range of environments.

Related Work

Akbarpour et al. (2020) provide the nearest precursor to this paper, studying the same contagion process in a similar random graph model. Given the expense involved in gathering network data, the authors ask if we should rather follow a simple alternative: seed at random, but place a few more seeds. The authors measure performance using the expected fraction of a large population that gets infected after seeding a finite number of individuals—all that

[4]This includes not just economics (Banerjee et al., 2013; Beaman et al., 2021), but also computer science (Kempe et al., 2003; Chen et al., 2009), marketing (Hill et al., 2006; Iyengar et al., 2011), and healthcare (Kim et al., 2015).

[5]For instance, network studies in developing economies rely on expensive ground surveys to map the social graph (e.g. Banerjee et al., 2013; Beaman et al., 2021), which makes it difficult to scale up interventions, while data from online social networks can miss important offline connections and include spurious links.

matters for the analysis is whether a giant component exists and whether it gets seeded or not. Each random seed is an independent chance to hit the giant component—if the giant component contains a fraction $1-p$ of the population, and we seed k individuals, the probability we hit it is $1-p^k$. Using this metric, seeding a few more individuals at random is almost always more cost effective than careful targeting: the best possible targeting policy might achieve $1-p$ for sure, while randomly seeding k individuals yields $1-p$ with probability $1-p^k$ and 0 otherwise. For large enough k, additional seeds make essentially no impact—in examples, there is often little gain from seeding more than a dozen individuals. Moreover, by this metric, no policy achieves non-zero contagion in the non-viral case.

I provide a complementary analysis, allowing us to compare policies in the non-viral case where targeting can make a large difference in the number of infections we generate per seed. In the viral case, I offer a more subtle contribution. When a giant component exists, I take as given that all agents within it get infected—as Akbarpour et al. (2020) show, we do not need very many seeds to be almost certain of hitting it. I ask instead: beyond the giant component, what is the marginal impact of another seed? Although multipliers are typically quite low in this case, individuals in the giant component may be systematically different from those outside it, and targeting can help us reach those we would otherwise miss.

The computer science literature often refers to targeted seeding as "influence maximization." In a seminal contribution, Kempe et al. (2003) define the problem both for a simple contagion process, as studied in the present paper, and for the "linear threshold model," in which an individual gets infected only if a high enough fraction of neighbors are infected. While identifying an optimal set of seeds in a given graph is an NP-hard problem, the authors provide heuristic algorithms that achieve provable performance bounds. In large networks, even these heuristics are too slow, and later work has sought simpler algorithms based on local, rather global, influence measures. See Chen et al. (2013) for a survey of this literature. Recognizing the importance of data limitations, more recent papers study a planner who is

uncertain about the probability of transmission across each link, contributing new algorithms that achieve robust performance bounds (Chen et al., 2016; Wilder et al., 2017). However, these models still require perfect knowledge of the links along which transmission is possible, and the planner's computational problem still scales with the size of the network. By employing a random graph model, my analysis produces a much lower dimensional problem, simultaneously mitigating both issues.[6] Other authors have studied a natural heuristic that also uses far less information: seed a random *neighbor* of a randomly chosen individual (See, for instance, Seeman and Singer, 2013; Kumar and Sudhir, 2019; Chin et al., 2022). This leverages selection via the friendship paradox to target higher-degree individuals. In Section 8, I discuss how my approach compares with this method and how we can combine the two to obtain even better performance.

A large and growing body of empirical work demonstrates the usefulness of targeted seeding. In an influential example, Banerjee et al. (2013) gathered detailed network information on a set of 75 rural Indian villages, and a microfinance lending program subsequently targeted 43 of these for expansion. Selected village leaders attended information sessions and were asked to inform others in their community about the lending program. The authors then tracked microfinance participation over time. Their data support a simple contagion model, and participation was substantially higher when the selected leaders were more central. Beaman et al. (2021) similarly study the diffusion of an agricultural technology in 200 villages in Malawi. After collecting network information, the authors oversaw an RCT in which two seeds in each village received training in pit planting and crop residue management. In contrast with Banerjee et al. (2013), their data are more consistent with complex contagion—farmers are more likely to adopt new practices if multiple neighbors do so—but again, training more central farmers has a large positive effect on subsequent adoption.

[6]There is also work in computer science studying contagion in random graphs—these papers largely focus on properties of networks that enable large cascades (e.g. Watts, 2002).

In each of the above cases, the interventions relied on expensive surveys to map the social network in each village, which makes it hard to scale up.[7] Kim et al. (2015) and Banerjee et al. (2019) demonstrate effective heuristics that address this problem. In an effort to increase the use of multivitamin supplements and chlorinated water in Honduras, Kim et al. (2015) gave free samples and information to small seed sets in several villages. The authors assigned each village to one of three treatments: targeting at random, targeting based on degree centrality, and targeting nominated friends of random villagers. This last method attempts to leverage the friendship paradox to cheaply find more central individuals, and indeed this treatment substantially outperforms the other two in their data. In a similar study, Banerjee et al. (2019) directly ask Indian villagers: who is good at spreading information? In their experiment, targeting these nominated individuals with information about immunization camps led to significant gains in childhood immunization relative to random targeting.

Closer to the present paper, Breza et al. (2020) demonstrate a method to study network spillovers in which they represent the graph as a "latent distance model." The crucial simplification is that one can estimate the necessary parameters using aggregated relational data—this means asking a random set of individuals questions like "how many of your friends have trait x?" Given these answers, and a basic vector of covariates for all individuals, they estimate a model in which each link is drawn as a conditionally independent Bernoulli trial. The authors subsequently replicate findings in earlier studies that used far more detailed, and expensive, network data. In a more recent contribution, Alidaee et al. (2020) demonstrate a fast method for non-parametric estimation of network structure from aggregated relational data when the underlying network's adjacency matrix has low rank. My analysis applies a similar approach to diffusion and targeted seeding, as opposed to peer effects.

[7] Beaman et al. (2021) explicitly acknowledge this problem, and they explore the possibility of using information on geography to proxy for network centrality. Unfortunately, this method of targeting performed poorly in their data.

2 Contagion in a Random Network

A contagion spreads through social ties in a large (countably infinite) population.[8] Some individuals are initially exposed, while others get exposed if and only if a neighbor is infected. Each person has one of finitely many types in the set Θ, and $T \in \Delta(\Theta)$ is the distribution of types. I often abuse notation, using capital letters for both distributions and random variables that follow them. Upon exposure, a type t individual gets infected with probability $\alpha_t \in [0,1]$—this is a standard SIR model in which type t is susceptible with probability α_t.

A multi-type configuration model generates the network linking these individuals. Person i with type t has degree $d_i \sim D_t \in \Delta(\mathbb{N})$, and each neighbor has a type drawn independently according to $Z_t \in \Delta(\Theta)$. Define $p_t = \mathbb{P}(T = t)$, $\mu_t = \mathbb{E}[D_t]$, and $q_s^t = \mathbb{P}(Z_t = s)$. We require

$$p_t \mu_t q_s^t = p_s \mu_s q_t^s$$

for every pair $t, s \in \Theta$—the expected number of type s neighbors of type t individuals equals the expected number of type t neighbors of type s individuals. Intuitively, a configuration model takes a uniform random draw from the set of all graphs that are consistent with these summary statistics.[9] In large networks, the local structure closely approximates an infinite tree, which facilitates analysis using recursive formulas. In Sadler (2020), I motivate this model as the limit of a sequence of finite random graphs, and an interested reader can find the corresponding technical details there. In this paper, I work directly with the limit object.

A planner chooses individuals to initially expose. Our planner can observe each individual's type and target according to types. Seeding a fraction s_t of type t individuals means

[8] This could be anything that spreads from person to person—a piece of information, awareness of a new technology, or even a habit or social norm.

