Universitat Autònoma de Barcelona

IDEA Master Program

Term Paper for course "Networks"

Diffusion and Seeding with Discrete Information Set

March 13, 2022

Bakbergen Ryskulov

README

This report comes along with the all required material including code and dataset to replicate the simulation results. The simulation process was implemented in R, using "igraph" external package and custom functions. All the required code is well-documented at the top of each .R file.
Warning: The implementation of the model in paper is done with the purpose of prototyping, so the computation complexity is not optimized. Functions do not use parallelization and mainly rely on *for-loops* for better intuitive understanding of the diffusion process.

Simulation technical requirements:
Programming language: **R**
External package: **igraph**, **tictoc**

The report includes the following files:

- **main.R** : main script to run the simulation
- **info_measure.R** : function which calculates the statistics used to assess the outcomes of the given seeding strategy
- **mc_simulate.R** : function which runs a Monte Carlo simulation for a diffusion process with discrete information set described in the paper
- **random_seeding.R**: function which randomly draws a pair of nodes to seed and returns an average of their outcomes
- **simulate_and_plot.R**: function which runs a Monte Carlo simulation for a given pair of nodes and plots igraph network, marking the seeded nodes with red color
- **village2.csv**: the dataset used in the simulation

Abstract

In this paper, I model the diffusion process where the node's information set is not restricted to be binary, but contains some subset of the initial information diffused. As a *preliminary result* from Monte Carlo simulation, I find that although the diffusion process cannot be characterized as "simple contagion", it is optimal to seed central and far apart nodes. However, this result could be driven by high clustering property of the chosen network and further research is required.

1. Introduction

Spreading information in the given community is a big challenge for policymakers, businesses and organizations. Better understanding of the information diffusion process helps us to find optimal ways and channels to spread information with minimum costs. While the most popular channels are broadcasting and seeding, the seeding strategy still remains commonly used due to its low costs, especially in the developing countries. Therefore, finding the best people to seed the initial information remains an important question for policymakers.

The central question of this paper is to find the optimal seeding strategy when information is complex[1] and can be broken into discrete parts. There are already some research done that helps us understand better the diffusion processes in networks. Banerjee et al. (2013)[1] study the diffusion of microfinance in the Indian villages, trying to disentangle contagion and adoption by estimating a structural model with information effects and endorsement effects. They do so by modelling the diffusion process and assigning the nodes to two different binary states. In the reduced-form estimation they find that eigenvector centrality matters most for maximizing diffusion and in the structural estimation they find that endorsement effect are absent. However, since the paper takes information status as a binary status, it doesn't help us to analyze the diffusion in complex information setting.

In another paper, Banerjee et al. (2018)[2] study the choice of a dissemination strategy, when it affects engagement in social learning. The main idea is that people have reputational concerns in asking questions, which leads them to engage less in social learning and therefore, it decreases the diffusion of information. Its relevance to this paper is that it studies the case of India's demonetization, when people need to ask question to comprehend the information. For that reason, in the experimental part of the paper, the information is a list of facts. Although the paper yields interesting results between broadcasting and seeding strategies, it doesn't explicitly model the information as a set and doesn't study the diffusion under this specific setting. Still paper of Banerjee et al. (2018)[2] gives a motivational case, when modelling the information as a set can be

[1] By complex, I mean that information can consist from different parts, which can be facts, rules, etc.

practically beneficial.

There are some papers in the literature, which depart from this simple notion of information state. For example, Jackson et al. (2019)[3] model a diffusion process when information is noisily relayed from person to person. The main result is that when there is uncertainty about mutation rates, optimizing learning requires either capping depth (how many times information is relayed), or if that is not possible, limiting breadth (the number of relay chains accessed) by capping the number of people to whom someone can forward a message. However, although the information is allowed to be noisily transmitted, it is still assumed to be simple and node's information set is a binary state.

Regarding the more complex diffusion processes, the paper of Beaman et al. (2021)[4] shed more light on our question. In their work they find that targeting central farmers based on network data is important to accelerate the diffusion process. The possible explanation in favor of this centrality is that diffusion process they study is governed by complex contagion. The theoretical and simulation analysis of my paper adapts some methodology from their work.

In this paper, I model the diffusion process where the node's information set is not restricted to be binary and contains some subset of the initial information diffused[2]. In this setting, the information is transmitted partially, i.e. randomly only part of the information can be transmitted further with some probability. The main motivation of the model is to find the optimal seeding strategy under such discrete information set setting.

