Assessing the effect of social structure on network diffusion

Candidate Number : 1029205

This manuscript was compiled on March 27, 2019

The rate of propagation in a diffusion process on a network depends on the initial "injection points", or nodes initially affected. In particular, it has been shown that Eigenvector Centrality is empirically correlated to the resulting number of future affected nodes. However, this dependence is also governed by more global network characteristics such as its connectedness or modularity. Here we use the data set gathered in the paper "The Diffusion of Microfinance" (BANERJEE et al. Science, 2013, vol. 341, no 6144, p. 1236498.) constituted of social networks of 43 villages in South India that have been exposed to Microfinance. We identify and measure the potential "stratification" of networks and elicit a strategy for choosing initial affected nodes that take advantage of this stratification. We then express and implement a model of diffusion to assess the importance of this disposition on the outcome of the number of affected nodes. We find that our strategy can, in certain cases of high modular networks, increase the number of adopters by 30% compared to a strategy that only take eigenvector centrality as decision criterion.

Microfinance | Diffusion Model | Social Networks | Modularity

Network diffusion captures the underlying mechanism of how events propagate throughout a complex network. No matter the context of such diffusion, technology diffusion (1), epidemic spreading (2) or propagation of knowledge in social networks (3), the fundamental question regarding those processes is how network structure governs the characteristics of the dissemination such as its outcome or its rate of diffusion.

This dependence is a complex relationship governed by multiple underlying factors. While this dynamical process has both, not necessarily balanced, evolution factors based on peer influence and homophily (4), a correlation between the ratio of affected nodes and the eigenvector centrality of the initial nodes emerges among empirical data (5). However, this type of relationship does not take into account the subjacent structure of the network. Indeed, we can intuitively think that the odds for a node to be affected depend on its connectedness with injection points. Similarly, we can conceive that a "tight" community announces a freer flow of information. Hence, more global network characteristics such as its connectedness, modularity or assortativity could have consequent influence on the outcome of a diffusion process, that is not necessarily captured by more local centrality measures.

We use empirical networks to, first, identify certain community structures among the data, then, elicit a diffusion model and a strategy for choosing initial nodes and, finally, assess the effect of stratification on the outcome of the spreading.

Background

Our study relies on the paper "The Diffusion of microfinance" written by Banerjee et al. (5), taking the same data set and extending its analysis. The data set consists of 43 social networks corresponding to 43 villages that have been exposed to microfinance, as well as individual household characterisitics. In this context, the authors exploit the variation of node characteristics in the "injection points" accross villages to derive a correlation between certain centrality measures and the resulting participation ratio in the villages. In particular, a high correlation between eigenvector centrality (a centrality that measure a node "influence" (6)) of nodes and participation ratio is uncovered. Furthermore, the paper conceives two structural diffusion models that allow to separate the effect of peer-influence and homophily in the decision making process of microfinance adoption. These models can also assess the effect of microfinance adoption on the information diffusion process. The paper estimated that participants were more than four times more likely to pass information than non-participants, but that non-participants were still responsible for one-third of information passing. They also concluded that the effect of peer-influence on already informed individuals is negligible with regard to microfinance adoption.

Aral et al. (4) performed a similar study on the distinction between influence-based contagion and homophily-driven diffusion in dynamic networks of instant-messaging networks , where individuals could adopt a mobile service application. The study concluded that, using matched sample estimation, homophily is predominantly responsible for the perceived contagion.

Data

The data were gathered in parallel to the deployment of a microfinance program run by Bharatha Swamukti Samsthe (BSS) in 43 villages rural southern Karnataka. The procedure for introducing microfinance in a village unfolds in this manner: First "leaders" are chosen according to their perceived diffusion potential, manifested by the nature of their job or role in the village, such as teachers, shopkeepers, or leaders of self-help groups. Once these leaders are selected, they are informed about the financial program, and are asked to help in presenting the program to their peers, through meetings and word of mouth. The data regarding the households of each village were gathered with questionnaires. Though incomplete, these data represent roughly 46% of the total number

Significance Statement

Diffusion on networks appears in a wide range of applications. In particular, the problem of influence maximization, or how to select the initial infected nodes in order to maximize the scope of the spreading, is a central challenge. Here we analyze the case study of the introduction of microfinance in villages in South India to quantify the effect of the stratification of a network on the diffusion process, thanks to a propagation model that we develop.

