Optimal Network-Based Targeting for Technology Adoption in Developing Countries

Preliminary and Incomplete - Please do not Cite without Permission

Aranya Chakraborty*

May 2022

Abstract

How do we use existing social ties to improve the adoption of a new technology? I explore optimal network-based targeting when the expected benefit associated with the new technology varies at the household level. In particular, I focus on two types of targeting: targeting households 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 agents 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 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.

JEL Codes: D83, O13, O33, Q16

Keywords: Targeting, Social Network, Technology Adoption, Agriculture

1 Introduction

Technology adoption in agriculture has been recognized as a driving force of economic growth via its effect on structural transformation (Bustos et al., 2016). However, the

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

adoption of modern technologies has been low in developing regions, especially in Sub-Saharan Africa (Bold et al., 2017). Information constraints are argued to be one of the key reasons behind such phenomenon (Magruder, 2018). 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. 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. In such a scenario, the literature recommends targeting agents central to the network (Beaman et al., 2021a). 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. In particular, I focus on the situation where the new technology can be more beneficial to some agents than others. The benefits can vary across agents due to several possible reasons. The agents can differ in terms of their education, skills, and ability that affect how much they can learn about a new technology and use it in practice. They can also vary in terms of other characteristics, e.g., land quality (for agriculture), size of operation (for both farm and firm households), access to infrastructure (such as road and irrigation facilities), access to other technologies. For my purpose, I consider heterogeneity in benefits to be dependent on the existing network structure. In doing so, I essentially assume the existing network structure to be representative of the sorting of agents with respect to their observable and unobservable characteristics. 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 probability of adoption.

My study makes three contributions to the existing literature. First, I provide evidence (both theoretical and empirical) that the success of network-based targeting strategies depends on the population level heterogeneity. Diffusion of information via networks is the key to increasing technology adoption (Besley and Case, 1993; Foster and Rosenzweig, 1995; Conley and Udry, 2010; Krishnan and Patnam, 2013). In recent years, there has been a number of studies focusing on the role of networks in the diffusion of technologies.[1] 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., 2021a; Akbarpour et al., 2021). However, these studies assume existing network ties to be the only factor characterizing diffusion. Thus, for the purpose of diffusion, households are assumed to be homogeneous in terms of other characteristics. In the current study, I consider the population to

[1]See Cheng (2021) for a review of the existing literature.

be heterogeneous in terms of the benefits they get from the new technology, with this heterogeneity having a direct effect on the effectiveness of targeting strategies. In such a scenario, I show evidence that 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 population-level heterogeneity in terms of the benefits from the new technology. Considering population level heterogeneity in social learning itself is not new (e.g., Munshi, 2004; Bandiera and Rasul, 2006; Conley and Udry, 2010).[2] However, to the best of my knowledge, the current study is the first to consider the consequences of population level heterogeneity on network-based 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.[3] Thus, simplifying assumptions are made such that the learning involves the characteristics that are similar for all the agents and not the characteristics that differentiate them. This assumption helps us to focus on a problem where the agents are collectively trying to uncover some hidden characteristics of interest (e.g., in the theoretical models of Acemoglu et al., 2008 and Golub and Jackson, 2010). In many scenarios, however, agents do face heterogeneous benefits in adopting a new technology (Suri, 2011). For example, in agriculture the performance of some practices depend on the quality of land.[4] Thus, farmers vary in terms of the benefits they derive in adopting those practices depending on the quality of their land. The consequences of heterogeneity in benefits on network-based targeting have not been formally studied in the existing literature. The current study attempts to close this gap.

Finally, I provide policy directions for network-based targeting when the population is heterogeneous. In particular, I argue in favor of targeting early adopter households when the heterogeneity is high and central households when the heterogeneity is low.[5] This contributes to the literature that focuses on understanding the characteristics and impact of

[2]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.

[3]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.

[4]In Munshi (2004), the adoption of new rice varieties is sensitive to growing conditions. Tjernström (2017) shows that soil quality heterogeneity affects farmers' ability to learn from their peer's experimentation with the new technology. Pit planting studied in BenYishay and Mobarak (2018) and Beaman et al. (2021a) requires flat land.

[5]Early adopter households are defined here as the households that are more likely to adopt a new technology given homogeneous cost, similar to the definition of natural early adopters in Catalini and Tucker (2017).

opinion leaders in diffusing new knowledge. In this literature, studies like Maertens (2017) and Miller and Mobarak (2015) show the learning to be more effective when the opinion leaders are in some way *superior* than their followers. On the other hand, BenYishay and Mobarak (2018) show that communicators who share a group identity with the farmers or face comparable agricultural conditions, do a better job at convincing farmers to adopt a new technology. Feder and Savastano (2006) takes a middle ground in arguing that the most effective opinion leaders are superior to the followers, but not excessively so. My study contributes to this debate from a network-based targeting perspective.

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[6] 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 it is 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 et al., 2020). The two-stage decision process is also something that can be found in the existing literature (Chandrasekhar et al., 2018).

Based on my theoretical model, I use simulations to evaluate the relative importance of different targeting strategies and to generate testable hypotheses.[7] I test these predictions combining two different data sources from Malawi. The first one is the replication data (Beaman et al., 2021b) from a randomized controlled trial (RCT) conducted by Beaman, BenYishay, Magruder, and Mobarak (2021a) (henceforth, BBMM). The second dataset is the Agricultural Extension Services and Technology Adoption Survey (henceforth, AESTAS) data (IFPRI, 2021a,b) collected by International Food Policy Research Institute (IFPRI). 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 are only realized after adoption, they cannot be factored into the targeting strategies. I attempt to solve this issue by using AESTAS data to estimate

[6] DeGroot learning refers to a social learning process whereby agents form beliefs/ opinions as a weighted average of the beliefs/ opinions of people they are connected to (including themselves). Here the weights correspond to how much the agents are influenced by one another. It is a heuristic, as agents do not account for the interdependence of beliefs between each of the people they are connected to (Barnett-Howell and Mobarak, 2021). More information on this type of learning can be found in Chapter 8.3 of Jackson (2010).

[7] The use of simulations is not new to the network literature. For example, Bala and Goyal (1998) use simulations to generate spatial and temporal patterns of adoption when individuals learn from their neighbours; Acemoglu et al. (2011) uses simulations to show that innovations might spread further across networks with a smaller degree of clustering. Similar to Beaman et al. (2021a), I use them to understand the effectiveness of targeting strategies a few periods down the line.

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 these estimates in the BBMM data to calculate the household level likelihood of adoption. BBMM data is used as their experiment relies on exploiting the centrality of seeds to improve adoption of a technology suitable for my analysis, thus including all other information that I need. Once I calculate households' probability of adoption in the BBMM data using estimates from the AESTAS sample, I exploit both the village-level and experimental variations in the BBMM data 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 the likelihood of adoption should perform better. The intuition behind such result lies in the characteristics of the central seeds in a network.[8] 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 favor of my hypothesis. Exploring village-level variations in the BBMM data, I show that the positive effect of seeds' centrality on the adoption of pit planting decrease with increase in village-level heterogeneity in terms of probability of adoption. Simultaneously, the negative effect of seeds' probability of adoption decreases with increased village-level heterogeneity. Weaker, but similar results are found when I shift my focus to exploring experimental variations.

The remainder of this article is organized as follows. In Section 2, I present the theoretical framework of my analysis. Section 3 presents the simulations that help me form the main hypotheses for this study. In Section 4, I discuss the hypotheses, my empirical strategy for testing them, and the data I use in the process. Section 5 presents and discusses my empirical results. Finally, in Section 6, I summarize my findings and make concluding remarks.

2 Theoretical Framework

I consider a choice problem that requires learning in a social network. The problem is that of technology adoption when the expected benefits associated with a new technology vary at the household level. In particular, expected benefits are such that the technology dominates existing technologies for only a sub-section of the population. On top of this,

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

the true expected benefits with the technology are initially unknown to 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 DeGroot (1974) and DeMarzo et al. (2003)). I focus on the scenario where the underlying state is time-varying, similar to Acemoglu et al. (2008). Like Banerjee et al. (2021) my model considers both informed and uninformed agents, where agents decide whether to get informed about the new technology.[9] In addition, I consider the possibility that agents are heterogeneous in terms of their distribution of payoffs associated with the new technology.

2.1 The Theoretical Model

Consider a two-stage decision process where in the first stage the households 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 π^T, whereas 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 ω_{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 (more details on this below).

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 household 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. This means that there is enough heterogeneity in the population such that for some households 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 ω_{it} at period t. Every

[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 et al. (2018). Which also forms the micro-foundation of Beaman et al. (2021a).

[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.

period, an uninformed household[12] has the option to become informed by applying effort $e_{it} \in \{0, 1\}$. Households put effort only once, i.e., if $e_{i\tau} = 1$, $e_{it} = 1\ \forall t \geq \tau$. If $e_{it} = 1$, the household learns the true distribution $p_i^*(\omega_{it})$ at cost η_i. The cost of learning is incurred only once - the first time the household gets informed. If $e_{it} = 0$, no effort cost is incurred and the household 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 $\hat{p}_{it} = \sum_{j \in \mathcal{I}} G_{ij} p_{jt-1}$ denotes household i's approximation based on others' opinion following the DeGroot averaging. The DeGroot averaging is rational here as the true distributions are positively correlated between the households such that: $p_i^* = \sum_{j \in \mathcal{I}} G_{ij} p_j^*$.

The belief of household i at period t is determined by the following process:

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

Thus, uninformed households use DeGroot averaging to approximate the true distribution with the help of their peers. On the other hand, informed households 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 households need to be informed before they adopt. As can be seen below, this assumption helps me explicitly capture the point when the households stop seeking information from their peers.

I assume the households to be risk-neutral and myopic. The assumption of risk-neutrality is for simplification purposes only, as it allows us to 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 households 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 households 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 households help me focus on a static model instead of a dynamic one. More importantly, if the households 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.

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

[12] I make no assumptions on the initial number of informed households. In reality, whether or not a household is informed might depend on their education, skills, and abilities. As I will argue in the next-subsection, from a policy-perspective, I am interested in the scenario where the majority (if not all) of the households are uninformed about the new technology to begin with.

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 households make this decision.

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

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

Finally, I assume the following timeline of decision making:

1. Every period, uninformed household i decides whether or not to get informed. Informed households do not need to get informed as they already know their true probability distribution.

