Optimal Network Based Targeting in Improving Technology Adoption: Evidence from Malawi

Preliminary and Incomplete - Please do not Cite

Aranya Chakraborty*

Abstract

How do we use existing social ties to improve the adoption of a new technology if for some agents the new technology is riskier than the others? I explore optimal network-based targeting strategies for such a scenario. In particular, I focus on two types of targeting: targeting people central to the network and targeting based on the likelihood of adoption. Through the lens of my theoretical framework, simulations suggest that the optimal targeting strategy relies on the underlying heterogeneity in the population. If the heterogeneity is high in terms of the applicability of the new technology, targeting based on the likelihood of adoption performs better than centrality-based targeting. Conversely, centrality-based targeting works better if the population is more homogeneous. I use secondary data sources from Malawi to test this hypothesis. My results show support in favor of my theoretical predictions.

Technology adoption has long been recognized as a driving force of economic development (see Besley and Case, 1993). Diffusion of information via network is the key to increase technology adoption (Foster and Rosenzweig, 1995; Conley and Udry, 2010). In the recent years, there has been a growing number of studies focusing on the role of networks in the diffusion of technologies. These studies can be broadly categorized into two groups. The first group of studies focus on presenting the evidence in favor of network effects (e.g., Bandiera and Rasul, 2006; Conley and Udry, 2010). The second group explores the most

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

effective way to use social networks to improve technology adoption (e.g., Banerjee et al., 2013; BenYishay and Mobarak, 2018). The current study falls under the second category.

The relevant policy question is: how do we use existing social ties to improve adoption of a new technology? The literature argue that the answer depends on the underlying diffusion process (Beaman et al., 2021). In particular, if the underlying diffusion process is of *simple contagion*, where information diffuses with a random probability from one agent to another, using existing social ties to improve adoption may not be optimal. This is because collecting information on existing social ties is costly. If information diffuses with a random probability from one agent to another, given enough time it will reach everyone irrespective of the initial *seeds*[1]. If time is of the essence, increasing the number of randomly chosen entry points will do the trick (Akbarpour, Malladi and Saberi, 2021). However, if the underlying diffusion process is of *complex contagion*, where 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. Under the second scenario, the literature recommends targeting agents central to the network (Beaman et al., 2021). The recommendation, however, is based on the underlying assumption that the agents are homogeneous in terms of the risk they face in adopting the new technology. What happens if for some agents the new technology is more risky than the others? In other words, if a technology is more beneficial to a sub-section of the population, even if the underlying diffusion process is of *complex contagion*, should we still target people that are central to the network?

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

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

as they have notable policy implications.

For this purpose, I form a theoretical model of DeGroot learning based on the works of Golub and Jackson (2010) and Banerjee et al. (2021)[2]. Based on this model, 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 risk distributions. In such settings, targeting based on likelihood of adoption should perform better. The intuition behind such result lies in the characteristics of the central seeds in a network. 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.

I test the predictions of my model combining two different data sources from Malawi. First one is the replication data from a randomized controlled trial (RCT) conducted by Lori Beaman, Ariel BenYishay, Jeremy Magruder and Ahmed Mushfiq Mobarak (2021) (henceforth, BBMM). The second dataset is the Agricultural Extension Services and Technology Adoption Survey (henceforth, AESTAS) data collected by International Food Policy Research Institute (IFPRI). Reduced form results show evidence in favour of my hypothesis.

I contribute to the growing body of literature that focuses on using existing network ties to improve technology adoption in the developing countries. Banerjee et al. (2013) is one of the earliest development studies that explores the possibility of improving diffusion by varying seeding strategies. The researchers focus on detailed network data on the diffusion of a micro-finance product collected from 43 Indian villages. Their reduced form results show support in favor of targeting based on *eigen-vector centrality*[3]. Their structural form results suggest that both participants and non-participants of the micro-finance

[2]DeGroot type learning is used in all the canonical models of information aggregation in the development literature. There is also evidence in favor of it in the literature (see Chandrasekhar, Larreguy and Xandri, 2020).

[3]*Eigen-vector centrality* is a recursive measure of centrality that captures the importance of a household based on how influential their neighbours are (Jackson, 2010).

product help in its transmission. Banerjee et al. (2021) develop a generalized DeGroot framework that builds on the findings of Banerjee et al. (2013). Beaman et al. (2021) builds their micro-foundation on the model developed in Banerjee et al. (2021) to study the relationship between targeting strategies and the underlying diffusion process. All these studies, however, used detailed network data which is costly to collect. From a more practical point of view, Banerjee et al. (2019) explore identifying central agents through the use of gossips. Similarly, BenYishay and Mobarak (2018) explore the role of incentives for the seeds on social learning. A detailed review of this literature can be found in Cheng (2021).

This stream of literature, however, consider agents to be homogeneous in terms of the risk they face in a new technology. The heterogeneity is assumed to be in their cost of adoption. But these heterogeneous costs are assumed to be known by the agents and thus do not require learning[4]. In other words, simplifying assumptions are made such that the learning involves the variable that takes a common value for all the agents and not the variable they differ in terms of. This assumption helps us to focus on a problem where the agents are collectively trying to uncover some hidden parameter of interest. In many scenarios, however, agents do face heterogeneous risks in adopting a new technology. In agriculture, for example, the performance of some practices depend on the quality of land[5]. Thus, farmers vary in terms of the risks they face in adopting those practices. In this study, I show that this heterogeneity can have serious implications for the effectiveness of targeting based on networks.

Considering population heterogeneity in social learning itself is not new in the literature. 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

[4]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.

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

veteran neighbours about the use of fertilizers for pineapple production in Ghana. However, to my best knowledge, the current study is the first to consider the consequences of population heterogeneity on targeting strategies.

More formally, I make the following contributions to the literature:

1. I provide a theoretical framework based on the existing literature that helps formally capture the heterogeneity in the risk of adoption. This model also helps to endogenize the findings of Banerjee et al. (2013) that adopters are more likely to diffuse a technology than the non-adopters.

2. My study provides empirical evidence in support of my theoretical findings.

3. I provide policy directions for targeting when the population is heterogeneous in terms of the risk they face from a new technology.

The remainder of this article is organized as follows. In section 1 I present the theoretical framework of my analysis. Section 2 presents the results from the simulations that helps me form my main hypothesis. Section 3 discusses the data sources used in the empirical analysis and, section 4 focuses on discussing the empirical framework for the analysis. In section 5, I present and discuss my results. Finally, section 6 summarizes the results and presents the concluding remarks.

1 Theoretical Framework

I consider a choice problem that requires learning in a social network. Specifically the problem is of technology adoption when the risk associated with the new technology varies at the household level. So, the new technology is not '*good*' for everyone in the population. On top of this, the risk associated with the technology is initially unknown to the households and must be learned.

Similar to Golub and Jackson (2010), I consider agents to have an initial opinion and involve in DeGroot learning (developed in DeMarzo, Vayanos and Zwiebel, 2003). I distinguish between informed and uninformed agents similar to Banerjee et al. (2021)[6]. I also consider the possibility that people are heterogeneous in terms of their distribution of risk and need to be informed before adoption.

1.1 The Theoretical Model

Consider a decision problem where the agents make an irreversible technology adoption decision between whether to stick to a traditional technology, or adopt a new technology. The traditional technology has a sure payoff of π^T, where the new technology provides a payoff of $\pi^N(\omega_{it})$ that depends on the state of the world parameter $\omega_{it} \in \Omega$. The state of the world parameter 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. I assume that the true distributions are positively correlated between households according to the existing network structure. This makes the social learning possible in this framework. If this is not the case, households are better off learning on their own, without taking any help from their connections. I also assume that $\forall it$, $\exists \omega_{it}, \omega'_{it} \in \Omega$ such that $\pi^N(\omega_{it}) \geq \pi^T \geq \pi^N(\omega'_{it})$; i.e., for each agent i and period t, there exist states of the world such that the payoff from the new technology is higher (lower) than the old technology. Finally, $\exists i, j \in \mathcal{I}$ such that $\int_{\omega_{it}\in\Omega} p_i^*(\omega_{it})\pi^N(\omega_{it}) - c_i \geq \pi^T$ and $\int_{\omega_{jt}\in\Omega} p_j^*(\omega_{jt})\pi^N(\omega_{jt}) - c_j \leq \pi^T$, where $\mathcal{I}$ denote the set of all households and c_i is the cost of new technology for household i. Which means that there is enough heterogeneity in the population such that for some agents the net expected benefits of adopting the new technology with respect to the traditional technology is greater than zero, while for others it is less than zero. This last assumption ensures that the new technology is 'good' for only a fraction of households in the population.