[9] More formally, consider a sequence of graphs with the number of nodes n approaching infinity, and generate each via a uniform random draw among those that match an exogenously given n-vector of degrees and types. As $n \to \infty$, we need to assume that the fraction of each type, the empirical degree distribution for each type, and the empirical neighbor type distributions converge appropriately.

exposing each such individual independently with probability s_t. Write $\mathbf{s} = (s_1, s_2, ..., s_\Theta)$ for the vector of seeded fractions. The total seeded fraction of the population is then

$$s := \mathbf{p} \cdot \mathbf{s} = \sum_{t \in \Theta} p_t s_t.$$

Similarly, write $\pi_t(\mathbf{s})$ for the fraction of type t individuals that are eventually infected, write $\boldsymbol{\pi}(\mathbf{s})$ for the corresponding vector, and write $\pi(\mathbf{s}) := \mathbf{p} \cdot \boldsymbol{\pi}(\mathbf{s})$ for the fraction of the total population that gets infected. The **seed multiplier** for type t is the marginal impact on π of an additional type t seed when s is small:

$$\beta_t := \lim_{\substack{\mathbf{s} \to \mathbf{0} \\ \mathbf{s} \gg \mathbf{0}}} \frac{1}{p_t} \frac{\partial \pi(\mathbf{s})}{\partial s_t}. \tag{1}$$

In this limit, we are taking the seeded fraction to zero while s_t remains strictly positive for each type t. The multiplier β_t measures the return from targeting type t individuals—each additional type t seed generates β_t additional infections on average.[10]

3 Diffusion in the Configuration Model

How many infections does a particular seed cause? A susceptible individual gets infected if and only if there is a path, passing only through other susceptible individuals, connecting her to a seed. Hence, each seed infects a connected component in a percolated graph: to answer our question, we must remove those who are not susceptible and study component sizes in the subgraph that remains.

Towards this end, Sadler (2020) shows that a *characteristic branching process* describes

[10]Implicit in the definition is a particular order of limits—we first take the size of the population to infinity while seeding a constant fraction of each type before taking the seeded fraction to zero. This ensures a deterministic outcome, though the assumption is stronger than necessary. All results hold in any limit model in which the number of seeds of each type grows without bound as the population does so.

the local structure of the configuration model, allowing us to assess any individual's chance of infection using a simple recursion. Imagine conducting a breadth first search of the graph starting from a random type t individual. The number of neighbors follows the distribution D_t, and each of these has a type drawn independently according to Z_t. Going forward, each type s individual has a number of additional neighbors drawn according to D'_s—the *forward distribution*—with

$$\mathbb{P}(D'_s = k) = \frac{(k+1) \cdot \mathbb{P}(D_s = k+1)}{\mu_s}, \tag{2}$$

and each of these draws a type according to Z_s. The forward distribution makes a standard adjustment for the friendship paradox. A vertex at the end of a random edge has a higher degree than a random vertex on average because a degree k vertex has k chances to show up. Hence, high-degree individuals are more likely to appear as neighbors, in proportion to their degrees. The forward distribution also removes the link through which an individual gets exposed. For convenience, I define the expected forward degree for type t:

$$\mu'_t = \mathbb{E}[D'_t] = \frac{1}{\mu_t}\mathbb{E}[D_t(D_t - 1)] = \frac{\mathrm{Var}[D_t]}{\mu_t} + \mu_t - 1. \tag{3}$$

The probability that an individual gets infected—and hence the fraction who get infected—is her susceptibility times her exposure probability. An individual gets exposed either if she is a seed or if a neighbor is infected. We can recursively compute the probability that a type t neighbor is infected, without our focal individual exposing her, as

$$\begin{aligned} y_t(\mathbf{s}) &= \alpha_t \left(s_t + (1 - s_t) \sum_{k=0}^{\infty} \mathbb{P}(D'_t = k) \left(1 - (1 - \mathbf{y}(\mathbf{s}) \cdot \mathbf{q}^t)^k\right) \right) \\ &= \alpha_t \left(1 - (1 - s_t) \frac{g'_t\left(1 - \mathbf{y}(\mathbf{s}) \cdot \mathbf{q}^t\right)}{\mu_t} \right). \end{aligned} \tag{4}$$

in which $g_t(x) := \mathbb{E}[x^{D_t}] = \sum_{k=0}^{\infty} \mathbb{P}(D_t = k) x^k$ is the probability generating function for the degree distribution D_t. A type t neighbor is susceptible with probability α_t, and with

probability s_t this neighbor is a seed. With probability $1 - s_t$, she requires another neighbor to expose her. The value $\mathbf{y}(\mathbf{s}) \cdot \mathbf{q}^t$ is the probability that any single neighbor is infected, so $1 - (1 - \mathbf{y}(\mathbf{s}) \cdot \mathbf{q}^t)^k$ is the probability that at least one out of k neighbors is infected. The probability that our focal individual gets infected, if she has type t, is then

$$\begin{aligned}\pi_t(\mathbf{s}) &= \alpha_t \left(s_t + (1 - s_t) \sum_{k=0}^{\infty} \mathbb{P}(D_t = k) \left(1 - (1 - \mathbf{y}(\mathbf{s}) \cdot \mathbf{q}^t)^k\right) \right) \\ &= \alpha_t \left(1 - (1 - s_t) g_t \left(1 - \mathbf{y}(\mathbf{s}) \cdot \mathbf{q}^t\right)\right). \end{aligned} \tag{5}$$

Our individual gets seeded with probability s_t, and a neighbor exposes her with probability $1 - g_t\left(1 - \mathbf{y}(\mathbf{s}) \cdot \mathbf{q}^t\right)$. Because the generating functions $\{g_t\}_{t \in \Theta}$ are convex, the system (4) has a unique solution $\mathbf{y}(\mathbf{s})$ whenever $\mathbf{s} \gg 0$, leading to a unique $\boldsymbol{\pi}(\mathbf{s})$ defined in (5).

The extent of diffusion depends crucially on how many susceptible individuals we find at each step moving away from a seed—computing this provides key intuition. Define

$$m_{ts} = \alpha_s q_s^t \mu_t, \quad \text{and} \quad m'_{ts} = \alpha_s q_s^t \mu'_t. \tag{6}$$

The value m_{ts} is the average number of susceptible type s individuals who are neighbors of a random type t individual—if type t seed gets infected, she exposes μ_t neighbors on average, a fraction q_s^t of whom are type s, and a fraction α_s of these are susceptible. Similarly, the value m'_{ts} is the average number of type s individuals who get infected in any subsequent step after we infect a type t neighbor. Write M and M' for the corresponding $\Theta \times \Theta$ matrices.

If we seed a type t individual, the expected number of infections after k steps is

$$\alpha_t \left(1 + \mathbf{e}_t^\intercal M \left(I + M' + ... + (M')^{k-1}\right) \mathbf{1}\right),$$

in which $\mathbf{e}_t$ is a coordinate vector with a 1 in entry t, and $\mathbf{1}$ is a vector of ones—the leading 1

corresponds to the seeded individual, the term with the matrix M covers infections one step away, the term with MM' covers infections two steps away, and so on. Taking $k \to \infty$, we have $M\left(I + M' + ... + (M')^{k-1}\right) \to M(I - M')^{-1}$ as long as the spectral radius $\rho(M')$ is less than one. If $\rho(M') > 1$, the series grows without bound. This distinguishes the **non-viral** case, in which $\rho(M') < 1$, from the **viral** case, in which $\rho(M') > 1$.[11] In the non-viral case, classic results on branching processes imply that the solution to (4) converges to $\mathbf{y} = \mathbf{0}$ as $\mathbf{s} \to \mathbf{0}$—a small set of seeds induces little contagion. In the viral case, the solution converges to a strictly positive $\boldsymbol{\zeta} \in [0,1]^{\Theta}$—a fraction ζ_t of type t individuals gets infected.[12]

4 Selecting a Small Set of Seeds: The Non-Viral Case

Recall the matrices M and M' from (6) with entries $m_{ts} = \alpha_s q_s^t \mu_t$ and $m'_{ts} = \alpha_s q_s^t \mu'_t$. Assume the spectral radius of M' is $\rho(M') < 1$. Substantively, this means that each infected neighbor causes in expectation less than one additional infection, and a small set of seeds cannot produce a large cascade. A first result tells us how to compute the seed multipliers.

Theorem 1. *In the non-viral case with $\rho(M') < 1$, the seed multiplier for type t is*

$$\beta_t = \alpha_t \left(1 + \mathbf{e}_t^{\intercal} M (I - M')^{-1} \mathbf{1}\right). \tag{7}$$

Proof. The proof of Theorem 1 entails differentiating (4) and (5) and simplifying the resulting expressions. See the Appendix for details. □

We can readily interpret the expression (7). The entries of M count the expected number of infections a seed induces among its neighbors of each type. As $(I - M')^{-1} = \sum_{k=0}^{\infty} (M')^k$, we see that the term in parentheses sums the expected number of infections among individuals

[11] The boundary case $\rho(M') = 1$ is similar to the non-viral case, but the seed multipliers are infinite.
[12] See Athreya and Ney (1972) for the relevant results on branching processes.

at each step away. A seed's marginal value is exactly the susceptibility α_t times one plus the expected number of additional infections. An important feature is that seed multipliers do not depend on fine details of the degree distributions. Recall from (3) that we can compute the expected forward degree μ'_t using the mean and variance of D_t. A type's expected forward degree is increasing in the variance of D_t—the friendship paradox drives this finding as high-degree individuals are overrepresented as neighbors in proportion to their degrees, and higher variance enhances this selection effect.