2. To decide, uninformed households collect information on beliefs from their peers $j \in \mathcal{I}$. These beliefs p_{jt-1} were formed in the last period (some informed, some uninformed). Using their peers' beliefs from the last period, the uninformed household i uses DeGroot averaging to calculate $\hat{p}_{it} = \sum_{j \in \mathcal{I}} G_{ij} p_{jt-1}$.

3. On the basis of $\hat{p}_{it}$, they decide whether or not to become informed.

4. If they do not get informed ($e_{it} = 0$), their new belief is formed to be equal to the DeGroot average ($p_{it} = \hat{p}_{it}$), and next period they repeat from 1. If they get informed ($e_{it} = 1$), they now know their true probability distribution (p^*_i) and makes adoption decision on the basis of that. The true probability distribution also becomes their belief from next period onward ($p_{is} = p^*_i \; \forall s \geq t$).

2.2 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 ω_H), and the other where the new technology has a lower payoff than the traditional one (denoted ω_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 if and only if:

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

That is, if and only if 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 if and only if:

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

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) is not satisfied 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 households, the following condition must be true:

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

Equation (6) implies that for households that 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, that $\bar{p}^*_{jH} + \bar{\eta}_j \geq p^*_{jH} \geq \bar{p}^*_{jH}$. Then even if p^H_{jt} is equal to p^*_{jH}, household j will end up not getting informed about the technology. Hence they will not adopt the technology, even if it is profitable for them to do so. This is due to the positive cost of learning (η_j). This feature is similar to the models of Chandrasekhar et al. (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. In particular, the equilibrium depends on the households' initial beliefs. If everyone except household i is informed, DeGroot averaging in this set-up will help household i to correctly decide whether or not to get informed. The problem, however, arises when most households are uninformed. Of particular interest is the situation when $p^H_{it} \approx 0 \ \forall it$. This occurs when everyone believes that, for them, the new technology yields a 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.

Network-based targeting can help in such scenario. A policy can be designed to exogenously target some households (seeds) to be informed for improving adoption. The informed household i will learn about their p^*_{iH} at period t, which will get household j to update their $\hat{p}^H_{jt+1}$ if j puts positive weight on i's opinion. This will in turn cause household k to update their $\hat{p}^H_{kt+2}$ 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 targeting strategy. In other words, following the initial seeding strategy, the outcomes will vary depending on the path of information diffusion. In such a scenario, for any given targeting strategy simulations help in attaining the outcomes. These outcomes can then help in understanding the relative effectiveness of different targeting strategies.

In the next section, I measure the relative performance of two types of such targeting strategies using simulations. In doing so, I consider the networks of households that face the decision problem described in this section. I focus on the scenarios where initially $p^H_{it} \approx 0 \ \forall it$, and thus the need for targeting. My simulations provide testable implications that are taken to the data in subsequent sections.

3 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 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.

3.1 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

[13]Here, similar to section 2.2, I am assuming two states of the world: either success or failure with the new technology.

their states of the world are independently drawn every 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 the true probabilities of success are shown in Figure 1.

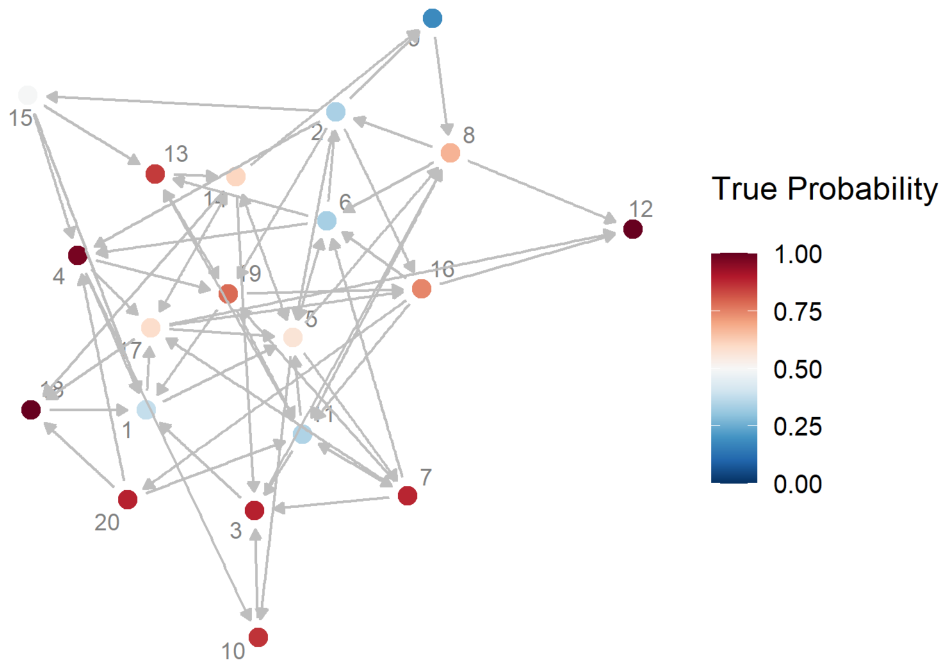


Figure 1: Distribution of True Probability within the network

Consider the scenario where, prior to any interventions, everyone believes their probabilities of success with the new technology to be zero ($p_{it}^{H} = 0\ \forall it$). Under such a 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 an intervention would be to ensure that the households that should adopt the technology under perfect information, end up adopting it. 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* captures 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 households adopt if and only if a certain threshold of their connections adopt (Beaman et al., 2021a). I consider probability-based targeting as an alternative to this. The probability-based targeting strategy is to seed households that have the highest true probabilities of success with the new technology (i.e., highest p^*_{iH}s in the network). These are the households who are more likely to adopt a technology given a homogeneous cost of learning for everyone and hence are considered to be the early adopters here (definition of early adopters similar to Catalini and Tucker, 2017). I consider this strategy for two reasons. First, it is the extreme opposite of the centrality-based targeting strategy. Whereas the centrality-based strategy relies on households that are similar to the average for diffusion, the probability-based strategy does the opposite by relying on households that are more likely than average to adopt a technology. Second, there is a debate in the existing literature regarding whether opinion leaders should be somewhat *superior* than there followers for effective diffusion of new knowledge (Feder and Savastano (2006); Miller and Mobarak (2015)). Through the lens of this debate, probability-based targeting seems to be a natural alternative for centrality-based targeting.

The centrality-based targeting strategy is to seed households that are central to the network. 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 an eigen-vector based measure of centrality (consult Appendix E 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 Appendix A.[14] The robustness of my results with respect to an eigen-vector based measure is important for two reasons. First, there is evidence in the existing literature in favor of targeting using eigen-vector based measures of centrality (e.g., Banerjee et al. (2013); Beaman et al. (2021a)). 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

[14] For a more detailed description of network centrality measures, consult section 2.2.4 of Jackson (2010) and Bloch et al. (2021).

the betweenness centrality measure is more simplistic which allows me to explain my theoretical results more intuitively. On top of this, alternating between centrality measures help me establish the robustness of my theoretical and empirical results.

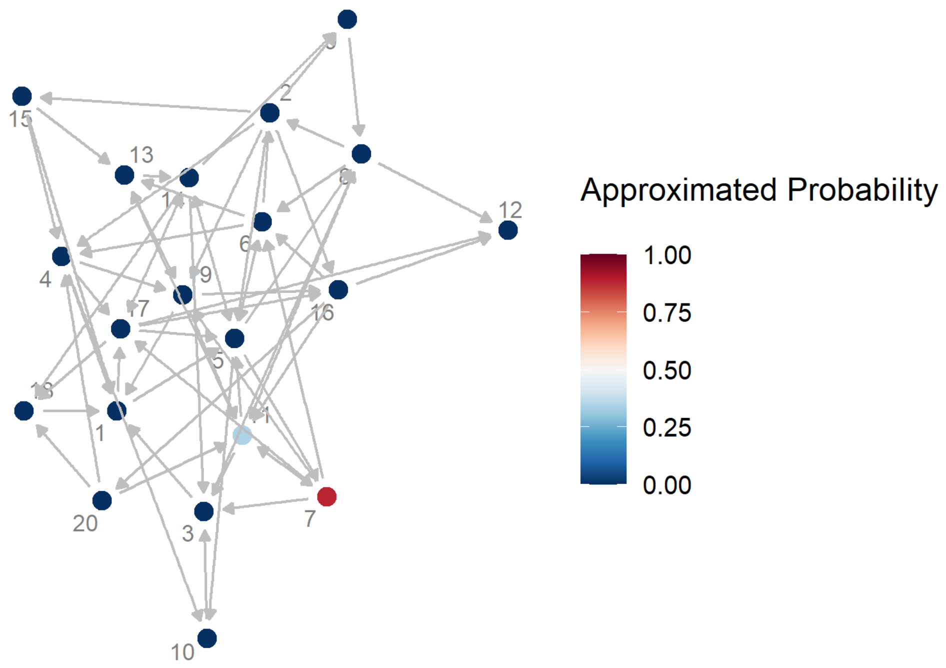


Figure 2: Seeding based on Centrality

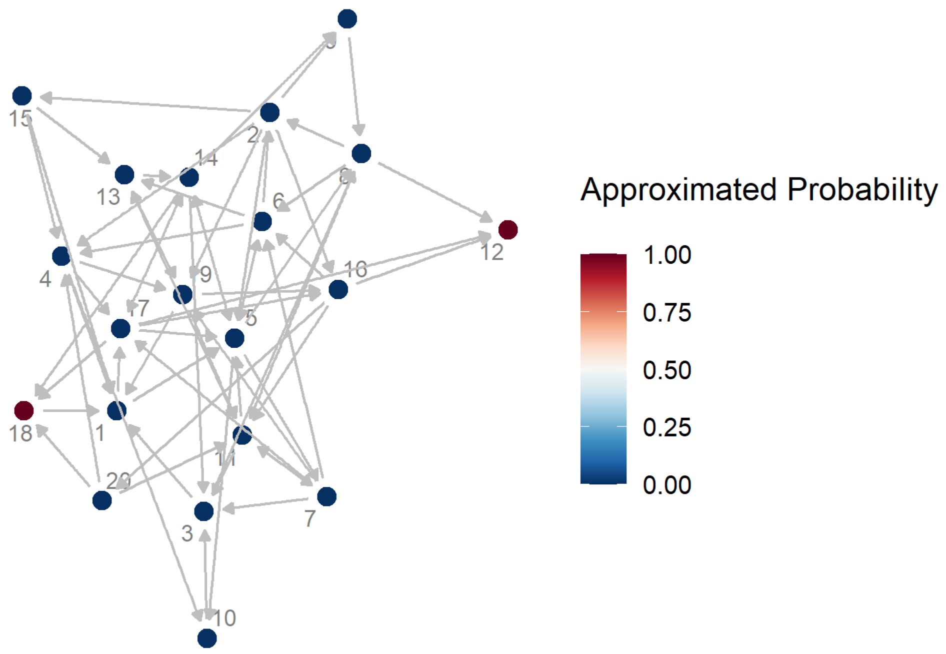


Figure 3: Seeding based on Probability

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 a 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 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 4.4. The seeds here are not so well connected in the network and represent early adopters.