The household i believe the distribution of ω_{it} to be $p_{it}(\omega_{it})$ for them at period t. Every period, an uninformed agent has the option to become informed by putting effort $e_{it} \in \{0, 1\}$. Agents put effort only once, i.e., if $e_{i\tau} = 1$, $e_{it} = 1 \ \forall t \geq \tau$. If $e_{it} = 1$, the agent

[6] 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.

learns the true distribution $p_i^*(\omega_{it})$ at cost $\eta(e_{it}) = \eta_i$. Cost are also incurred only once - the first time the agent gets informed. If $e_{it} = 0$, no effort cost is incurred and the agent uses DeGroot averaging to approximate the true distribution. Let T denote the $n \times n$ weighted, directed, and non-negative influence matrix ($n = |\mathcal{I}|$), where $T_{ij} \geq 0$ represents the weight i places on j's opinion (with $\sum_{j \in \mathcal{I}} T_{ij} = 1$). Then $\hat{p}_{it}(\omega_{it}) = \sum_{j \in \mathcal{I}} T_{ij} p_{jt-1}(\omega_{jt-1})$ denotes household i's approximation based on others' opinion following the DeGroot averaging. The belief of agent i at period t is thus determined by the following process:

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

Finally, I assume the agents to be risk-neutral and myopic.

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

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

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

   Only uninformed agents make this decision.

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

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

   Similar to the decision to be informed, adoption is also irreversible. Note that I assume that only informed agents adopt (this is different from assuming that all informed agents adopt).

1.2 Implications of the Model

Consider the situation when there are only two states of the world: one where the new technology has a higher payoff than the traditional one (denoted 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} denote the true probability that for household i the new technology has a higher payoff than the traditional one. Suppose p^H_{it} 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 $\hat{p}^H_{it}$. Here $\hat{p}^H_{it}$ denotes the households' approximation of p^*_{iH} based on their network, following DeGroot averaging. Under this simplified scenario, let me now try to solve the model following backward induction.

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

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

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

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

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

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

Suppose for household j, $\bar{p}^*_{jH} + \bar{\eta}_j \geq p^*_{jH} \geq \bar{p}^*_{jH}$. Then even if p^H_{jt} is equal to p^*_{jH}, the household j will end up not getting informed about the technology. Hence they will not adopt the technology, even if its profitable for them to do so. This is due to the positive cost of learning (η_j). This feature is similar to the models of Chandrasekhar, Golub and Yang (2018) and Banerjee et al. (2018), where social stigma of information seeking can stop people from learning.

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

2 Simulations

In this section I consider networks of households whose true risk associated with a new technology are imperfectly correlated. In particular, the risks are positively related according to the existing network structure. Under such scenario, I first demonstrate the potential problem for a centrality based targeting strategy with a specific example. Then I simulate 1000 such networks to analyze whether the problem will persist on average, and compare the centrality based targeting with a probability based targeting strategy (defined below). I observe that the answer depends on the level of heterogeneity in a region, in terms of the applicability of the new technology.

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

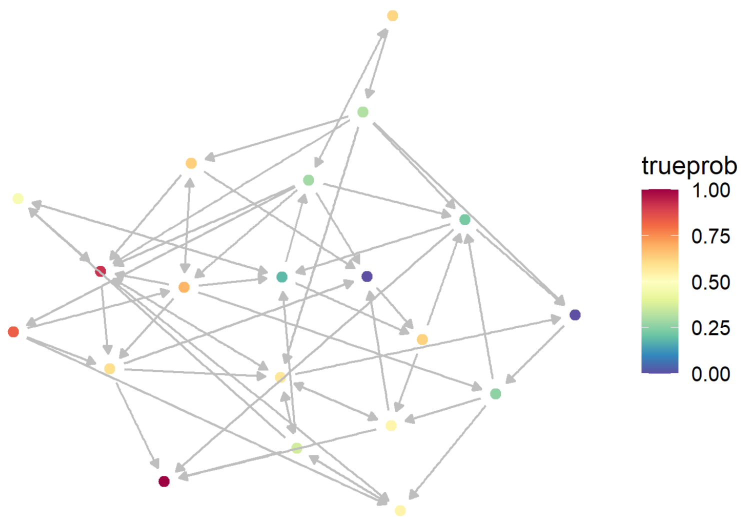


Figure 1: Distribution of True Probability within the network

Consider the scenario where, prior to any interventions, everyone believe their probabilities of success with the new technology to be zero ($p^H_{it} = 0 \ \forall it$). An intervention is then required with the objective of making every household, that should adopt under perfect information, learn about their true risk with the new technology. The ultimate policy goal behind such intervention would be to ensure the households, that should adopt the technology under perfect information, end up adopting. The efficiency of a targeting strategy can be measured as:

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

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

For the purpose of this example, consider the threshold probability of learning (i.e. $(\bar{p}^*_{iH} + \bar{\eta}_i)$) to be 0.3 for every household. In other words, if the true success probability of a household is more than 30%, 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 efficient for 70% of all households to get informed in this network.

The centrality based targeting strategy is to seed people that are central to the network. This is done based on some global measure of centrality. For the particular example here and the subsequent analysis in this section, I will consider betweenness centrality. The main results turn out to robust for another measure of centrality (consult appendix for more details)[7]. The probability based targeting strategy is to seed people that have the highest true probabilities of success with the new technology (i.e., highest p^*_{iH}s in the network). These are the people who are more likely to adopt a technology given the same probability threshold of learning for everyone and hence are considered to be the early adopters here. For each type of strategy, I will consider seeding only two households, similar to BBMM.

The following figure 2 capture seeding based on centrality based targeting. As central seeds are well connected in the network, by construction they represent the average households. Similarly, figure 3 capture seeding based on probability based targeting. The seeds here are not so well connected in the network and represent early adopters.

[7]For a discussion on different measures of centrality, consult Chapter 2 of Jackson (2010). In particular, the betweenness centrality measure captures how important a household is in terms of connecting the other households.

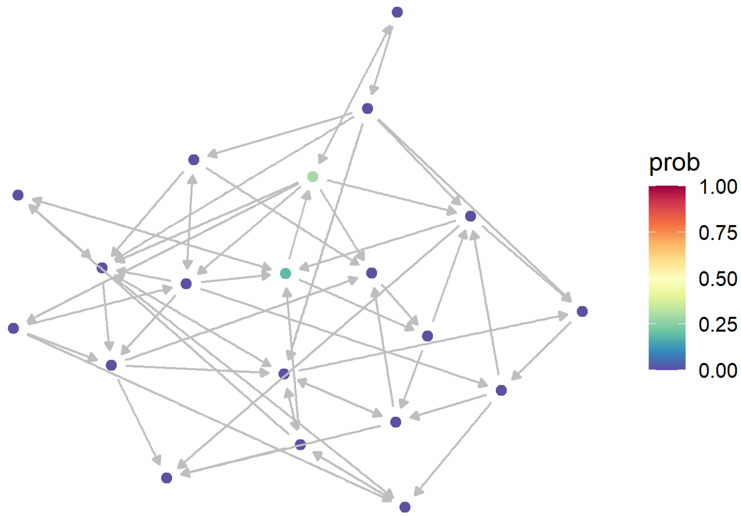


Figure 2: Seeding based on Centrality

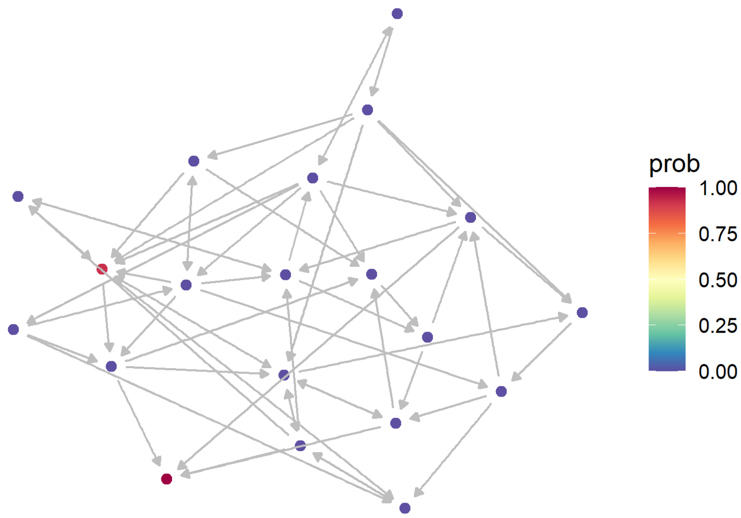


Figure 3: Seeding based on Probability

After the initial seeding, I let the diffusion take place for 10 periods, according to the diffusion process described in *Theoretical Framework* 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. Therefore, the centrality based targeting strategy fails miserably in this scenario.