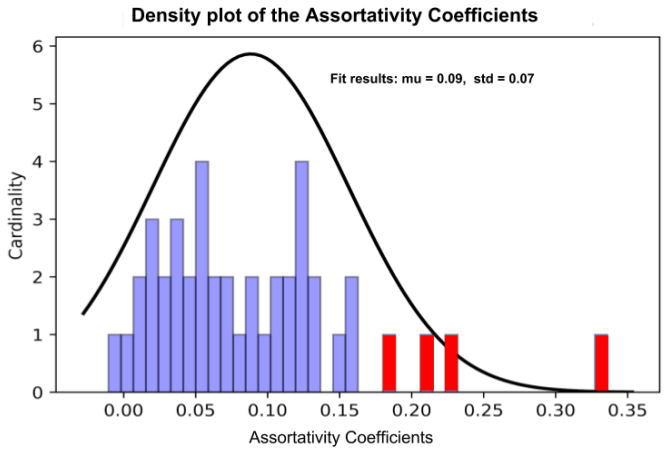

Fig. 1. Density plot of the assortative coefficients with regard to microfinance adoption in every village. Four particular villages stand out : village 30, 28, 4 and 11.

of household per village.

Finally, the final data consist of 43 undirected adjacency matrices, where two nodes are adjacent if and only if at least one of the individual represented by these nodes mentions the other as a contact in the questionnaire. Furthermore, we obtain household characteristics consisting of the religion adopted, the caste or subcaste, the rooftype, room and bed number, electricity type of ownership, latrine type of ownership, type of accomodation ownership, and whether or not the individual has been selected as leader and/or has adopted microfinance.

Evidence for Modularity and social structure

The main challenge is to identify the influence of the stratification of the social structure of the network over the number of adopters after a certain period of time. Our first approach was to identify clues for alignments of groups of microfinance adopters and non-adopters. To do this, we computed the assortativity coefficient (7) with regard to the node attribute of microfinance adoption for each village. The resulting density plot is showed in Figure 1. This plot indicates a significant tendency for villagers that have the same behaviour towards microfinance adoption to have relationship with each other. Four villages show particular evidence for this phenomenon : village 30, 28, 4 and 11, in order of magnitude of the assortativity coefficient.

Having this scheme in mind, we further investigated the community structure underlying this homophily. We took the following approach : we computed modularity-based communities of the networks, using Python and the Louvain Algorithm available in the library "Community". Then, we assessed whether groups of adopters and non-adopters form robust communities by two different ways.

The first method involves statistical hypothesis testing : For every village, we associated each individual with two variables. The first variable indicates the class of the cluster of the individual. The second binary variable indicates whether or not the villager has adopted microfinance. Then, we performed a Chi-Square test of independence on those two variables. For each village, a low p-value indicates that the null hypothesis - the two variables are independent - is unlikely to be true, hence suggesting a sparse distribution of ratio adopters/non-adopters in each cluster. This points out robust communities

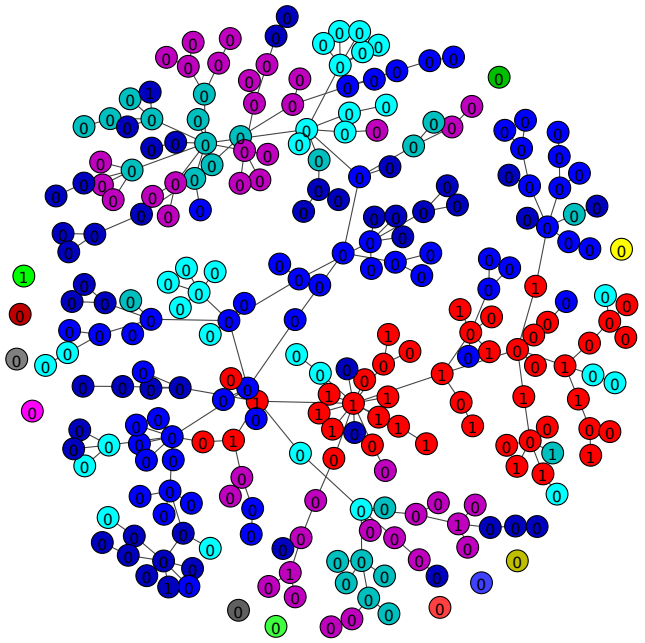

Fig. 2. Maximum spanning tree of village 30, together with its colored clusters and nodes labeled according to their adoption status

