# Optimal Network Based Targeting for Technology Adoption in Developing Countries

*Preliminary and Incomplete - Please do not Cite*

Aranya Chakraborty*

**Abstract**

How do we use existing social ties to improve the adoption of a new technology? I explore optimal network-based targeting when the true expected benefit associated with the new technology varies at the household level. In particular, I focus on two types of targeting: targeting people central to the network and targeting based on the likelihood of adoption. I develop a theoretical framework where initially uninformed agents engage in DeGroot learning to decide whether or not to get fully informed about a new technology. Conditional on being fully informed, they then decide whether or not to adopt the new technology. The model predicts the possibility of low information equilibria where nobody will adopt the new technology even if it's efficient for some of them to do so. This may happen when people are not optimistic about the prospect of the new technology. Targeting is needed to improve adoption in this context. My simulations suggest that the optimal targeting strategy in such a scenario relies on the underlying heterogeneity in the population. If the heterogeneity is high in terms of the applicability of the new technology, targeting based on the likelihood of adoption performs better than centrality-based targeting. Conversely, centrality-based targeting works better if the population is more homogeneous. I test these predictions using data from Malawi. My results show support in favor of my theoretical model. I argue that in designing targeting strategies for technology adoption we should pay particular attention to the characteristics of the population.

*Department of Economics, McGill University; Canada. E-mail: aranya.chakraborty@mail.mcgill.ca.

Technology adoption has long been recognized as a driving force of economic development (see Besley and Case, 1993). How do we use existing social ties to improve adoption of a new technology? The literature argue that the answer depends on the underlying diffusion process (Beaman et al., 2021). If information diffuses only if a certain threshold of each agents' connections are informed, targeting based on existing social ties may be required for widespread adoption. The literature recommends targeting agents central to the network (Beaman et al., 2021). The recommendation, however, is based on the underlying assumption that the agents are homogeneous in terms of the benefits they get by adopting the new technology. What happens if for some agents the new technology is more beneficial than the others?

In this paper I study optimal network based targeting strategies for improving technology adoption, when agents are heterogeneous in terms of the applicability of the new technology. In particular, I focus on the situation where the new technology can be more beneficial to some agents than the others. I explore whether the optimal network based targeting strategies vary as I vary the degree of heterogeneity within the network. More specifically, I concentrate on the relative performance of two type of strategies: targeting based on centrality and targeting based on likelihood of adoption.

My study makes four contributions to the existing literature. First, I provide evidence (both theoretical and empirical) that the success of network-based targeting strategies depend on the population level heterogeneity. Diffusion of information via network is the key to increase technology adoption (Foster and Rosenzweig, 1995; Conley and Udry, 2010). In the recent years, there has been a number of studies focusing on the role of networks in the diffusion of technologies. A growing proportion of these studies explore the most effective way to use social networks to improve technology adoption (e.g., Banerjee et al., 2013; BenYishay and Mobarak, 2018). A few of these studies explore the role of underlying diffusion process in designing the most effective targeting policies (e.g., Beaman et al., 2021; Akbarpour, Malladi and Saberi, 2021). However, these studies assume population to be homogeneous in terms of the benefits they get from the new technology. In the current study I show evidence that in scenarios where this assumption is not true, the optimal targeting strategies may differ from the ones prescribed in the existing literature. In particular, the effectiveness of a targeting policy will vary depending on the population level heterogeneity in terms of the benefits from the new technology. Considering population level heterogeneity in social learning itself is not new in the literature (e.g., Munshi, 2004; Bandiera and Rasul, 2006; Conley and Udry, 2010)[1]. However, to the best of my knowledge, the current study is the first to consider the consequences of population level heterogeneity on targeting strategies.

Second, my theoretical framework helps formalize the scenario where agents learn from their network about a technology that is more beneficial to some of them than the others. Existing studies consider technologies to be equally beneficial to everyone. The adoption may still differ due to heterogeneity in costs. But these heterogeneous costs are assumed to be known by the agents and thus do not require learning[2]. Thus, simplifying assumptions are made such that the learning involves the variable that takes a common value for all the agents and not the variable they differ in terms of. This assumption helps us to focus on a problem where the agents are collectively trying to uncover some hidden parameter(s) of interest. In many scenarios, however, agents do face heterogeneous benefits in adopting a new technology. For example, in agriculture the performance of some practices depend on the quality of land[3]. Thus, farmers vary in terms of the benefits they get in adopting those practices depending on the quality of their land. The consequences of this is not formally studied in the existing literature. The current study aspire to close this gap.

Third, I am able to use existing data for testing the predictions of my model. This is obviously done based on some strong assumptions. One of the reasons existing studies made simplifying assumptions on the structure of heterogeneity in the population, is the difficulty in observing heterogeneity in benefits beforehand. As the benefits only realize

[1]Using the data from Indian Green Revolution, Munshi (2004) finds that information flows are weaker for rice growers than wheat growers as rice growing regions are more heterogeneous. Bandiera and Rasul (2006) observe network effects on technology adoption to vary based on the number of adopters in the network for sunflower production in Mozambique. Conley and Udry (2010) finds that only novice farmers learn from their veteran neighbours about the use of fertilizers for pineapple production in Ghana.

[2]Even if the heterogeneous costs are not known to the agents. There is no possibility of learning from the network as these costs are not assumed to be correlated within network. There can still be a possibility of learning-by-doing.

[3]For example, pit planting studied in Beaman et al. (2021) requires flat land.

after the adoption, they cannot be factored into the targeting strategies. I attempt to solve this issue by using additional data to estimate adoption conditional on observable demographics. This way I am able to categorize the population in terms of their propensity to adopt a new technology. I use this estimates with experimental data from Malawi (more details below) to test the predictions of my theory[4]. To my best knowledge, this is the first study that does this in the context of network based targeting.

Finally, I provide policy directions for network-based targeting when the population is heterogeneous. In particular, I argue in favour of targeting early adopters when the heterogeneity is high[5].

I develop a theoretical framework where economic agents participate in a two-stage decision process. In the first stage, uninformed agents engage in DeGroot learning to decide whether or not to get fully informed about a new technology. In the second stage, conditional on being fully informed they decide whether or not to adopt the technology. This framework helps me formalize a scenario where pessimism regarding the prospect of a new technology will lead to its low adoption, even if its efficient for many agents to adopt. The structure of the model is based on the works of Golub and Jackson (2010) and Banerjee et al. (2021). Similar to these studies, I also consider DeGroot type learning as it is used in all the canonical models of information aggregation in the development literature. There is also empirical evidence in favor of it (see Chandrasekhar, Larreguy and Xandri, 2020). The two-stage decision process is also something that can be found in the existing literature (Chandrasekhar, Golub and Yang, 2018).

Based on my theoretical model, I use simulations to evaluate the relative importance of different targeting strategies. This helps me generate testable hypotheses that I take to the data[6].

I test the predictions of my model combining two different data sources from Malawi. First one is the replication data from a randomized controlled trial (RCT) conducted by Lori Beaman, Ariel BenYishay, Jeremy Magruder and Ahmed Mushfiq Mobarak (2021)

[4]Comments on how common this is in Economic research.

[5]Description of the literature that argues in favour of targeting early adopters, noting that the contribution is made here for network-based targeting.

[6]Comments on how common simulations are in the literature.

(henceforth, BBMM). The second dataset is the Agricultural Extension Services and Technology Adoption Survey (henceforth, AESTAS) data collected by International Food Policy Research Institute (IFPRI). The AESTAS dataset helps me estimate the likelihood of adoption based on some observable demographics. I use these estimates in the BBMM dataset to calculate household level likelihood of adoption. It should be noted that the calculation relies on observable demographics only and are not specific to any particular technology. I exploit the experimental set-up of BBMM, together with my calculations of household level likelihood of adoption to test my hypotheses.

My simulations indicate that the relative performance of different targeting strategies depend on the degree of heterogeneity in a network. Centrality based targeting strategies should be less effective in settings where the agents vary significantly in terms of their true benefits from adopting a technology. In such settings, targeting based on likelihood of adoption should perform better. The intuition behind such result lies in the characteristics of the central seeds in a network[7]. As central seeds are, by definition, the most well connected people in a network, they represent the average network characteristics. In a setting where a new technology is applicable to only a certain sub-section of the population, targeting based on centrality becomes more likely to fail in reaching the population of interest.

Reduced form results show evidence in favour of my hypothesis[8].

The remainder of this article is organized as follows. In section 1, I present the theoretical framework of my analysis. Section 2 presents my simulations that forms the basis of my hypotheses. In Section 3, I discuss my empirical strategy together with the data sources. Section 4 presents and discusses my empirical results. Finally, in section 6, I summarize my findings and make concluding remarks.

[7]In the network literature, information entry points are termed *seeds*.

[8]Detailed empirical results needed here.

# 1 Theoretical Framework

I consider a choice problem that requires learning in a social network. Specifically the problem is of technology adoption when the true net expected benefit associated with the new technology varies at the household level. In particular, the new technology is not '*better*' for everyone in the population with respect to an existing technology, even under homogeneous cost of adoption. On top of this, the true expected benefit with the technology is initially unknown to the households and must be learned.

Similar to Golub and Jackson (2010), I consider agents to have an initial opinion and involve in DeGroot learning (developed in DeMarzo, Vayanos and Zwiebel, 2003). I distinguish between informed and uninformed agents similar to Banerjee et al. (2021)[9]. In addition, I consider the possibility that people are heterogeneous in terms of their true probability distribution associated with the new technology.

## 1.1 The Theoretical Model

Consider a two-stage decision process where in the first stage the agents decide whether or not to make an irreversible investment to learn about an available new technology. Then, conditional on making that investment, in the second stage they decide whether to stick to a traditional technology, or adopt the new technology[10]. The traditional technology has a sure payoff of $\pi^T$, where the new technology provides a payoff of $\pi^N(\omega_{it})$ that depends on the state of the world parameter $\omega_{it} \in \Omega$. The state of the world parameter $\omega_{it}$ is drawn independently at each period $t$ according to the true distribution $p_i^*(\omega_{it})$ for household $i$. Therefore, the draws are not correlated over time within household and between households[11]. I assume that the true distributions are positively correlated between households according to the existing network structure.

[9]In Banerjee et al. (2021), uninformed agents have empty beliefs and informed agents can be partially or fully informed. In contrast, I assume uninformed agents to have an initial opinion (this includes partially informed agents) and informed agents to be fully informed.

[10]This two-step decision process is similar to the model presented in Chandrasekhar, Golub and Yang (2018). Which also forms the micro-foundation of Beaman et al. (2021).

[11]The assumption of draws not being correlated over time within household helps me abstract away from learning-by-doing, where households are observing the draws over time and updating their beliefs accordingly. The assumption of draws not being correlated over time between households constrains the ways the households can learn from each other.

I also assume that $\forall it$, $\exists \omega_{it}, \omega'_{it} \in \Omega$ such that $\pi^N(\omega_{it}) \geq \pi^T \geq \pi^N(\omega'_{it})$; i.e., for each agent $i$ and period $t$, there exist states of the world such that the payoff from the new technology is higher (lower) than the old technology. Finally, $\exists i, j \in \mathcal{I}$ such that $\int_{\omega_{it} \in \Omega} p_i^*(\omega_{it})\pi^N(\omega_{it}) - c_i \geq \pi^T$ and $\int_{\omega_{jt} \in \Omega} p_j^*(\omega_{jt})\pi^N(\omega_{jt}) - c_j \leq \pi^T$, where $\mathcal{I}$ denote the set of all households and $c_i$ is the cost of new technology for household $i$. Which means that there is enough heterogeneity in the population such that for some agents the net expected benefits of adopting the new technology with respect to the traditional technology is greater than zero, while for others it is less than zero. This last assumption ensures that the new technology is 'better' for only a fraction of households in the population.