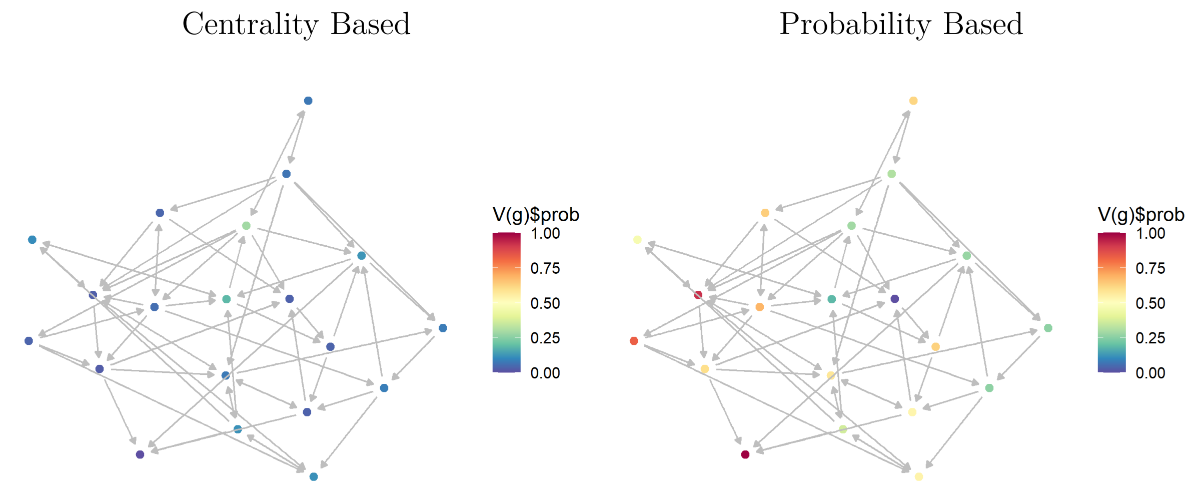


Figure 4: Performance of seeds after 10 periods

In this particular example, however, the p^*_{iH}s, although correlated according to the network's influence matrix, has a high variance. 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 1000 networks (given by their respective influence matrices), I vary the correlation level of p^*_{iH}s. The objective of this exercise is to note the relative performances of centrality and probability based targeting strategies over varying heterogeneity of the networks.

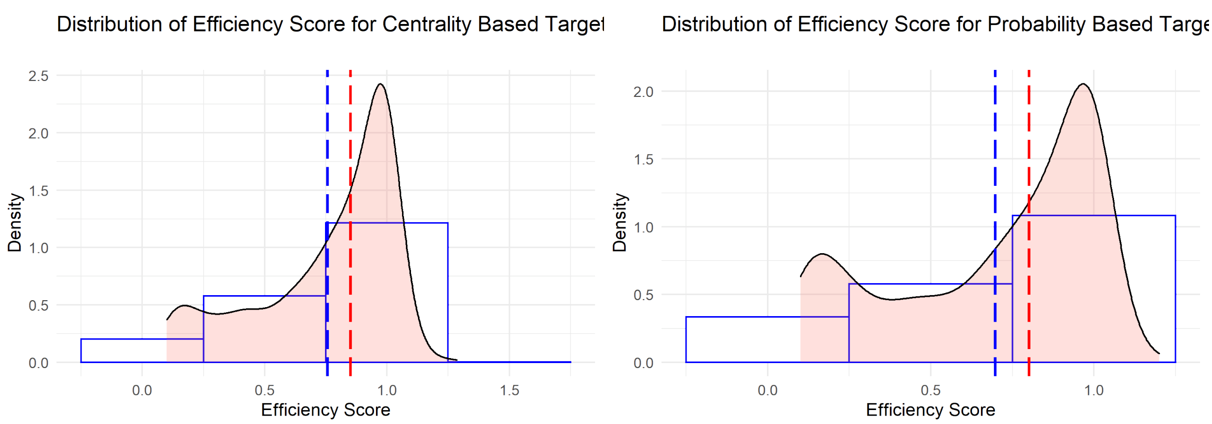
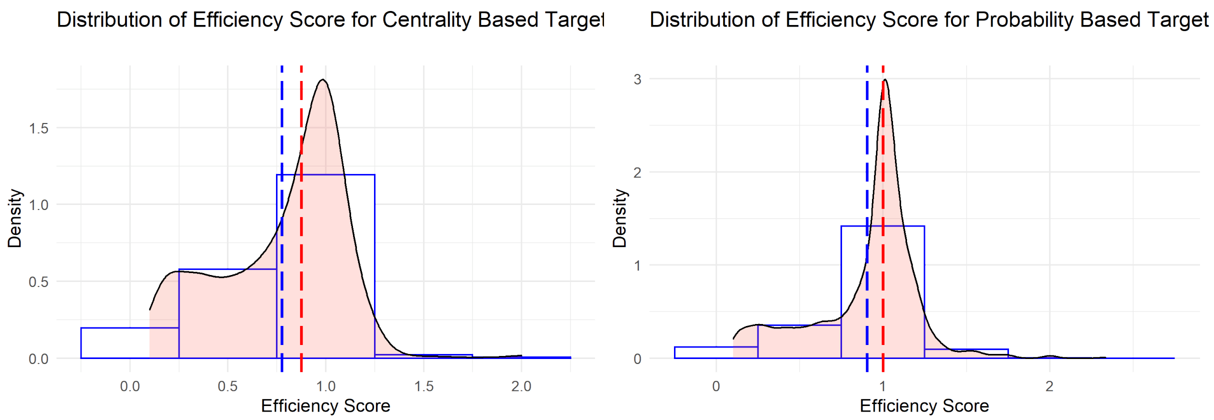


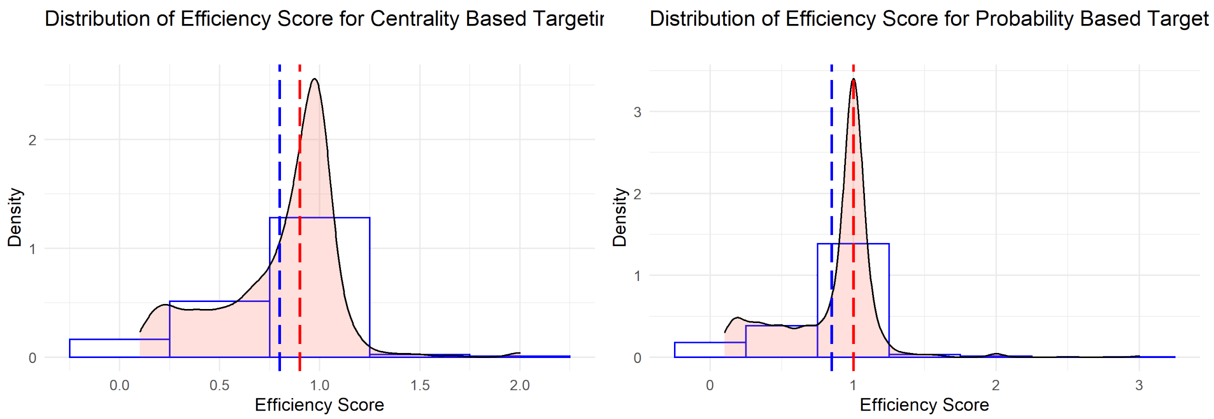
Figure 5: Distribution of efficiency scores when probabilities are perfectly correlated

Figure 5 represents the benchmark case of perfect correlation. For this figure and the next, the blue dashed lines represent the mean and the red dashed lines represent the median of each distribution. In this figure, centrality based targeting seems to have a thinner left tailed distribution than the probability based targeting. As can be seen from Table 1, when p^*_{iH}s are perfectly correlated, the mean efficiency score of centrality based targeting is 0.76. Which is higher than 0.70, the mean efficiency score of probability based targeting in this case. Even in terms of the median efficiency score, centrality based targeting (0.85) outperforms the probability based targeting (0.80). Both efficiency score distributions has similar variances. Thus, in the case of perfect correlation of p^*_{iH}s, the centrality based targeting performs better than the probability based targeting. This is the expected result. When p^*_{iH}s are perfectly correlated, my model reduces to a model of complex contagion where everyone faces the same risk. This is similar to the micro-foundation model of BBMM that assumes complex diffusion. Hence, the results are in favor of centrality based targeting.

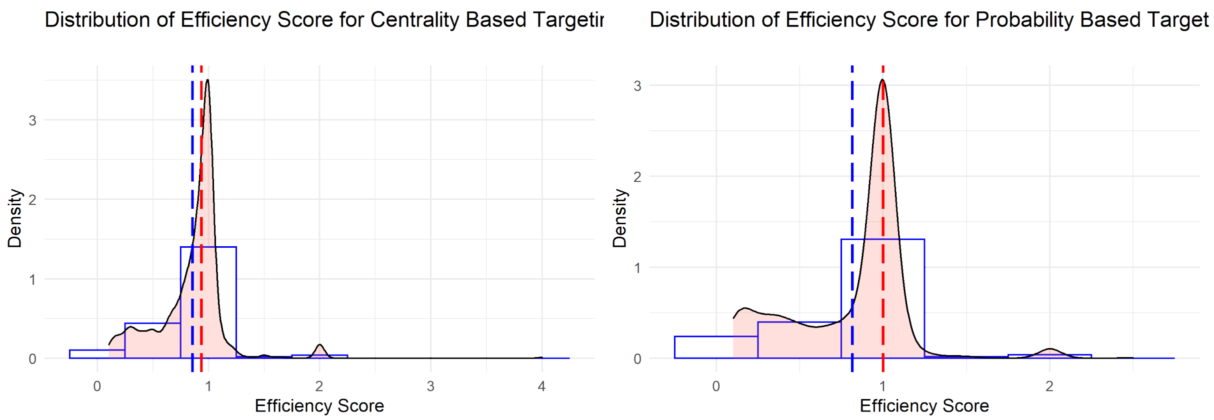
Having established the benchmark performance of the two targeting strategies, I move to explore the performances over different correlation levels of p^*_{iH}s. Figure 6 presents the efficiency score distributions for three different levels of heterogeneity in p^*_{iH}s. If there is a low correlation 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 centrality based targeting both in terms of mean (0.90 vs 0.78) and median (1 vs 0.88) efficiency scores, with the same variance (0.10). As the level of correlation increases to a medium level, I observe that the gaps are closing. Centrality based targeting has a mean of 0.80, compared to a mean of 0.85 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 still 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 (0.85 vs 0.82). In terms of the median efficiency, however, the probability based targeting (1) is still higher than centrality based targeting (0.93), although the difference is getting smaller. In this particular case, probability based targeting efficiency scores have a higher variance (0.13) than their centrality based targeting counterpart (0.10). All the results are also available in Table 1.