of adopters or non-adopters in this village. The literature often use 0.01 or 0.05 as significance levels to accept or reject the null hypothesis. The resulting density plot of the p-values, displayed on a log scale of base 10, is given in Figure 3. Again, we can see that one particular village, the number 30, stands out by its very low p-value of 10^{-14}. Other villages show the same inclination, such as, in order of low p-value, 28, 4, 41, 33, 3, 40, 17 or 22. If we look more precisely at the distribution of adopters ratio within the clusters in each of those villages, we can clearly confirm separation between communities of adopters and non-adopters. For instance in village n°30 we obtain six non-trivial clusters of equivalent size, with the following proportion of adopters : 0.48, 0.02, 0, 0.04, 0.05, and 0.04. We computed the maximum spanning tree of the graph representing village 30 and plotted it. The result is displayed in Figure 2. In the Figure, we can clearly see that the adopters (labeled with the number 1), are mostly gathered on a single cluster, the one represented by the colour red. We obtain the same tendency for the other selected villages. We compared these values to a null case : we took these villages, computed a randomized version of these networks with the configuration model, implemented on networkx. This randomized version allows us to obtain a network with the same degree sequence over nodes, along with their adopter status, but with edges randomly drawn. Then, we ran the same community detection algorithm used previously to find potential clusters. We then computed the p-values according to the same independence test. While our initial villages displayed p-values inferior to 10^{-6}, the p-values we find in our randomized versions are comprised between 0.2 and 1, well above the usual significance level for hypothesis testing.

The second method involves using a distance between partitions based on information theory called "Variation of Information", introduced by *Meilă* (8). This distance can be interpreted as a "measure of the amount of information lost

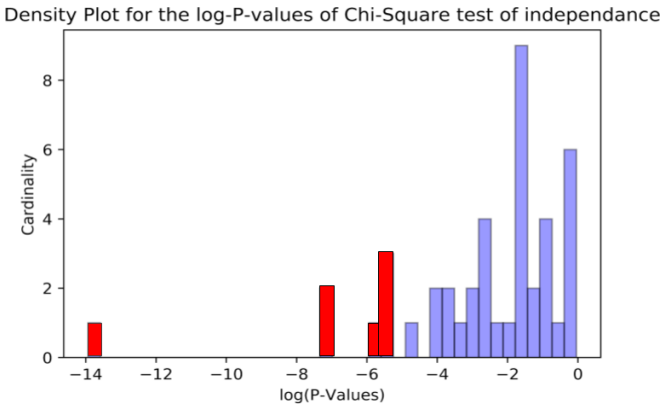

Fig. 3. Density plot of the p-values.

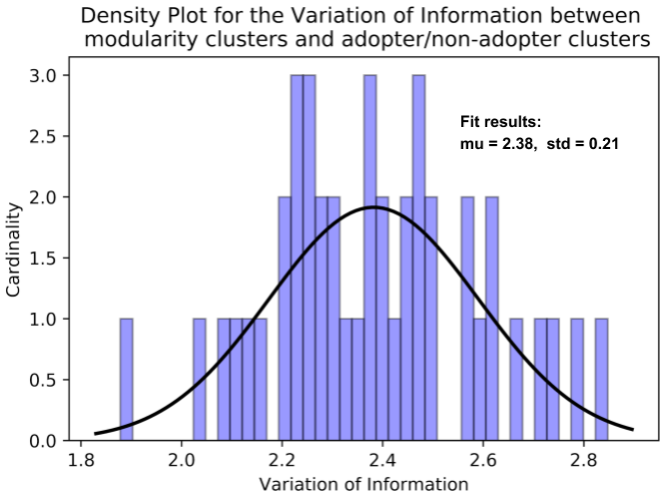

Fig. 4. Density plot Variation of information

and gained in changing from clustering C to clustering C'" (8).