For some given set of parameters and for one real network, Monte Carlo simulation yields contrary results to my analytical expectations, i.e. simulation suggests that it is optimal to seed central and far apart nodes as in "simple contagion" case. This result could be driven due to high clustering property of the chosen network for simulation.

The rest of the paper is organized as follows. In the Section 2, I formally characterize the diffusion process and general environment. Moreover, I discuss the outcomes of diffusion in terms of "simple" and "complex" contagion processes. In the Section 3, I provide the necessary information for simulating the model given the set of parameters. Specifically, I provide descriptive statistics of chosen network, the methodology of simulation and preliminary results. Finally, in Section 4, I briefly discuss the possible extensions of the model.

2. Model

2.1 Diffusion process

In this model, I assume that the information is shared via word-of-mouth in oral or written form using social media services or not. The main assumption of the model is that

[2] including empty set, if the node hasn't received any information in the diffusion process

the information diffused by the policymakers can be discretely splitted and the agents can randomly transmit some of its part.

I see it as a realistic assumption, because the information policymakers share can be complex and people may forget some of its parts. Some of the examples are list of measures that individuals should take into account in Covid time, list of changes in the rules passed by the local government, steps to undertake in order to participate in local elections, etc.

For the sake of simplicity, I model the initial information diffused denoted as s_0 as the set of numbers from 1 to 10. Intuitively, it can be thought as a set of rules the policymaker want to diffuse into a community.

$$s_0 = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\} \tag{1}$$

At each time t the individual i have the information v_i^t, which is the subset of the initial information diffused:

$$v_i^t \subset s_0 \tag{2}$$

At initial state $t = 0$, some nodes are chosen as primary nodes and seeded the full information s_0. This nodes are denoted as $P = \{i \in N : i \text{ is primary node}\}$, where N is the set of all nodes in the network. At each period t, each node i with probability p transmits the signal s_{ij}^t to the node j. The node that shares the information can share information that she already knows, so

$$s_{ij}^t \subset v_i^t \tag{3}$$

Moreover, the nodes at each time t share only the subset of the information they have. For the most simple case, I assume that it is some fixed fraction k of knowledge the node has. If the node i transmits the signal to the node j, this node cannot transmit the signal anymore in the future periods.

The receiving node updates its knowledge by aggregating its previous information and newly received information from all linked nodes, i.e. by taking union of all information it has received. It is reasonable to assume, because in conversations people know the topic of discussion. Given the topic of conservation, people can easily recollect the missing parts of the information if they're provided with them.

$$v_i^t = \bigcup_j s_{ij}^t \cup v_i^{t-1} \tag{4}$$

I also define the total number of nodes that has some part of information as C^t and total amount of information in the network as I^t, which is simply the total number of elements each node has at the period t of the simulation.

$$I^t = \sum_i \mid v_i^t \mid \quad \text{for any t} \tag{5}$$

The diffusion process can be summarized as follows:

At $\mathbf{t} = 0$

- Subset of nodes are selected as primary nodes, which are defined as a set P.
- Primary nodes are seeded with the full information, i.e. $v_i^0 = s_o$ for i $\in P$.

Iterate for each time $\mathbf{t} \in \{1, 2, ..., T\}$

- Informed nodes can transmit to their connected nodes the k fraction of their current knowledge with probability p. The information can go through any given edge only once in the whole simulation.
- The receiving nodes update their knowledge by aggregating all the knowledge they receive. They keep their current information and add to it the information they receive in the current period.

To keep the model simple, I seed only two nodes with the information. The goal in the seeding strategy is straightforward: maximize the total information the people have using different seeding strategies. One of the initial hypothesis to test is how close the optimal seeding in this setting to the seeding with standard centrality measures such as eigenvector centrality.

2.2 Analytical Expectations

There is a bulk of theoretical and empirical research papers on choosing the optimal seeding. As it was mentioned in the introduction section, Banerjee et al. (2013)[1] provide evidence in favor of choosing the central nodes based on randomized control trials. Akbarpour, Malladi and Saberi (2020)[5] show in their theoretical work that adding a few additional seeds leads to more diffusion than targeting central nodes in the network.