The household $i$ has beliefs $p_{it}(\omega_{it})$ over the distribution of $\omega_{it}$ at period $t$. Every period, an uninformed agent has the option to become informed by putting effort $e_{it} \in \{0, 1\}$. Agents put effort only once, i.e., if $e_{i\tau} = 1$, $e_{it} = 1$ $\forall t \geq \tau$. If $e_{it} = 1$, the agent learns the true distribution $p_i^*(\omega_{it})$ at cost $\eta_i$. The cost of learning is incurred only once - the first time the agent gets informed. If $e_{it} = 0$, no effort cost is incurred and the agent uses DeGroot averaging to approximate the true distribution. Let $G$ denote the $n \times n$ weighted, directed, and non-negative influence matrix ($n = |\mathcal{I}|$), where $G_{ij} \geq 0$ represents the weight $i$ places on $j$'s opinion (with $\sum_{j \in \mathcal{I}} G_{ij} = 1$). Then $\hat{p}_{it}(\omega_{it}) = \sum_{j \in \mathcal{I}} G_{ij} p_{jt-1}(\omega_{jt-1})$ denotes household $i$'s approximation based on others' opinion following the DeGroot averaging. The belief of agent $i$ at period $t$ is thus determined by the following process:

$$p_{it}(\omega_{it}) = e_{it}(p_i^*(\omega_{it})) + (1 - e_{it})\hat{p}_{it}(\omega_{it}) \tag{1}$$

In other words, uninformed agents use DeGroot averaging to approximate the true distribution with the help of their peers. On the other hand, informed agents can actually observe the true distribution and hence do not need to approximate it anymore. Their effort level determine whether they are informed or uninformed. In addition, I assume that agents need to be informed before they adopt. As can be seen below, this assumption helps me explicitly capture the point when the agents stop seeking information from their peers.

Finally, I assume the agents to be risk-neutral and myopic. The assumption of riskneutrality is for simplification purpose only. This is so that I can focus solely on the expected values, without the need to think about the variation around it. As the new technology is assumed to be riskier than the traditional technology here, risk-averse agents may find it less attractive. As such, the net benefit of the new technology would be less than the one perceived by a model where the agents are risk neutral. This can easily be accommodated in the current model by dividing the expected payoff of the new technology by its variance. Such an exercise would not change the main results of the model. The assumption of myopic agents help me focus on a static model instead of a dynamic one. More importantly, if the agents are not myopic they may wait until their peers get informed before they decide whether or not to get informed themselves. This may lead to a more complicated scenario where everyone is waiting for their peers to get informed first. Such a scenario is beyond the scope of this paper[12].

Under the above assumptions, an agent's adoption decision is a two step process:

1. First they decide whether or not to get informed, based on the following rule:

$$e_{it} = \begin{cases} 1 \; if \;\; \int_{\omega_{it} \in \Omega} \hat{p}_{it}(\omega_{it})\pi^N(\omega_{it}) - c_i - \pi^T \geq \eta_i \\ 0 \; otherwise \end{cases} \tag{2}$$

   Only uninformed agents make this decision.

2. Conditional on being informed, they decide whether or not to adopt the new technology:

$$Adopt_{it} = \begin{cases} 1 \; if \;\; \int_{\omega_{it} \in \Omega} p_i^*(\omega_{it})\pi^N(\omega_{it}) - c_i \geq \pi^T \\ 0 \; otherwise \end{cases} \tag{3}$$

## 1.2 Implications of the Model

Consider the situation when there are only two states of the world: one where the new technology has a higher payoff than the traditional one (denoted $\omega_H$), and the other

[12]Comments on the literature that makes similar assumption for developing country agriculture.

where the new technology has a lower payoff than the traditional one (denoted $\omega_L$). Thus $\Omega = \{\omega_H, \omega_L\}$. Let $p^*_{iH} := p^*_i(\omega_H)$ denote the true probability that for household $i$ the new technology has a higher payoff than the traditional one. Suppose $p^H_{it} := p_{it}(\omega_H)$ is household $i$'s belief of $p^*_{iH}$ at period $t$. Then, following (1), $p^H_{it}$ is equal to $p^*_{iH}$ if the household is informed, otherwise it is equal to $\hat{p}^H_{it} := \hat{p}_{it}(\omega_H)$. Here $\hat{p}^H_{it}$ denotes the households' approximation of $p^*_{iH}$ based on their network, following DeGroot averaging. Under this simplified scenario, I can now solve the model following backward induction.

In step 2, conditional on being informed, the household decides whether or not to adopt the new technology. The household will adopt the new technology iff:

$$\begin{aligned} & p^*_{iH}\pi^N(\omega_H) + (1 - p^*_{iH})\pi^N(\omega_L) - c_i \geq \pi^T \\ & \Rightarrow p^*_{iH} \geq \frac{c_i + (\pi^T - \pi^N(\omega_L))}{(\pi^N(\omega_H) - \pi^N(\omega_L))} = \bar{p}^*_{iH} \text{ (say)} \end{aligned} \tag{4}$$

That is, iff the true probability of success with the new technology ($p^*_{iH}$) is higher than a threshold ($\bar{p}^*_{iH}$), it is profitable for the household to adopt the new technology. Given this condition for adoption in step 2, in step 1 the household $i$ will choose to get informed at time $t$ iff:

$$\begin{aligned} & p^H_{it}\pi^N(\omega_H) + (1 - p^H_{it})\pi^N(\omega_L) - c_i - \pi^T \geq \eta_i \\ & \Rightarrow p^H_{it} \geq \frac{c_i + (\pi^T - \pi^N(\omega_L))}{(\pi^N(\omega_H) - \pi^N(\omega_L))} + \frac{\eta_i}{(\pi^N(\omega_H) - \pi^N(\omega_L))} = \bar{p}^*_{iH} + \bar{\eta}_i \text{ (say)} \end{aligned} \tag{5}$$

From (4) and (5), it is clear that if, for household $i$, $p^H_{it}$ is equal to $p^*_{iH}$, and they choose to get informed in step 1, they will also adopt the technology in step 2. Conversely, if (4) is not satisfied, then (5) should not satisfy if the diffusion of information is efficient. In other words, under fully efficient information diffusion, only those who would adopt the technology in step 2 would end up getting informed in step 1. Thus, for these people, the following condition must be true:

$$p^*_{iH} \geq \bar{p}^*_{iH} + \bar{\eta}_i \tag{6}$$

Equation (6) implies that for people who end up adopting the technology, it must be so that their true probability of success justifies the cost of seeking information ($\bar{\eta}_i$) on top of their threshold probability of adoption ($\bar{p}^*_{iH}$). Suppose for household $j$, $\bar{p}^*_{jH} + \bar{\eta}_j \geq p^*_{jH} \geq \bar{p}^*_{jH}$. Then even if $p^H_{jt}$ is equal to $p^*_{jH}$, the household $j$ will end up not getting informed about the technology. Hence they will not adopt the technology, even if its profitable for them to do so. This is due to the positive cost of learning ($\eta_j$). This feature is similar to the models of Chandrasekhar, Golub and Yang (2018) and Banerjee et al. (2018), where social stigma of information seeking can stop people from learning.

From the above discussion, it is clear that there are multiple possible equilibria for this model. One equilibrium that is of particular interest to me is when $p^H_{it} \approx 0 \; \forall it$. This is when everyone believes that, for them, the new technology yields lower payoff than the traditional one with certainty. In such a scenario, nobody will adopt the new technology even it may be efficient for some to do so. The role of a policy intervention is then to exogenously make some agents (seeds) informed to improve adoption. The informed agent $i$ will learn about their $p^*_{iH}$, which will get agent $j$ to update their $\hat{p}^H_{jt}$ if $j$ puts positive weight on $i$'s opinion. This will in turn get agent $k$ to update their $\hat{p}^H_{kt+1}$ if $k$ puts positive weight on $j$'s opinion, and so on. The outcome of this intervention in terms of technology adoption, a few periods down the line, will depend on the initial seeding strategy. In the next section, I focus on measuring the relative performance of two types of such seeding strategies.

## 2 Simulations

In this section, I consider networks of households whose true probabilities associated with a new technology are imperfectly correlated. In particular, the probability distributions are positively related according to the existing network structure. Under such a scenario, I first demonstrate the potential problem for a centrality based targeting strategy with the example of a specific network. Then I simulate 200 such networks to analyze whether the problem will persist on average and compare the centrality based targeting with a probability based targeting strategy (defined below). As a benchmark, I run the simulations first for a scenario where the true probabilities are perfectly correlated before moving on to the case where the correlation is imperfect. For the latter, I observe that the relative performance of targeting strategies depends on the level of heterogeneity in the population.

## 2.1 An illustrative example

I start with the example of a specific network that has 20 households. The households are heterogeneous with respect to their true probability of success associated with a new technology (represented by the $p^*_{iH}$s)[13]. These probabilities matter for the households as their states of the world are independently drawn at each period. The probabilities are correlated according to the existing network structure (given by the network's influence matrix). This introduces the possibility of learning from the network. The distribution of true probabilities of success are shown in Figure 1.

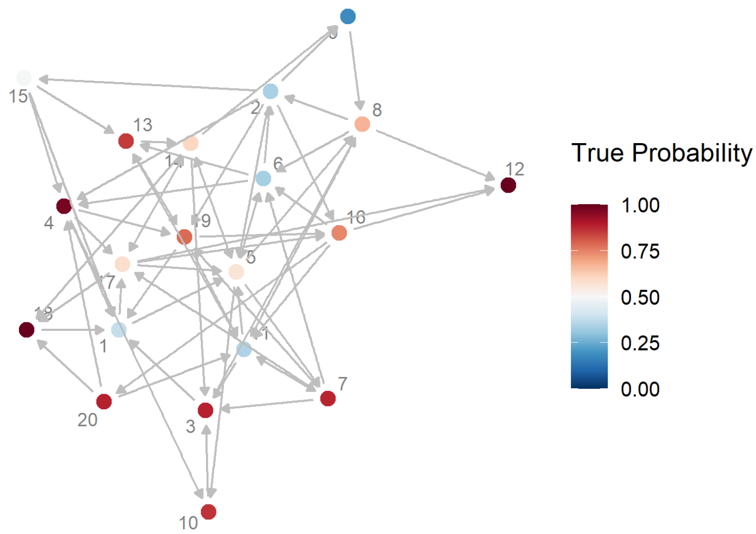



Figure 1: Distribution of True Probability within the network

Consider the scenario where, prior to any interventions, everyone believe their probabilities of success with the new technology to be zero ($p^H_{it} = 0 \; \forall it$). Under such scenario, households will not adopt the technology even if it is optimal for some of them to do so. An intervention is then required to improve adoption. The objective behind such intervention would be to ensure the households that should adopt the technology under perfect

[13]Here, similar to section 1.2, I am assuming two states of the world: either success or failure with the new technology.

information, end up adopting. The efficiency of a targeting strategy can be measured as:

$$\text{Targeting Efficiency} = \frac{\text{\% of informed households}}{\text{\% of informed households under full efficiency}} \tag{7}$$

where *% of informed households* capture the fraction of households that choose to get informed within some periods of implementing the targeting strategy (i.e. they satisfy equation (5)); the *% of informed households under full efficiency* is the fraction of households that should get informed as they would adopt the technology under perfect information (i.e. they satisfy equation (6)).