For this study, and each village, we compared the variation of information between the clusters found by modularity maximization and the partition of the village between adopters and non-adopters. According to its interpretation, a "small" distance should signal that the clusters describe well the decision of adopting microfinance or not. We obtained Figure 5. According to this approach, we find that the modularity-based partition is a good representation of the separation between adopters and non-adopters for villages 41, 30, 37, 6, 43, 29. We recover the first two first villages that already stood out in the previous results.

Contagion Model and Strategy

Our goal is to show that in certain cases of highly modular networks, eigenvalue centrality is not sufficient enough to optimize diffusion. For this study we selected villages 30, 41, 4 and 28 which seem to stand out in the previous results. For each of those villages, we computed the proportion of initial leaders that adopted microfinance and the final ratio of adopters in every modularity-based cluster. We then performed a linear regression between those two values. We found a slope of 0.818 and a p-value of 0.000, shown in Figure 5. We performed the same regression, this time with the total number of leaders

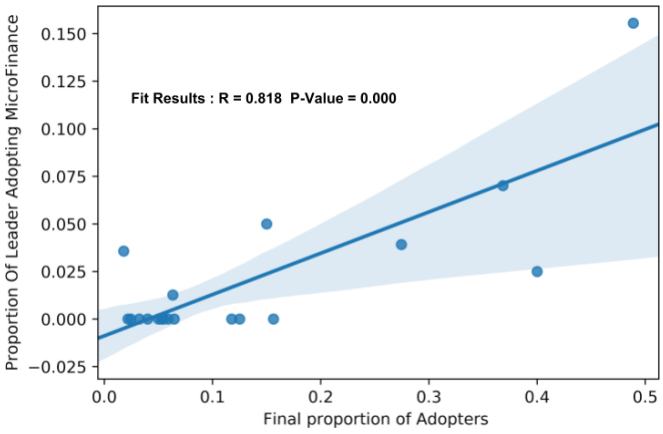

Fig. 5. Linear model between proportion of adopting leaders and resulting number of adopters in each cluster of the villages selected

(that has, or not, adopted microfinance), we found a slope of 0.452 with a p-value of 0.052. Those results suggest that, in the identified villages, a more uniform distribution of the selected leaders among the modularity-based clusters could improve the result. We will further justify this conclusion later.

Hence we formulate the following strategy: given the number of leaders that we can inform in each village, we first allocate leaders proportionally to the size of each modularity-based cluster. Since eigenvector centrality is correlated to future success of diffusion (5), we, then, whithin each cluster, choose the individuals with the highest eigenvector centralities. This strategy should both take advantage of the stratification of the social network and the node importance of leaders. This strategy makes the assumption that the intrinsic characteristics of individuals has a negligible effect on the diffusion process. Even if the statistical studies point toward this hypothesis, as we will see later, this is a considerable assumption in a real-life context.

To test this strategy, we implemented the same diffusion process as in the paper of Banerjee et al. (5), based on information spreading, that we describe here.

1. A group of leaders is selected, according to the predefined strategy, and informed about the possibility to adopt microfinance. They are told to inform their surrounding about this program.

2. Each leader decides whether he wants to participate to the program or not, depending on its intrinsic characteristics.

3. For each leader and each of his neighbour, the leader spreads the word with a certain probability depending on his adoption status.

4. Now the process repeats the following scheme during a predefined time span (counted here in number of trimesters during which the study has been carried out).

   (a) Newly informed individuals take decision about whether to participate to the program or not.

   (b) All informed individuals spread the word to each of their neighbours with a certain probability depending on their adoption status.