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 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.

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 the p^*_{iH}s. The objective of this exercise is to note the relative performances of centrality and probability-based targeting strategies over varying degrees 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 the 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.

Centrality-Based　　　　　　　　Probability-Based

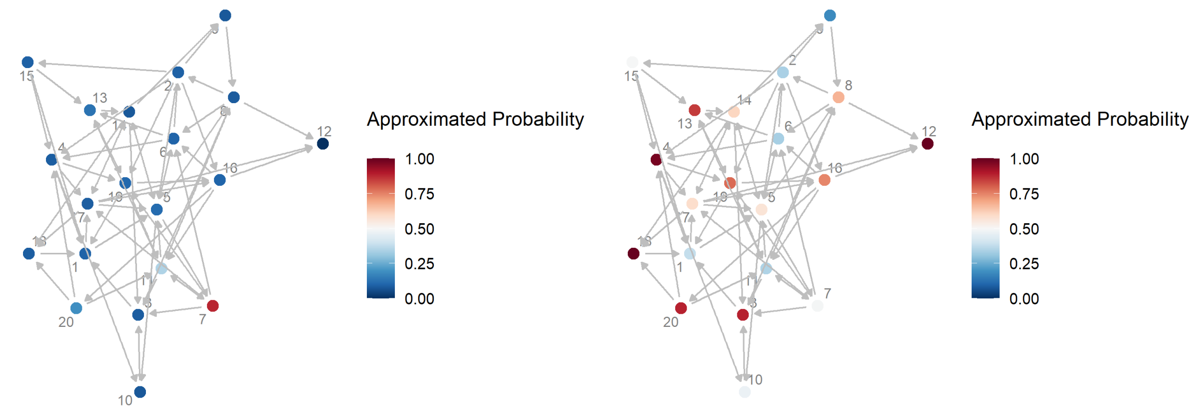


Figure 4: Performance of seeds after 10 periods

3.2 Targeting Homogeneous Networks

Centrality-Based　　　　　　　　Probability-Based

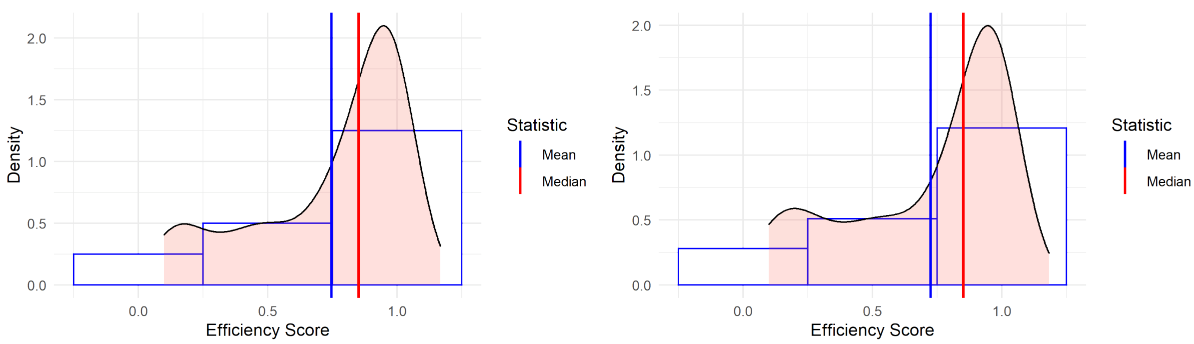


Figure 5: Distribution of efficiency scores when probabilities are perfectly correlated

Figure 5 represents the benchmark case of perfect correlation with respect to the p^*_{iH}s. The last column of table 1 presents 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 the probability-based measure (detailed results are provided in Appendix E).

These results are what is expected in perfectly homogeneous networks. When the p^*_{iH}s are perfectly correlated, my model reduces to one 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 favor 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	0.75
	Median	0.75	0.90	0.95	0.85
	Variance	0.16	0.10	0.09	0.08
Probability-Based	Mean	0.93	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

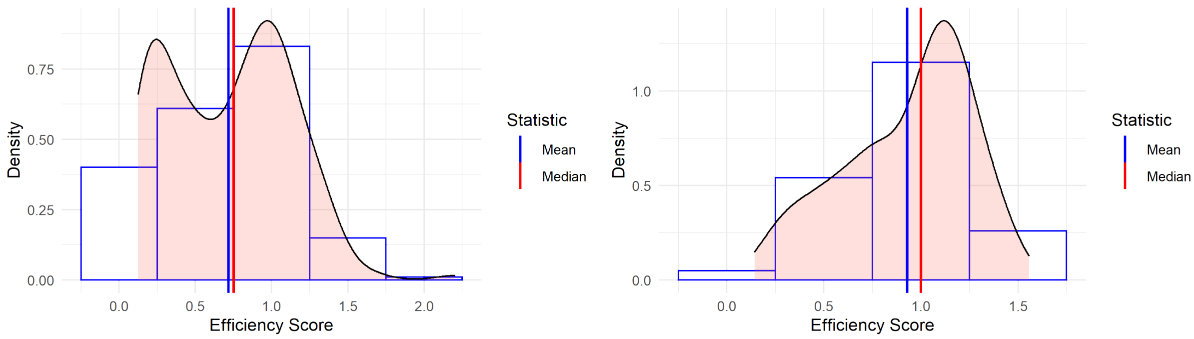
Notes: Simulations on varying levels of correlation are all done for 200 networks, each containing 20 households. However, upon generation of the true probabilities, some networks are dropped as they contained 0% of informed households under full efficiency.

3.3 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.

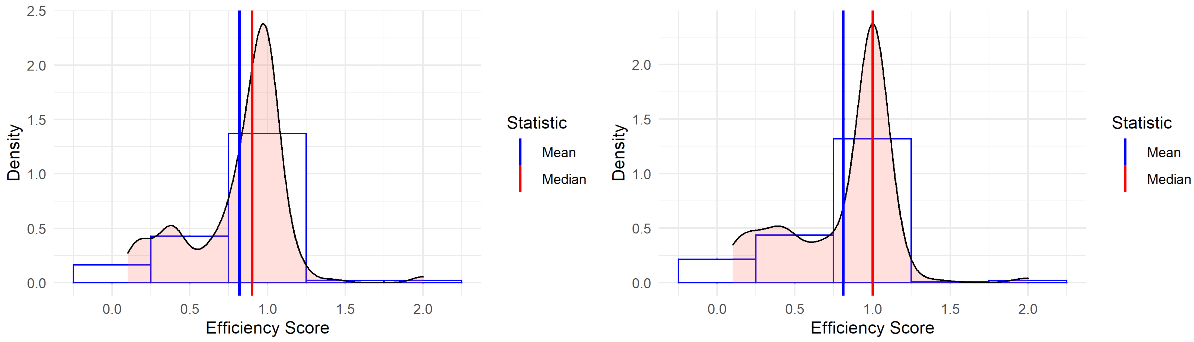
[15] Complex diffusion models assumes that information diffuses to a household if and only if a certain threshold of the household's connections get informed. A more detailed description of different models of diffusion and their use in Development and Agricultural Economics literature can be found in Breza et al. (2019) and Barnett-Howell and Mobarak (2021).

Centrality-Based | Probability-Based



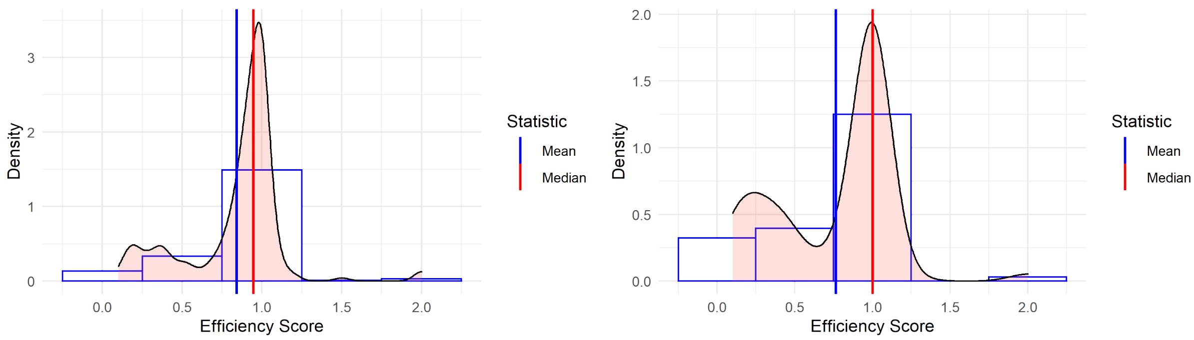
Low Level of Correlation

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 (0.93 vs 0.72) and median (1 vs 0.75) efficiency scores, with a lower variance (0.11 vs 0.16). 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 minimal. 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.

3.4 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 heterogeneity is high, there is a clear distinction between the households that should adopt the technology and the ones that should not. In that scenario, targeting households that are central to the network and hence, representative of the average household, is not efficient. Targeting households that are more representative of the households who should adopt the technology is more efficient. On the contrary, when the heterogeneity is low, everybody in the network are similar in terms of the applicability of the new technology. Targeting households 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 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, whereas the results in the last column do not.

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, 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, and larger networks (discussed in Appendix E). 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 heterogeneity within the network increases.