Low Level of Correlation



Medium Level of Correlation



High Level of Correlation

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

Table 1: Simulation Results

Strategy	Statistic	Level of Correlation Low	Medium	High	Perfect
Centrality Based	Mean	0.78	0.80	0.85	0.76
	Median	0.88	0.90	0.93	0.85
	Variance	0.10	0.09	0.10	0.08
Probability Based	Mean	0.90	0.85	0.82	0.70
	Median	1	1	1	0.80
	Variance	0.10	0.11	0.13	0.10
	Observations	998	992	973	1000

Note: Simulations on varying levels of correlation are all done for 1000 networks. However, upon generation of the true probabilities, some networks have to be dropped as they contained 0% of informed households under full efficiency.

From the above analysis, I observe that the relative performance of the two targeting strategies considered here depend on the level of heterogeneity within the network in terms of the applicability of the new technology. When the heterogeneity is high, there is a clear distinction between the people that should adopt the technology and the ones that should not. In that scenario, targeting people that are central to the network and hence, representative of the average household, is not efficient. Targeting households that are more representative of the people who should adopt the product works better. On the contrary, when the heterogeneity is low, everybody in the network are similar in terms of the applicability of the new technology. Targeting people that are central to the network, and thus representative of the average household, is a good strategy. This is because as the heterogeneity approaches to zero, targeting based on probability converges to random assignment (as everyone has the same probability of adoption). Then, similar to the existing literature, centrality based targeting outperforms random assignment given the same number of seeds. It is also worth noting that as the heterogeneity approaches to zero, we converge to the world of perfect correlation. Thus, the last panel of figure 6 is similar to figure 5. The actual numbers differ though. This is because the probabilities in figure 6 takes the network's influence matrix into account, the probabilities in figure 5 do not.

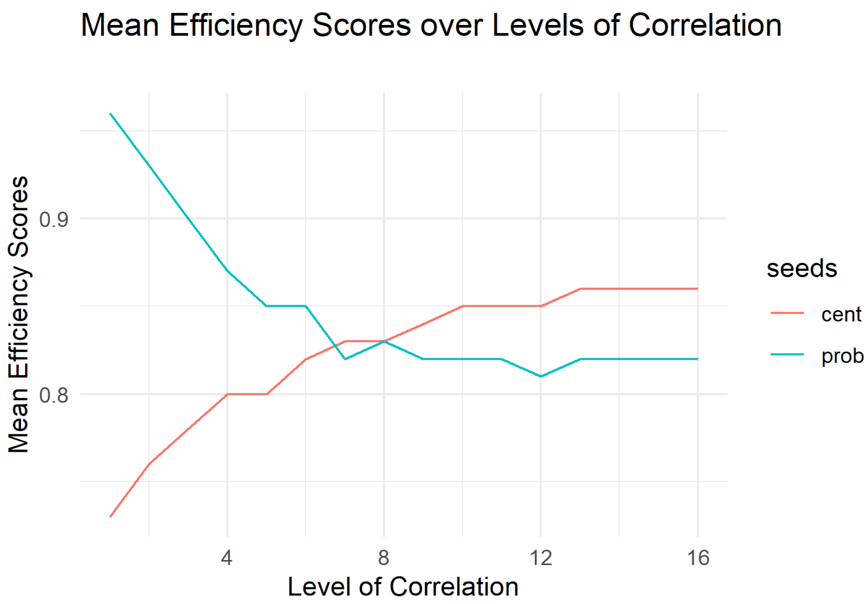


Figure 7: Mean efficiency scores over increasing levels of correlation

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

Figures 8 and 9 represent the median and variance of efficiency scores over a range of different levels of correlation. The results of figure 8 are not robust to alternative definition of centrality (details in Appendix C). Figure 9 do not show any particular patterns that can be tested.

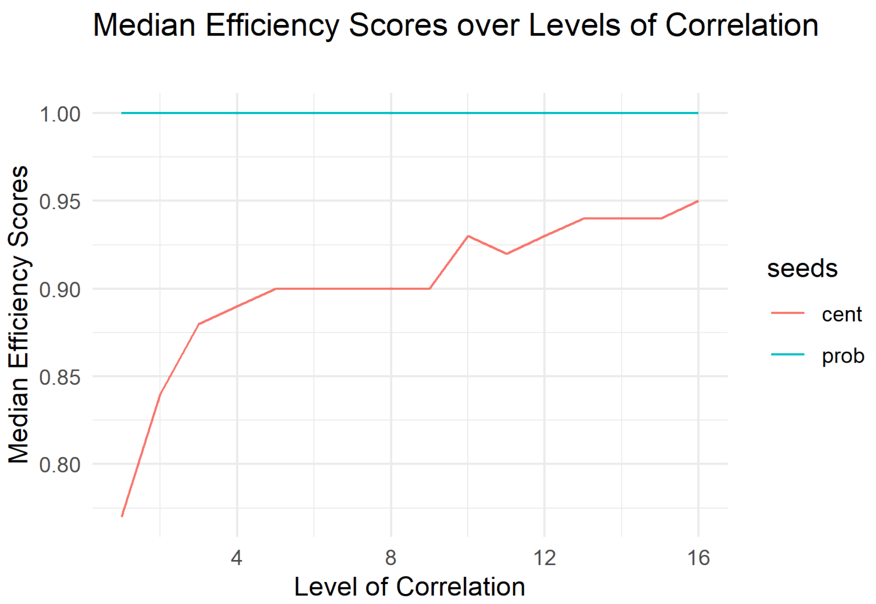


Figure 8: Median efficiency scores over increasing levels of correlation

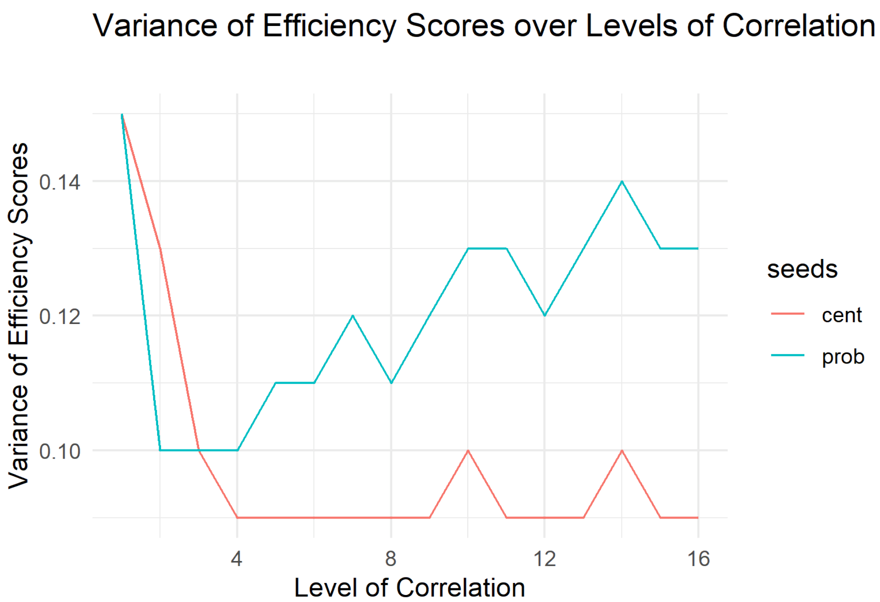


Figure 9: Variance of efficiency scores over increasing levels of correlation

3 Data Sources

For my empirical analysis, I use data from two different sources:

- Replication data of BBMM.

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

3.1 Replication data of BBMM

BBMM used Randomized Controlled Trial (RCT) to promote *Pit Planting* (PP) and *Crop Residue Management* (CRM) for Maize farmers in Malawi. CRM is not a 'new' technology in these areas, PP is. Hence I expect my model predictions to be valid for PP, and not valid for CRM. The researchers seeded 200 villages from 3 Malawian districts with semi-arid climates (Machinga, Mwanza, and Nkhotakota) with 2 'seed' farmers each, to induce widespread social learning. The intervention was to train these seed farmers on PP and CRM, with the training remaining the same across different treatment arms. The villages got equally divided into 4 treatment groups:

1. **Complex Contagion:** Seeding done assuming the underlying diffusion process to be of complex contagion. Both the chosen seeds turned out to be central in the network.