The multiplier values give a sense of how much we might gain from targeting, but all that ulimately matters for a targeting strategy is having the correct ranking. Equation (7) is simple enough to compute that one can readily perform a sensitivity analysis by varying the entries of M and M'—continuity ensures there is always at least some robustness to small errors in estimated parameters. In the special case in which all types have the same susceptibility, it is enough to compare the row sums of the matrix $M(I-M')^{-1}$ to determine if one type has a higher multiplier than another. As a consequence, any mismeasurement that proportionately inflates or deflates our estimates of μ_t and μ'_t has no impact on the estimated rankings—as long as missing or spurious links do not appear in our data more commonly for one type than another, we can expect to get the right ranking.

Corollary 1. *In the non-viral case with $\rho(M') < 1$, suppose that $\alpha_t = \alpha$ for each type t. The seed multiplier for type t is larger than that for type s if and only if the sum of entries in row t of the matrix $M(I-M')^{-1}$ is larger than the sum of entries in row s.*

Proof. This is immediate from Theorem 1. □

To gain some intuition for what makes an attractive target, it helps to look at a model with fewer parameters. Define

$$q_t^* := \frac{p_t \mu_t}{\sum_{s \in \Theta} p_s \mu_s}, \tag{8}$$

the fraction of all links that connect to type t individuals. If there were no bias in the interactions between types, we would have $\mathbf{q}^t = \mathbf{q}^*$ for each type $t \in \Theta$—I call $\mathbf{q}^*$ the *unbiased neighbor type distribution*. In a *simple homophily model*, the neighbor type distribution $\mathbf{q}^t$ is a mixture between the unbiased distribution $\mathbf{q}^*$ and a point mass on type t.

Definition 1. *In a* ***simple homophily model*** *with parameter $h \in [0,1]$, we have $\mathbf{q}^t = (1-h)\mathbf{q}^* + h\mathbf{e}_t$ for each type $t \in \Theta$.*

The parameter h is the inbreeding homophily. If $h = 0$, every type has the unbiased neighbor distribution $\mathbf{q}^*$. If $h = 1$, the different types are completely isolated from one another. In between, individuals may have neighbors of all types, but they have an own-type bias. Within this model, we can more easily compare types.

Proposition 1. *In the non-viral case with a simple homophily model, the seed multiplier for type t is*

$$\beta_t = \alpha_t \left(1 + \mu_t \frac{\alpha_t h + (1-h)\boldsymbol{\alpha}^\intercal (I - M')^{-1}\mathbf{q}^*}{1 - \alpha_t h \mu_t'}\right). \tag{9}$$

Proof. See Appendix. □

In a simple homophily model, a type's multiplier is increasing in its susceptibility, expected degree, and expected forward degree. If homophily is low, so $h \approx 0$, the forward degree plays no role, and we can rank types according to the product of susceptibility and expected degree $\alpha_t \mu_t$. As we increase h, the forward degree μ_t' becomes more important. This leads to subtler comparisons because the type with the highest degree need not have the highest forward degree. For instance, suppose one type always has degree 4 and another has degree 1 or 6 with equal probability, giving an expected degree of $3.5 < 4$. The first type has a forward degree of 3, but selection via the friendship paradox means the second type has an average forward degree of $4\frac{2}{7}$. While the first type is a better target when homophily is low, the second is better when homophily is high. If $\alpha_t h \mu_t'$ is very close to 1, the difference

in the corresponding multipliers can become arbitrarily large. In general, homophily favors types with higher variance in their connections.

Figure 1 illustrates potential gains from targeting in a simple example.[13] In this figure, there are two equally prevalent types 0 and 1 with susceptibilities $\alpha_0 = \alpha_1 = 0.5$, homophily $h = 0.75$, Poisson degree distributions, and average degrees $\mu_1 = 2\mu_0$—type 1 has twice as many neighbors on average as type 0. The lines plot seed multipliers β_0 (blue) and β_1 (red) as functions of μ_0. To better interpret this graph, imagine we are seeding 50 individuals in a population of 1000. If $(\mu_0, \mu_1) = (0.75, 1.5)$, then we can expect random seeding to generate 67 infections, while targeting the type with the highest multiplier should lead to 89 infections. Closer to the viral threshold, with $(\mu_0, \mu_1) = (0.9, 1.8)$, our 50 seeds should produce 114 infections if placed randomly versus 169 if targeted. Notice that as we move further to the right, past the viral threshold, the multiplier β_1 falls below β_0. This happens because high-degree individuals get exposed through the giant component with high probability, so seeding them becomes redundant—I discuss this in more detail in Section 6.

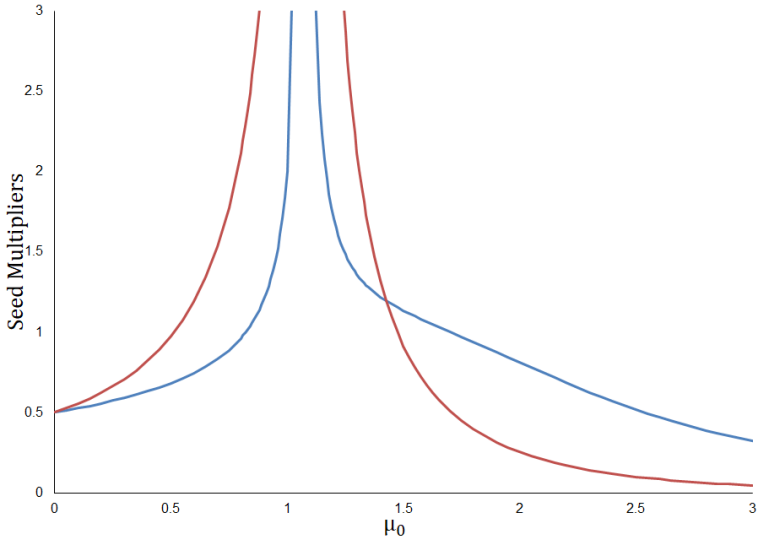

Figure 1: Multipliers β_0 (blue line) and β_1 (red line) as a function of μ_0.

[13]Note that past the peak, we are in the viral case. This portion of the figure is based on theoretical results in Section 6.

5 A Practical Roadmap

How might an applied economist or government agency use this method in practice? We need a way to (i) define and identify types, (ii) estimate network parameters, and (iii) estimate susceptibility. I discuss two broad approaches. One approach entails making assumptions about how observable attributes (e.g., sex, race, education level, religion, profession, etc.) map to types and then estimating parameters for these pre-defined types. Alternatively, one can assume types are latent and jointly estimate types and network parameters. Either way, it suffices to have aggregated relational data for a random sample of the population. Such data consist of a tuple $(a_i, d_{i1}, d_{i2}, ..., d_{iK})$ for each individual i in the sample, in which $a_i \in \{1, 2, ..., K\}$ represents the observable attributes of individual i, and d_{ik} is the number of i's neighbors with attribute k.[14] One also needs the attribute a_i for each individual i in the population of potential targets.

In the first approach, estimation is straightforward. A type $t \subseteq \{1, 2, ..., K\}$ is a predefined collection of attribute values, and we can use standard estimators:

$$\hat{\mu}_t = \frac{1}{n_t} \sum_{i\,:\,a_i \in t} \sum_{k=1}^{K} d_{ik}, \quad \hat{\mu}'_t = \frac{1}{\hat{\mu}_t n_t} \sum_{i\,:\,a_i \in t} \left(\sum_{k=1}^{K} d_{ik} \right)^2 - 1, \quad \hat{q}^t_s = \frac{\sum_{i\,:\,a_i \in t} \sum_{k \in s} d_{ik}}{\sum_{i\,:\,a_i \in t} \sum_{k=1}^{K} d_{ik}}.$$

With parameter estimates in hand, an analyst can compute the matrices M and M', apply the formula in Theorem 1, and seed individuals whose attributes correspond to the type with the highest multiplier. The main drawback to this approach is that the predefined types may be misspecified, leading to biased estimates of the multipliers. In Section 8, I discuss what kinds of mistakes we might expect in this case.