For the purpose of this example, consider the threshold probability of learning (i.e. $(\bar{p}^*_{iH} + \bar{\eta}_i)$ in (5)) to be 0.5 for every household. Thus, if the true success probability of a household is more than 50% the household should get informed under full efficiency. Given the distribution of true probabilities of success shown in Figure 1, it turns out that it is then efficient for 70% of all households to get informed in this network.

For my analysis, I will focus on two types of targeting strategies: centrality based and probability based. For each type of strategy, I will consider seeding only two households, similar to BBMM. Centrality based targeting is considered since there is evidence in favor of its success in the existing literature (Banerjee et al., 2013). This is particularly true for the diffusion process described here, where the households adopt iff a certain threshold of their connections adopt (Beaman et al., 2021). I consider probability based targeting as an alternative to this. The probability based targeting strategy is to seed people that have the highest true probabilities of success with the new technology (i.e., highest $p^*_{iH}$s in the network). These are the people who are more likely to adopt a technology given the same probability threshold of learning for everyone and hence are considered to be the early adopters here. I consider this strategy since it is the extreme opposite of the centrality based targeting strategy. Where centrality based strategy relies on people that are similar to the average for diffusion, probability based strategy does the opposite in relying on people that are more likely than average to adopt a technology[14].

The centrality based targeting strategy is to seed people that are central to the network.

[14] Comments on the existing literature that argues in favor of targeting early adopters for faster diffusion.

For the particular example here and the subsequent analysis in this section, I will consider betweenness centrality. Betweenness centrality of a household captures how important the household is in terms of connecting to other households. The results of my analysis are robust with respect to different measures of centrality- particularly, an eigen-vector based measure of centrality (consult appendix D for detailed results). Eigen-vector based centrality measures take into account not only the connectivity of a household to other households, but also the importance of their connections in terms of their respective connections. A more formal definition of different centrality measures can be found in the section 2.2.4 of Jackson (2010). Robustness of my results with respect to eigen-vector based measures is important for two reasons. First, there is evidence in the existing literature in favor of targeting based on eigen-vector centrality (e.g., Banerjee et al. (2013); Beaman et al. (2021)). Second, for my empirical analysis I use eigen-vector centrality as the primary measure of centrality. However, for my primary theoretical analysis here I use betweenness centrality instead. This is because betweenness centrality measure is more simplistic and because of that it allows me to explain my theoretical results more intuitively. On top of this, going back and forth between centrality measures help me establish the robustness of my theoretical and empirical results.

Figure 2 captures the initial seeding where the targeting is based on betweenness centrality. Here I seed households numbered 7 and 11. In other words, households numbered 7 and 11 are exogenously made aware of their respective probabilities of success with the new technology. These particular households are chosen because among all the households in this network, they are most important in terms of connecting to other households. This can be partly seen by following the arrows in the figure (the arrows represent connection, with the direction being the direction of the connection). It can be observed right away that where one of the seeds (household number 7) has a high true probability of success, the same cannot be said about the other seed (household number 11). This is because as central seeds are well connected in the network, by construction they represent the average households and not the early adopters. This will have consequences for the final performance of this targeting strategy.

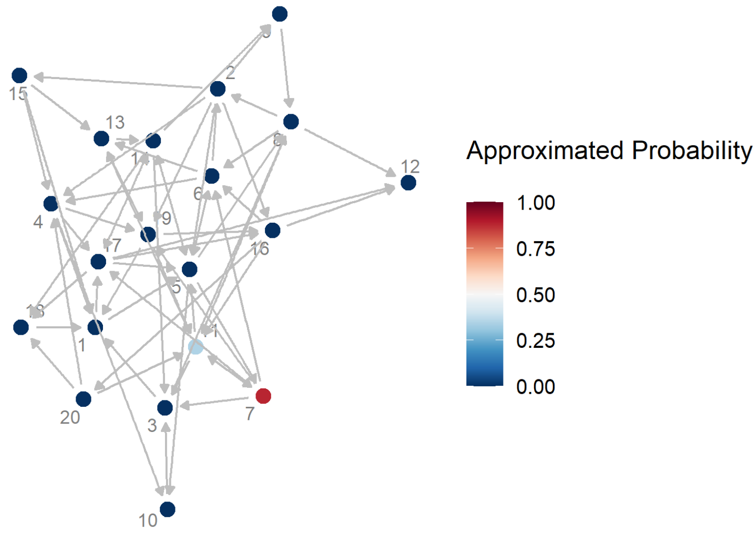



Figure 2: Seeding based on Centrality

Figure 3, on the other hand, captures initial seeding based on probability based targeting. The seeded households are numbered 12 and 18. They are picked as they have the highest true probability of success with the new technology among all the households in this network. This is obviously much easier to do theoretically. In practice, we may not have the information needed to identify these households. This is something I will focus in section 3.4. The seeds here are not so well connected in the network and represent early adopters.

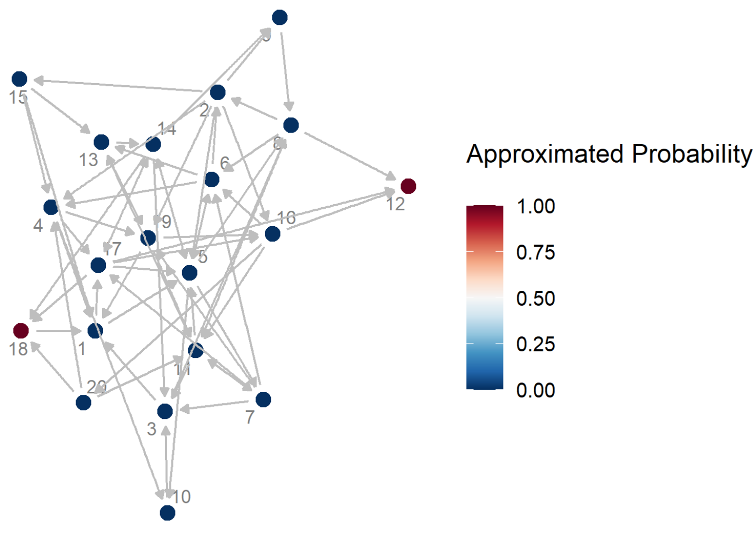



Figure 3: Seeding based on Probability

After the initial seeding, I let the diffusion take place over 10 periods, according to the diffusion process described in the last section. The performance of both targeting strategies at the end of the 10 periods are presented in figure 4. In this particular scenario, probability based seeds perform better than their centrality based counterparts. Comparing these performances with the distribution of true probabilities of success within the network, I observe that centrality based seeds manage to convince only 10% households to get informed about their true probabilities of success, where the probability based seeds convince 90% of households. Given that 70% of households should have gotten informed under full efficiency, this means 14.3% and 128.6% targeting efficiency for centrality based and probability based seeds following (7). Therefore, the betweenness centrality based targeting strategy fails miserably in this scenario. It is worth noting that if the targeting is done on the basis of an eigen-vector based centrality instead, centrality based targeting performs much better. In fact the efficiency of a eigen-vector based targeting strategy would be 114.3% in this scenario. Which is still worse than a probability based targeting strategy, but much better than a betweenness centrality based targeting. This gives me more reason to check the robustness of my results with respect to an eigen-vector based centrality measure.

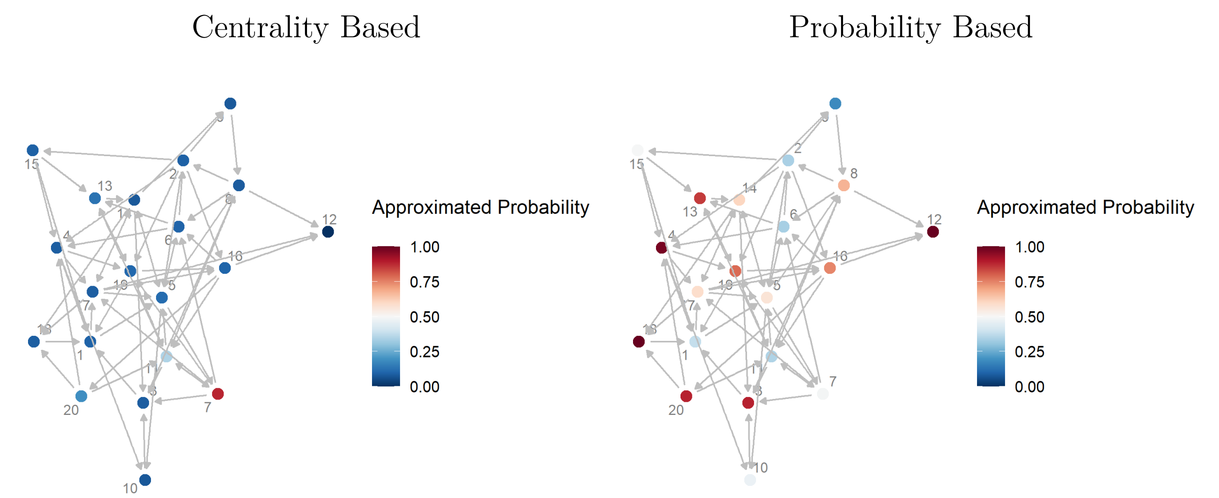



Figure 4: Performance of seeds after 10 periods

It should also be noted that in this particular example the $p^*_{iH}$s are highly heterogeneous. They are correlated according to the network's influence matrix, but their variation within the network is very high. In terms of the applicability of the new technology, this represents that the households are highly heterogeneous in this network.

In what follows, I vary the degree of this heterogeneity. In particular, for a set of same 200 networks (given by their respective influence matrices), I vary the correlation level of $p^*_{iH}$s. The objective of this exercise is to note the relative performances of centrality and probability based targeting strategies over varying degree of population level heterogeneity. However, before doing that I would like to establish the results for the benchmark case of perfect correlation between $p^*_{iH}$s. This is the case where $p^*_{iH}$s are the same all the households in a network. These networks are thus homogeneous in terms of the applicability of the new technology.

## 2.2 Targeting Homogeneous Networks

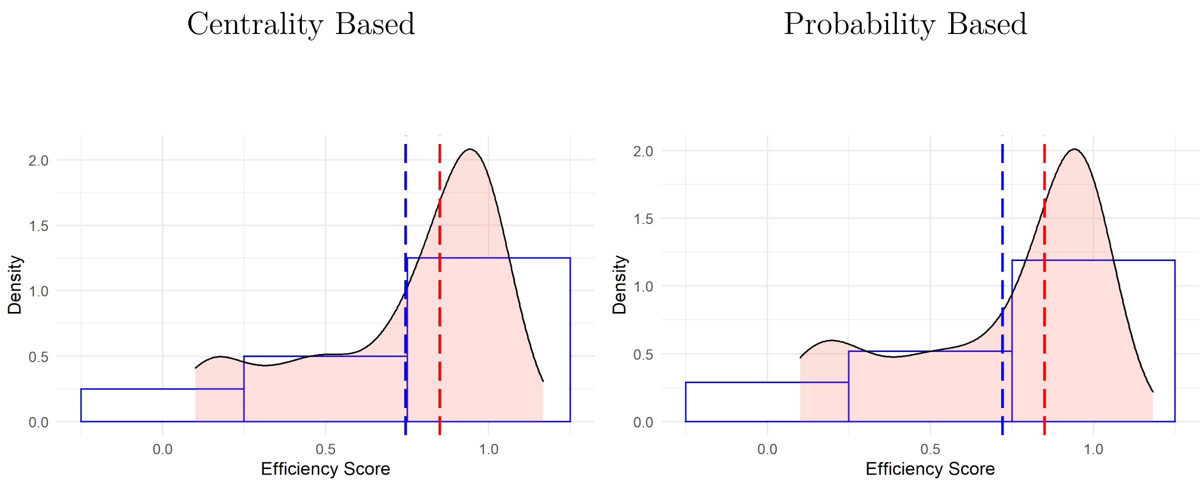



Figure 5: Distribution of efficiency scores when probabilities are perfectly correlated

Figure 5 represents the benchmark case of perfect correlation with respect to the $p^*_{iH}$s. The last column of table 1 represents the corresponding results. I am interested in the distribution of efficiency scores after 10 periods of simulations over 200 networks. In this case, probability and betweenness centrality based targeting strategies perform equally well. In fact, betweenness centrality based targeting performs slightly better in terms of higher mean and lower variance. It should be noted that in this case an eigen-vector centrality based measure completely outperforms probability based measure (detailed results in Appendix D).