2. **Simple Contagion:** Seeding done assuming the underlying diffusion process to be of simple diffusion. As a result, one seed turned out to be central, while the other was in the periphery.

3. **Geo:** Seeding done solely based on geographic proximity. The seeds turned out to be near each other (in terms of geography), but not central (in terms of the network data).

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

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

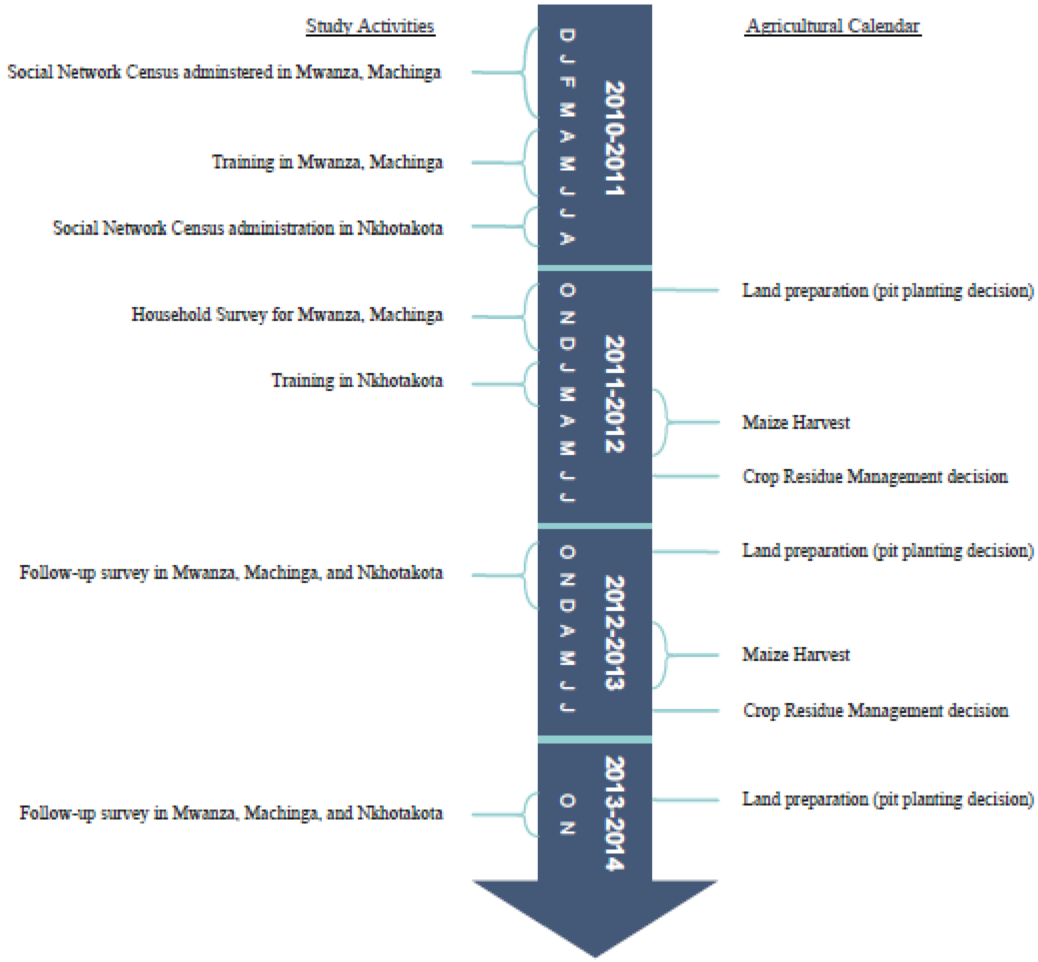


Figure 10: Project Timeline of Beaman et al. (2021)

The authors attempted randomly surveying a panel of around 30 households in each village, which involved all the seed and shadow farmers[8], as well as 22-24 other farmers.

[8] 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.

They managed to collect 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, that consists of 59 study villages, they collected only two rounds of survey data (in 2012 and 2013). The first round of survey, that was conducted a few months after the training of the seed farmers, attempted to capture some baseline characteristics and knowledge levels (of PP and CRM) of the households. 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 thrice for Machinga and Mwanza, and twice for Nkhotakota. On the other hand, since CRM is used after harvest, the CRM adoption decision was only observed twice for Machinga and Mwanza, and once for Nkhotakota. Figure 10, taken from the online appendix of BBMM, presents the full timeline of the project. More details on the intervention and sampling can be found in their paper.

3.2 AESTAS data

AESTAS is a nationally representative household survey conducted by International Food Policy Research Institute (IFPRI). Objective of this survey was to monitor the lead farmer program in Malawi[9]. The survey covered all 29 districts of Malawi[10], except Likoma. The data was collected in two waves: wave 1 in 2016 and wave 2 in 2018. The publicly available version of the survey dataset (citation needed) contain information from three different types of interviews:

1. **Household Interviews:** Random sample of around 10 households were selected for interview from randomly selected sections[11] within each district. Stratification was done based on whether or not the household had a lead farmer (LF). Per section,

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

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

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

up to two households with LFs were selected. A total of around 299 sections were surveyed. The same households were interviewed in the two waves with very small level of attrition (around 4%). Around 3000 households were covered in wave 1, with 2880 being re-interviewed in wave 2. For each household, both household head and their spouses were interviewed. The survey collected information on household level technology adoption, awareness, exposure; access to extension services; as well as socioeconomic and household characteristics.

2. **Lead Farmer (LF) Interviews:** Around 531 LF households were selected for household interviews. These households were additionally interviewed with a 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 number of lead farmers, type of training they received, number of projects, and other community characteristics.

For the purpose of this study, I will only be using the information collected through the household interviews. In particular, I am mostly interested in 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. More details 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).

4 Empirical Framework

Using the data sources described in the last section, the objective of this study is to empirically test the following hypothesis derived from my theoretical framework using simulations.

Hypothesis: 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

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

1. Centrality based seeding

2. Probability based seeding

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

Then I can use the following reduced-form regression:

$$Y_v = \alpha_0 + \alpha_1 Centrality_v + \alpha_2 Probability_v + \alpha_3 Heterogeneity_v + \alpha_4 Centrality_v \times Heterogeneity_v + \alpha_5 Probability_v \times Heterogeneity_v + u_v \quad (8)$$

Here Y_v denote some adoption related outcome for village v. $Centrality_v$ and $Probability_v$ dummies indicate whether the village got assigned to centrality or probability based seeding. $Heterogeneity_v$ measures the village level heterogeneity in terms of the applicability of a new technology. My hypothesis states that $\alpha_4 < 0$ and $\alpha_5 > 0$. u_v captures the random error term in the regression. I do not, however, have access to data from this ideal experiment. I use the replication data from BBMM together with the survey data from AESTAS conducted by IFPRI.

4.2 The Identification Strategy

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

$$Y_v = \beta_0 + \beta_1 Seed\ Centrality_v + \beta_2 Seed\ Probability_v + \beta_3 Heterogeneity_v + \beta_4 Seed\ Centrality_v \times Heterogeneity_v + \beta_5 Seed\ Probability_v \times Heterogeneity_v + \epsilon_v \quad (9)$$

I calculate average centrality of the seeds at the village level by using eigen-vector centrality of the seeded households at the baseline. This information is available in the replication data of BBMM. I proxy for average probability of adoption for the seeds at the village level by a predicted adoption variable that I calculate, at the baseline, using the estimates from another regression. This predicted adoption variable is calculated conditional on some observable household demographics. The other regression uses the survey data from AESTAS. The next sub-section provides details on this regression. I use the coefficient of variation of the same predicted adoption variable 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). Following my hypothesis, I expect $\beta_4 < 0$ and $\beta_5 > 0$. The random error of the regression is captured by ϵ_v.

The experiment in BBMM randomly assigned villages into a particular seeding strategy (for more details, consult *Data Sources*). This means, by construction, the seed characteristics were randomly selected. Hence, I argue that $Seed\ Centrality_v$ and $Seed\ Probability_v$ in (9) are exogenous sources of variation. The experimental design, however, ensures that some villages will have more central seeds than the other. Thus, not accounting for the treatment status in the regression can lead to some sort of omitted variable bias. For this reason, I include the treatment dummies in my set of control variables for the regression. The $Heterogeneity_v$ variable is the coefficient of variation of the predicted adoption at the village level. The predicted adoption is calculated conditional on some observable household demographics. Thus, if households from certain demographics are more likely to adopt a new technology, not controlling for these demographics in the regression can lead to omitted variable bias in identifying β_3 in (9). However, my coefficients of interest are β_4 and β_5. If the $Seed\ Centrality_v$ and $Seed\ Probability_v$ are exogenous, then even with an endogenous $Heterogeneity_v$ the interaction terms' coefficients can be identified without any bias.