However, as Beaman et al. (2021)[4] note, these results hold if three conditions are satisfied:

1. The diffusion process is charactectized by "simple contagion", i.e. agents are infected or receive information after a single exposure to someone else already infected
2. Time period for adaption is sufficiently long
3. Social interaction within the network is frequent

In my theoretical setting, it is reasonable to assume the latter two conditions. In particular, we can assume that time period is sufficiently long, because the object of diffusion is information. This is not true, for example, for technology diffusion. The second condition largely depends on the information being diffused. The information which is not important for the community can quickly stop spreading simply because people stop talking about it. So even given sufficiently long time, the information won't spread through well-connected network. However, the matter of concern of this paper is the information diffused by policymakers. For that reason, we can assume that information is important enough and most likely to be discussed or spoken in conversations. We can also reasonably assume that social interaction within the network is frequent, since spreading the information bears low cost.

The problematic conditions is the first one. In the "simple contagion" models exposure to more informed nodes increases only the probability of being informed, since the node can take only a binary state: informed or not informed. In our model, we cannot assume the "simple contagion", because for a given node the exposure to more informed nodes increases not only the chances to be informed, but also to enlarge its current information set. Therefore, seeding the information far apart to minimize redundancy is the same area can be not an efficient strategy.

Beaman et al. (2021)[4] develop a threshold model to capture the diffusion process which can be characterized as "complex contagion", where sharing the informed connections can increase the chances of technology adoption. In our case, it is difficult to characterize the diffusion process with dicrete information set as "complex contagion", at least in Beaman et al.'s perspective. But, intuitively, the general implications of "complex contagion" can apply.

The core of the diffusion in our model is that an informed node can transmit the k fraction of information it possesses, and this signal is randomly drawn for its each neighbor. In other words, at time t an informed node i can pass different sets of information to its connected nodes j and l. It implies that if these nodes j and l interact in the future period, they can get the missing part of the information from each other, or to be more precise, one of them can obtain[3]. This feature of the model makes it more attractive to seed initial nodes close to each other. Because once these nodes start transmitting some fraction of the initial information, their neighbors can infer the missing parts of information from each other. This can be thought as sparkles in the network and seeding nodes close to each other make this sparkle stronger and more difficult to fade away.

To sum up, the expectation is that seeding a pair of central nodes close to each other yields higher diffusion and higher total information set I^t, which is formulated in the following proposition:

Proposition 1 *In a diffusion process with a discrete information setting, it is optimal to seed a pair of central nodes close to each other.*

3. Simulation

The easy way to see and follow how diffusion process goes is simulation. In this part, I describe the main aspects of the simulation process.

[3] because for now we assume information transmission only to one direction

3.1 Data

The standard way for simulations is to run them on random graphs for benchmarking purposes. But most of them require distributional assumptions on degrees and other parameters of the model, which enlarges our spaces of assumptions. The simpliest random graph "Erdos-Reny" is not a good candidate as well due to its low clustering property. As we discussed in the previous section, clustering drives the diffusion in our model.

To run the simulation I chose one of the 75 villages in rural southern Karnataka, a state in India, which was collected by a NSF-Funded Project led by Banerjee et al. (2013)[6]. I chose this network, because it represents a real network and this kind of villages are usually a target for policymakers. The network has 195 nodes (N) and 674 edges (M). The plot of the network is shown in the figure 1. The network is relatively small which makes simulation easier. I provide the code for this paper, and simulation can be easily applied to other networks.