This is what is expected in perfectly homogeneous networks. When $p^*_{iH}$s are perfectly correlated, my model reduces to a model of complex diffusion where everyone gets the same benefits from adopting the new technology[15]. This is similar to the micro-foundation model of BBMM. Hence, the results are in favor of centrality based targeting.

Table 1: Simulation Results

| | | Level of Correlation | | | |
|---|---|---|---|---|---|
| Strategy | Statistic | Low | Medium | High | Perfect |
| Centrality Based | Mean | 0.72 | 0.82 | 0.84 | 0.74 |
| | Median | 0.77 | 0.90 | 0.95 | 0.85 |
| | Variance | 0.17 | 0.10 | 0.09 | 0.08 |
| Probability Based | Mean | 0.94 | 0.81 | 0.76 | 0.72 |
| | Median | 1 | 1 | 1 | 0.85 |
| | Variance | 0.11 | 0.11 | 0.14 | 0.09 |
| | Observations | 200 | 197 | 192 | 200 |

*Note:* Simulations on varying levels of correlation are all done for 200 networks, each containing 20 households. However, upon generation of the true probabilities, some networks have to be dropped as they contained 0% of informed households under full efficiency.

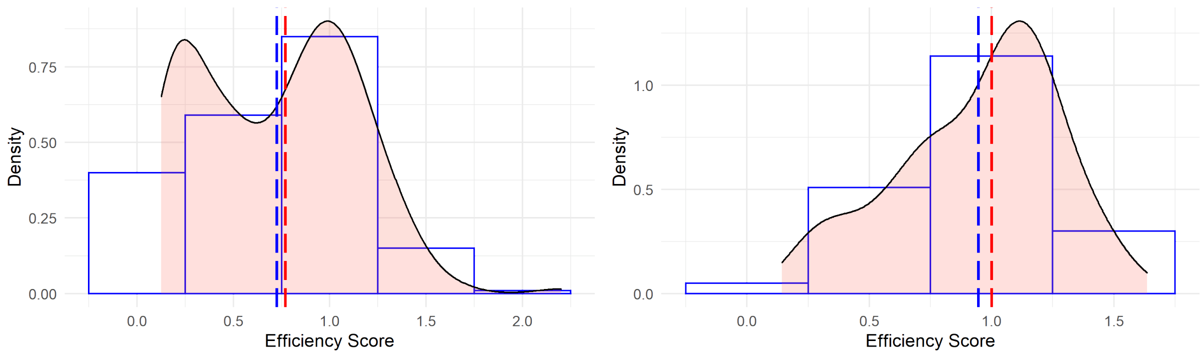
## 2.3 Targeting Heterogeneous Networks

What happens if the households in a network differ in terms of the benefits they get from adopting a new technology? In this subsection, I explore the performances of targeting strategies over varying degree of heterogeneity within network in terms of the applicability of a new technology. Similar to the last subsection, I focus on the distribution of efficiency scores after 10 periods of simulations over 200 networks. Table 1 presents the results for three different levels of correlation of $p^*_{iH}$s between households within networks. Figure 6 represents the corresponding distributions.

[15]Complex diffusion models assumes that information diffuses to a household iff a certain threshold of the household's connections get informed.

Centrality Based

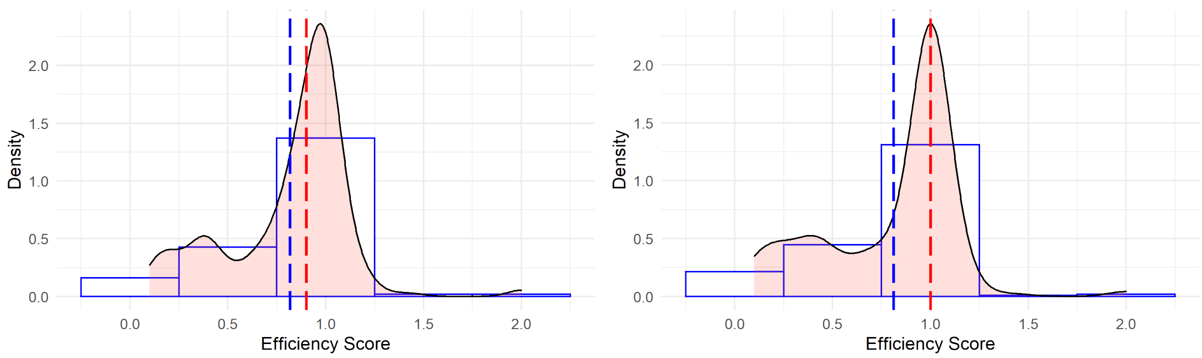
Probability Based




Low Level of Correlation

Centrality Based

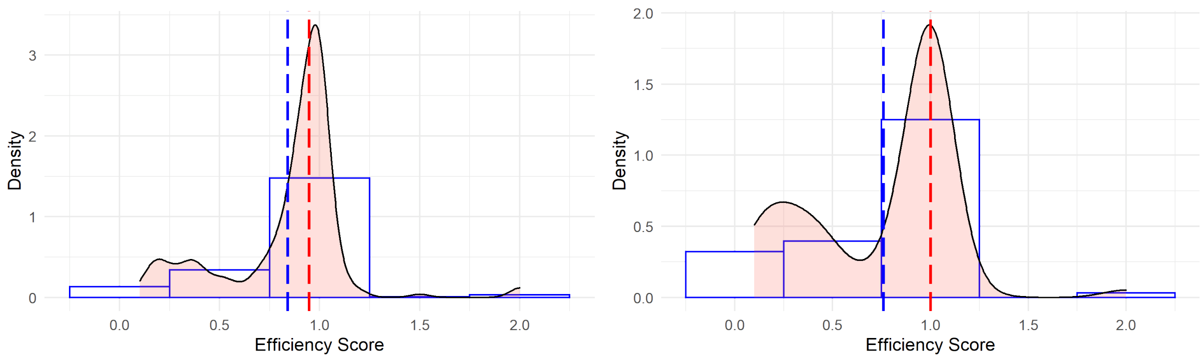
Probability Based




Medium Level of Correlation

Centrality Based

Probability Based




High Level of Correlation

Figure 6: Distribution of efficiency scores when probabilities are imperfectly correlated

If there is a low correlation between the $p^*_{iH}$s, the networks heterogeneity is high in terms of the applicability of the new technology. In this scenario, the probability based targeting outperforms the betweenness centrality based targeting both in terms of mean (0.94 vs 0.72) and median (1 vs 0.77) efficiency scores, with a lower variance (0.11 vs 0.17). As the level of correlation increases to a medium level, I observe that the gaps are closed. Centrality based targeting has a mean of 0.82, compared to a mean of 0.81 for the probability based targeting. The median efficiency scores are 0.90 for centrality based targeting and 1 for probability based targeting. The variances are similar. Finally, I consider the results with a high level of correlation between the $p^*_{iH}$s that represents a low heterogeneity of the networks in terms of the applicability of the new technology. Here, centrality based targeting outperforms the probability based targeting in terms of mean efficiency (0.84 vs 0.76). In terms of the median efficiency, however, the probability based targeting (1) is still higher than centrality based targeting (0.95), although the difference is getting really small. In this particular case, probability based targeting efficiency scores have a higher variance than their centrality based targeting counterpart (0.14 vs 0.09). All the results are also available in Table 1.

## 2.4 Discussion

From the above analysis, I observe that the relative performance of the two targeting strategies considered here depend on the level of heterogeneity within the network in terms of the applicability of the new technology. When the heterogeneity is high, there is a clear distinction between the people that should adopt the technology and the ones that should not. In that scenario, targeting people that are central to the network and hence, representative of the average household, is not efficient. Targeting households that are more representative of the people who should adopt the technology works better. On the contrary, when the heterogeneity is low, everybody in the network are similar in terms of the applicability of the new technology. Targeting people that are central to the network, and thus representative of the average household, is a good strategy. This is because as the heterogeneity approaches to zero, targeting based on probability converges to random assignment (as everyone has the same probability of adoption). Then, similar to the existing literature, centrality based targeting outperforms random assignment given the same number of seeds. It is also worth noting that as the heterogeneity approaches to zero, we converge to the world of perfect correlation. Thus, the last column of Table 1 is qualitatively similar to the penultimate column. The actual numbers differ though. This is because the results in the penultimate column take the networks' influence matrices into account, the results in the last column do not.

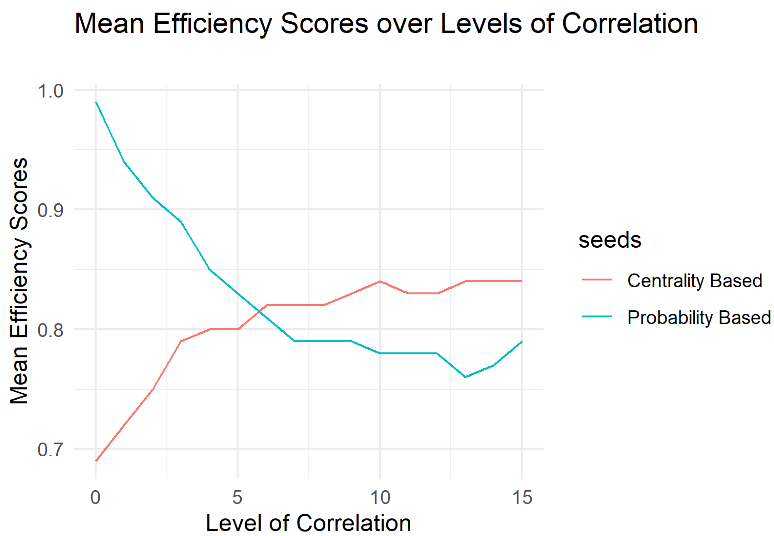



Figure 7: Mean efficiency scores over increasing levels of correlation

Figure 7 represents the mean efficiency scores over a range of different levels of correlation. The main result discussed above holds for this wider range of observations. As the correlation increases, hence heterogeneity decreases, centrality based targeting performs better and probability based targeting performs worse. This result is weaker but robust to an eigen-vector based measure of centrality (discussed in Appendix D). These results help form the main hypotheses of my study which I take to the data in the subsequent sections. For the purpose of my empirical strategy, it helps to think of the two targeting strategies with respect to a benchmark. In particular, consider the benchmark of selecting two seeds randomly. I expect the efficiency scores of a random assignment to remain the same over different heterogeneity levels. This is because random assignment is independent to the level of correlation, and hence should not be affected by it. Thus, with respect to this benchmark, I expect probability based targeting to perform better and centrality based targeting to perform worse, as the heterogeneity within the network increase.