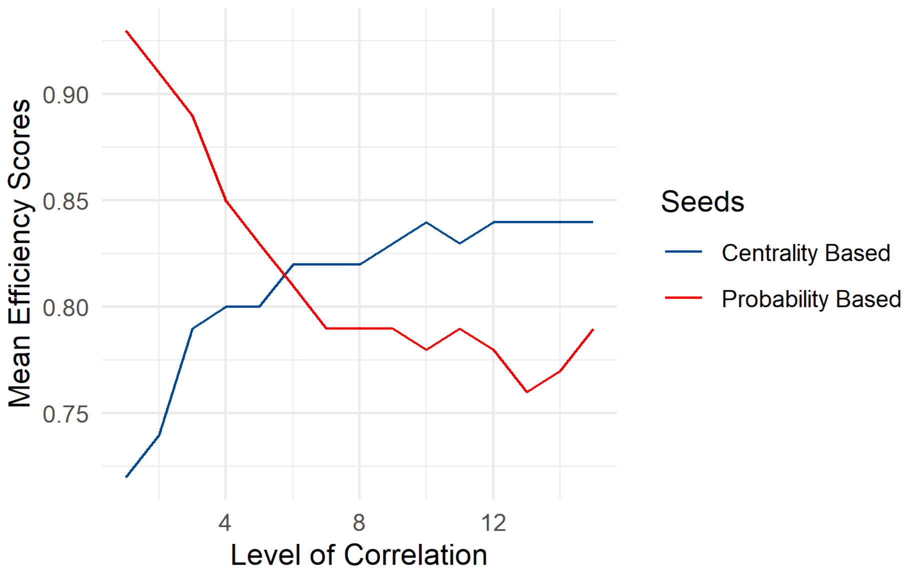


Figure 7: Mean efficiency scores over increasing levels of correlation

Figure 8 graphs the median efficiency scores over a range of different levels of correlation. Here, probability-based targeting always outperforms centrality-based targeting. The result is weakly robust for a different measure of centrality, however, it does not hold for larger networks. The fact that relative performance of the targeting strategies vary for mean and median measures indicate that there may be a distributional shift across different levels of correlation. Figure 9 captures the distributions by focusing on the variance of efficiency scores over different levels of correlation. It can be seen that for low level of correlations, both targeting strategies have high variance. This drops and becomes stagnant as the level of correlation increases. These results are robust to a different measure of centrality and larger networks.

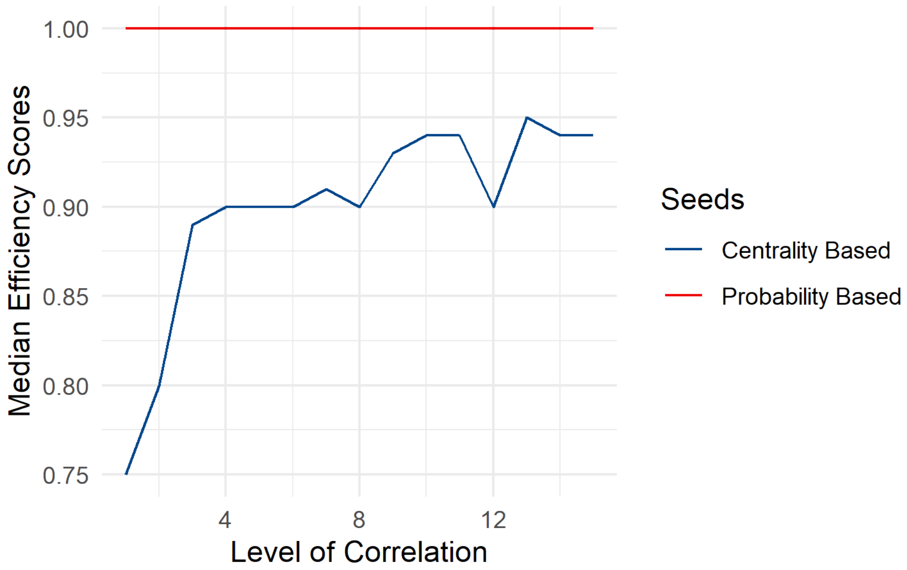


Figure 8: Median of efficiency scores over increasing levels of correlation

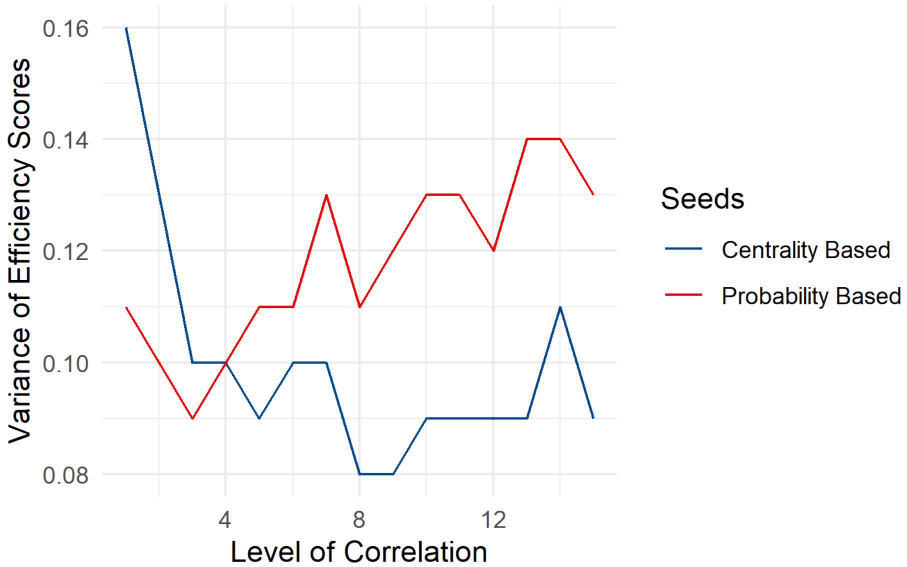


Figure 9: Variance of efficiency scores over increasing levels of correlation

In terms of the efficiency scores, a higher mean and lower variance is desirable. A higher mean indicates better performance on average. A lower variance is an indicator of lower deviation around the average performance. Here, the results show that the centrality-based strategy has higher variation at low levels of correlation, compared to probability-based targeting. However, for high levels of correlation, I observe the opposite. Thus, the results are similar to that of the mean measure. That is, the centrality-based strategy is *better* for high levels of correlation; the probability-based strategy is *better* for low levels of correlation. In the rest of this paper, I take this result to the data.

4 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.

4.1 The Ideal Experiment

Before diving into the data and identification strategies used in this study to validate my theoretical findings, it may be useful to think about the ideal experimental set-up for this purpose. In an experimental set-up, I 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 run the following reduced-form regression:

$$\begin{aligned} Y_v = \alpha_0 + \alpha_1 Centrality\ Based_v + \alpha_2 Probability\ Based_v + \alpha_3 Heterogeneity_v \quad (8) \\ + \alpha_4 Centrality\ Based_v \times Heterogeneity_v + \alpha_5 Probability\ Based_v \times Heterogeneity_v + u_v. \end{aligned}$$

Here Y_v denotes some adoption related outcome variable of interest, for village v. $Centrality\ Based_v$ and $Probability\ Based_v$ dummies indicate whether the village got assigned to either centrality or probability-based seeding strategy. $Heterogeneity_v$ is the village-level Coefficient of Variation (CV) of the probability of adoption. This captures the village level heterogeneity in terms of the applicability of a new technology. Finally, u_v is a random error term in the regression. My hypotheses state that $\alpha_4 < 0$ and $\alpha_5 > 0$.

4.2 Data Sources

I do not have access to the data from the ideal experiment described above. Hence, I use the replication data from BBMM together with the survey data from AESTAS conducted by IFPRI. In this subsection, I describe these datasets before proceeding to the description of my identification strategies in the next subsection.

4.2.1 Replication data of BBMM

BBMM conducted a 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 consisted in training the seed farmers on PP and CRM, with the training remaining the same across different treatment arms. The villages were equally divided into 4 experimental 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.

It is important to note that this experimental set-up focuses on selecting seed households solely based on their centrality measures. From the perspective of information diffusion, households were assumed to be homogeneous otherwise. This is the key assumption I relax for my analysis. I consider households to be heterogeneous in terms of their expected benefits from the new technology, with this heterogeneity having a direct effect on the diffusion of information for a given seeding strategy.

The researchers first collected the social network census data in 2010-11, 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 matched these 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 was allocated to. Once the training was complete, they conducted household survey to collect data on farming techniques, input use, yields, assets, and other characteristics.

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 used for land preparation, the households' adoption decision of PP was observed three times 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. More details on the intervention and sampling can be found in their paper.

The objective of the data was to assess the effectiveness of different centrality-based targeting strategies on the adoption of pit planting. For that purpose, detailed data was collected on household-level adoption decisions over different survey rounds. The replication package also includes information on household-level measures of centrality that were used to select seeds under different experimental interventions. The former helps me form the main dependent variables for my analysis, while the latter helps by providing

[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.

the information required to assess the centrality of seed households given the experimental set-up. In addition to this set of information, my analysis requires the assessment of seed households in terms of their ex-ante probability of adoption. This information is not available in the data as Beaman et al. (2021a) do not consider the population-level heterogeneity in designing targeting strategies. For this purpose, I turn to the following dataset.

4.2.2 AESTAS data

AESTAS is a nationally representative household survey conducted by International Food Policy Research Institute (IFPRI). The 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 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 the 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 separate 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 the

[17]More details on the lead farmer program can be found in Khaila et al. (2015).

[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.

number of lead farmers, type of training they received, number of projects, and other community characteristics.

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).

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 D for details on the construction of these indices). I use these indices as proxies for probability of adoption.

4.3 Identification Strategies

I now turn to discuss the identification strategies of my empirical analysis. The empirical analysis uses both between and within treatment group variations in the BBMM sample. The objective of this analysis is to use the experimental set-up of BBMM to test the predictions of my simulations. In doing so, contrary to BBMM which focus on comparing the effectiveness of different centrality-based targeting strategies, I focus on the comparison of the centrality-based vis-à-vis probability-based targeting strategies for varying level of population heterogeneity.

In this subsection, I will first focus on discussing how I explore the overall village-level variations in the data. These variations are not experimentally induced. Thus, identification using them requires some assumptions that are also discussed. Next, I focus on the identification using between treatment group variations. Both identification strategies require calculating probability of adoption at the household level. For that purpose, I use the survey data from AESTAS. The next sub-section provides details on that.