4.3 Strategy for Approximating Probabilities of Adoption

For (9), I need to calculate $Seed\ Probability_v$. It is the average probability of adopting a new technology for the seeds. The probability of adopting a new technology needs to be calculated for other households as well. The later 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 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 an $Adoption\ Index_{it}$ for each household i at time period t. Details on the construction of this index can be found in Appendix B. Once calculated, $Adoption\ Index_{it}$ is used in the following regression:

$$Adoption\ Index_{it} = \gamma_1 X_{1it} + \gamma_2 X_{2it} + \mu_{it} \tag{10}$$

where X_{1it} are covariates that are available in the replication data from BBMM, X_{2it} are some other covariates available in the AESTAS dataset, included for improvement in estimation. The random error in the regression is captured by μ_{it}. For simplicity here, I consider an ordinary least square (OLS) specification. I check the robustness of my results with respect to different specifications in Appendix C.

Once (10) is estimated using the AESTAS data, I use estimated γ_1 (say, $\hat{\gamma_1}$) and X_{1it} in the BBMM sample to calculate $\hat{\gamma_1} X_{1it}$. I call this new variable *Predicted Adoption*, which is used as a proxy for the households' probability of adopting a new technology. The objective of this variable is to capture household level variation in terms of the applicability of a new technology.

5 Results and Discussion

I begin by presenting the results for regression (10). I use the estimated coefficients of this regression to calculate the *Predicted Adoption* variable. As discussed above, I use this variable as a proxy for the households' probability of adopting a new technology. The distribution of this variable, both for the AESTAS and BBMM data are presented and discussed here. Next sub-section provides descriptive statistics that help me understand the distributions of some key variables. These are *Seed Centrality*, *Seed Probability*, *Heterogeneity*, as well as the main outcome variables. Finally, in the last sub-section, I present the main empirical results of this study.

5.1 Approximating Probabilities of Adoption

Table 2 presents the results of regression (10) with two different dependent variables. These are *Adoption Index* and a similarly calculated measure *Usage Index*. Details of both their calculations are in Appendix B. The results do not vary much by outcome variable, or by whether or not the household controls are used. Here the sample of interest is the households surveyed in AESTAS.

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. However, the effect size is lower than that of the wealth indicators. The results are mostly highly significant with or without the household controls (with the exception of one variable: *No. of Children*). Columns (3) and (4) present the results for *Usage Index*, with and without the full set of household controls. The results here are very similar to the ones observed for *Adoption Index*, with the exception of the coefficients for housing material.

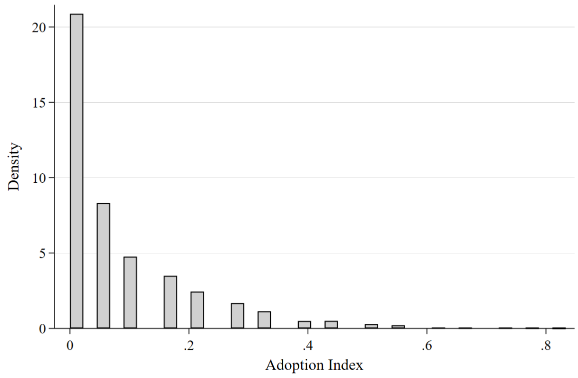
Figure 11 focuses on comparing the actual and predicted adoption indices for the AESTAS sample. For calculating predicted adoption, I use the results from column (2) of Table 2. The calculation is done conditional on four variables: *No. of Adults*, *Housing*, *Livestock* and *Assets*. The variable *No. of children* is excluded in the calculation as it has an insignificant coefficient. The in-sample prediction helps me compare the actual variation of adoption index with its predicted counterpart. The predicted variable has a standard deviation of 0.02, which is much lower than the 0.13 standard deviation of the original in the baseline. However, the objective here is not to perfectly predict adoption indices, it is to capture part of the variation in adoption indices conditional on some variables that are also observable in the BBMM replication dataset. The predicted adoption variable also need to be re-scaled such that it is between 0 and 1, which is the range of the adoption index. This also make the predicted adoption variable more comparable to a probability of success with the new technology, which it proxies for in my analysis. The predicted

Table 2: OLS Regression Results for Adoption and Usage Indices

	Adoption Index		Usage Index	
Variables	(1)	(2)	(3)	(4)
No. of Adults	0.008***	0.005**	0.011***	0.008***
	(0.002)	(0.002)	(0.002)	(0.002)
No. of 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 *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 cows, goats, sheep, pigs, and chickens. 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).

variable has a mean of 0.01, where the mean of the adoption index in the baseline is 0.09. Thus, I re-scale the variable by adding 0.1 to it, which makes the mean of the re-scaled predicted adoption 0.11. This is more comparable to the mean of the actual adoption index in the baseline.



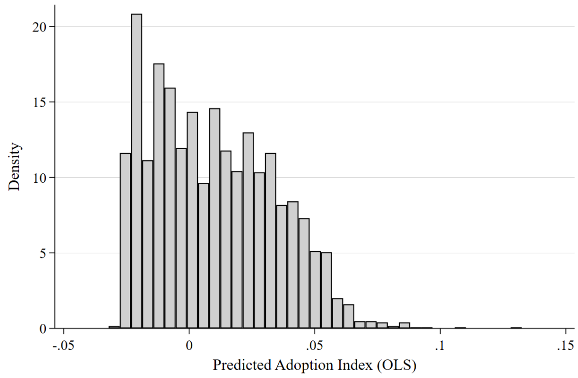


Figure 11: Actual and Predicted Adoption Index

Once done with the in-sample comparison, I use the estimated coefficients of *Adoption Index* from column (2) of Table 2 to calculate the predicted adoption for the BBMM sample. Figure 12 presents the *Predicted Adoption* variable calculated at the baseline. The calculation included the re-scaling mentioned above. In the baseline, the variable has a mean of 0.11 with a standard deviation of 0.02. This is exactly the mean and standard deviation of the re-scaled predicted adoption variable in the AESTAS data.

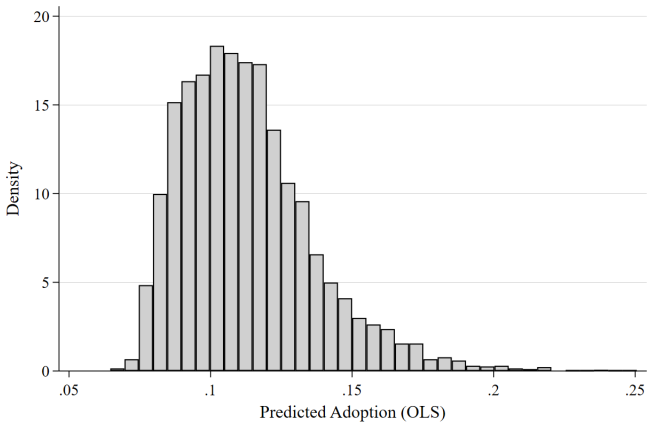


Figure 12: Predicted Adoption at the Baseline

5.2 Descriptive Statistics

In this section, I focus on the distributions of my main explanatory variables and outcomes of interest. I begin, in figure 13, by looking at the distribution of mean eigen-vector centrality for the village seeds at the baseline. I present it by different treatment groups.

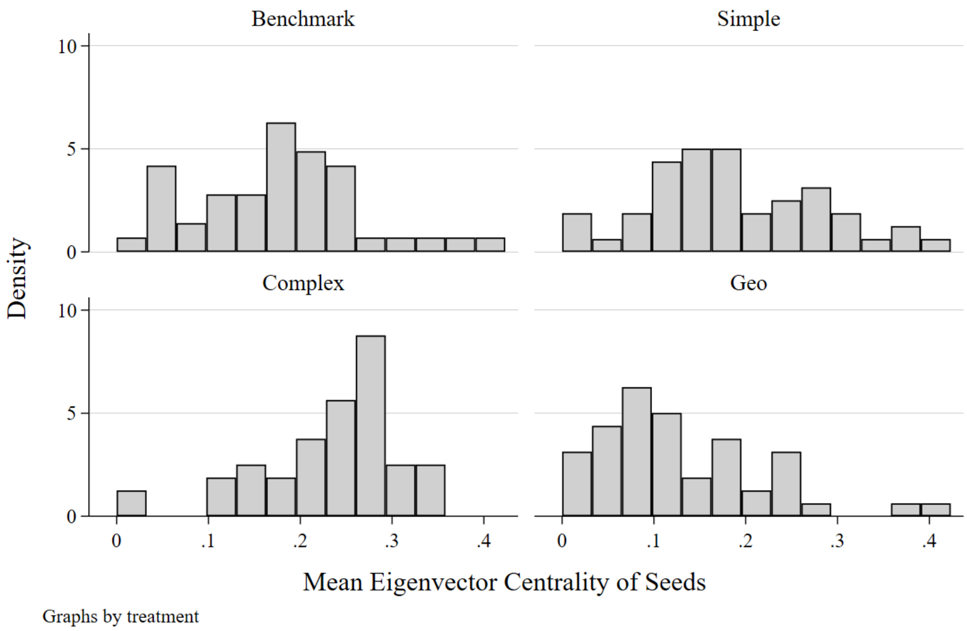


Figure 13: Baseline Mean Eigen-vector Centrality by Treatment Groups

By the design of the experiment, complex seeds are constructed to have the highest average centrality. This is because, as argued in BBMM, if the underlying model of diffusion is of complex contagion, it is optimal to seed only central households. 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 diffusion. BBMM also argues that geo seeds should be less central as, by design, they have less than average land (which is a measure of less than average wealth) and hence less likely to be well connected. In practice, I observe that baseline mean of eigen-vector centrality is highest for complex seeds at 0.23. It is 0.19 for the simple seeds, which is a bit higher than the benchmark mean of 0.18. Geo seeds indeed have the lowest average eigen-vector centrality of 0.13. However, in terms of the standard deviation, they are all similar and within 0.08-0.10. The mean values are comparable to the ones reported in the Table 1 of BBMM. I turn next to the distribution of average predicted adoption of village seeds, presented in figure 14. The calculations are done at the baseline by treatment groups. In general, no particular pattern is expected here. The micro-foundation of BBMM do not predict any differences in adoption, by treatment groups, for the seeds. In practice, however, some treatment groups have more central seeds than the other. The central seeds may differ in some particular demographics and the calculation of predicted adoption is conditional on those demographics. This may lead to some treatment groups having higher predicted adoption than the others. This do not turn out to be the case for this sample, though. All treatment groups have similar mean (within 0.11-0.13) and standard deviation (within 0.02-0.03) of their seeds' predicted adoption in the baseline.