Taking a more sophisticated approach, an analyst could estimate latent types and network parameters simultaneously. The multi-type configuration model is quite similar to stochastic block models studied in the statistics literature, and existing methods for com-

[14] As Breza et al. (2020) argue, such data are far easier to gather than more granular network information.

munity detection in stochastic block models readily adapt to our setting.[15] Importantly, we can compute likelihoods using only aggregated relational data, so methods such as that in Amini et al. (2013) apply more or less off-the-shelf.[16] These methods output probabilities with which the individuals belong to each latent type, and one can target those who are most likely to be a type with a high multiplier.

Filling in susceptibilities requires that we either make assumptions or run a pilot study. To run a pilot, we could seed a sample of individuals and track whether they take up whatever we are trying to spread—given identified types, just take the number of infections divided by the number of seeds of each type. If we can also track subsequent contagion to immediate neighbors, it becomes possible to estimate network parameters without separate data on network connections. This kind of "tree-based" or "nomination-based" sampling is a common way to gather data on a network, and there are well-known methods for estimating network models based on this type of sample (e.g. Li et al., 2023). Depending on the size of the pilot study, there may well be significant uncertainty around important parameters, and it seems prudent to compute multipliers for a range of parameter values.

[15] In a standard stochastic block model, a link exists between individuals of types t and s independently with some probability p_{ts}. A sparse stochastic block model is a special case of the multi-type configuration model in which degree distributions are mixtures of Poisson distributions. However, methods in the statistics literature apply not only to the standard stochastic block model but also to the "degree-corrected stochastic block model," which allows more flexibility in the degree distributions. A few key references include Bickel et al. (2013), who establish consistency and asymptotic normality of a maximum likelihood estimator, Amini et al. (2013), who demonstrate a more computationally efficient method using a pseudo-likelihood, and Wang and Bickel (2017), who offer an approach for model selection.

[16] There are two minor complications. First, we cannot freely partition the population into types because in order to target seeds later, we need a partition that is measurable with respect to observable attributes—this can readily be incorporated as a constraint. The second complication is that statistics papers often assume we have data for the entire population, rather than a sample. However, the pseudo-likelihood used in Amini et al. (2013) is additively separable across individuals, so it is straightforward to compute the corresponding pseudo-likelihood for our sample. Tabouy et al. (2020) offer a more comprehensive treatment of estimating stochastic block models under various types of sampling.

6 The Viral Case

If each infected neighbor causes more than one additional infection on average—if $\rho(M') > 1$—then the network of susceptible individuals has a "giant component." This component contains a positive fraction of the population, even as $n \to \infty$, so the probability that we avoid seeding it goes to zero. As Akbarpour et al. (2020) show, even a small number of randomly placed seeds are likely to hit this component, so the value of targeting comes primarily from "last mile" adoption. Once we infect the giant component, the marginal value of another seed depends on connections between individuals who lie outside it. These individuals are systematically different, but we can still use the recursive formulas in Section 3 to assess their impact on contagion.

We need to account for multiple selection effects—those outside the giant component of susceptible individuals differ in their degrees, susceptbility, and neighbor type distributions. Taking a limit as $\mathbf{s} \to \mathbf{0}$ in (4), the vector $\boldsymbol{\zeta}$, in which ζ_t is the probability that a type t neighbor is part of the giant component, is the largest solution in $[0,1]^\Theta$ to the system

$$\zeta_t = \alpha_t \left(1 - \frac{g_t'(1 - \boldsymbol{\zeta} \cdot \mathbf{q}^t)}{\mu_t}\right), \quad t \in \Theta. \tag{10}$$

The dot product $\boldsymbol{\zeta} \cdot \mathbf{q}^t$ computes the probability that a neighbor of a type t agent links to the giant component, so if a type t agent has k neighbors, she is isolated from the giant component with probability $(1 - \boldsymbol{\zeta} \cdot \mathbf{q}^t)^k$. A type t individual is therefore part of the giant component with probability

$$\delta_t := \alpha_t \left(1 - \sum_{k=1}^{\infty} \mathbb{P}(D_t = k)(1 - \boldsymbol{\zeta} \cdot \mathbf{q}^t)^k\right) = \alpha_t \left(1 - g_t(1 - \boldsymbol{\zeta} \cdot \mathbf{q}^t)\right), \tag{11}$$

and the degree distribution conditional on isolation from the giant component $D_{\zeta,t}$ satisfies

$$\mathbb{P}(D_{\zeta,t}=k) := \frac{\mathbb{P}(D_t=k)(1-\boldsymbol{\zeta}\cdot\mathbf{q}^t)^k}{\sum_{\ell=0}^{\infty}\mathbb{P}(D_t=\ell)(1-\boldsymbol{\zeta}\cdot\mathbf{q}^t)^\ell} = \frac{\mathbb{P}(D_t=k)(1-\boldsymbol{\zeta}\cdot\mathbf{q}^t)^k}{g_t(1-\boldsymbol{\zeta}\cdot\mathbf{q}^t)}. \tag{12}$$

The average degree and forward degree among these individuals are[17]

$$\mu_{\zeta,t} = \sum_{k=1}^{\infty} k\cdot\frac{\mathbb{P}(D_t=k)(1-\boldsymbol{\zeta}\cdot\mathbf{q}^t)^k}{g_t(1-\boldsymbol{\zeta}\cdot\mathbf{q}^t)} = \frac{g_t'(1-\boldsymbol{\zeta}\cdot\mathbf{q}^t)}{g_t(1-\boldsymbol{\zeta}\cdot\mathbf{q}^t)}(1-\boldsymbol{\zeta}\cdot\mathbf{q}^t), \quad \text{and} \tag{13}$$

$$\begin{aligned}\mu'_{\zeta,t} &= \sum_{k=1}^{\infty} k\cdot\mathbb{P}(D'_{\zeta,t}=k) = \sum_{k=1}^{\infty}\frac{k(k+1)\cdot\mathbb{P}(D_{\zeta,t}=k+1)}{\mu_{\zeta,t}} \\ &= \sum_{k=1}^{\infty} k(k+1)\cdot\frac{\mathbb{P}(D_t=k+1)(1-\boldsymbol{\zeta}\cdot\mathbf{q}^t)^{k+1}}{\mu_{\zeta,t}g_t(1-\boldsymbol{\zeta}\cdot\mathbf{q}^t)} = \frac{g_t''(1-\boldsymbol{\zeta}\cdot\mathbf{q}^t)}{g_t'(1-\boldsymbol{\zeta}\cdot\mathbf{q}^t)}(1-\boldsymbol{\zeta}\cdot\mathbf{q}^t)\end{aligned} \tag{14}$$

respectively.

To account for selection in susceptibility and neighbor types, we compute

$$\alpha_{\zeta,t} = \frac{\alpha_t-\zeta_t}{1-\zeta_t} \quad \text{and} \quad q^t_{\zeta,s} = q^t_s\frac{1-\zeta_s}{1-\boldsymbol{\zeta}\cdot\mathbf{q}^t}.$$

The mass α_t of susceptible type t potential neighbors includes a mass ζ_t who link to the giant component, and $\alpha_{\zeta,t}$ conditions on their removal. Likewise, a type s neighbor links to the giant component with probability ζ_s, and $q^t_{\zeta,s}$ reweights the neighbor type distribution for type t accordingly.[18] Analogous to (6), define

$$m_{\zeta,ts} = \alpha_{\zeta,s}q^t_{\zeta,s}\mu_{\zeta,t} \quad \text{and} \quad m'_{\zeta,ts} = \alpha_{\zeta,s}q^t_{\zeta,s}\mu'_{\zeta,t}, \tag{15}$$

giving respectively the average number of new type s infections a type t seed can generate among neighbors and in further steps, conditioned on having no link to the giant component.

[17] Note this subsumes the non-viral case as substituting $\boldsymbol{\zeta}=\mathbf{0}$ yields $\mu_{\zeta,t}=\frac{g_t'(1)}{g_t(1)}=\mu_t$ and $\mu'_{\zeta,t}=\frac{g_t''(1)}{g_t'(1)}=\mu'_t$.

[18] Again, note that these reduce to the non-viral case $\alpha_{\zeta,t}=\alpha_t$ and $q^t_{\zeta,s}=q^t_s$ if $\boldsymbol{\zeta}=\mathbf{0}$.