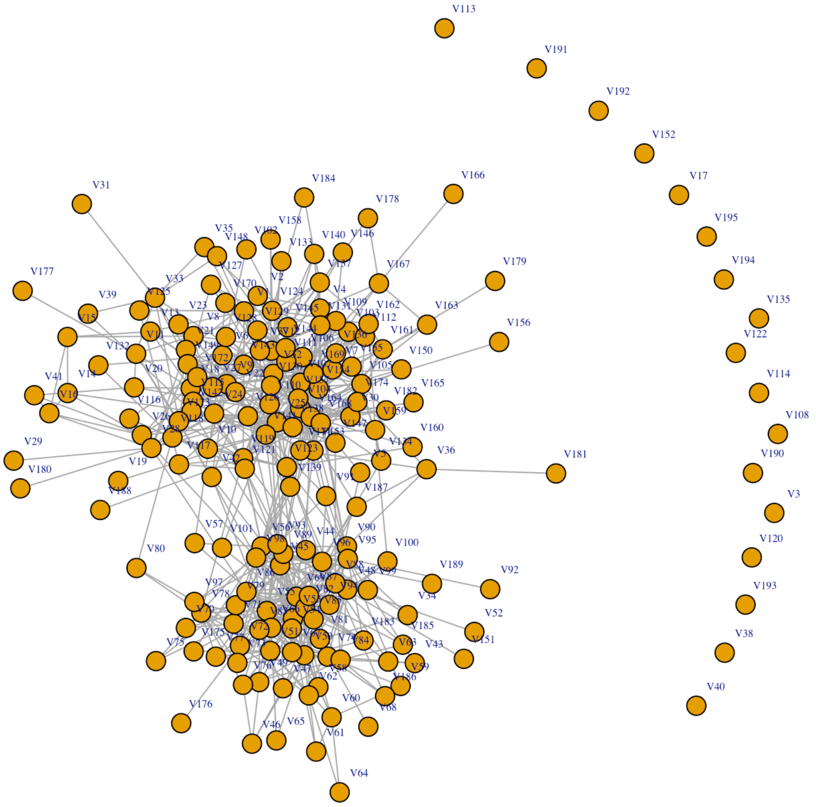

Figure 1: Plot of the village network

3.2 Simulation methodology

The simulation itself is close in its methodology to Beaman et al. (2021)[4]. I run Monte Carlo simulation given randomness present in the model. In particular, there are two sources of randomness:

- Randomness in p, i.e. the probability that an informed node i will transmit the information to uninformed node j
- The information set that node i transmits to j, i.e. the objects that are in a set s_{ij}^t

For the sake of simplicity, I study seeding of two nodes. The other parameters of the simulation model are as follows:

- $T = 5$. The number of periods of the simulation
- $p = 0.5$. The probability that an informed node i will transmit the information to uninformed node j
- $k = 0.7$. The fraction of information transmitted
- $S = 500$. The number of iterations for Monte Carlo simulation

The simulation is done for the following seeding strategies:

1. Random seeding
2. Eigen-central nodes (close and far apart pairs)
3. Search the optimal seeding through running the simulation for all possible pair of nodes.

In the simulation process, I keep track of the following statistic measures listed below. I chose these measures because they are the most informative and other measures can be inferred from them. For example, if we want to obtain average information percentage of the network, we can simply multiply I^T by maximum possbile amount of total information ($N\times \mid s_0 \mid$), which for the chosen village equals $195 \times 10 = 1950$.

1. C^T: The total number of informed nodes at T
2. I^T: Total amount of information the network has at T
3. σ_v^T: The standard deviation of cardinality of information sets v_i^t at T
4. J^T: The total amount of information the nodes acquired after being already informed, i.e. having $\mid v_i^t \mid > 0$

3.3 Simulation results

In random seeding I run Monte Carlo simulation for 500 randomly drawn pair of nodes and average their results. The total number of combinations is equal to $18,915$. The results can be be seen in the 2nd column denoted as "Random" of the Table 1.

To find eigen-central nodes, I compute eigen-centrality measure for every node. I pick the node with the highest eigen-centrality measure (node 88) and then using it create 6 pairs. First three pairs are chosen as the pair of node 88 and another eigen-central node, which is **close** to node 88 and have the high eigen-centrality. These pairs are visualized on the Figure 2. The relative centrality of the node and the distance between a pair of nodes can be inferred from the plot. Another 3 pairs of node are created as the pair of node 88 and another eigen-central node, which is **far** from node 88 and have the high eigen-centrality. These pairs which a bit farther apart from each other are visualised on the Figure 3. The results of Monte Carlo simulation for eigen-central nodes are reported in the Table 1.

It is evident that random seeding yields the worse results, but it can be useful as a benchmark to compare with other seeding methods. On average, the measure statistics of close eigen-central nodes are pretty similar to far eigen-central nodes. However, it seems like far eigen-cental nodes yield a slightly higher total information I^T. This results is surprising and in contrast with my analytical expectation in the Section 2.

The possible explanation is that the chose network is highly-clustered with clustering coefficient of 0.2279. In the first period after seeding, high-clustering secures that the sparkle is strong enough to boost the diffusion process afterwards.