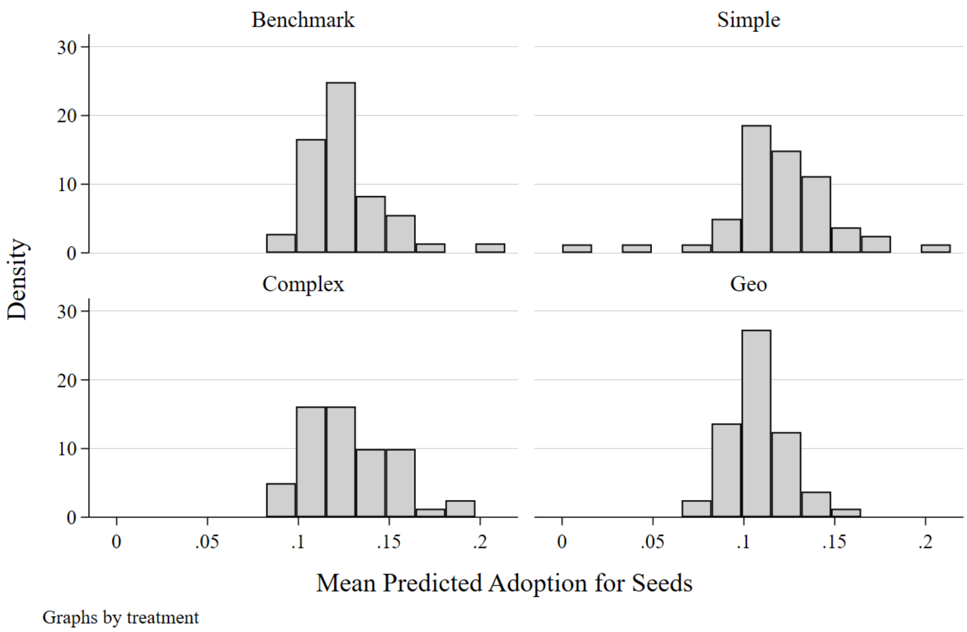


Figure 14: Baseline Mean Predicted Adoption by Treatment Groups

In figure 15, I look at the coefficient of variation of predicted adoption index, calculated at the baseline. In the baseline, the villages has a mean coefficient of variation of 0.19 for predicted adoption index. Standard deviation of the variable is 0.04. I use this variable to capture the village level heterogeneity in the applicability of a new technology. The variable range from 0.12 in the 1st percentile to 0.32 in the 99th percentile. Hence it has the variation I need for identification purposes.

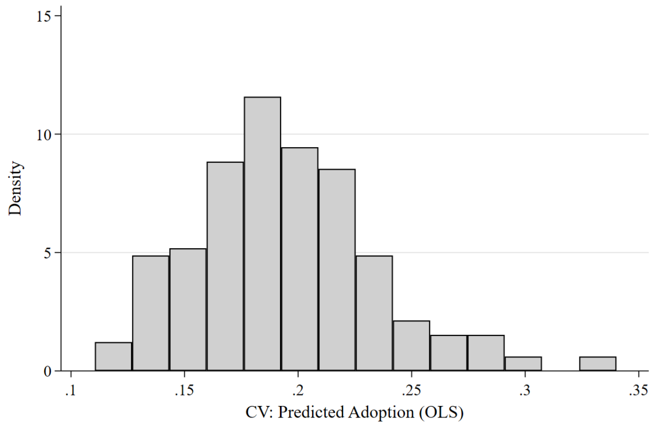
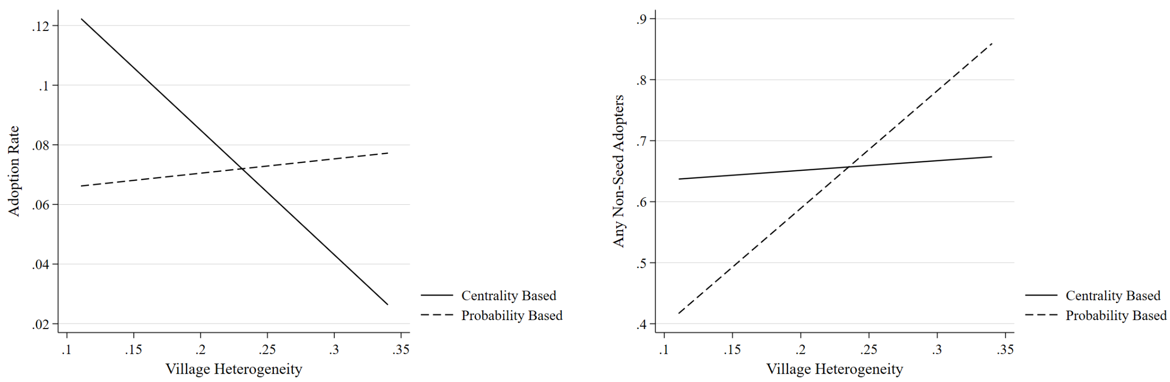


Figure 15: Baseline Village Level CV of Predicted Adoption

Finally, I turn to figure 16, which is similar to figure 7, calculated for the BBMM sample. For this figure, I consider the outcome variables of my analysis. These are $Adoption\ Rate$ and $Any\ Non-Seed\ Adopters$. Both these measures correspond to typical farmers in a village, i.e., the farmers that were not selected as *seed* or *shadow* farmers in the experiment. $Adoption\ Rate$ captures the proportion of typical farmers per village that adopted the pit planting in each agricultural season. $Any\ Non-Seed\ Adopters$ is a dummy variable that captures whether the village had at least one typical farmer adopting pit planting in an agricultural season. The same variables are used for the village-level analysis in BBMM.



Adoption Rate

Any Non-Seed Adopters

Figure 16: Outcomes of Different Seeding Strategies with respect to Village Heterogeneity

In figure 16 I present the outcome variables, for the second and third agricultural seasons, over varying level of village heterogeneity. The figure distinguishes between seeds that can be considered either *centrality based* or *probability based*. For the purpose of this figure, *centrality based* seeds are defined to be the seeds that have higher than median level of mean eigen-vector centrality in the baseline. Similarly, *probability based* seeds are defined here to have higher than median level of predicted adoption in the baseline. Based on my simulations, I expect the *centrality based* seeds to perform worse, and the *probability based* seeds to perform better, as the village heterogeneity increases. For the variable *Adoption Rate*, this is the pattern I observe. However, for the variable $Any\ Non-Seed\ Adopters$, both type of seeds seem to perform better with increasing level of village heterogeneity. Although, in this case, the performance of *centrality based* seeds improve at a much slower rate than the performance of *probability based* seeds. This, however, do not take into account the village-level heterogeneity in terms of other variables. In defining the *centrality based* and *probability based* seeds as dummy variables, the figures also fail to capture the full heterogeneity of these seeds' in terms of their centrality and probability measures. In the next sub-section, I present the reduced form results of my analysis that test my hypothesis more formally.

5.3 Reduced Form Results

Finally, I come to my main empirical results. Table 3 presents the results of regression (9). Columns (1) and (2) present the results for the outcome variable *Adoption Rate*, with and without the full set of village-level controls. Baseline average adoption of pit planting is only 2.6% (with a 6% standard deviation). In the subsequent years, adoption seem to increase with average centrality of the seeds. For an average village in the baseline, centrality of the seeds alone lead to 21-22.5% adoption rate. This is considering the baseline mean eigen-vector seed centrality of 0.18. However, once the baseline mean heterogeneity of the average village is taken into account (which is 0.19), the effect reduces to only 1.2-3% adoption rate. Thus, the impact of the central seeds on the adoption rate decreases with village heterogeneity. Average villages in the baseline also have a mean predicted adoption index of 0.12 for the seeds. As can be seen from the second row of columns (1) and (2), seeding based on likelihood of adoption alone has a negative impact on adoption rate. The effect size is a decrease of around 37-47.6%. Once the baseline mean heterogeneity of the average village is taken into account, however, the effect reduces to a decrease of 7-9% adoption rate. This supports my hypothesis that the impact of seeding based on likelihood of adoption improves with village heterogeneity. All these results are significant at 5% level. The village level heterogeneity alone leads to a 17.5-26.8% decrease in the adoption rate for the average village in the baseline. However, this effect is not significant once the village-level characteristics are controlled for.