4.3.1 Exploring village level variations

Given the selection of seeds in the BBMM experiment, I calculate the seeds' average centrality and probability of adoption. This information is used in the following regression:

$$
\begin{aligned}
Outcome_{vt} &= \beta_0 + \beta_1 Centrality_v + \beta_2 Probability_v + \beta_3 Heterogeneity_v \\
&+ \beta_4 Centrality_v \times Heterogeneity_v + \beta_5 Probability_v \times Heterogeneity_v + \lambda X_v + \zeta_t + \epsilon_{vt}.
\end{aligned} \tag{9}
$$

$Outcome_{vt}$ denotes some adoption related outcome for village v at time t. Similar to BBMM, I focus on the outcomes in years 2 and 3. The outcome variables used for my analysis are discussed in the next section. $Centrality_v$ represents the average centrality of the seeds for village v, at the baseline. I calculate this by using eigen-vector centrality of the seed households at the baseline. The centrality measures are pre-calculated by BBMM and available in their replication data. $Probability_v$ represents the average probability of adoption for the seeds for village v, at the baseline. I proxy for probability of adoption using predicted adoption and usage indices. I calculate these indices at the baseline, conditional on some observable household demographics. The calculation uses estimates from another regression, which will be discussed in the next subsection. 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). It is important to 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 $\beta_5 > 0$. I control for baseline village level characteristics (X_v), as well as year fixed effects (ζ_t). The random error of the regression is captured by ϵ_{vt}.

The calculation of outcome variables exclude the seeded households. $Centrality_v$ and $Probability_v$ are calculated using the information on the seeded households. I assume that seed characteristics are exogenous to the outcome variables. This is a reasonable assumption as the village-level outcomes do not take seeded households into consideration. I assume that, conditional on these village level controls, $Heterogeneity_v$ is also exogenous in (9). In any case, my coefficients of interest are β_4 and β_5. If $Centrality_v$ and $Probability_v$ are exogenous, then even with an endogenous $Heterogeneity_v$ the interaction terms' coefficients can be identified without any bias. Finally, not accounting for the treatment status in the regression can lead to omitted variable bias if there is some measurement error in calculating $Centrality_v$. This is because, the experimental design ensures that some villages will have more central seeds than the other. In Appendix E, I check the robustness of my results with respect to including the treatment dummies. As my results remain robust, in the next section I present them without the treatment dummies.

4.3.2 Exploring between treatment group variations

To explore between treatment group variations, I use the following specification:

$$\begin{aligned}
Outcome_{vt} &= \theta_0 + \theta_1 Centrality_v + \theta_2 Probability_v + \theta_3 Heterogeneity_v \\
&+ \xi_b Centrality_v \times Heterogeneity_v + \xi_c Centrality_v \times Heterogeneity_v \times Complex_v \\
&+ \xi_s Centrality_v \times Heterogeneity_v \times Simple_v + \xi_g Centrality_v \times Heterogeneity_v \times Geo_v \\
&+ \phi_b Probability_v \times Heterogeneity_v + \phi_c Probability_v \times Heterogeneity_v \times Complex_v \\
&+ \phi_s Probability_v \times Heterogeneity_v \times Simple_v + \phi_g Probability_v \times Heterogeneity_v \times Geo_v \\
&+ \gamma X_v + \rho_t + \eta_{vt}.
\end{aligned} \tag{10}$$

Specification (10) is similar to specification (9), except the interactions of $Centrality_v \times Heterogeneity_v$ and $Probability_v \times Heterogeneity_v$ with treatment dummies. Here, ξ_b captures the interaction between seed centrality and village level heterogeneity for the benchmark treatment group. ξ_c, ξ_s, and ξ_g captures how that interaction changes compared to the benchmark for complex, simple, and geo treatment groups. Similarly, ϕ_b captures the interaction between seed probability and village level heterogeneity for the benchmark treatment group. ϕ_c, ϕ_s, and ϕ_g captures how that interaction changes compared to the benchmark for complex, simple, and geo treatment groups. Thus, for example for complex treatment group, the effect of $Centrality_v \times Heterogeneity_v$ on the outcome variable is $(\xi_b + \xi_c)$; the effect of $Probability_v \times Heterogeneity_v$ on the outcome variable is $(\phi_b + \phi_c)$. I expect the impact of $Centrality_v \times Heterogeneity_v$ to be negative and the effect of $Probability_v \times Heterogeneity_v$ to be positive, within different treatment groups. However, using this specification, I am more interested in exploring between group variations. Thus, main coefficients of interest in this specification are: $\xi = \{\xi_b, \xi_c, \xi_s, \xi_g\}$ and $\phi = \{\phi_b, \phi_c, \phi_s, \phi_g\}$.

For a treatment group that has the same level of heterogeneity as the benchmark, I expect outcomes to be positively related with centrality and negatively related with probability. Thus, given the heterogeneity of a group and the probability of the seeds, moving to higher (lower) centrality seeds helps diffuse the technology to more (less) households. Similarly, given the heterogeneity of a group and the centrality of the seeds, moving to higher (lower) probability seeds diffuse the technology to less (more) households. This is because central seeds are more representative of the population and hence are more connected. Given the same level of heterogeneity and probability of adoption, more (less) average households will improve (decrease) adoption through higher (lower) diffusion. On the other hand, probability-based seeds are early adopters that are less similar to the average population. Thus, given the same level of heterogeneity and centrality, less (more) average households will decrease (improve) adoption through lower (higher) diffusion.

If the treatment group is less heterogeneous than the benchmark, I expect seeds with higher centrality to perform better and seeds with higher probability to perform worse. In this case, however, my theory do not have any prediction for seeds with less centrality and probability. This is because, for treatment groups that are less heterogeneous than benchmark, the effect of having seeds with less centrality or probability depends on the relative effect of heterogeneity and centrality or probability. Similarly, for treatment groups having higher heterogeneity than the benchmark, I expect seeds with lower centrality to perform better and seeds with lower probability to perform worse. In this case my theory do not have any prediction for seeds with more centrality and probability.

As an example, let us consider the complex treatment group. If this group has the same level of heterogeneity as the benchmark, I expect outcomes to be positively related with centrality measure and negatively related with probability measure. Thus, ξ_c is positive (negative) if complex treatment group has more (less) central seeds than the benchmark. Similarly, ϕ_c is negative (positive) if complex treatment group has higher (lower) seed probability of adoption than the benchmark. Now, if the complex treatment group is less heterogeneous than the benchmark (captured by the means of $Heterogeneity_v$ at the group level), I expect the following:

- If they have more central seeds than the benchmark: positive ξ_c; less central seeds than benchmark: depends on the relative effects of drop in centrality and heterogeneity.
- If they have seeds with higher probability than the benchmark: negative ϕ_c; seeds with lower probability than benchmark: depends on the relative effects of drop in probability and heterogeneity.

Similarly, if the complex treatment group is more heterogeneous than the benchmark, I expect:

- If they have less central seeds than the benchmark: positive ξ_c; more central seeds than benchmark: depends on the relative effects of increase in centrality and heterogeneity.
- If they have seeds with lower probability than the benchmark: negative ϕ_c; seeds with higher probability than benchmark: depends on the relative effects of increase in probability and heterogeneity.

Like the last specification, here also I control for baseline village level characteristics, as well as year fixed effects. As the main coefficients of interest use interactions with the treatment dummies, no additional assumption are needed. This is because the selection into different treatment groups is randomly assigned in the experiment.

4.4 Strategy for Approximating Probabilities of Adoption

For (9), I need to calculate $Probability_v$, 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 D. Once calculated, the indices are used in the following regression model:

$$Adoption/Usage\ Index_{it} = f(X_{it}; \mu_{it}), \tag{11}$$

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 μ_{it} captures the random error in the regression. In my preferred specification, I consider function $f(\cdot)$ to be linear (thus the 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 proxies for the households' probability of adopting a new technology.

5 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. In doing so, I present the relevant regression results and discuss the assumptions needed for using these results for the rest of my analysis. The 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.

5.1 Approximating Probabilities of Adoption

I start by comparing key baseline demographic information across datasets. This is presented in Table 2. The comparison is important as it helps me understand how results derived using the AESTAS data map into the BBMM data. The five variables chosen 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 PCA (Principal Component Analysis) scores. This is not surprising given that AESTAS focused on a nationally representative sample of farmers in Malawi, where BBMM focused only on the Maize farmers.

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

Notes: 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., 2021a)

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 the 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.

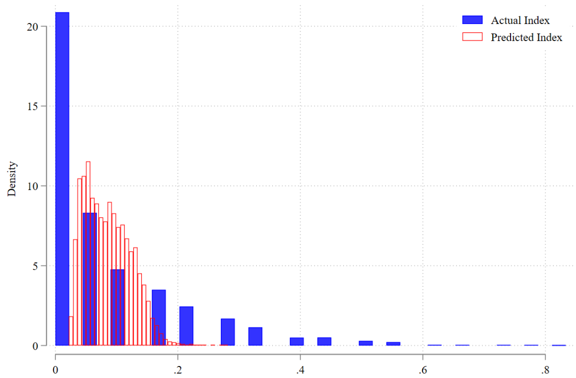
Table 3: OLS Regression Results for Adoption and Usage Indices

Variables	Adoption Index (1)	(2)	Usage Index (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).

The results are mostly highly significant irrespective of whether the household controls are included. Columns (3) and (4) present the results for Usage Index, with and without the full set of household controls. The results are similar to that of Adoption Index, with the exception of the coefficient related to housing PCA score. The main takeaway from these results is that the coefficients are mostly similar with or without the full set of household controls. Thus for calculating the predicted indices, I use the estimates without the household controls. Also, the results vary a bit depending on whether the dependent variable is Adoption or Usage Index. Hence, I use both for my analysis.

Panel A: Adoption Index



Panel B: Usage Index

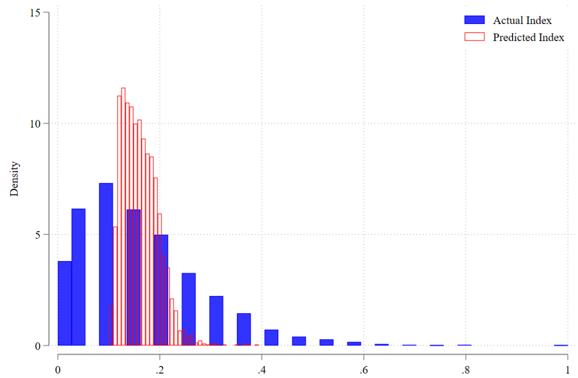


Figure 10: Actual and Predicted Adoption and Usage Indices

Figure 10 focuses on comparing the actual and predicted indices for the AESTAS sample. 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 the quality of prediction (actual has mean of 0.163 with a standard deviation of 0.122, where predicted has a mean of 0.162 and a standard deviation of 0.035). Thus, the predictions are doing a good job in predicting the mean but only capture 1/3rd of the actual variation. This is not surprising given that the predictions are made based on only a few observable demographics.