Write M_ζ and M'_ζ for the corresponding matrices.

Theorem 2. *In the viral case with $\rho(M') > 1$, the seed multiplier for type t is*

$$\beta_t = \alpha_t g_t(1 - \boldsymbol{\zeta} \cdot \mathbf{q}^t) \left(1 + \mathbf{e}_t^\intercal M_\zeta (I - M'_\zeta)^{-1} \mathbf{1}\right). \tag{16}$$

Proof. See Appendix. □

Theorem 2 has exactly the same interpretation as Theorem 1. The terms of the sum $M_\zeta(I - M'_\zeta)^{-1} = M_\zeta \sum_{k=0}^{\infty} (M'_\zeta)^k$ count the expected number of infectious paths from individuals of each type to those *who would not otherwise get infected*. Relative to Theorem 1, we have the additional factor $g_t(1 - \boldsymbol{\zeta} \cdot \mathbf{q}^t)$—with probability $1 - g_t(1 - \boldsymbol{\zeta} \cdot \mathbf{q}^t)$ a type t seed would have gotten exposed anyway, so this portion is redundant. This factor is generally smaller for types that have more connections.[19] Moreover, the selection captured by M_ζ and M'_ζ skews further in favor of seeding low-degree types because the highest degree members of high-degree types are disproportionately selected out. In contrast with our earlier findings, the best targets in the viral case are typically those with few connections who we are unlikely to reach through the giant component—we saw this earlier in the example of Figure 1.

In the viral case, seeding based on these multipliers makes sense if we have already infected the giant component, or if we are seeding enough individuals so that we are almost sure to infect it. However, with a small number of seeds, we may face a meaningful tradeoff between infecting the giant component and reaching more individuals outside this component. The values δ_t computed in (11) allow us to properly assess this tradeoff—each type t seed infects the giant component with probability δ_t. Types with high values of δ_t will typically have lower multipliers β_t because they are more likely to be part of the giant component and are therefore less likely to help infect individuals outside of it. Hence, whether a given type is a

[19] From the definition of the generating function, one can readily verify that a first-order stochastic dominant shift upward in a type's degree distribution leads to uniformly lower values of g_t.

better seed in practice depends on subtle comparisons.[20]

Suppose that we seed m_t individuals of type t, and the total population has size n. The giant component contains roughly $n(\boldsymbol{\delta} \cdot \mathbf{p})$ individuals and gets seeded with probability

$$1 - \prod_{t \in \Theta} (1 - \delta_t)^{m_t}.$$

Therefore, the expected number of infections we get from this seeding strategy is

$$n(\boldsymbol{\delta} \cdot \mathbf{p}) \left(1 - \prod_{t \in \Theta} (1 - \delta_t)^{m_t} \right) + \mathbf{m} \cdot \boldsymbol{\beta}. \tag{17}$$

The first term captures the contribution of the giant component, and the second term captures the contribution of small components. Notice that if the population is larger, the contribtion of the first term is more important, but as we increase the number of seeds, the marginal impact of another seed on this term goes down. Because this marginal impact goes down at an exponential rate, seeding based on multipliers is optimal even for fairly modest seed budgets. However, we should note that, as Akbarpour et al. (2020) argue, the gains from targeting in the viral case are likely quite small in any event.

To get some sense for how many seeds we need before the second term dominates, consider a numerical example based on the earlier plot in Figure 1.[21] If type 0 has 2.15 neighbors in expectation, and type 1 has 4.3, then we have $\beta_0 \approx 0.72$, $\beta_1 \approx 0.19$, $\delta_0 \approx 0.195$, and $\delta_1 \approx 0.405$. The giant component covers approximately 30% of the population, and the low-degree type is half as likely as the high-degree type to be in the giant component, but it triggers more than three times as many infections on average outside this component. In a population with $10,000$ individuals, we should optimally seed only the low degree type once

[20] If it is feasible, one might prefer to spend only a fraction of the available seed budget at first, focusing on types with high δ_t, and switch to seeding types with high β_t after observing a viral cascade.

[21] Recall this plot is based on a simple homophily model with $\alpha_0 = \alpha_1 = 0.5$, $h = 0.75$, and degrees that are drawn from a Poisson distribution.

we have more than about 50 seeds. If there are $100,000$ individuals, seeding only the low degree type is optimal once we have more than 65 seeds.

7 Modeled Networks versus Real Networks

To provide a proof-of-concept, I simulate contagion using the actual village networks from Banerjee et al. (2013). I present two exercises, one directly based on the individual village networks and a second that involves creating cross-village links to form a larger network. In the first exercise, I estimate seed multipliers based on 2000 iterations for each village, and I compare these values to multipliers I compute using the configuration model. The data contain adjacency matrices for 75 villages, with populations ranging from around 300 to more than 1800 individuals. There is also demographic information on a subset of individuals in each village. I restrict the data to include only individuals for whom demographic data are available, resulting in networks of size ranging from approximately 100 to 400 individuals. I model each village as a network with two-types, men and women, assuming $\alpha_m = \alpha_f = 1$, and I use as input the empirical mean and variance of the degree distributions for each.

For this modeling exercise, I parameterize the degree distribution in the configuration model,[22] and I introduce a method to vary network density and account for clustering. I approximate each empirical degree distribution using a negative binomial distribution that matches the empirical mean and variance. A negative binomial distribution requires two parameters $r > 0$ and $\rho \in (0, 1)$, and it has probability mass function[23]

$$P_k = \frac{\prod_{i=r}^{r+k-1} i}{k!}(1-\rho)^r \rho^k.$$

[22] In the non-viral case, multipliers only depend on the first two moments of the degree distributions, but in the viral case, they depend on the entire degree distribution through the generating function.

[23] If r is a positive integer, this distribution corresponds to the number of successes that occur before r failures in a sequence of i.i.d. Bernoulli trials with success probability p. If $r = 1$, this is the geometric distribution.

The mean of the distribution is $\mu = \frac{r\rho}{1-\rho}$, and the variance is $\sigma^2 = \frac{r\rho}{(1-\rho)^2}$. Given μ and σ^2, we can obtain the parameters r and ρ as $r = \frac{\mu^2}{\sigma^2-\mu}$ and $\rho = 1 - \frac{\mu}{\sigma^2}$.[24] The probability generating function is

$$g(s) = \left(\frac{1-\rho}{1-\rho s}\right)^r,$$

and the forward distribution follows a negative binomial with parameters $r+1$ and ρ—we have $\frac{g'(s)}{\mu} = \left(\frac{1-\rho}{1-\rho s}\right)^{r+1}$.

In order to vary the effective network density, I allow infected individuals to expose each neighbor with some probability $\gamma < 1$—this is equivalent to deleting each link in the configuration model independently with probability $1-\gamma$. If $g_t(s)$ is the probability generating function for the original degree distribution D_t, the generating function for the thinned distribution is $g_t(1-\gamma+\gamma s)$, and we can apply the same formulas. However, because real networks exhibit clustering—those who share a common neighbor are often linked themselves—the configuration model will systematically overestimate diffusion without further adjustment. Suppose i gets infected and exposes j, who then gets infected and exposes k. If i and k are linked, this latter exposure is redundant, and we should not count it. We can make a simple correction using the clustering coefficient C, deleting each forward link with probability C. The generating function for the forward distribution becomes $\frac{g_t'(1-\gamma(1-C)+\gamma(1-C)s)}{\mu_t}$.

Putting this all together, to obtain our modeled multipliers, we solve the system

$$\zeta_t = 1 - \left(\frac{1-\rho_t}{1-\rho_t\left(1-\gamma(1-C)\boldsymbol{\zeta}\cdot\mathbf{q}^t\right)}\right)^{r_t+1}, \tag{18}$$

in which ρ_t and r_t are obtained from the empirical mean and variance as described above.[25]

[24] The negative binomial distribution requires $\sigma^2 > \mu$. This is typically true for real networks as the degree distributions have relatively heavy tails, but it is not true for every type in every village in the data. When this condition is violated, I approximate the degree distribution as a Poisson distribution with the same expectation as the empirical degree distribution—this was done for 3 village-type pairs out of a total of 150 such pairs.