Table 3: Village level Regression of Adoption Outcomes (Pit Planting)

	Adoption Rate		Any Non-Seed Adopters	
Variables	(1)	(2)	(3)	(4)
Eigen-vector Centrality of Seeds	1.235**	1.146**	1.404	1.831
(= $Seed\ Centrality_v$)	(0.563)	(0.522)	(1.247)	(1.199)
Predicted Adoption of Seeds	-3.968**	-3.120**	-11.460**	-6.336
(= $Seed\ Probability_v$)	(1.722)	(1.454)	(4.525)	(4.485)
CV of Predicted Adoption	-1.411**	-0.921	-4.779*	-0.205
(= $Heterogeneity_v$)	(0.647)	(0.584)	(2.590)	(2.667)
$Seed\ Centrality_v \times Heterogeneity_v$	-5.529**	-5.597**	-6.889	-11.690*
	(2.541)	(2.374)	(6.518)	(6.493)
$Seed\ Probability_v \times Heterogeneity_v$	18.070**	14.700**	53.190***	30.010
	(7.370)	(6.416)	(18.450)	(18.530)
Baseline Mean	0.026	0.026	0.346	0.346
(Standard Deviation)	(0.060)	(0.060)	(0.476)	(0.476)
Village-level Controls	No	Yes	No	Yes
Observations	324	324	324	324
R-squared	0.090	0.208	0.049	0.213

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

Columns (3) and (4) present the results for the outcome variable $Any\ Non{-}Seed\ Adopters$, with and without the full set of village-level controls. In the baseline, on average 34.6% villages have a typical farmer that adopted pit planting (with a standard deviation of 47.6%). Based on the data from subsequent years, centrality of the seeds alone can increase the average villages' probability of having at least one non-seed adopter by 25.5-33.3%.

Again, once the baseline mean heterogeneity of the average village is taken into account, the impact size reduces to an decrease of 7% to an increase of 1.7%. Most of these results are not statistically significant. The signs of the estimated coefficients, however, are as per my hypothesis that the positive impact of seed centrality decreases with village heterogeneity. When village controls are not taken into account (column (3)), I observe that seeding based on likelihood of adoption alone makes the average baseline village 137.5% less likely to have at least one typical farmer adopter. However, once the baseline mean heterogeneity of the average village is taken into account, the effect size reduces to a decrease of 16.5% only. This result is significant at the 5% level. Although the effect signs remain the same, both the results become insignificant in column (4) once the village-level controls are taken into account. The signs support my hypothesis that the impact of probability based seeding improves with village heterogeneity. Finally, the results suggest that the village level heterogeneity alone can make the baseline average village 4-90.8% less likely to have a non-seed adopter. The effect is not significant once the village-level characteristics are controlled for, and only significant at 10% level otherwise.

6 Summary and Concluding Remarks

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

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

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

References

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

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

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

Banerjee, Abhijit, Arun G Chandrasekhar, Esther Duflo, and Matthew O Jackson. 2019. "Using Gossips to Spread Information: Theory and Evidence from Two Randomized Controlled Trials." *The Review of Economic Studies*, 86(6): 2453–2490.

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

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

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

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

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

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

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

Cheng, Hoi Wai Jackie. 2021. "Factors Affecting Technological Diffusion Through Social Networks: A Review of the Empirical Evidence." *The World Bank Research Observer*. lkab010.

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

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

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

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

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

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

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

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

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

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

Appendices

A Technical Details on the Simulation Method

B Construction of Adoption Indices

C Robustness Checks

C.1 Simulations

Robustness 1: Simulation Results

		Level of Correlation			
Strategy	Statistic	Low	Medium	High	Perfect
Centrality Based	Mean	0.93	0.95	0.98	0.98
	Median	1	1	1	1
	Variance	0.09	0.08	0.06	0.01
Probability Based	Mean	0.90	0.85	0.82	0.70
	Median	1	1	1	0.80
	Variance	0.10	0.11	0.13	0.10
	Observations	998	992	973	1000

Note: Simulations on varying levels of correlation are all done for 1000 networks. However, upon generation of the true probabilities, some networks have to be dropped as they contained 0% of informed households under full efficiency.

C.2 Empirical Results

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

Variables	Adoption Rate (1)	(2)	Any Non-Seed Adopters (3)	(4)
Eigen-vector Centrality of Seeds	0.875*	0.787*	0.553	0.792
(= $Seed\ Centrality_v$)	(0.496)	(0.467)	(1.253)	(1.147)
Predicted Usage of Seeds	-2.800	-1.963	-5.822	-0.853
(= $Seed\ Probability_v$)	(1.840)	(1.558)	(4.823)	(4.513)
CV of Predicted Usage	-1.185	-0.663	-3.137	1.617
(= $Heterogeneity_v$)	(0.855)	(0.755)	(3.448)	(3.338)
$Seed\ Centrality_v \times Heterogeneity_v$	-4.244*	-4.361*	-2.806	-7.275
	(2.488)	(2.341)	(7.440)	(6.880)
$Seed\ Probability_v \times Heterogeneity_v$	14.830*	11.390	33.380	8.384
	(8.726)	(7.659)	(22.910)	(21.330)
Village-level Controls	No	Yes	No	Yes
Observations	324	324	324	324
R-squared	0.057	0.184	0.035	0.201

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

Robustness 3: Village level Regression of Adoption Outcomes (Pit Planting)

Variables	Adoption Rate (1)	(2)	Any Non-Seed Adopters (3)	(4)
Closeness Centrality of Seeds	0.640**	0.534**	0.829	0.794
(=$Seed\ Centrality_v$)	(0.306)	(0.258)	(0.617)	(0.566)
Predicted Adoption of Seeds	-2.814**	-2.205*	-10.160**	-4.907
(=$Seed\ Probability_v$)	(1.262)	(1.140)	(4.411)	(4.456)
CV of Predicted Adoption	-0.578	-0.274	-3.346	1.185
(=$Heterogeneity_v$)	(0.569)	(0.559)	(3.024)	(2.935)
$Seed\ Centrality_v \times Heterogeneity_v$	-2.755**	-2.554**	-4.479	-5.723*
	(1.395)	(1.224)	(3.276)	(3.224)
$Seed\ Probability_v \times Heterogeneity_v$	13.120**	10.250**	47.710***	22.600
	(5.422)	(5.001)	(17.760)	(18.040)
Village-level Controls	No	Yes	No	Yes
Observations	324	324	324	324
R-squared	0.096	0.203	0.050	0.212

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

Robustness 4: Village level Regression of Adoption Outcomes (Pit Planting)

Variables	Adoption Rate (1)	(2)	Any Non-Seed Adopters (3)	(4)
Seed Centrality Dummy	0.049**	0.047*	0.087	0.050
(= $Cent_v$)	(0.023)	(0.026)	(0.077)	(0.077)
Seed Adoption Dummy	-0.009	-0.010	-0.114	-0.114
(= $Prob_v$)	(0.022)	(0.022)	(0.077)	(0.075)
Adoption CV Dummy	0.016	0.026*	0.099	0.159*
(= Het_v)	(0.016)	(0.015)	(0.089)	(0.089)
$Cent_v \times Het_v$	-0.038	-0.071**	-0.043	-0.157
	(0.027)	(0.031)	(0.108)	(0.102)
$Prob_v \times Het_v$	0.006	0.021	0.069	0.128
	(0.027)	(0.027)	(0.108)	(0.102)
Village-level Controls	No	Yes	No	Yes
Observations	324	324	324	324
R-squared	0.030	0.182	0.049	0.220

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

Robustness 5: Village level Regression of Adoption Outcomes (Crop Residue Management)

Variables	Adoption Rate (1)	(2)	Any Non-Seed Adopters (3)	(4)
Eigen-vector Centrality of Seeds	0.677*	0.604	0.201	-0.053
(=$Seed\ Centrality_v$)	(0.395)	(0.433)	(1.400)	(1.281)
Predicted Adoption of Seeds	-0.911	-0.651	1.964	0.409
(=$Seed\ Probability_v$)	(1.696)	(1.856)	(5.017)	(4.900)
CV of Predicted Adoption	0.332	0.560	-0.043	-1.224
(=$Heterogeneity_v$)	(0.935)	(1.088)	(2.627)	(2.726)
$Seed\ Centrality_v \times Heterogeneity_v$	-3.012	-2.640	-3.598	-0.981
	(1.983)	(2.090)	(6.928)	(5.769)
$Seed\ Probability_v \times Heterogeneity_v$	4.038	2.130	-0.877	-0.939
	(7.759)	(8.447)	(24.63)	(24.09)
Village-level Controls	No	Yes	No	Yes
Observations	133	133	133	133
R-squared	0.029	0.059	0.043	0.178

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