Based on the above discussion, it is clear that I need to make some strong assumptions to use the estimates from Table 3 for predicting probability of adoption in the BBMM data. In particular, I assume the following:

- **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.

5.2 Descriptive Statistics

Table 4 focuses on describing key baseline characteristics in the BBMM sample. The last column of this table represents overall village level variations. This is the variation exploited in the regression specification (9). The first four columns of the table represent within treatment group variations. Regression specification (10) uses the variations between these four groups.

Table 4: Baseline Village-level Sample Characteristics

	Treatment Status				
Variable	Benchmark	Complex	Simple	Geo	Overall
Adoption Rate (PP)	0.018	0.030	0.029	0.029	0.026
	(0.035)	(0.063)	(0.060)	(0.077)	(0.060)
Any Non-Seed Adopters (PP)	0.300	0.340	0.320	0.420	0.345
	(0.463)	(0.479)	(0.471)	(0.499)	(0.477)
Eigen-vector Centrality of Seeds†	0.178	0.235	0.187	0.129	0.182
	(0.090)	(0.077)	(0.096)	(0.090)	(0.096)
Predicted Adoption Index of Seeds‡	0.110	0.114	0.101	0.082	0.101
	(0.034)	(0.036)	(0.041)	(0.025)	(0.036)
Predicted Usage Index of Seeds‡	0.184	0.186	0.172	0.158	0.175
	(0.031)	(0.032)	(0.042)	(0.024)	(0.035)
CV of Predicted Adoption Index	0.389	0.378	0.379	0.366	0.378
	(0.069)	(0.077)	(0.075)	(0.062)	(0.071)
CV of Predicted Usage Index	0.193	0.188	0.185	0.180	0.187
	(0.039)	(0.040)	(0.037)	(0.033)	(0.038)
Observations	50	50	50	50	200

Notes: † Contains 44 observations for the benchmark treatment group, 49 observations for the other treatment groups. ‡ Contains 48 observations for the complex treatment group. 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.

The first two rows present the main outcome variables of my analysis. Adoption Rate (PP) captures the proportion of *typical* farmers per village that adopted pit planting in each agricultural season. Here, *typical* farmers correspond to the farmers that were not selected as *seed* or *shadow* farmers in the experiment. Any Non-Seed Adopters (PP) is a dummy variable that captures whether the villages had at least one *non-seed* farmer adopting pit planting in an agricultural season. I focus on these two outcome variables as they are also used in the village-level analysis of BBMM. The baseline data suggests an adoption rate of around 2-3% across treatment arms. Also, only 30-42% villages had at least one *non-seed* farmer adopting pit planting in the baseline. These numbers suggest low adoption of pit planting in the baseline, providing an ideal setting to test the predictions of my theoretical analysis. According to my theoretical framework, this low adoption in the baseline is due to the pessimism regarding the prospect of pit planting. Hence, this is also a setting where targeting is indeed required to improve adoption.

The next three rows of table 4 focus on average seed characteristics. The average of seeds' characteristics is taken, whenever the information on both seeds are available (for 138 villages). Otherwise it is the characteristic of the only seed for which the data is available (for 53 villages). Eigen-vector Centrality of Seeds capture average eigen-vector centrality of the selected seeds at the baseline.[20] For this purpose, I use the eigen-vector centrality values that are pre-calculated and available in the BBMM replication dataset.

By the design of the experiment, complex seeds have the highest average centrality. As argued in BBMM, this is because it is optimal to seed only central households when the underlying model of diffusion is of complex contagion. Similarly, simple seeds are expected to have relatively less average centrality than complex seeds as it is optimal to seed one central and one peripheral household when the underlying diffusion process is of simple contagion. BBMM also argue that geo seeds should be less central as they have less than average land by design (which is a measure of less than average wealth), and hence are less likely to be well connected. This is indeed what I observe in the baseline. In terms of the average eigen-vector centrality of the seeds, the simple seeds are not statistically different than the benchmark seeds. However, both complex and geo seeds are statically different than the benchmark (at respectively 1% and 5% level of significance).

I use predicted adoption and usage indices as proxies for probability of adoption. Depending on which proxy is used, the probabilities differ but the ranking over different treatment groups remain the same. Here, complex and benchmark seeds have the highest probabilities of adoption. This is followed by the simple seeds, with the geo seeds having the least baseline probability of adoption. In terms of both measures of probability of adoption, there are no statistically significant differences between benchmark, complex, and simple seeds. However, geo seeds are statistically different than their benchmark counterparts (at 1% level of significance). In terms of the CV of predicted adoption and usage indices, complex and simple treatment groups are similar and slightly less than their benchmark counterparts. These differences are not statistically significant. But, the geo treatment group is significantly less heterogeneous than the benchmark group (at the 10% level).

Before proceeding to my main empirical results in the next sub-section, let me focus on

[20]Formal definition of eigen-vector centrality can be found in Appendix A.

figure 11. Figure 11 presents the main outcome variables over varying degrees of village-level heterogeneity. Here, village-level heterogeneity is represented by the CV of predicted adoption index. The figure considers seeds to be centrality-based and/or probability-based. For this figure, centrality-based seeds are defined to be the seed household(s) that have higher than the median level of mean eigenvector centrality at baseline. Similarly, probability-based seed household(s) are defined here to have higher than the median level of predicted adoption index in the baseline. Given the seeding strategy of BBMM, the seed household(s) of a village can fall under either, none, or both of the categories. Based on my simulations, I expect the centrality-based seeds to perform worse, and the probability-based seeds to perform better, as village heterogeneity increases. However, I only expect to see this pattern in years 2 and 3 after the interventions. In year 1, since no treatment effect should have been realized yet, the same pattern should not be observed.[21]

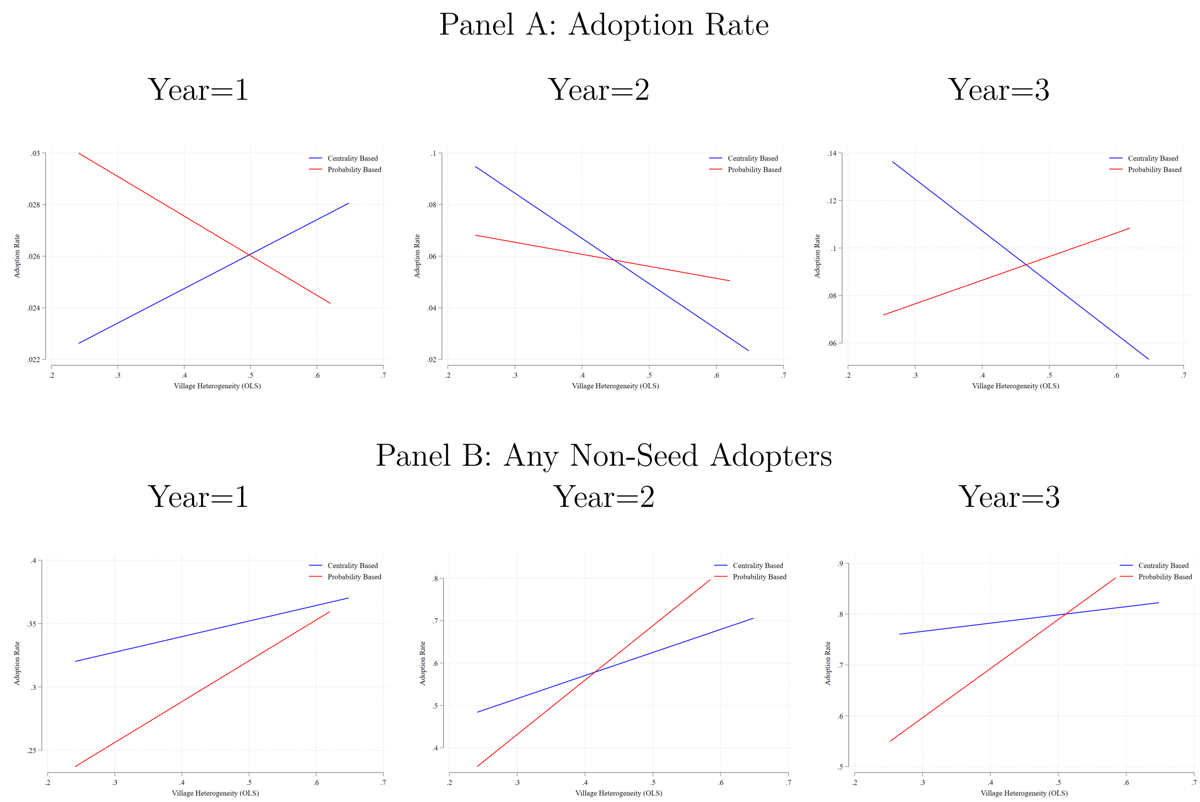


Figure 11: Outcomes for Different Seeding Strategies with respect to Village Heterogeneity

This is essentially the pattern I observe. In years 2 and 3, as village-level heterogeneity increases, the performance of centrality-based seeds decreases compared to their probability-based counterpart. The opposite is true for probability-based seeds compared to centrality-based seeds. On the contrary, the opposite pattern can be observed in year 1 for the adoption rate. For the dummy variable on any non-seed adopters, however, I observe

[21]Since the treatment was given just a few months before the household survey in year 1, diffusion would not have been realized yet. Thus, my regression results assess the effect on the outcome variables from years 2 and 3 only. The same approach is taken by Beaman et al. (2021a).

that in year 1 the gap between centrality-based and probability-based seeds is closing with an increase in village-level heterogeneity. But, the performance of centrality-based seeds remains better than their probability-based counterpart for all levels of village-level heterogeneity.

Although informative, the descriptive figures do not take into account 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 village level variations 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 hypotheses more formally.

5.3 Reduced Form Results

Table 5 focuses on exploring village level variations. Subsequently, Table 6 presents results exploring between treatment group variations. For both these tables, I proxy for probability of adoption using predicted adoption index. The results are similar using the predicted usage index as the proxy. Thus, they are not presented here to avoid repetitions. These results can be found in Appendix E.