Table 1: Simulation results for each seeding

Seeding / Statistic	Random	Close eigen (88, 55)	Close eigen (88, 86)	Close eigen (88, 69)	Far eigen (88, 123)	Far eigen (88, 141)	Far eigen (88, 169)
C^T	153.46	159.91	159.90	160.22	160.56	160.81	160.75
I^T	1298.98	1417.45	1414.93	1432.17	1444.03	1462.60	1463.69
σ_v^T	3.5985	3.6652	3.6596	3.6625	3.6559	3.6584	3.6613
J^T	267.28	260.04	260.79	252.69	249.70	233.18	232.63

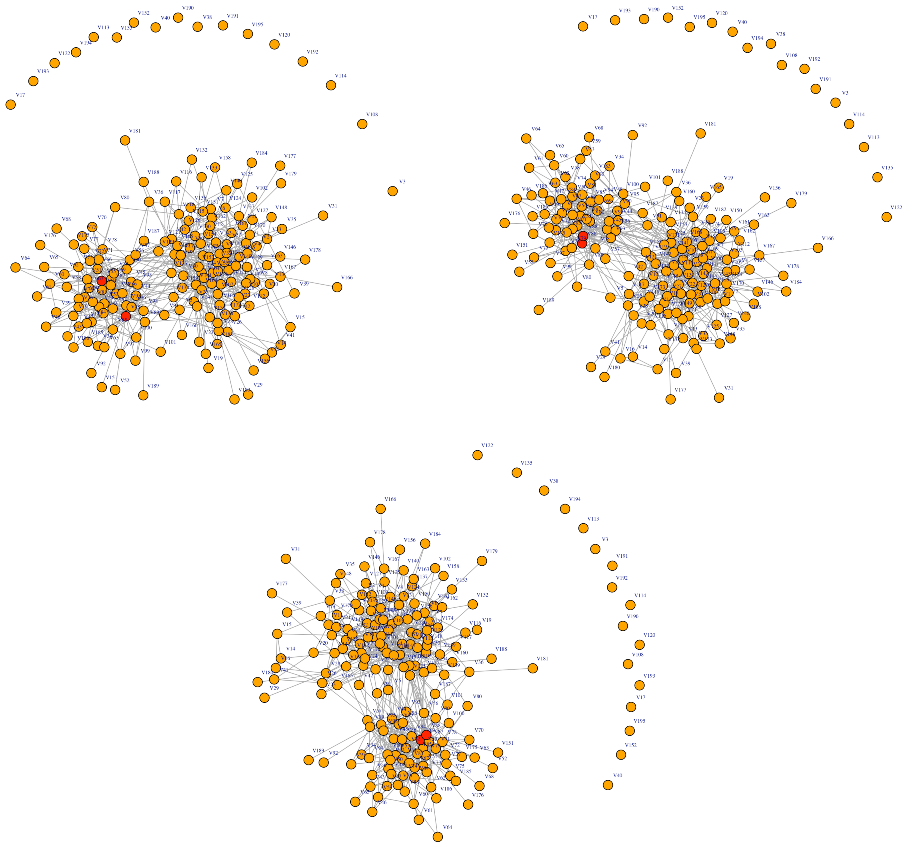

Figure 2: Plot of eigen-central close pair of nodes

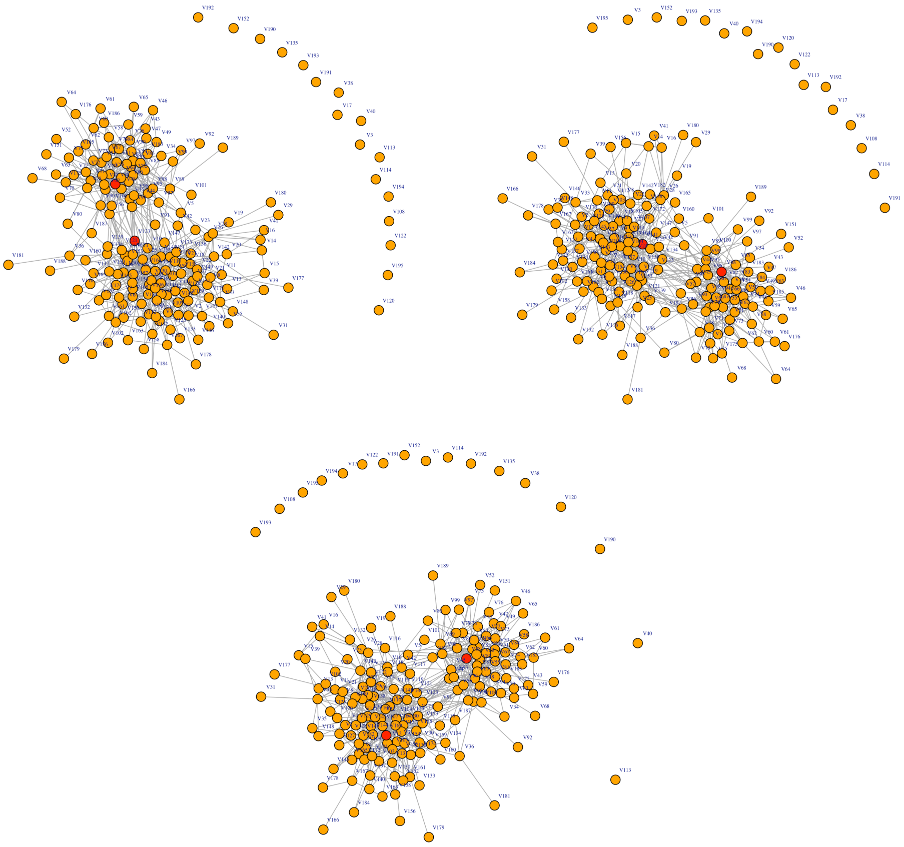

Figure 3: Plot of eigen-central far pair of nodes

To find the optimal seeding strategy it is necessary to run brute-force exhaustive search and analyze the properties of the pairs, which yield the best outcomes. There are $18,915$ possible combinations of pairs. Monte Carlo simulation for one pair of nodes takes approximately 7.1 seconds on my *Intel i5 8th generation* processor, so running brute-force search will take me approximately 37 hours. For this reason, I leave this computational exercise for the future.

4. Extensions

There are several important extensions of the model and simulation methodology. I discuss them briefly below.

For simulation, it is important to run the diffusion process for different type of networks. In the previous section, I got the results that are contrary to my analytical expectations. The possible explanation is the property of the chosen network, in particular, the high clustering. In the future analysis, I would like to simulate the model on other real-world networks with different network properties.

As for the model, it is interesting to incorporate heterogeneity. It can be implemented by introducing heterogeneity to random parameters such as p and k. The assumption would be that some individuals are more probable to transmit information or transmit the higher fraction of information they possess. The question of interest is how much diffusion process is dependent on heterogeneity of individuals. Would the diffusion process fade away quickly if it encounters low-type individuals at the initial periods?