Figures 8 and 9 represent the median and variance of efficiency scores over a range of different levels of correlation.

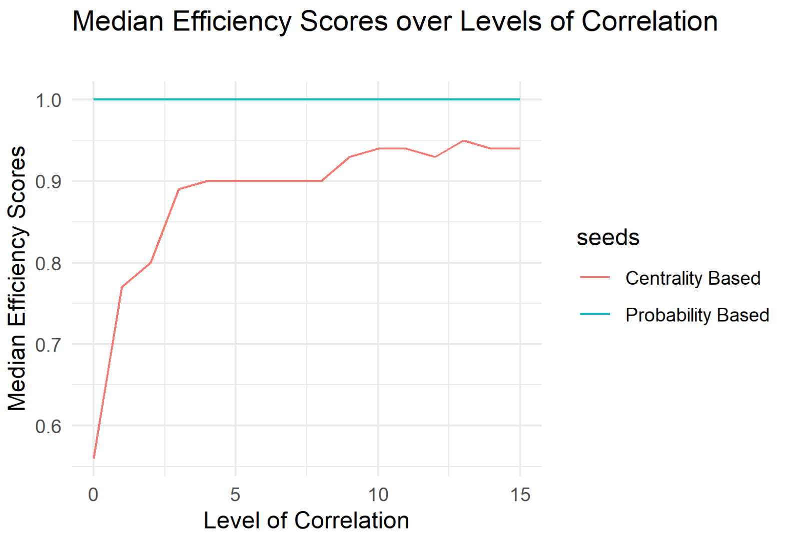



Figure 8: Median efficiency scores over increasing levels of correlation

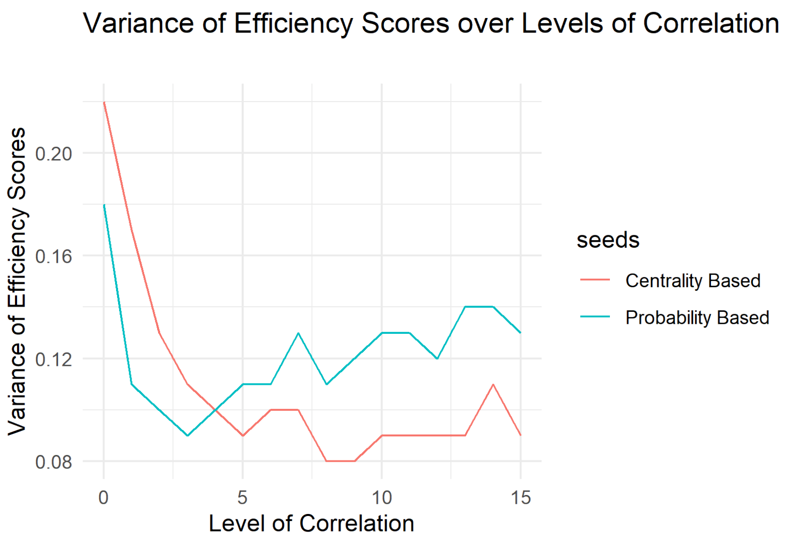



Figure 9: Variance of efficiency scores over increasing levels of correlation

# 3 Empirical Framework

My next objective is to empirically test the following hypotheses derived from my theoretical framework using simulations:

**Hypotheses:** The performance of centrality based targeting and probability based targeting strategies depend on the level of heterogeneity in a region in terms of applicability of a new technology. More specifically:

1. As the level of heterogeneity in terms of the applicability of a new technology increases in a region, central seeds perform worse than the benchmark.

2. As the level of heterogeneity in terms of the applicability of a new technology increases in a region, probability based seeds perform better than the benchmark.

## 3.1 The Ideal Experiment

In the next subsection, I will discuss the identification strategy used in this research to test the above hypotheses. Before getting into that, I think it will be useful to think about the ideal experimental set-up that can help test the above hypotheses in the most straight-forward fashion. In an experimental set-up, I will need to randomly allocate regions/villages into three types of seeding strategies:

1. Centrality based seeding

2. Probability based seeding

3. Random seeding (which will serve as a benchmark/ control group)

Then I can use the following reduced-form regression:

$$\begin{aligned} Y_v = \alpha_0 + \alpha_1 Centrality_v + \alpha_2 Probability_v + \alpha_3 Heterogeneity_v \\ + \alpha_4 Centrality_v \times Heterogeneity_v + \alpha_5 Probability_v \times Heterogeneity_v + u_v \end{aligned} \tag{8}$$

Here $Y_v$ denote some adoption related outcome for village $v$. $Centrality_v$ and $Probability_v$ dummies indicate whether the village got assigned to either centrality or probability based seeding strategy. $Heterogeneity_v$ measures the village level heterogeneity in terms of the applicability of a new technology. Finally, $u_v$ captures the random error term in the regression. My hypotheses state that $\alpha_4 < 0$ and $\alpha_5 > 0$.

However, I do not have access to the data from this ideal experiment. Hence, I use the replication data from BBMM together with the survey data from AESTAS conducted by IFPRI. In the following subsection, I describe these datasets before moving on to the description of my identification strategy.

## 3.2 Data Sources

I use data from two different sources:

- Replication data of BBMM.
- AESTAS data collected by International Food Policy Research Institute (IFPRI).

### 3.2.1 Replication data of BBMM

BBMM used Randomized Controlled Trial (RCT) to promote *Pit Planting* (PP) and *Crop Residue Management* (CRM) for Maize farmers in Malawi. The sample on the use of CRM is very limited. Thus, similar to the main analysis of their study, I focus on Pit Planting only. I also do not expect my predictions to be valid for CRM. This is because CRM is not a *new* technology in the sampled areas, where PP is. The researchers seeded 200 villages from 3 Malawian districts with semi-arid climates (Machinga, Mwanza, and Nkhotakota) with 2 'seed' farmers each, to induce widespread social learning. The intervention was to train these seed farmers on PP and CRM, with the training remaining the same across different treatment arms. The villages got equally divided into 4 treatment groups:

1. **Complex Contagion:** Seeding done assuming the underlying diffusion process to be of complex contagion. Under the assumption of this diffusion process information diffuses only if a certain threshold of each households' connections get informed. As a result of this assumption, both the chosen seeds were central in the network.

2. **Simple Contagion:** Seeding done assuming the underlying diffusion process to be of simple diffusion. Under the assumption of this diffusion process information diffuses with a random probability from one household its connections. As a result of this assumption, one seed household was central while the other seed household was in the periphery.

3. **Geo:** Seeding done solely based on geographic proximity. As a result, the seeds were near each other (in terms of geography), but not central (in terms of the network data).

4. **Benchmark (control):** Extension agents selected two seeds like they usually do.

The researchers first collected the social network census data, before any intervention or household survey took place. The census elicited names of people each respondent consults when making agricultural decisions, information on household composition, socioeconomic characteristics of the household, general agriculture information, and work group membership information. They used this responses with the village listing to identify links. They considered individuals linked if either party named each other (undirected network) or if they are part of the same household. Based on this network information, the researchers used simulations to identify seeds according to complex and simple diffusion processes, and geographic proximity to optimize diffusion after four periods. For each of the 200 villages, they identified 2 seeds for each type of prospective treatment (except for benchmark seeds, which were identified by the extension agents only for the control villages). The next step was to randomly allocate villages to one of the four treatment groups and selecting seeds for training based on the treatment group the village got allocated to. Once the training is complete, they conducted household survey to collect data on farming techniques, input use, yields, assets, and other characteristics.

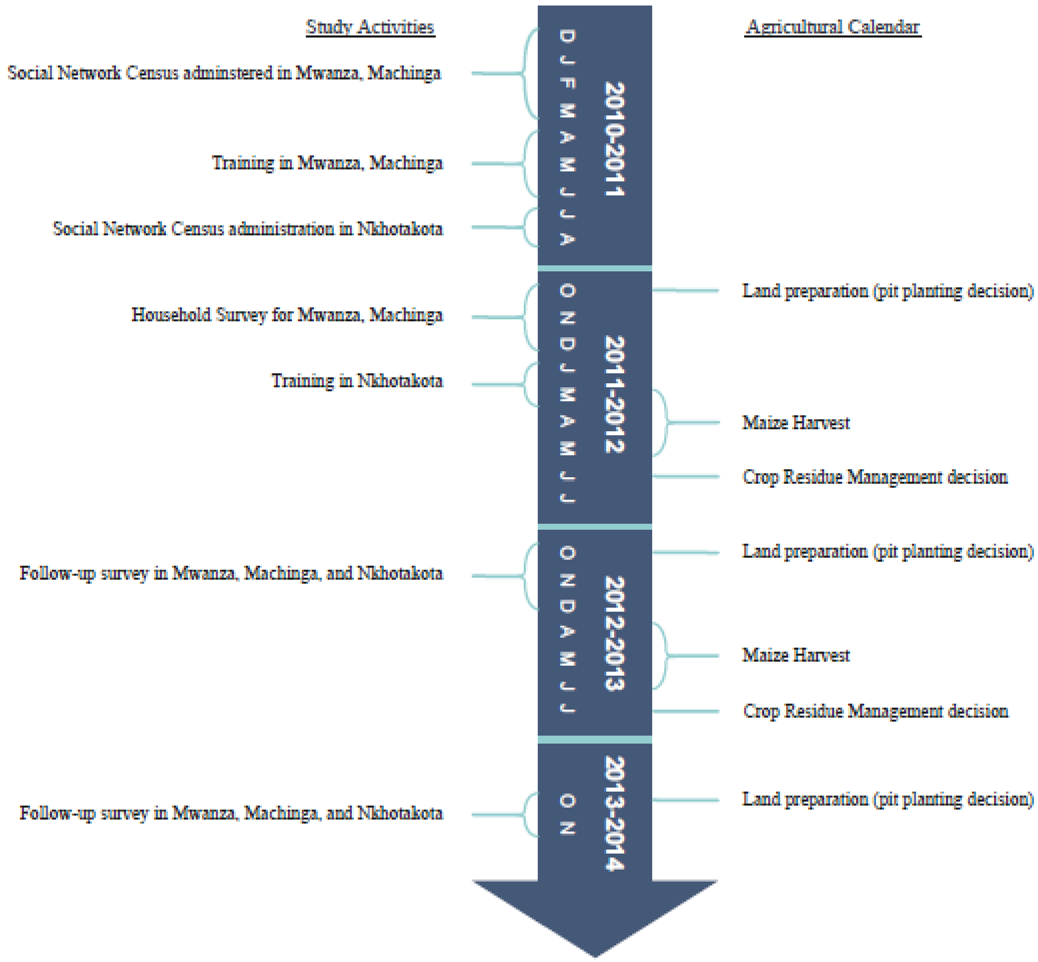



Figure 10: Project Timeline of Beaman et al. (2021)

The authors randomly surveyed a panel of approximately 30 households per village. This involved all the seed and shadow farmers[16], as well as 22-24 other farmers. They collected information on approximately 5600 households from the 200 villages. In 2 districts (Machinga and Mwanza) that consists of 141 study villages, they collected three rounds of survey data in 2011, 2012, and 2013. Due to unanticipated delays in project funding, in the third district (Nkhotakota) they could only start the operation in 2012. Hence, for the third district with 59 study villages, they collected only two rounds of survey data (in 2012 and 2013). The first round of survey was conducted a few months after the training of the seed farmers. This round attempted to capture some baseline characteristics and knowledge levels of the households regarding both PP and CRM. Every survey round was conducted at the start of the agricultural season, after the land preparation. As PP is

[16]A shadow farmer is a seed farmer chosen by the simulation, assuming some underlying diffusion model, but was not seeded as the village got randomly assigned to seeding based on a different diffusion model.

used for land preparation, the households' adoption decision of PP was observed thrice for Machinga and Mwanza, and twice for Nkhotakota. On the other hand, since CRM is used after harvest, the CRM adoption decision was only observed twice for Machinga and Mwanza, and once for Nkhotakota. Thus, the sample on the use of CRM is limited. Figure 10, taken from the online appendix of BBMM, presents the full timeline of the project. More details on the intervention and sampling can be found in their paper.