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

Variables	Adoption Rate (1)	(2)	Any Non-Seed Adopters (3)	(4)
Eigen-vector Centrality of Seeds	1.173**	0.917*	1.181	1.235
($=Centrality_v$)	(0.581)	(0.467)	(1.439)	(1.332)
Predicted Adoption Index of Seeds	-2.973**	-2.140	-8.019**	-3.344
($=Probability_v$)	(1.467)	(1.318)	(3.257)	(3.233)
CV of Predicted Adoption Index	-0.296	-0.157	-0.928	0.506
($=Heterogeneity_v$)	(0.208)	(0.214)	(1.079)	(1.053)
$Centrality_v \times Heterogeneity_v$	-2.625**	-2.131**	-2.851	-3.299
	(1.324)	(1.066)	(3.777)	(3.562)
$Probability_v \times Heterogeneity_v$	6.715**	4.762*	18.480***	7.562
	(3.131)	(2.796)	(6.997)	(7.073)
Village-level Controls	No	Yes	No	Yes
Observations	324	324	324	324
R-squared	0.080	0.180	0.049	0.169

Notes: * 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, and district fixed effects.

For Table 5, the main coefficients of interest are those corresponding to the interactions of $Heterogeneity_v$, with $Centrality_v$ and $Probability_v$. As argued in the last section, I expect the coefficient of $Centrality_v \times Heterogeneity_v$ to be negative and the coefficient of $Probability_v \times Heterogeneity_v$ to be positive. Columns (1) and (2) present the results for Adoption Rate, with and without the village level controls. Here, both coefficients of interest are of the desired sign and highly significant. The results show that for a completely homogeneous village, a 1 standard deviation increase in eigen-vector centrality leads to 9.17%-11.73% increase in adoption rate. This is a 352.69%-451.15% increase compared to the baseline mean adoption rate of 2.6%. However, for villages with heterogeneity at the level of baseline mean, the effect drops to an increase of only 1.11%-1.81%. Which is a 42.69%-69.62% increase compared to the baseline mean adoption rate. Similarly, 1 standard deviation increase in predicted adoption decreases adoption rate by 8.56%-11.89% for a homogeneous village. Which is a decrease of 329.23%-457.31%, compared to the mean of adoption rate at the baseline. However, for villages having heterogeneity at the level of baseline mean, the effect drops to a decrease of 1.36%-1.74% only. This is a much smaller decrease of 52.31%-66.92% compared to the baseline mean of adoption rate.

The results for Any Non-Seed Adopters are presented in columns (3) and (4), with and without the village level controls. Although the coefficients of interest are of the desired sign, they are mostly insignificant. The results show that for completely homogeneous villages, 1 standard deviation increase in eigen-vector centrality leads to 11.81%-12.35% increase in the probability of having at least one non-seed adopters. Compared to the baseline mean of 34.5% for the variable Any Non-Seed Adopters, this is an increase of 34.23%-35.80%. But, for villages with heterogeneity at the level of baseline mean, the effect drops to being between a 0.12% decrease and 1.03% increase in the probability. This is between a drop of 0.35% and an increase of 2.99%, compared to the baseline mean of the dependent variable. On the other hand, 1 standard deviation increase in predicted adoption decreases the probability of having at least one non-seed adopters by 13.38%-32.08% for a homogeneous village. Compared to the baseline mean of the variable Any Non-Seed Adopters, this is a decrease of 38.78%-92.99%. For villages with heterogeneity at the level of baseline mean, however, the effect drops to a probability decrease of 1.94%-4.13% only. Which is a drop of only 5.62%-11.97% compared to the baseline mean of the dependent variable.

Table 6 focuses on exploring between treatment group variations. Here, I am interested in the coefficients of $Centrality_v \times Heterogeneity_v$ and $Probability_v \times Heterogeneity_v$, across different treatment groups. Before getting into that, it is important to note that the sign of $Centrality_v \times Heterogeneity_v$ is negative and the sign of $Probability_v \times Heterogeneity_v$ is positive within different treatment groups. This shows the validity of my hypotheses within treatment groups. This is in line with the results of table 5, which focuses on exploring village level variations without controlling for the treatment

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	0.775**	0.633*	1.704	1.638
(=$Centrality_v$)	(0.377)	(0.366)	(1.538)	(1.477)
Predicted Adoption Index of Seeds	-2.362***	-1.579*	-10.420***	-5.947*
(=$Probability_v$)	(0.899)	(0.888)	(3.666)	(3.582)
CV of Predicted Adoption Index	-0.321	-0.150	-0.923	0.417
(=$Heterogeneity_v$)	(0.267)	(0.262)	(1.087)	(1.058)
$Centrality_v \times Heterogeneity_v$	-2.423**	-2.237**	-6.693	-6.574
	(1.045)	(1.008)	(4.258)	(4.064)
$Centrality_v \times Heterogeneity_v \times Complex$	0.657	0.665	4.328**	3.756**
	(0.485)	(0.469)	(1.977)	(1.894)
$Centrality_v \times Heterogeneity_v \times Simple$	0.416	0.428	1.078	0.431
	(0.499)	(0.484)	(2.033)	(1.953)
$Centrality_v \times Heterogeneity_v \times Geo$	2.026***	1.942***	0.103	-0.070
	(0.545)	(0.542)	(2.221)	(2.185)
$Probability_v \times Heterogeneity_v$	5.881***	4.104**	22.970***	12.350
	(2.040)	(2.019)	(8.314)	(8.142)
$Probability_v \times Heterogeneity_v \times Complex$	-0.155	-0.232	-1.275	-0.679
	(0.851)	(0.823)	(3.469)	(3.320)
$Probability_v \times Heterogeneity_v \times Simple$	-0.121	-0.110	1.941	3.511
	(0.887)	(0.862)	(3.617)	(3.476)
$Probability_v \times Heterogeneity_v \times Geo$	-2.588***	-2.562***	-0.391	0.538
	(0.949)	(0.932)	(3.870)	(3.759)
Village-level Controls	No	Yes	No	Yes
Observations	324	324	324	324
R-squared	0.133	0.224	0.113	0.222

Notes: * 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, and district fixed effects.

status. Table 6 notes the differences in the coefficients of $Centrality_v \times Heterogeneity_v$ and $Probability_v \times Heterogeneity_v$, across different treatment groups. Some of these differences are statistically significant, some are not. This is in line with the between treatment group variations noted in table 4. However, the signs are all consistent with my discussion in the last section.

Columns (5) and (6) present the results for Adoption Rate, with and without the village level controls. The results show that for a completely homogeneous village, 1 standard deviation increase in eigen-vector centrality leads to 6.33%-7.75% increase in adoption rate. This is a 243.46%-298.08% increase compared to the baseline mean adoption rate of 2.6%. However, for benchmark villages having heterogeneity at the level of baseline benchmark mean, the effect drops to a decrease of 1.68%-2.37%. Which is a 64.62%-91.15% decrease compared to the baseline mean adoption rate. The negative effect of heterogeneity, on the impact of eigen-vector centrality on adoption rate, is insignificantly lower for complex and simple treatment groups compared to the benchmark. However, for geo treatment group the effect is significantly lower compared to the benchmark. For geo villages having heterogeneity at the level of baseline geo mean, the effect of 1 standard deviation increase in eigen-vector centrality is a 5.25%-6.30% increase in adoption rate. Which is a 201.92%-242.31% increase compared to the baseline mean adoption rate. Thus, the effect is in the same direction for all the treatment groups, but significantly different for the geo treatment. Similarly, 1 standard deviation increase in predicted adoption index decreases adoption rate by 6.32%-9.45% for a homogeneous village. Which is a decrease of 243.08%-363.46%, compared to the mean of adoption rate at the baseline. However, for benchmark villages having heterogeneity at the level of baseline benchmark mean, the effect drops between a decrease of 0.30% and an increase of 0.07%. This is a much smaller effect of between 11.54% decrease and 2.69% increase, compared to the baseline mean of adoption rate. The positive effect of heterogeneity, on the impact of predicted adoption index on adoption rate, is insignificantly lower for complex and simple treatment groups compared to the benchmark. However, for geo treatment group the effect is significantly lower compared to the benchmark. For geo villages having heterogeneity at the level of baseline geo mean, the effect of 1 standard deviation increase in predicted adoption index is a 4.06%-4.63% decrease in adoption. Which is a 156.15%-178.08% decrease compared to the baseline mean adoption rate. Again, the effect is in the same direction for all the treatment groups, but significantly different for the geo treatment.

The results for Any Non-Seed Adopters are presented in columns (7) and (8), with and without the village level controls. For this outcome variable, the effect also is in the same direction for all the treatment groups. The results show that for completely homogeneous villages, 1 standard deviation increase in eigen-vector centrality leads to 16.38%-17.04% increase in the probability of having at least one non-seed adopters. Compared to the baseline mean of 34.5% for the variable Any Non-Seed Adopters, this is an increase of

47.48%-49.39%. But, for benchmark villages having heterogeneity at the level of baseline benchmark mean, the effect drops to being around a 9% decrease of probability. This is a 26.09% drop compared to the baseline mean of the dependent variable. The negative effect of heterogeneity, on the impact of eigen-vector centrality on the probability of having any non-seed adopters, is insignificantly lower for simple treatment group compared to the benchmark. It is insignificantly different for geo treatment group. However, for complex treatment group the effect is significantly lower compared to the benchmark. For complex villages having heterogeneity at the level of baseline complex mean, the effect of 1 standard deviation increase in eigen-vector centrality is a 5.73%-8.10% increase in the probability of having at least one non-seed adopters. Which is a 16.61%-23.48% increase compared to the baseline mean for the variable Any Non-Seed Adopters. On the other hand, 1 standard deviation increase in predicted adoption index decreases the probability of having at least one non-seed adopters by 23.79%-41.68% for a homogeneous village. Compared to the baseline mean of the variable Any Non-Seed Adopters, this is a decrease of 68.96%-120.81%. However, for benchmark villages having heterogeneity at the level of baseline benchmark mean, the effect drops to a decrease of 4.57%-5.94%. Which is a drop of only 13.25%-17.22% compared to the baseline mean of the dependent variable. The positive effect of heterogeneity, on the impact of predicted adoption index on adoption rate, is insignificantly lower for complex treatment group; higher for simple treatment group; and different for geo treatment group, compared to the benchmark.