In the model, we assumed uni-directional transmission of information. However, it is also reasonable to assume that people can talk, discuss and exchange the information. So, the natural extension would be simultaneous information exchange. Specifically, if any given edge is activated given probability p and both of the nodes are informed, then each of them transmit some fraction of information and each of them aggregates the received information.

In the seeding strategy, we looked only on seeding a pair of nodes. It would be also interesting to study the optimal seeding strategy if we seed more initial nodes, because it usually the case that policymakers 5-6 key nodes. Would the general results still hold in this case?

Bibliography

[1] Banerjee A. et al. “The Diffusion of Microfinance”. In: *Science* 341.6144 (2013), p. 1236498. DOI: 10.1126/science.1236498. eprint: https://www.science.org/doi/pdf/10.1126/science.1236498. URL: https://www.science.org/doi/abs/10.1126/science.1236498.

[2] Banerjee A. et al. *When Less is More: Experimental Evidence on Information Delivery During India’s Demonetization*. Working Paper 24679. National Bureau of Economic Research, June 2018. DOI: 10.3386/w24679. URL: http://www.nber.org/papers/w24679.

[3] Jackson O. M., Malladi S., and McAdams D. *"Learning through the Grapevine: The Impact of Noise and the Breadth and Depth of Social Networks"*. Working Paper. Available at SSRN, Mar. 2019. URL: https://ssrn.com/abstract=3269543.

[4] Beaman L. et al. “"Can Network Theory-Based Targeting Increase Technology Adoption?"”. In: *American Economic Review* 111.6 (June 2021), pp. 1918–43. DOI: 10.1257/aer.20200295. URL: https://www.aeaweb.org/articles?id=10.1257/aer.20200295.

[5] Akbarpour M., Malladi S., and Saberi A. *"Just a Few Seeds More: Value of Network Information for Diffusion"*. Working Paper. Available at SSRN, Aug. 2020. DOI: http://dx.doi.org/10.2139/ssrn.3062830. URL: https://ssrn.com/abstract=3062830.

[6] Abhijit Banerjee et al. *The Diffusion of Microfinance*. Version V9. 2013. DOI: 10.7910/DVN/U3BIHX. URL: https://doi.org/10.7910/DVN/U3BIHX.