Table 1. Fit results of a logistic regression between the acceptance probability and the covariates

Characteristics	Value
Intercept	-0.5015
RoofType :	
- 1	0.5625
- 2	0.339
- 3	0.080
- 4	0.232
- 5	-0.240
Religion :	
- Christianity	-0.277
- Hinduism	-0.612
- Islam	0.387
Caste :	
- General	-0.244
- Minority	0.702
- OBC	-0.061
- SCHEDULE CASTE	0.031
- SCHEDULE Tribe	-0.455
Electricity :	
- None	-0.316
- Government	0.094
- Private	-0.280
Latrine :	
- Common	-0.514
- None	0.236
- Owned	-0.223
Own rent :	
- Given by Government	-0.136
- Owned	-0.098
- Owned but shared	-0.099
- Rented	-0.169

Banerjee et al. (5) concluded that an individual's decision is mostly affected by its intrinsic characteristics rather than peer-effect (influence of the proportion of individuals that has adopted microfinance in the surrounding of an informed individual on its decision making). Aral et al. (4) drew the same conclusion on their own data set whose context is similar to ours. These conclusions justify a information-based model. For that matter, we fitted a logistic regression among the leaders taking household characteristics as covariates, using the scikit-learn library in Python. As in (5) we used the dataset consituted of the leaders to perform the fit of the model, since there is absolutely no peer effect on these individuals, who take decision immediately after being informed. We took the rooftype, religion, type of caste, electricity type of ownership, latrine type of ownership, and type of rent as covariates for the model. The results are summarized in table 1. From this table we can notice that the four first rooftypes, the Islam religion, being in a Minority Caste and having no latrine influences positively the decision of adopting microfinance among the leaders. Applied to the whole dataset, we find statistics presented in Table 2. We also find that 95% of the dataset has an acceptance probability included in $[0.117659, 0.52232]$. Regarding the diffusion of information in our model, Banerjee et al. Banerjee et al. (5) found in their model that the probability for an adopter to inform a neighbour about microfinance is $q^P = 0.450$. For a non-adopter this probability shrinks down to $q^N = 0.095$.

Table 2. Statistics of the acceptance probability of microfinance adoption over the whole data set

Quantity of interest	Mean	Variance	Min	Max
Acceptance Probability	0.2610296	0.00964	0.07996	0.8324259

Table 3. Leaders accepting probability whithin clusters that align with non-adopters status according to the logit model

Village n°	Mean	Variance	#Leaders	Ratio Adopting Leaders
- 30	0.2979	0.00258	86	0.0232558
- 28	0.2075	0.0044	22	0
- 4	0.248289	0.0085	115	0
- 41	0.2915	0.0043	58	0.05

One important phenomenon to lean on is whether the "Cold Start" (the fact that, in certain clusters, there is a disproportion between the number of injection points and the resulting number of adopters) encountered in multiple clusters among the selected villages is purely due to randomness or intrinsic characteristics, in the former case our strategy is justified and should work better on average whereas in the latter case the alignment between clusters and adoption is unavoidable. In that perspective, we took the "outliers" of our regression analysis : we manually took the set of clusters in the selected villages in which there is a high proportion of initial leaders introduced to microfinance, and a low resulting proportion of final adopters at the end of the study. We ended up with 9 clusters. We then computed thanks to the previous fitted logistic regression model the acceptance probability for each leader selected within those clusters, as well as the ratio of leaders that have adopted microfinance. We obtained Table 3.

From this table of results, we can see a clear dissociation between the predicted probability of acceptance and the effective number of leaders adopting microfinance. This is particularly conspicuous for village 4, with 115 selected leaders in 3 different clusters of outliers in which no leader has accepted microfinance, whereas their intrinsic characteristics predict a 25% chance of acceptation on average if one look at the whole data set across villages.

This result comforts us in the choice of our strategy, indeed there is no evidence to suggest that certain available, intrinsic, indicators could discriminate certain individuals over others among the identified clusters to be eligible as leader. According to the available amount of knowledge, the conclusions drawn from the previous linear regressions are valid, increasing the proportion of leaders in the clusters should, *a priori*, increase the resulting number of adopters. To reinforce this conclusion, we also looked at the assortativity coefficents with regard to every covariate for village 30, and we found no assortativity coefficient, when defined (for instance, every individual in village 30 have adopted the Hindu religion, the coefficient is, thus, undefined), that exceeds 3.10^{-3}.