[25] The presented results use C equal to the global clustering coefficient for each village. In the viral case, this system has multiple solutions as $\boldsymbol{\zeta} = \mathbf{0}$ is always a solution—the maximal solution is the right one to

We then compute

$$\mu_{\zeta,t} = \frac{r_t \rho_t \gamma (1 - \boldsymbol{\zeta} \cdot \mathbf{q}^t)}{1 - \rho_t \left(1 - \gamma \boldsymbol{\zeta} \cdot \mathbf{q}^t\right)}, \quad \mu'_{\zeta,t} = \frac{(r_t + 1) \rho_t \gamma (1 - C)(1 - \boldsymbol{\zeta} \cdot \mathbf{q}^t)}{1 - \rho_t \left(1 - \gamma (1 - C) \boldsymbol{\zeta} \cdot \mathbf{q}^t\right)} \tag{19}$$

along with $\alpha_{\zeta,t} = \frac{\alpha_t - \delta_t}{1 - \delta_t}$ and $\mathbf{q}^t_\zeta = q^t_s \frac{1 - \zeta_s}{1 - \boldsymbol{\zeta} \cdot \mathbf{q}}$, and we apply Theorem 2. To estimate seed multipliers through simulation, I seed 3% of the population and divide the total number of infections by the number of seeds, averaging over 2000 iterations.[26] I present results for two transmission rates: setting $\gamma = 0.1$ places most villages in the non-viral case, and setting $\gamma = 0.25$ places most villages in the viral case.

Ex-ante, we should expect some differences between outcomes in small graphs and predictions using large network limits. First, in small graphs, the largest component should cover more of the population than the model predicts, which tends to increase multipliers in the non-viral case (leading the model to underestimate the multipliers) and decrease them in the viral case (leading the model to overestimate the multipliers). To understand why, think about the chance an individual has to *avoid* having any neighbors in a large component. If there are m other individuals who *are not* in this component, and our individual links with one of them, there are now only $m-1$ such individuals to whom she can link next. With each subsequent link, the chance of connecting to a large component increases, and this effect is greater when the graph is smaller. A similar effect leads to multiplier compression—in small graphs, the differences between multipliers for different types are smaller.

Figure 2 presents results assuming a 10% transmission rate—for each of the villages included in the analysis, the corresponding model is non-viral.[27] Each dot represents a

use.

[26] If the network contains a component covering more than 25% of the population, I treat this as a giant component and exclude it from the multiplier estimate.

[27] The plot includes data for 65 of the villages. I exclude those containing components that cover more than 20% of the individuals—these villages are near the viral threshold, and the model produces unrealistically large multipliers. However, even in these cases the comparison between men and women is still directionally correct.

village, with the calculated multiplier shown on the vertical axis and the simulated multiplier shown on the horizontal axis. The black line is the 45 degree line—if the model were a perfect fit, every dot would be on this line. The left plot shows the multipliers for men, and the right shows the multipliers for women. There is clearly a strong correlation between the calculated and simulated multipliers, and the model correctly predicts that men have a higher multiplier in all but one village. Particularly for the women, the model tends to underestimate multipliers, but as explained above we should expect bias in this direction.

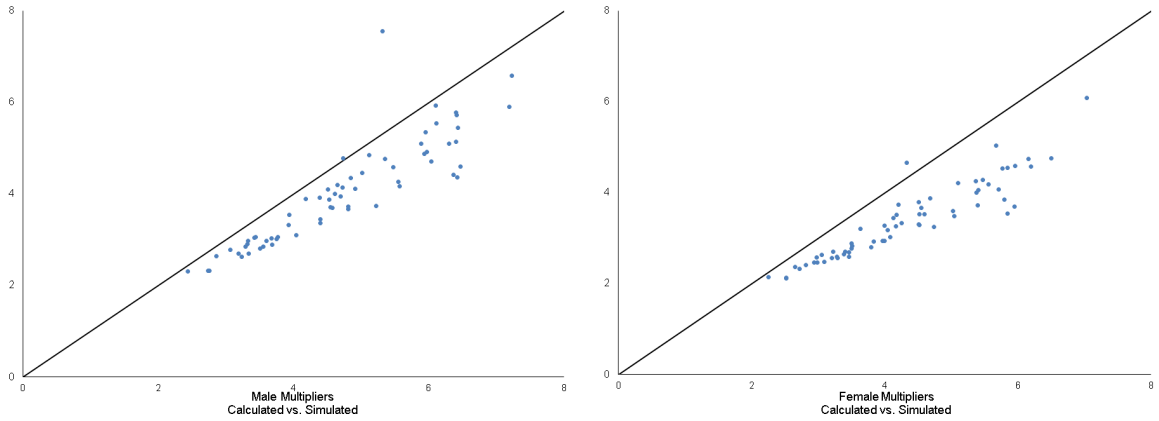

Figure 2: Calculated versus simulated multipliers for men (left) and women (right), assuming 10% transmission rate.

Figure 3 presents analogous plots assuming a 25% transmission rate—most of the village models are now in the viral case.[28] We again see a strong correlation between the model and the simulations, with the model now overestimating the multipliers on average. Because men have higher degrees, they are more likely than women to get exposed through the giant component and tend to be redundant as seeds. The model correctly predicts that seeding lower degree *women* generates greater diffusion in all but two of the villages. Table 1 reports average values of the multipliers across all villages in the analysis. Looking at differences in the multipliers between men and women, we see significant compression in the simulated results, which again we should expect in small graphs.

[28]The plot includes data for 67 of the villages—I have again excluded those near the viral threshold.

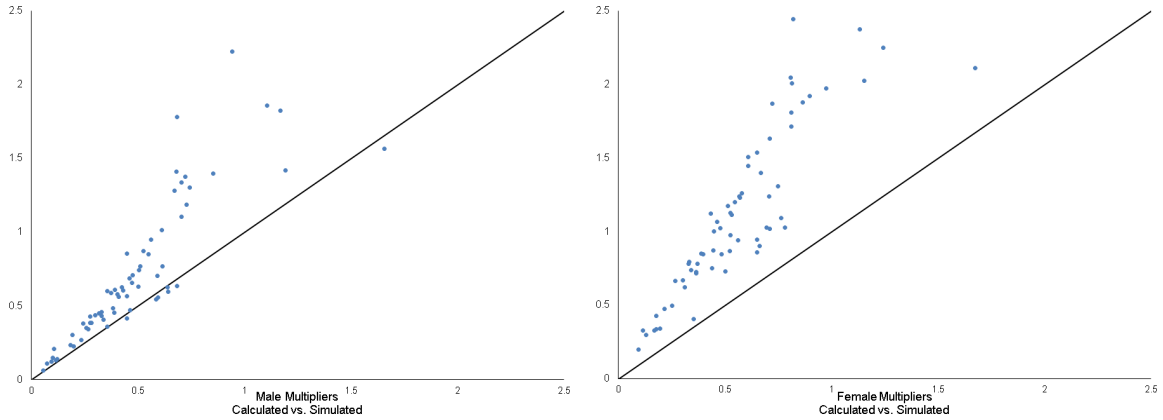

Figure 3: Calculated versus simulated multipliers for men (left) and women (right), assuming 25% transmission rate.

	Simulated β_m	Simulated β_f	Calculated β_m	Calculated β_f
$\gamma = 0.1$	4.71	4.32	4.01	3.39
$\gamma = 0.25$	0.486	0.569	0.710	1.15

Table 1: Average seed multipliers.

Figure 4 plots the calculated size of the giant component $\pi(\mathbf{0})$ versus the size of the largest component in the simulations. The model slightly underestimates the size of this component, which again we should expect in small graphs. Overall, the limit model performs quite well even though the village networks contain only a few hundred vertices each.

Do we get better point estimates if the graph is larger? Can we expect meaningful gains from targeting in practice? A second exercise suggests a positive answer to both questions. I connect the networks for the 10 largest villages by adding random links, and I fit a model in which types describe sex and caste.[29] A transmission rate of 6% puts us firmly in the non-viral case, and table 2 presents computed and simulated multipliers for each type as well as for random seeding. In this hybrid of real and random networks, the model correctly identifies the two most effective types to seed (male with general or missing caste and male

[29]For each pair of individuals who are not in the same village, I add a link independently with a probability that depends on the individuals' types. Probabilities were chosen to keep neighbor type distributions constant, though this increased the average degree by approximately 13%. The combined network contains 3559 individuals with an average degree of 10.98. As the added links are completely random, I do not make a clustering correction for the computed multipliers.