### 3.2.2 AESTAS data

AESTAS is a nationally representative household survey conducted by International Food Policy Research Institute (IFPRI). Objective of this survey was to monitor the lead farmer program in Malawi[17]. The survey covered all 29 districts of Malawi[18], except Likoma. The data was collected in two waves: wave 1 in 2016 and wave 2 in 2018. The publicly available version of the survey dataset (citation needed) contain information from three different types of interviews:

1. **Household Interviews:** Random sample of around 10 households were selected for interview from randomly selected sections[19] within each district. Stratification was done based on whether or not the household had a lead farmer (LF). Per section, up to two households with LFs were selected. A total of around 299 sections were surveyed. The same households were interviewed in the two waves with very small level of attrition (around 4%). Around 3000 households were covered in wave 1, with 2880 being re-interviewed in wave 2. For each household, both household head and their spouses were interviewed. The survey collected information on household level technology adoption, awareness, exposure; access to extension services; as well as socioeconomic and household characteristics.

2. **Lead Farmer (LF) Interviews:** Around 531 LF households were selected for household interviews. These households were additionally interviewed with a sepa-

[17]More details needed on the lead farmer program.

[18]The survey considered Mzimba district to be divided into North and South, and Lilongwe district to be divided into East and West.

[19]Sections are geographical units in Malawi that are one level lower than districts.

rate semi-structured module within the household survey, in the first wave. These interviews collected information on the LFs characteristics, activities, roles, expectations, incentives, challenges, suggestions for improvement, support received from agricultural extension development officers (AEDOs), support received from other organizations etc.

3. **Community Interviews:** In addition to the household surveys, 2-4 village per community leaders were interviewed in each community. This survey was done in both waves. The objective was to collect community level information like number of lead farmers, type of training they received, number of projects, and other community characteristics.

For the purpose of this study, I will only be using the information collected through the household interviews. In particular, I am mostly interested in the data on household level technology adoption. Two different types of technology adoption information were collected in this survey:

1. Self-reported adoption for a list of pre-determined technologies and practices. This covered both agricultural and food processing practices.

2. Self-reported plot-level usage for a list of pre-determined agricultural technologies and practices.

This information help me create adoption indices that are crucial to my analysis (see Appendix C for more details). More information on the survey and associated sampling can be found in Ragasa and Niu (2017), Niu and Ragasa (2018), Ragasa (2020), and Ragasa et al. (2021).

## 3.3 The Identification Strategy

Instead of conducting the ideal experiment described above, I take the seeding strategy of BBMM as given. Given the selection of seeds in their experiment, I calculate the seeds' average centrality and probability of adoption which I use to run the following regression:

$$\begin{aligned} Y_v = \beta_0 &+ \beta_1 Seed\ Centrality_v + \beta_2 Seed\ Probability_v + \beta_3 Heterogeneity_v \\ &+ \beta_4 Seed\ Centrality_v \times Heterogeneity_v + \beta_5 Seed\ Probability_v \times Heterogeneity_v + \epsilon_v \end{aligned} \tag{9}$$

I calculate average centrality of the seeds at the village level by using eigen-vector centrality of the seeded households at the baseline. This information is available in the replication data of BBMM. I proxy for average probability of adoption for the seeds at the village level by predicted adoption and usage indices that I calculate at the baseline, using the estimates from another regression. These indices are calculated conditional on some observable household demographics. In calculating them I use the survey data from AESTAS. The next sub-section provides details on this. I use the coefficient of variation of the same adoption and usage indices at the village level to capture village level heterogeneity in terms of the applicability of a new technology. This is represented by the $Heterogeneity_v$ variable in (9). Note that, both the probability of adoption and the related coefficient of variation are proxied by variables that are calculated conditional on observable demographics. These variable are therefore not particular to any technology. Instead, they represent whether the household is likely to adopt *any* new technology conditional on observable characteristics. Following my hypothesis, I expect $\beta_4 < 0$ and $\beta_5 > 0$. The random error of the regression is captured by $\epsilon_v$.

The experiment in BBMM randomly assigned villages into seeding strategies. This means, the seed characteristics were randomly selected by construction. Hence, $Seed\ Centrality_v$ and $Seed\ Probability_v$ in (9) are exogenous sources of variation. However, the experimental design ensures that some villages will have more central seeds than the other. Thus, not accounting for the treatment status in the regression can lead to omitted variable bias if the treatment status itself affects the outcome variable controlling for the particular measure of centrality used. For this reason, I include the treatment dummies in my set of control variables for the regression. The $Heterogeneity_v$ variable is the coefficient of variation of the predicted adoption and usage indices at the village level. The predicted adoption and usage indices are calculated conditional on some observable household demographics. Thus, if households from certain demographics are more likely to adopt a new technology, not controlling for these demographics in the regression can lead to omitted variable bias in identifying $\beta_3$ in (9). However, my coefficients of interest are $\beta_4$ and $\beta_5$. If the $Seed\ Centrality_v$ and $Seed\ Probability_v$ are exogenous, then even with an endogenous $Heterogeneity_v$ the interaction terms' coefficients can be identified without any bias.

## 3.4 Strategy for Approximating Probabilities of Adoption

For (9), I need to calculate $Seed\ Probability_v$. It is the average probability of adopting a new technology for the seeds. The probability of adopting a new technology needs to be calculated for other households as well. The latter will help me capture the village level heterogeneity in terms of the applicability of the technology (denoted $Heterogeneity_v$ in (9)). However, BBMM did not collect any information about this probabilities directly, as their micro-foundation assumed the probabilities to be the same for all households. Hence, I need to find a way to be able to approximate these probabilities conditional on the observable characteristics of the households surveyed in their study.

This is where I use the data from AESTAS. The data contains information on technology adoption and household characteristics. It surveys nationally representative set of farmers in Malawi on a universe of technologies that includes the technologies covered in BBMM. I use this information on the universe of technologies to calculate $Adoption\ Index_{it}$ and $Usage\ Index_{it}$, for each household $i$ at time period $t$. Details on the construction of this index can be found in Appendix C. Once calculated, the indices are used in the following regression model:

$$Adoption/Usage\ Index_{it} = f(X_{it}; \mu_{it}) \tag{10}$$

where $X_{it}$ are household demographics that are available in the both the replication data from BBMM, as well as the AESTAS data. The term $\mu_{it}$ captures random error in the regression. In my preferred specification, I consider function $f(\cdot)$ to be linear (thus estimation uses ordinary least square). However, I check the robustness of my results with respect to non-linear specifications.

I use the estimations of this model to construct Adoption and Usage indices conditional on the $X_{it}$s available in the BBMM dataset. I use these variables as a proxies for the households' probability of adopting a new technology.

# 4 Results and Discussion

In this section, I present the empirical results of my analysis. The first subsection focuses on discussing how I approximate adoption probabilities using AESTAS data. Here, I present the relevant regression results and discuss the assumptions needed for using these results for the rest of my analysis. Next sub-section focuses on exploring key variables in the BBMM data via descriptive statistics. In the last sub-section, I present the main empirical results of this study.

## 4.1 Approximating Probabilities of Adoption

I start by comparing key baseline demographic information across datasets. This is presented in Table 2. These five variables are chosen as they are available in both AESTAS and BBMM data. In terms of the mean and median, both datasets are similar in the number of adults and children in the household. However, the BBMM sample is slightly richer than its AESTAS counterpart. This can be seen by comparing the mean and median of standardized housing, livestock and assets principal component analysis (pca) scores.

Table 3 presents the relevant estimation results for this subsection. Here, I estimate the Adoption and Usage Indices conditional on the demographics presented in Table 2. This is done using AESTAS data. Columns (1) and (2) present the results for $Adoption\ Index$, with and without the full set of household controls. Wealthier households that has their houses made of more modern materials (as opposed to traditional materials), own more livestock, and more assets, has a higher adoption index than their poorer counterparts. Adoption indices turn out to be higher for families with more adults and children as well.

Table 2: Baseline Demographics Across Datasets

| | | Variables | | | | |
|---|---|---|---|---|---|---|
| Dataset | Statistic | Adults | Children | Housing | Livestock | Assets |
| AESTAS | Mean | 2.14 | 3.00 | -0.09 | -0.03 | -0.03 |
| | (SD) | (1.00) | (2.00) | (0.98) | (0.99) | (1.00) |
| | Median | 2.00 | 3.00 | -0.29 | -0.40 | -0.29 |
| | Skewness | 2.59 | 1.00 | 0.54 | 3.74 | 0.31 |
| | Kurtosis | 16.76 | 6.01 | 2.01 | 26.67 | 1.79 |
| | Observations | 2820 | 2820 | 2803 | 2820 | 2820 |
| BBMM | Mean | 2.36 | 2.77 | -0.02 | 0.02 | 0.09 |
| | (SD) | (0.95) | (1.86) | (0.99) | (1.02) | (1.03) |
| | Median | 2.00 | 3.00 | -0.24 | -0.31 | -0.10 |
| | Skewness | 1.21 | 0.76 | 2.48 | 4.64 | 1.24 |
| | Kurtosis | 5.43 | 4.43 | 8.70 | 35.03 | 5.64 |
| | Observations | 5384 | 5407 | 5382 | 5407 | 5407 |

*Note:* The variables *Adults* and *Children* represent number of adults and children in a household, respectively. The variables *Housing*, *Livestock*, and *Assets* were standardized first principal components. For the AESTAS sample: *Housing* includes information on materials walls are made of, roof materials, and floor materials. Each of the three variables are coded to be 0- Traditional, 1- Modern. *Assets* includes the number of bicycles, radios and cell phones the household owns. *Livestock* includes the number of sheep, goats, chickens, cows, and pigs. For the BBMM sample: *Housing* includes information on materials walls are made of, roof materials, floor materials and whether the household has a toilet. *Assets* includes the number of bicycles, radios and cell phones the household owns. *Livestock* is an index including the number of sheep, goats, chickens, cows, pigs, guinea fowl, and doves. (footnote 1 from Table A5 of Beaman et al., 2021)

Table 3: OLS Regression Results for Adoption and Usage Indices

| | Adoption Index | | Usage Index | |
|---|---|---|---|---|
| Variables | (1) | (2) | (3) | (4) |
| Adults | 0.008*** | 0.005** | 0.011*** | 0.008*** |
| | (0.002) | (0.002) | (0.002) | (0.002) |
| Children | 0.003*** | 0.002 | 0.003*** | 0.002** |
| | (0.001) | (0.001) | (0.001) | (0.001) |
| Housing | 0.009*** | 0.007*** | 0.003 | 0.002 |
| | (0.002) | (0.002) | (0.002) | (0.002) |
| Livestock | 0.010*** | 0.005* | 0.014*** | 0.009*** |
| | (0.003) | (0.003) | (0.002) | (0.002) |
| Assets | 0.024*** | 0.017*** | 0.020*** | 0.014*** |
| | (0.002) | (0.002) | (0.002) | (0.002) |
| Constant | 0.061*** | 0.053*** | 0.130*** | 0.119*** |
| | (0.006) | (0.009) | (0.005) | (0.007) |
| Baseline Mean | 0.088 | 0.088 | 0.138 | 0.138 |
| (Standard Deviation) | (0.129) | (0.129) | (0.115) | (0.115) |
| Household Controls | No | Yes | No | Yes |
| Observations | 5610 | 5606 | 5610 | 5606 |
| R-squared | 0.096 | 0.149 | 0.085 | 0.123 |

*Notes:* * p<0.10, ** p<0.05, *** p<0.01. Robust standard errors clustered at the section level are in parentheses. All regressions use sample weights. The variables *Adults* and *Children* represent number of adults and children in a household, respectively. The variables *Housing*, *Livestock*, and *Assets* were standardized first principal components. *Housing* includes information on materials walls are made of, roof materials, and floor materials. Each of the three variables are coded to be 0- Traditional, 1- Modern. *Assets* includes the number of bicycles, radios and cell phones the household owns. *Livestock* includes the number of sheep, goats, chickens, cows, and pigs. Household Controls include: gender and age of household head, activity of household head (0- Non-Farmer, 1- Farmer), whether the household applied for a loan in the past, the households' time and risk preferences, and whether a household member is a lead farmer (LF).