Results

We implemented the previous diffusion process and applied it on the selected villages 30, 41, 4 and 28. For each village, we simulated 1000 diffusion processes in four different ways. First,

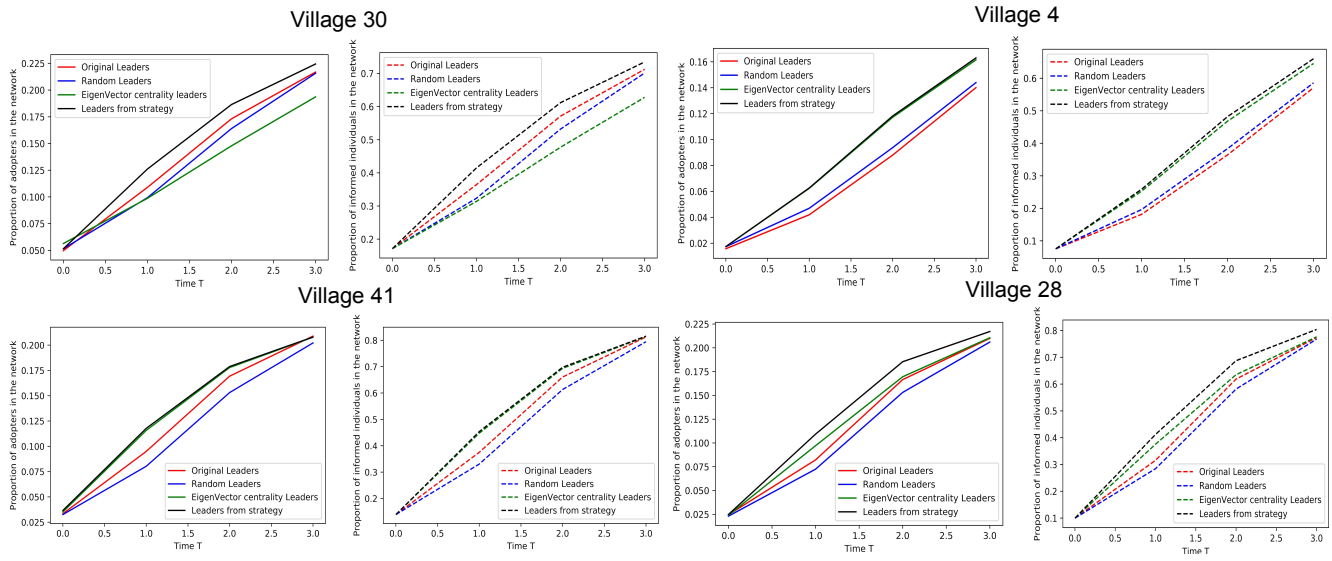

Fig. 6. Evolution of the proportion of adopters and informed individuals over time in village 30, 41, 4 and 28, and according to the information-based propagation model, for different strategies for the choice of initial leaders

Table 4. Performance mean maximum gain of our strategy compared to a strategy based solely on eigenvector centrality

Village n°	Max Gain (%)	Time Max Gain
- 30	27.8	1
- 41	4.2	2
- 4	0.9	3
- 28	12.6	1

we chose the initial leaders as the leaders originally selected in the paper. Then we chose the leaders according to our strategy based on a uniform ditribution of injection points through the modularity-based clusters, weighed by the size of the clusters, and then selected by their eigenvector centrality. Afterwards, we selected leaders randomly in the networks, and finally we picked leaders displaying the largest eigenvector centrality. We then plotted the curves showing the evolution of the mean proportion of informed individuals and individuals adopting microfinance over time. For $T = 3$ we can notice that already roughly 70% of the network is informed about microfinance in every village. Since we have implemented a model taking uniquely into account intrinsic characteristics of individuals in order to take decisions, it is irrelevant to look at the proportion of individuals adopting microfinance on the long run, since it is independant of the initial strategy as long as people are informed. Thus, we stopped at $T = 3$ in our simulation and looked at the rate of evolution of those quantities of interest. The plots are displayed in Figure 6.