These results show that for homogeneous villages more central seeds lead to improvements in adoption. Existing literature recognizes the role played by central households in improving diffusion, and subsequent adoption of a product. In their paper, Beaman et al. (2021a) uses the same data to show that more central seeds cause higher adoption. Seeds' centrality is found to be one of the main reasons for improved adoption of a microfinance product in India by Banerjee et al. (2013), and improved take-up of an insurance product in China by Cai et al. (2015). I add to this literature by providing evidence that the positive effect of seeds' centrality decreases as the target population becomes more heterogeneous. In addition, I show evidence in favor of an alternative probability-based seeding strategy to work better in such a scenario.

6 Summary and Concluding Remarks

I focus on optimal network-based targeting strategies for improving technology adoption when a new technology has more benefits to some agents than others. In particular, I assume this heterogeneity in benefits directly impacts the performance of targeting strategies. This deviates from the existing literature that considers such performances to be affected by existing social ties only. I present a theoretical model that helps formalize such a scenario. This adds to the theoretical literature that considers households to be homogeneous in terms of what they need to learn about new technologies. Through the lens of my theoretical model, simulations help me shape the main hypotheses of my study. I hypothesize that the relative performance of different targeting strategies depends on the level of heterogeneity in the population. This is heterogeneity in terms of the expected benefits of adopting a technology. In particular, I expect centrality-based targeting to perform worse as the heterogeneity increase but, targeting based on the probability of adoption to perform better in such a 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 new technology, I use the AESTAS dataset also collected from Malawi. Reduced form results lend support in favor of my hypotheses. Exploring village-level variations, I show that the positive effect of seeds' centrality on the adoption of pit planting decrease with an increase in village-level heterogeneity in terms of probability of adoption. Simultaneously, the negative effect of seeds' probability of adoption decreases with an increase in village-level heterogeneity. Although weaker, I found similar results when I shifted my focus to exploring the experimental variations of BBMM.

The reduced form analysis is based on a series of assumptions that are discussed. The main challenge in targeting based on the probability of adoption is the fact that probabilities depend on benefits that realize only after the adoption takes place. I attempt to solve this issue by using additional data to predict adoption conditional on observable demographics. A better approach would be to collect additional data on the same households that are making the adoption decisions. For that purpose, and also for a clearer identification, a randomized controlled trial that mimics the ideal experiment discussed in this paper is more suitable. A randomized controlled trial of such a nature can also help me disentangle the effects of centrality and probability of seeds. This, along with a more structural approach can help separately identify the effects of targeting strategies discussed here. This is an exciting avenue for future research.

In terms of policy, my results suggest that network-based targeting may require more than identifying central households within a social network. More specifically, I argue for the need to have an understanding of possible heterogeneity in benefits across households. This adds to the existing literature that highlights the importance of central agents for targeting policies (Beaman et al. (2021a)) and focuses on cost-effectively identifying these households (Banerjee et al. (2019)). This is, of course, if a new technology is such that there can be sufficient heterogeneity in the population in terms of its applicability. In practice, this demands more information than the requirement for just identifying central households. Thus, it increases the cost of targeting. This may make it more attractive to randomly seed more households following the approach proposed by Akbarpour et al. (2021). A proper cost-benefit analysis is required for that purpose, which is beyond the scope of this paper.

Alternatively, policies can focus on promoting technologies that have more homogeneous benefits across households. Focus group discussions, widely used for designing interventions, can help identify such technologies. Interventions can also be directed towards making households more homogeneous in terms of benefits from new technology. For example, consider a new seed variety that would be equally beneficial to all households in a region if they have equal access to irrigation facilities or have similar land quality. In such a scenario, policy interventions can focus on increasing the households' access to irrigation facilities or providing support to improve the quality of their lands. Improvements can also be made by promoting new network ties within a population, thus making the households more similar in terms of access to information needed to benefit from new technologies.

References

Acemoglu, D., A. Nedic, and A. Ozdaglar (2008): "Convergence of rule-of-thumb learning rules in social networks," in *2008 47th IEEE Conference on Decision and Control*, 1714–1720.

Acemoglu, D., A. Ozdaglar, and E. Yildiz (2011): "Diffusion of innovations in social networks," in *2011 50th IEEE Conference on Decision and Control and European Control Conference*, 2329–2334.

Akbarpour, M., S. Malladi, and A. Saberi (2021): "Just a Few Seeds More: Value of Network Information for Diffusion," *Unpublished Manuscript*.

Bala, V. and S. Goyal (1998): "Learning from Neighbours," *Review of Economic Studies*, 65, 595–621.

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

Banerjee, A., E. Breza, A. G. Chandrasekhar, and B. Golub (2018): "When Less is More: Experimental Evidence on Information Delivery During India's Demonetization," Working Paper 24679, National Bureau of Economic Research.

Banerjee, A., E. Breza, A. G. Chandrasekhar, and M. Mobius (2021): "Naïve Learning with Uninformed Agents," *American Economic Review*, 111, 3540–74.

Banerjee, A., A. G. Chandrasekhar, E. Duflo, and M. O. Jackson (2013): "The Diffusion of Microfinance," *Science*, 341, 1236498.

——— (2019): "Using Gossips to Spread Information: Theory and Evidence from Two Randomized Controlled Trials," *The Review of Economic Studies*, 86, 2453–2490.

Barnett-Howell, Z. and A. M. Mobarak (2021): "Chapter 84 - Social networks analysis in agricultural economies," in *Handbook of Agricultural Economics*, ed. by C. B. Barrett and D. R. Just, Elsevier, vol. 5 of *Handbook of Agricultural Economics*, 4613–4652.

Beaman, L., A. BenYishay, J. Magruder, and A. M. Mobarak (2021a): "Can Network Theory-Based Targeting Increase Technology Adoption?" *American Economic Review*, 111, 1918–43.

——— (2021b): "Data and Code for: Can Network Theory-based Targeting Increase Technology Adoption?" *Nashville, TN: American Economic Association.*

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

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

Bloch, F., M. O. Jackson, and P. Tebaldi (2021): "Centrality measures in networks," *Available at SSRN 2749124*.

Bold, T., K. C. Kaizzi, J. Svensson, and D. Yanagizawa-Drott (2017): "Lemon Technologies and Adoption: Measurement, Theory and Evidence from Agricultural Markets in Uganda," *The Quarterly Journal of Economics*, 132, 1055–1100.

Breza, E., A. Chandrasekhar, B. Golub, and A. Parvathaneni (2019): "Networks in economic development," *Oxford Review of Economic Policy*, 35, 678–721.

Bustos, P., B. Caprettini, and J. Ponticelli (2016): "Agricultural Productivity and Structural Transformation: Evidence from Brazil," *American Economic Review*, 106, 1320–1365.

Cai, J., A. De Janvry, and E. Sadoulet (2015): "Social Networks and the Decision to Insure," *American Economic Journal: Applied Economics*, 7, 81–108.

Catalini, C. and C. Tucker (2017): "When early adopters don't adopt," *Science*, 357, 135–136.

Chandrasekhar, A. G., B. Golub, and H. Yang (2018): "Signaling, Shame, and Silence in Social Learning," Working Paper 25169, National Bureau of Economic Research.

Chandrasekhar, A. G., H. Larreguy, and J. P. Xandri (2020): "Testing models of social learning on networks: Evidence from two experiments," *Econometrica*, 88, 1–32.

Cheng, H. W. J. (2021): "Factors Affecting Technological Diffusion Through Social Networks: A Review of the Empirical Evidence," *The World Bank Research Observer*, lkab010.

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

DeGroot, M. H. (1974): "Reaching a Consensus," *Journal of the American Statistical Association*, 69, 118–121.

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

Feder, G. and S. Savastano (2006): "The role of opinion leaders in the diffusion of new knowledge: The case of integrated pest management," *World Development*, 34, 1287–1300.

Foster, A. D. and M. R. Rosenzweig (1995): "Learning by Doing and Learning from Others: Human Capital and Technical Change in Agriculture," *Journal of Political Economy*, 103, 1176–1209.

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

IFPRI (2021a): "Agricultural Extension Services and Technology Adoption Survey, 2016," *Harvard Dataverse*.

——— (2021b): "Agricultural Extension Services and Technology Adoption Survey, 2018," *Harvard Dataverse*.

Jackson, M. O. (2010): *Social and Economic Networks*, Princeton University Press.

Khaila, S., F. Tchuwa, S. Franzel, and B. Simpson (2015): "The Farmer-to-Farmer Extension Approach in Malawi: A Survey of Lead Farmers," ICRAF Working Paper 189, World Agroforestry Centre.

Krishnan, P. and M. Patnam (2013): "Neighbors and Extension Agents in Ethiopia: Who Matters More for Technology Adoption?" *American Journal of Agricultural Economics*, 96, 308–327.

Maertens, A. (2017): "Who Cares What Others Think (or Do)? Social Learning and Social Pressures in Cotton Farming in India," *American Journal of Agricultural Economics*, 99, 988–1007.

Magruder, J. R. (2018): "An Assessment of Experimental Evidence on Agricultural Technology Adoption in Developing Countries," *Annual Review of Resource Economics*, 10, 299–316.

Miller, G. and A. M. Mobarak (2015): "Learning About New Technologies Through Social Networks: Experimental Evidence on Nontraditional Stoves in Bangladesh," *Marketing Science*, 34, 480–499.

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

Niu, C. and C. 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, C. (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, C., D. Mzungu, K. Kalagho, and C. Kazembe (2021): "Impact of interactive radio programming on agricultural technology adoption and crop diversification in Malawi," *Journal of Development Effectiveness*, 13, 204–223.

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

Suri, T. (2011): "Selection and Comparative Advantage in Technology Adoption," *Econometrica*, 79, 159–209.

Tjernström, E. (2017): "Learning from others in heterogeneous environments," *Unpublished Manuscript*.

Appendices

A Mathematical Definitions

B Details on the Simulation Method

C Detailed Simulation Results

D Construction of Adoption and Usage Indices

E Robustness Checks

E.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.

E.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 set of controls.
- Robustness with respect to different proxy for probability of adoption (Predicted Usage Index instead of Predicted Adoption Index).
- Robustness with respect to different measure of centrality.
- Robustness with respect to predicted adoption and usage indices calculated using different regression specification.