However, the effect size is lower than that of the wealth indicators. The results are mostly highly significant with or without the household controls. Columns (3) and (4) present the results for *Usage Index*, with and without the full set of household controls.

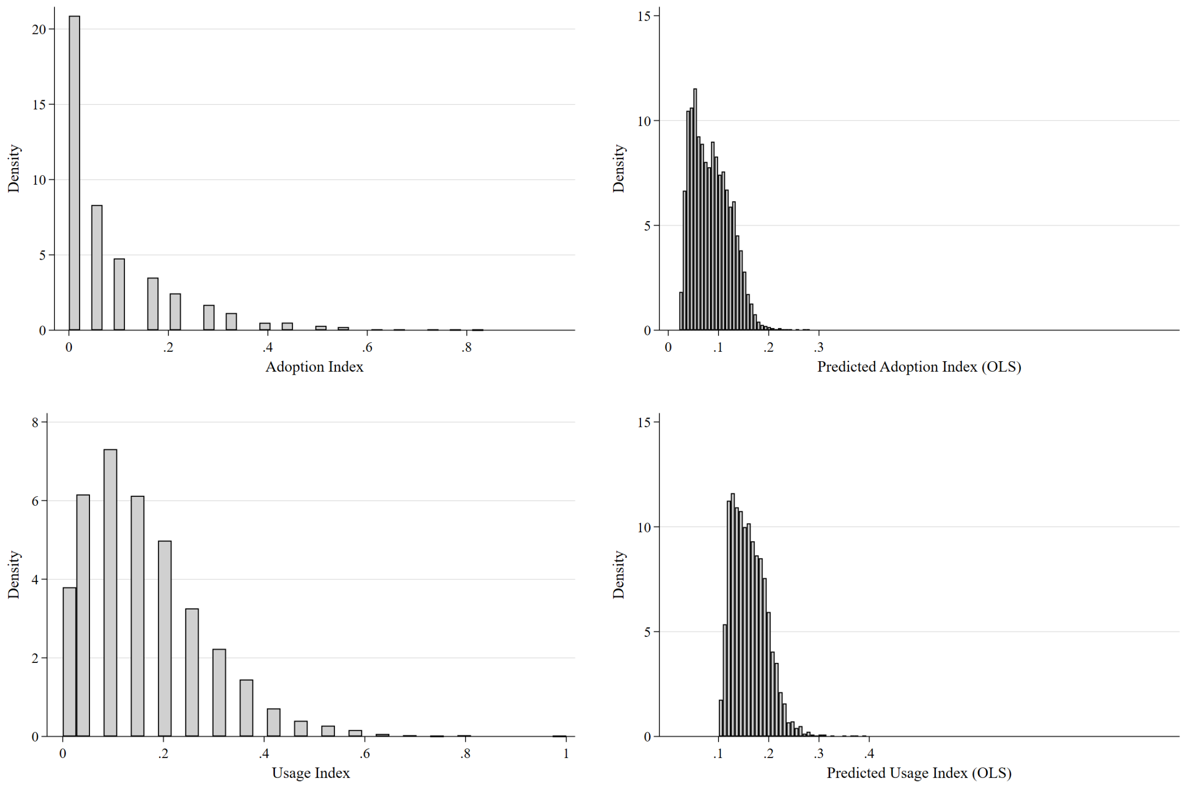



Figure 11: Actual and Predicted Adoption and Usage Indices

Figure 11 focuses on comparing the actual and predicted indices for the AESTAS sample. For calculating predicted indices, I use the estimates without household controls. It is important to note here that the estimates capture only a fraction of the actual variation. The actual adoption index has a mean of 0.085 with a standard deviation of 0.120. In comparison, its predicted counterpart has a mean of 0.086 with a standard deviation of 0.038 only. The numbers are similar for usage index in terms of quality of prediction (actual has mean 0.163 with standard deviation 0.122, predicted has 0.162 and standard deviation 0.035).Thus, predictions only represent 1/3 of the actual variation. This is not surprising given that the predictions are made based on a few observable demographics only.

I make the following assumptions in using estimates from Table 3 for predicting adoption and usage indices in the BBMM data:

- **Assumption 1:** Adoption and Usage indices are good proxies for probability of adoption.
- **Assumption 2:** The variation in adoption and usage indices that can be captured by the demographics described in Table 2, are sufficient for my analysis.
- **Assumption 3:** The mapping from the observable demographics to the adoption and usage indices are the same in the BBMM sample, as it is in the AESTAS sample.

## 4.2 Descriptive Statistics

This section focuses on describing key baseline characteristics in the BBMM sample.

Table 4: Baseline Village-level Sample Characteristics

| | Treatment Status | | | | |
|---|---|---|---|---|---|
| Variable | Complex | Simple | Geo | Benchmark | Overall |
| Adoption Rate (PP) | 0.030 | 0.029 | 0.029 | 0.018 | 0.026 |
| | (0.063) | ( 0.060) | (0.077) | (0.035) | (0.060) |
| Any Non-Seed Adopters (PP) | 0.340 | 0.320 | 0.420 | 0.300 | 0.345 |
| | (0.479) | (0.471) | (0.499) | (0.463) | (0.477) |
| Betweenness Centrality of Seeds | 0.235 | 0.187 | 0.129 | 0.178 | 0.182 |
| | (0.077) | (0.096) | (0.090) | (0.090) | (0.096) |
| Predicted Adoption Index of Seeds | 0.114 | 0.101 | 0.082 | 0.110 | 0.101 |
| | (0.036) | (0.041) | (0.025) | (0.034) | (0.036) |
| Predicted Usage Index of Seeds | 0.186 | 0.172 | 0.158 | 0.184 | 0.175 |
| | (0.032) | (0.042) | (0.024) | (0.031) | (0.035) |
| CV of Predicted Adoption Index | 0.378 | 0.379 | 0.366 | 0.389 | 0.378 |
| | (0.077) | (0.075) | (0.062) | (0.069) | (0.071) |
| CV of Predicted Usage Index | 0.188 | 0.185 | 0.180 | 0.193 | 0.187 |
| | (0.040) | (0.037) | (0.033) | (.039) | (.038) |
| Observations | 50 | 50 | 50 | 50 | 200 |

*Note:* Seed level measures are calculated using the average of two seeds, whenever the information on both seeds are available. Otherwise they reflect the information for one seed. Coefficient of Variations (CV) are calculated at the village level for the whole village. Adoption Rate and Any Non-Seed Adopters are calculated excluding seed or shadow farmers in a village.

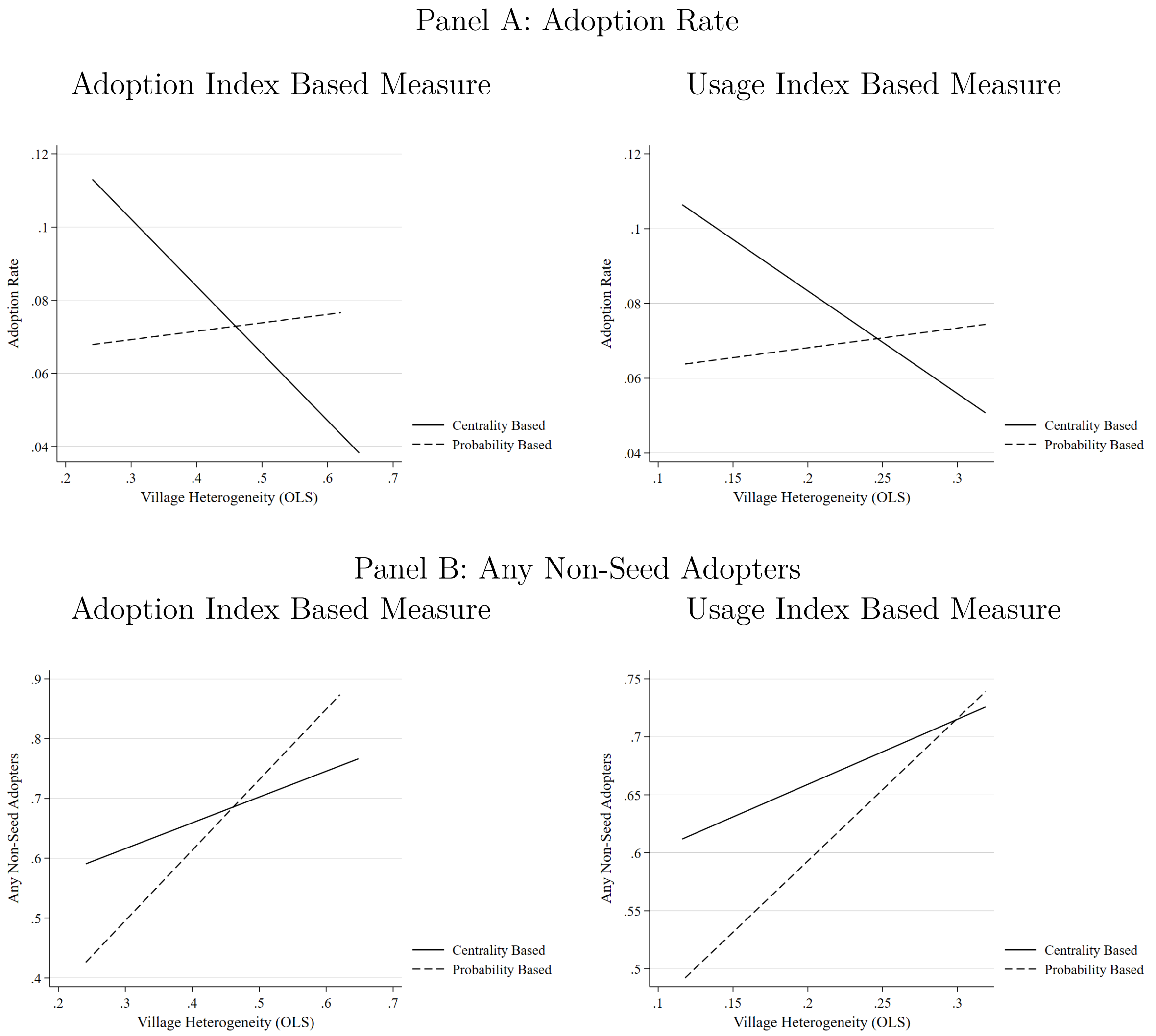



Figure 12: Outcomes for Different Seeding Strategies with respect to Village Heterogeneity

In figure 12, I present the main outcome variables over varying level of village heterogeneity. The figure distinguishes between seeds that can be considered either *centrality based* or *probability based*. For the purpose of this figure, *centrality based* seeds are defined to be the seeds that have higher than median level of mean eigen-vector centrality in the baseline. Similarly, *probability based* seeds are defined here to have higher than median level of predicted adoption (usage) in the baseline. Based on my simulations, I expect the *centrality based* seeds to perform worse, and the *probability based* seeds to perform better, as the village heterogeneity increases. For the variable *Adoption Rate*, this is the pattern I observe. However, for the variable $Any\ Non-Seed\ Adopters$, both type of seeds seem to perform better with increasing level of village heterogeneity. Although, in this case, the relative performance of *centrality based* seeds worsen as the village heterogeneity increases.