We obtained very promising results, in every village we can observe that our strategy is the best strategy on average. In village 30, whose clusters were the most aligned with community of adopers/non-adopters, we can see that selecting injection points according to the eigenvector centrality is even disadvantageous in comparison with a random selection of leaders. In village 41 and 4, selecting injection points according to the eigenvector centrality is very similar to our strategy, showing that, in our model, the effect of the stratification of the village is negligible. In every village, the results show that the original leaders selected in the data set and the leaders selected by our strategy perform better than a random strategy. To assess the potential gain of our strategy, or the effect of stratification on knowledge diffusion, we computed the maximum percentage gain that our strategy can provide in the proportion of individuals adopting microfinance, compared to the strategy that solely takes the best leaders according to their eigenvector centrality. We also noted the time at which this maximum is attained. The results are displayed in Table 4.

Like already noticed on the plots, network stratification has the largest effect on village 30. In this village, at time $T = 1$, there are almost 30% more microfinance adopters on average while adopting our strategy instead of solely looking at eigenvector centrality. We can state a similar conclusion regarding village 28 with a 12.6% gain at time $T = 1$.

Until know we have shown the effect of network stratification on diffusion on networks preselected *a posteriori* on the identification of different aggregation of adopters and non-adopters among clusters. One legitimate practical question from a decision maker point of view would be how to assess *a priori*, on a given network, the potential effect, if any, of network modularity. To explore this question, we implemented a function computing the modularity of a graph given its partition in clusters. We then plotted the maximal gain of our strategy over the strategy based on eigenvector centrality as defined above. For computing power reasons, we limited the number of diffusion to 250 for every village, in the computation of the mean ratio of adopters and informed nodes over time. We obtain Figure 7. A linear model predicts a slope $R = 0.452$ with a p-Value $P = 0.002$. The p-value indicates a certain confidence towards the positive relationship between the two quantities. However, as we will discuss in the next section, this relationship is to be carefully considered, since modularity in itself an intrinsic network characteristic, it is difficult to generalize to the entire set of networks and hence to elicitate quantitative criteria for the considered effect. Moreover mod-

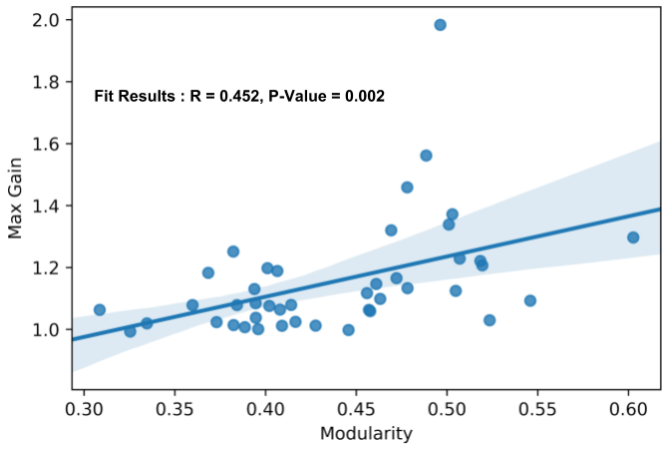

Fig. 7. Linear model between the maximal gain of our strategy over the modularity of each village

ularity, as quantity, can be tricky and arise from factors not associated with modularity, in a graphical sense.

Discussion

From our data set, we developed a scheme of analysis to identify networks that present high assortativity with respect to microfinance adoption, as well as villages that displayed high modularity with clusters showing notable alignment with respect to microfinance adoption/non-adoption. We uses information theory as well as statistical tests. From this preselection of villages, we inferred a strategy that takes advantage of this stratification of the network to optimize the choice of injection points for microfinance diffusion.

We find that a more uniform distribution, weighed by cluster size, of leaders among the modules of the preselected villages can display a potential maximum gain of 30% for village n°30 compared to the strategy based on eigenvector centrality, a quantity positively correlated with final ratio of adopters. This indicates a consistent effect of modularity on information flow, as taking certain leaders not based on their eigenvector centrality but according to their strategic placement among the modules of the network increases the diffusion rate.