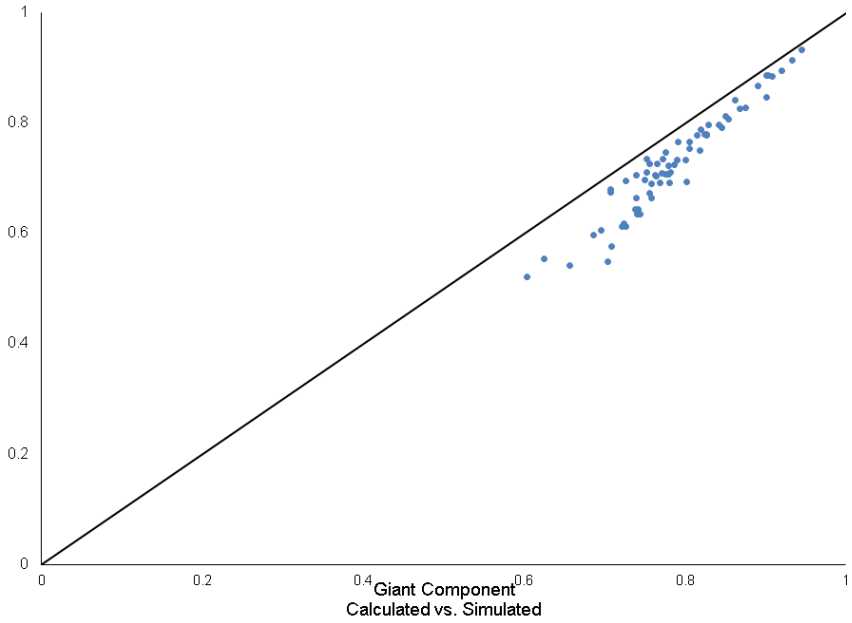

Figure 4: Calculated versus simulated size of largest component as a fraction of the population.

OBC). Moreover, the simulated multipliers indicate a potential gain of 23% more infections per seed relative to random targeting. Achieving this gain with a very coarse set of types suggests that the model can generate valuable targeting improvements at relatively low cost.

	Simulated β	Calculated β
Male, OBC	4.39	4.64
Male, Scheduled Caste	3.53	3.43
Male, Scheduled Tribe	4.16	3.68
Male, General or Missing	4.77	4.78
Female, OBC	3.72	3.96
Female, Scheduled Caste	3.11	3.00
Female, Scheduled Tribe	3.51	3.19
Female, General or Missing	3.92	3.82
Random Type	3.87	3.97

Table 2: Seed multipliers in a larger model.

8 Discussion

Model Misspecification

We might not have perfect data on the relevant types. If we sort individuals based on too many demographic attributes, including multiple types with essentially the same linking patterns, the only cost lies in noisier estimates of the relevant multipliers—the estimates are still unbiased, and with enough data we should reach correct conclusions. However, if we sort individuals based on too *few* attributes (i.e., we are missing important demographic variables), this can lead to biased estimates of the seed multipliers. How large an issue is this, and can we say anything about the direction of the bias?

To get some insight into these questions, suppose the true graph is drawn from a multi-type configuration model, but we have no data on types. Suppose further that all types have the same susceptibility α, and we are in the non-viral case with $\rho(M') < 1$. The only available policy is to seed at random, but we would like to know the corresponding seed multiplier. If we knew the true model, we could compute

$$\beta_t = \alpha \left(1 + \mathbf{e}_t^\intercal M(I - M')^{-1}\mathbf{1}\right)$$

for each type t and take an average

$$\overline{\beta} = \sum_{t\in\Theta} p_t \beta_t. \tag{20}$$

Since we do not observe the types, we might instead measure the average degree and forward degree in the graph

$$\overline{\mu} = \sum_{t\in\Theta} p_t \mu_t, \quad \overline{\mu}' = \sum_{t\in\Theta} \frac{p_t \mu_t}{\overline{\mu}} \mu_t', \tag{21}$$

and compute the corresponding estimate:

$$\hat{\beta} = \alpha \left(1 + \frac{\alpha \overline{\mu}}{1 - \alpha \overline{\mu}'} \right). \tag{22}$$

How does our estimate $\hat{\beta}$ differ from the true multiplier $\overline{\beta}$?

Recall the unbiased neighobr type distribution with $q_t^* = \frac{p_t \mu_t}{\sum_{s \in \Theta} p_s \mu_s}$. If there is any homophily relative to this benchmark, the true multiplier $\overline{\beta}$ is higher than our estimate $\hat{\beta}$.

Proposition 2. *Suppose $\alpha_t = \alpha$ for each type $t \in \Theta$, and $\rho(M') < 1$. If $q_s^t = q_s^*$ for all t and s, then $\overline{\beta} = \hat{\beta}$. If $q_t^t \geq q_t^*$ and $q_s^t \leq q_s^*$ for all $t \neq s$, then $\overline{\beta} \geq \hat{\beta}$.*

Proof. See Appendix. □

Since homophily is typical in real networks, we should expect that missing variables lead us to underestimate multipliers, at least in the non-viral case. Intuitively, when homophily increases, types with higher degrees become more valuable because their neighbors are more likely to have high degrees as well, and this positive selection gets magnified at each step of contagion—a quadratic increase two steps away, a cubic increase three steps away, etc. Low-degree types analogously become less valuable, but since polynomials with positive coefficients are convex, the redistribution leads to greater overall contagion.

One can similarly show that homophily increases the spectral radius of M', so the true model may include a giant component even if our estimated model is non-viral. If this happens, the true multipliers could be *smaller* than our estimates because they depend on the subgraph that excludes the giant component, which contains smaller components on average than we would expect. Nevertheless, we would underestimate the extent of contagion due to unexpected viral diffusion. If both the estimated model and the true model are viral, it becomes more difficult to make a clear comparison, but this case is empircally uncommon, and there is less need for careful targeting.

Comparing with Other Seeding Strategies

How does multiplier-based seeding compare with other targeting strategies? A primary motivation of this paper is that most existing methods require more information about a network than is often practical to obtain—I discuss an important exception at the end of this section. The most well-known approaches take as input a full adjacency matrix G in which links represent possible transmission channels. Since these techniques use so much more information about a network, we should expect them to produce greater contagion. The advantage of multiplier-based seeding lies primarily in data availability and cost—and to a lesser extent, in a lower demand for computing power. We trade off performance for broader applicability.

As discussed in Section 5, multiplier-based seeding requires gathering demographic variables for the target population, which we map to types based on a model of the network, and aggregated relational data for a sample of the population. In contrast with Kempe et al. (2003) and subsequent work in the same paradigm, we do not need any data on network connections for most individuals. Even within the sample used to fit the network model, we need not identify any particular links. Obtaining accurate data on demographic variables like gender, education level, or profession is often far easier than surveying network links.

Aside from data availability, computational complexity poses as challenge. Kempe et al. (2003) show that optimal seeding is NP-complete, and any heuristic we employ still needs to process the full adjacency matrix. The fastest algorithms with performance guarantees require at least $O\left((n+m)\log n\right)$ operations (Borgs et al., 2014), in which n is the number of vertices and m the number of edges. Moreover, the unstated constant can be quite large as algorithms typically require simulating a diffusion process over many iterations. Multiplier-based seeding has the potential to be far simpler. Computing the multipliers themselves is extremely easy—the most costly operations involve mutiplying and inverting $|\Theta| \times |\Theta|$ matrices, and this does not scale with population size. The real work is to fit the network

model in the first place. How demanding this is depends on how we estimate the model—using pre-defined types will be much simpler than using maximum likelihood to jointly estimate network parameters and types—and how large a sample we use. In any event, the computing work scales with the number of types and the sample size, not the population size, so we are guaranteed savings in sufficiently large networks.

Multiplier-based seeding is not the only method designed to use limited information. Another popular approach leverages selection via the friendship paradox: choose a random individual and seed a random *neighbor* of that individual. Instead of gathering aggregated relational data, this requires asking individuals to nominate specific neighbors. Because random neighbors tend to have more connections than random individuals, this offers an inexpensive way to select seeds with higher than average degrees. How does this compare with multiplier-based seeding? My analysis allows a straightforward answer—we can even assess a hybrid strategy in which we target random neighbors of a particular type of individual.