Which is what I hypothesize. However, this descriptive figures do not take into account the village-level heterogeneity in terms of other variables. In defining the *centrality based* and *probability based* seeds as dummy variables, the figures also fail to capture the full heterogeneity of these seeds' in terms of their centrality and probability measures. In the next sub-section, I present the reduced form results of my analysis that test my hypothesis more formally.

## 4.3 Reduced Form Results

Finally, I come to my main empirical results.

Table 5: Village level Regression 1 of Adoption Outcomes (Pit Planting)

| | Adoption Rate | | Any Non-Seed Adopters | |
|---|---|---|---|---|
| Variables | (1) | (2) | (3) | (4) |
| Eigen-vector Centrality of Seeds | 1.173** | 0.981* | 1.181 | 1.210 |
| (=$Seed\ Centrality_v$) | (0.581) | (0.517) | (1.439) | (1.396) |
| Predicted Adoption Index of Seeds | -2.973** | -2.087* | -8.019** | -3.832 |
| (=$Seed\ Probability_v$) | (1.467) | (1.226) | (3.257) | (3.337) |
| CV of Predicted Adoption Index | -0.296 | -0.093 | -0.928 | 0.669 |
| (=$Heterogeneity_v$) | (0.208) | (0.194) | (1.079) | (1.096) |
| $Seed\ Centrality_v \times Heterogeneity_v$ | -2.625** | -2.365** | -2.851 | -4.218 |
| | (1.324) | (1.158) | (3.777) | (3.714) |
| $Seed\ Probability_v \times Heterogeneity_v$ | 6.715** | 4.779* | 18.48*** | 8.921 |
| | (3.131) | (2.644) | (6.997) | (7.197) |
| Village-level Controls | No | Yes | No | Yes |
| Observations | 324 | 324 | 324 | 324 |
| R-squared | 0.080 | 0.190 | 0.049 | 0.210 |

*Note:* * p<0.10, ** p<0.05, *** p<0.01. Robust standard errors are in parentheses. All regressions include a constant term and year fixed effects. Village-level controls include percentage of village using pit planting at baseline, percentage of village using compost at baseline, percentage of village using fertilizer at baseline, village size, the square of village size, district fixed effects, and treatment dummies.

Table 6: Village level Regression 2 of Adoption Outcomes (Pit Planting)

| Variables | Adoption Rate (5) | (6) | Any Non-Seed Adopters (7) | (8) |
|---|---|---|---|---|
| Eigen-vector Centrality of Seeds | 0.999* | 0.891* | 0.984 | 1.047 |
| (=$Seed\ Centrality_v$) | (0.565) | (0.537) | (1.303) | (1.243) |
| Predicted Usage Index of Seeds | -2.174 | -1.502 | -4.599 | -0.645 |
| (=$Seed\ Probability_v$) | (1.410) | (1.213) | (3.317) | (3.274) |
| CV of Predicted Usage Index | -1.091 | -0.560 | -2.549 | 1.939 |
| (=$Heterogeneity_v$) | (0.805) | (0.712) | (2.905) | (2.998) |
| $Seed\ Centrality_v \times Heterogeneity_v$ | -4.481* | -4.438* | -4.874 | -7.720 |
| | (2.623) | (2.492) | (6.889) | (6.694) |
| $Seed\ Probability_v \times Heterogeneity_v$ | 10.33* | 7.660 | 23.13 | 4.858 |
| | (6.160) | (5.462) | (14.19) | (14.27) |
| Village-level Controls | No | Yes | No | Yes |
| Observations | 324 | 324 | 324 | 324 |
| R-squared | 0.063 | 0.186 | 0.037 | 0.203 |

*Note:* * p<0.10, ** p<0.05, *** p<0.01. Robust standard errors are in parentheses. All regressions include a constant term and year fixed effects. Village-level controls include percentage of village using pit planting at baseline, percentage of village using compost at baseline, percentage of village using fertilizer at baseline, village size, the square of village size, district fixed effects, and treatment dummies.

# 5 Summary and Concluding Remarks

In this study, I focus on optimal network based targeting strategies for improving technology adoption. Particularly when the new technology has more benefits to some agents than the others. Simulations help me form my hypothesis in such scenario, through the lens of my theoretical model. I hypothesized that the relative performance of different targeting strategies depend on the level of heterogeneity in the population. This is heterogeneity in terms of expected benefits of adopting the technology. In particular, I expect centrality based targeting to perform worse as the heterogeneity increase but, targeting based on probability of adoption to perform better in such scenario. I test the hypotheses using the replication data of BBMM collected from Malawi. To generate variation in the BBMM sample in terms of the applicability of a new technology, I use the AESTAS dataset also collected from Malawi. Reduced form results show support in favor of my hypothesis. On one hand, I observe that the positive impact of seed centrality decrease as village level heterogeneity increases. On the other hand, the increase in village level heterogeneity reduces the negative impact of seeds' with high probability of adoption.

In this reduced form analysis, I do not observe any positive impact of seeding based on likelihood of adoption. This is not surprising for the BBMM experiment, as the researchers focused on seeding based on centrality not probability. A more structural approach is needed to identify what would have happened if the seeding was done based on probability instead. For that purpose, and also for a clearer identification, a randomized controlled trial is more suitable.

In terms of the policy, the results underlined the importance of context in designing targeting strategies for technology adoption. The optimal policy requires taking into consideration the characteristics of the new technology together with the characteristics of the target population. If a new technology is such that there is sufficient heterogeneity in the population in terms of its applicability, targeting people central in the network may not be optimal. In such scenario, we may need to focus more on the sub-section of the population that is more likely to adopt. However, this may be more costly in practice as it require higher quality in data collection. Given the increase in the cost of targeting in such scenario, random seeding may turn out to be more attractive ex-ante. I leave that analysis for a future research.

# References


**Akbarpour, Mohammad, Suraj Malladi, and Amin Saberi.** 2021. "Just a Few Seeds More: Value of Network Information for Diffusion." *Unpublished Manuscript.*

**Bandiera, Oriana, and Imran Rasul.** 2006. "Social Networks and Technology Adoption in Northern Mozambique." *The Economic Journal*, 116(514): 869–902.

**Banerjee, Abhijit, Arun G. Chandrasekhar, Esther Duflo, and Matthew O. Jackson.** 2013. "The Diffusion of Microfinance." *Science*, 341(6144): 1236498.

**Banerjee, Abhijit, Emily Breza, Arun G Chandrasekhar, and Benjamin Golub.** 2018. "When Less is More: Experimental Evidence on Information Delivery During India's Demonetization." National Bureau of Economic Research Working Paper 24679.

**Banerjee, Abhijit, Emily Breza, Arun G. Chandrasekhar, and Markus Mobius.** 2021. "Naïve Learning with Uninformed Agents." *American Economic Review*, 111(11): 3540–74.

**Beaman, Lori, Ariel BenYishay, Jeremy Magruder, and Ahmed Mushfiq Mobarak.** 2021. "Can Network Theory-Based Targeting Increase Technology Adoption?" *American Economic Review*, 111(6): 1918–43.

**BenYishay, Ariel, and A Mushfiq Mobarak.** 2018. "Social Learning and Incentives for Experimentation and Communication." *The Review of Economic Studies*, 86(3): 976–1009.

**Besley, Timothy, and Anne Case.** 1993. "Modeling Technology Adoption in Developing Countries." *The American Economic Review*, 83(2): 396–402.

**Chandrasekhar, Arun G, Benjamin Golub, and He Yang.** 2018. "Signaling, Shame, and Silence in Social Learning." National Bureau of Economic Research Working Paper 25169.

**Chandrasekhar, Arun G, Horacio Larreguy, and Juan Pablo Xandri.** 2020. "Testing models of social learning on networks: Evidence from two experiments." *Econometrica*, 88(1): 1–32.

**Conley, Timothy G., and Christopher R. Udry.** 2010. "Learning about a New Technology: Pineapple in Ghana." *American Economic Review*, 100(1): 35–69.

**DeMarzo, Peter M., Dimitri Vayanos, and Jeffrey Zwiebel.** 2003. "Persuasion Bias, Social Influence, and Unidimensional Opinions*." *The Quarterly Journal of Economics*, 118(3): 909–968.

**Foster, Andrew D, and Mark R Rosenzweig.** 1995. "Learning by doing and learning from others: Human capital and technical change in agriculture." *Journal of political Economy*, 103(6): 1176–1209.

**Golub, Benjamin, and Matthew O. Jackson.** 2010. "Naïve Learning in Social Networks and the Wisdom of Crowds." *American Economic Journal: Microeconomics*, 2(1): 112–49.

**Jackson, Matthew O.** 2010. *Social and Economic Networks.* Princeton University Press.

**Munshi, Kaivan.** 2004. "Social learning in a heterogeneous population: technology diffusion in the Indian Green Revolution." *Journal of Development Economics*, 73(1): 185–213.

**Niu, Chiyu, and Catherine Ragasa.** 2018. "Selective attention and information loss in the lab-to-farm knowledge chain: The case of Malawian agricultural extension programs." *Agricultural Systems*, 165: 147–163.

**Ragasa, Catherine.** 2020. "Effectiveness of the lead farmer approach in agricultural extension service provision: Nationally representative panel data analysis in Malawi." *Land Use Policy*, 99: 104966.

**Ragasa, Catherine, and Chiyu Niu.** 2017. "The state of agricultural extension and advisory services provision in Malawi: Insights from household and community surveys." International Food Policy Research Institute (IFPRI).

**Ragasa, Catherine, Diston Mzungu, Kenan Kalagho, and Cynthia Kazembe.** 2021. "Impact of interactive radio programming on agricultural technology adoption and crop diversification in Malawi." *Journal of Development Effectiveness*, 13(2): 204–223.

# Appendices

## A Technical Details on the Simulation Method

## B Detailed Simulation Results

## C Construction of Adoption and Usage Indices

## D Robustness Checks

### D.1 Simulations

- Robustness with respect to using different measure of centrality.
- Robustness with respect to different number of households per village (50 vs 20).
- Robustness with respect to different measure of efficiency.

### D.2 Empirical Results

Calculation of predicted adoption and usage indices:

- Robustness with respect to non-linear regression specifications.
- Robustness of pooled regression with respect to regressions using observations from individual years.

Main regression:

- Robustness with respect to different measure of centrality.
- Robustness with respect to predicted adoption and usage indices calculated using different regression specification.