Our work is not without limitations. First, the data gathered are incomplete, since 46% of the households is taken into account, the network modularity of the empirical data, can be over or under estimated. Secondly, as we mentionned it in the previous section, it is hard to quantify the effect of network stratification on the characteristics of a knowledge diffusion process *a priori* without knowing any potential "alignment" between adopter and clusters. We showed in Figure 7 a positive correlation between the modularity and the potential gain in the proportion of adopters in our strategy compared to a strategy based on eigenvector centrality, hence modularity can serve as indicator for the potential effect if stratification. However, since our strategy relies on community detection based on modularity maximization, this indicator should be taken cautiously. Indeed, modularity is an intrinsic quantity for each network, and is not comparable to other network modularity. Hence, it cannot be generalized and therefore serve for the elicitation of a global description. This prudence is reinforced by the fact that the value of modularity can emerge in such a way that has nothing to do with the modularity of the network (9). Thirdly, the modularity landscape tends to be very uneven (10), with possibly very different partitions having similar value of modularity. This limitation, together with the issue of resolution limit of modularity (11), suggests that one should be careful and further inspect the resulting partitions. In particular, the issue of resolution limit could mask a hierarchical effect of stratification, which can be inspected by iterative community detection (11). Finally, our strategy for choosing the initial leaders only take the network structure into account, even if we have shown that in the most modular villages, no conspicuous assortativity is present, this is unlikely to be true in general, hence a more adaptative strategy could be of interest.

The limitations raised previously suggest a further theoretical and practical study of the robustness of our model over scale, since our study was limited to networks of similar size of the order of one hundred nodes, as well as the the elicitation of a resolution pipeline. Moreover, a more complex strategy that adapts to the "covariates importance" whithin clusters could be more appropriate for the choice of initial leaders. This could, for instance, be based on the acceptance probability resulting from a logit regression or a more complex model formulated *a priori*.

Materials and Methods

We used a distance based on information theory (8), given by

$$VI(C,C') = H(C) + H(C') - 2I(C,C') \quad [1]$$

Where, with K cluster in C:

$$P(k) = \frac{n_k}{n} \quad [2]$$

$$P(k,k') = \frac{|C_k \cap C'_{k'}|}{n} \quad [3]$$

$$H(C) = -\sum_{k=1}^{K} P(k)log(P(k)) \quad [4]$$

$$I(C,C') = \sum_{k=1}^{K}\sum_{k'=1}^{K'} P(k,k')log(\frac{P(k,k')}{P(k)P(k')}) \quad [5]$$

The acceptance probability of microfinance adoption is predicted by a logistic regression model $Y(X) = \frac{1}{1+exp(-\alpha - X\beta)}$. We dropped "bed n°" as well as "room n°" considered useless, and used dummy coding to exploit categorical data.

1. Chang SB, Lai KK, Chang SM (2009) Exploring technology diffusion and classification of business methods: Using the patent citation network. *Technological Forecasting and Social Change* 76(1):107–117.
2. Morris M (1993) Epidemiology and social networks: Modeling structured diffusion. *Sociological methods & research* 22(1):99–126.
3. Cowan R, Jonard N (2004) Network structure and the diffusion of knowledge. *Journal of economic Dynamics and Control* 28(8):1557–1575.
4. Aral S, Muchnik L, Sundararajan A (2009) Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences* 106(51):21544–21549.
5. Banerjee A, Chandrasekhar AG, Duflo E, Jackson MO (2013) The diffusion of microfinance. *Science* 341(6144):1236498.
6. Newman ME (2016) Mathematics of networks. *The new Palgrave dictionary of economics* pp. 1–8.
7. Newman ME (2002) Assortative mixing in networks. *Physical review letters* 89(20):208701.
8. Meilă M (2007) Comparing clusterings—an information based distance. *Journal of multivariate analysis* 98(5):873–895.
9. Guimera R, Sales-Pardo M, Amaral LAN (2004) Modularity from fluctuations in random graphs and complex networks. *Physical Review E* 70(2):025101.
10. Good BH, De Montjoye YA, Clauset A (2010) Performance of modularity maximization in practical contexts. *Physical Review E* 81(4):046106.
11. Fortunato S, Barthelemy M (2007) Resolution limit in community detection. *Proceedings of the National Academy of Sciences* 104(1):36–41.