Focusing on the non-viral case, we can adapt the proof of Theorem 1 to compute the average number of infections we get by targeting a random neighbor of a random individual. If we select a random neighbor of a type t individual, this neighbor has type s with probability q_s^t and, given type s, is susceptible with probability α_s. We need to sum two terms, one capturing diffusion that goes through our initial random individual and one that does not. Conditional on infecting our seed, we expose the initial random individual and get the corresponding diffusion—we just need to avoid double counting the seeded neighbor. Define

$$\mu_t^- = \mathbb{E}[D_t - 1 \,|\, D_t \geq 1],$$

the expected degree of a type t individual minus one, conditional on having at least one neighbor.[30] Replacing M in equation (7) with the matrix M^-, with entries $m_{ts}^- = \alpha_s q_s^t \mu_t^-$,

[30]When we seed a random neighbor, I assume we keep selecting a new random individual until we find one with a neighbor to seed.

gives us the expected number of infections that spread through our initial random individual, conditional on infecting the chosen neighbor. To get the first of our two terms, we multiply by $\mathbf{q}^t \cdot \boldsymbol{\alpha}$, the probability that a neighbor of a type t individual is susceptible. For the second term, we use the forward distribution to compute

$$\mathbf{e}_t^\intercal Q(I - M')^{-1}\mathbf{1},$$

in which the matrix Q has entry $q_{ts} = \alpha_s q_s^t$. This is just the second term in (7), replacing M with Q to account for the exposure of a single neighbor rather than μ_t of them in expectation.

Once we have the above calculation for each type, we can compute a multiplier for seeding a random neighbor of a random individual: take a weighted average over types with weights $\mathbf{p}$. Depending on the degree distributions, this may give a higher or lower multiplier than simply seeding the optimal type. However, if we have data to both fit the multi-type configuration model *and* identify neighbors, we can combine the two strategies: find the type for which seeding a random neighbor yields the highest number of infections on average, pick random individuals of that type, and seed their neighbors. Such a hybrid approach necessarily yields a higher multiplier than either method on its own.

9 Final Remarks

This paper delivers a simple approach for seeding a contagion. By representing the network as a random graph, we reduce both the dimension of the optimal seeding problem and the data required for implementation. Simulations on real networks provide support for key predictions and highlight potential gains we can realize at low cost. One issue not discussed thus far is any notion of fairness. Implicit in the analysis is that, at least when the seeding budget is small, a planner focuses her efforts on a single type. In some contexts, this

could lead to undesirable inequities between groups. Despite this issue, the seed multipliers still provide a fairness-minded planner with the ability to assess tradeoffs—we can quantify both the loss in total contagion and the distributional impact from alternative policies.

Though beyond the scope of the present paper, the ability to study more complex contagions—in which additional infections among neighbors increase one's chance of getting infected—is crucial for many applications, and incentives for transmission are also clearly important. While this leaves much work for the future, the broader conceptual point in this paper remains relevant: type-based targeting can deliver more tractable analytical results that require less data in practice. The recursive structure of the configuration model makes it especially well-suited for studying diffusion processes, and I am optimistic that this or a related model will help later research confront these more difficult questions.

References

Akbarpour, Mohammad, Suraj Malladi, and Amin Saberi (2020), "Just a Few Seeds More: Value of Network Information for Diffusion." Working Paper.

Alidaee, Hossein, Eric Auerbach, and Michael Leung (2020), "Recovering Network Structure from Aggregated Relational Data using Penalized Regression." Working Paper.

Amini, Arash, Aiyou Chen, Peter Bickel, and Elizaveta Levina (2013), "Pseudo-Likelihood Methods for Community Detection in Large Sparse Networks." *The Annals of Statistics*, 41, 2097–2122.

Athreya, Krishna and Peter Ney (1972), *Branching Processes*. Springer-Verlag, New York.

Banerjee, Abhijit, Arun Chandrasekhar, Esther Duflo, and Matthew Jackson (2013), "The Diffusion of Microfinance." *Science*, 341.

Banerjee, Abhijit, Arun Chandrasekhar, Esther Duflo, and Matthew Jackson (2019), "Using Gossips to Spread Information: Theory and Evidence from Two Randomized Controlled Trials." *Review of Economic Studies*, 86, 2453–2490.

Beaman, Lori, Ariel BenYishay, Jeremy Magruder, and Ahmed Mushfiq Mobarak (2021), "Can Network Theory-Based Targeting Increase Technology Adoption?" *American Economic Review*, 111, 1918–1943.

Bickel, Peter, David Choi, Xiangyu Chang, and Hai Zhang (2013), "Asymptotic Normality of Maximum Likelihood and Its Variational Approximation for Stochastic Block Models." *The Annals of Statistics*, 41, 1922–1943.

Borgs, Christian, Michael Brautbar, Jennifer Chayes, and Brendan Lucier (2014), "Maximizing Social Influence in Nearly Optimal Time." In *Proceedings of the 25th ACM-SIAM Symposium on Discrete Algorithms*, 946–957.

Breza, Emily, Arun Chandrasekhar, Tyler McCormick, and Mengjie Pan (2020), "Using Aggregated Relational Data to Feasibly Identify Network structure without Network Data." *American Economic Review*, 110, 2454–2484.

Chen, Wei, , Tian Lin, Zihan Tan, Mingfei Zhao, and Xuren Zhou (2016), "Robust Influence Maximization." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 795–804.

Chen, Wei, Laks Lakshmanan, and Carlos Castillo (2013), "Information and Influence Propagation in Social Networks." In *Synthesis Lectures on Data Management* (M. Tamer Özsu, ed.), 1–177, Morgan & Claypool Publishers.

Chen, Wei, Yajun Wang, and Siyu Yang (2009), "Efficient Influence Maximization in Social Networks." In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 199–208.

Chin, Alex, Dean Eckles, and Johan Ugander (2022), "Evaluating Stochastic Seeding Strategies in Networks." *Management Science*, 68, 1714–1736.

Goel, Sharad, Ashton Anderson, Jake Hofman, and Duncan Watts (2016), "The Structural Virality of Online Diffusion." *Management Science*, 62, 180–196.

Goel, Sharad, Duncan Watts, and Daniel Goldstein (2012), "The Structure of Online Diffusion Networks." In *Proceedings of the 13th ACM Conference on Electronic Commerce*, 623–638.

Hill, Shawndra, Foster Provost, and Chris Volinski (2006), "Network-Based Marketing: Identifying Likely Adopters via Consumer Networks." *Statistical Science*, 21, 256–276.

Iyengar, Raghuram, Christophe Van den Bulte, and Thomas Valente (2011), "Opinion Leadership and Social Contagion in New Product Diffusion." *Marketing Science*, 30, 195–212.

Kempe, David, Jon Kleinberg, and Eva Tardos (2003), "Maximizing the Spread of Influence through a Social Network." *KDD Conference Proceedings.*

Kim, David, Alison Hwang, Derek Stafford, Alex Hughes, James O'Malley, James Fowler, and Nicholas Christakis (2015), "Social Network Targeting to Maximise Population Behavior Change: a Cluster Randomised Controlled Trial." *Lancet*, 386, 145–153.

Kumar, Vineet and K. Sudhir (2019), "Can Friends Seed More Buzz and Adoption?" Cowles Foundation Working Paper Number 2178.

Leskovec, Jure, Ajit Singh, and Jon Kleinberg (2006), "Patterns of Influence in a Recommendation Network." In *Proceedings of the 10th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, 380–389.

Li, Tianxi, Elizaveta Levina, and Ji Zhu (2023), "Community Models for Networks Observed Through Edge Nominations." *Journal of Machine Learning Research*, 24, 1–36.

Sadler, Evan (2020), "Diffusion Games." *American Economic Review*, 110, 225–270.

Seeman, Lior and Yaron Singer (2013), "Adaptive Seeding in Social Networks." In *IEEE 54th Annual Symposium on Foundations of Computer Science*, 459–468.

Tabouy, Timothée, Pierre Barbillon, and Julien Chiquet (2020), "Variational Inference for Stochastic Block Models From Sampled Data." *Journal of the American Statistical Association*, 115, 455–466.

Wang, Rachel and Peter Bickel (2017), "Likelihood-Based Model Selection for Stochastic Block Models." *The Annals of Statistics*, 45, 500–528.

Watts, Duncan (2002), "A Simple Model of Global Cascades on Random Networks." *Proceedings of the National Academy of Sciences*, 99, 5766–5771.

Wilder, Bryan, Amulya Yadav, Nicole Immorlica, Eric Rice, and Milind Tambe (2017), "Uncharted but not Uninfluenced: Influence Maximization with an Uncertain Network." In *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems*, 1305–1313.