The Effect of Community Structure on Microfinance Diffusion

Venelin Martinov[a,1]

[a]Department, University of Oxford, Oxford, England, United Kingdom

This manuscript was compiled on March 17, 2022

We extend the results of The Diffusion of Microfinance by Banerjee et al. by showing that the internal community structure of the networks affects the diffusion of participation in a microfinance program and propose a way to choose initially informed nodes by taking this into account. To show this, we measure the assortativity of adopters and non-adopters, and the stability of the adopter community under the Louvain algorithm. We also show a large variance between the adoption rates of communities inside each network and a significant maximal community adoption rate when compared to random versions of the networks. We find good agreement between the measures and hence a relatively clear definition of significant internal community structures for the purposes of diffusion. We propose a strategy for choosing initially informed nodes based on this community structure and find that it performs well when compared to the original choices and when compared to a strategy inferred from the finding of the original paper.

community structure | network diffusion | microfinance adoption

Network diffusion is an important topic to many different disciplines. From the spreading of infectious illnesses to the propagation of failures in a computer network, there are many instances of the same theoretical problem, which are crucial to the understanding of various systems. The spread of ideas on social networks is also a relatively well-studied subtopic with many academic works documenting various instances of the phenomenon. In (1) the authors track the adoption of a mobile service in an instant messaging network and distinguish between homophily and influence spreading - the similar behaviour of similar nodes and the similar behaviour of neighbours. In (2) the authors consider the social networks of 43 rural Indian villages and information on their households and their participation in a loan program. The villagers can choose to take a loan of about 10000 rupees (around 200$) for which five of them are jointly liable. In each village the loan company initially informs a pre-defined set of "leaders" and the authors analysed how the importance of the initially informed leaders affects the spread of adoption of the program. They find that the eigenvector centrality of the leaders is a strong predictor of eventual take-up of the microfinance scheme.

We are dealing with the same dataset in this work, but are considering how the internal community structure of the networks affect the diffusion of participation in the program.

Establishing the Existence of Community Differences

We start by presenting the different measures of how much a network's communities differ in their adoption behaviour. As can be seen on Fig. 1, some of the networks exhibit very clear divides in their communities' adoption. To quantify the differences we calculate four different metrics by which to measure these:

- The assortativity coefficient of the adopter and non-adopter communities.
- The stability of the adopter community when fed as an initial partition of the network to the Louvain algorithm (3), compared to the stability of the same cluster of nodes when the underlying network is randomly rewired.
- The variance in the adoption rates of the communities within each network.
- The maximal adoption rate of a community of each network, compared to the maximal adoption rate of a random rewiring of the same network.

Assortativity Coefficient. The assortativity coefficient as defined in (4), is a measure of how much nodes tend to connect with nodes of their own group. It is given by $r = \frac{Tr\mathbf{e} - \|\mathbf{e}^2\|}{1 - \|\mathbf{e}^2\|}$, where $\mathbf{e}$ is the mixing matrix, whose entries e_{ij} are the fraction of the edges that connect nodes of type i to ones of type j. In this case the two types of nodes are adopters and non-adopters. A high value of the assortativity coefficient of the network tells us that adopters tend to be connected with other adopters and non-adopters tend to be connected with other non-adopters. A comparison with a random rewiring of the network, according to the configuration model, while keeping the adoption status and degrees of nodes shows the significance of the results.

Significance Statement

The network importance of the initially informed people has been identified as an explanation for the differences in diffusion spread in a previous paper. However, the authors do not consider the internal community structures of the networks and its effect. In this work, we show that this structure matters for the diffusion by measuring the imbalance of diffusion between the communities of a network in several ways. We also show a relatively good agreement between the metrics and a possible way to choose good initial nodes based on community considerations. The methods are applied to a dataset consisting of the social networks of 43 rural Indian villages, information on their households and their participation in a microfinance scheme.

Venelin Martinov analysed the data and wrote the report

The author declares no conflict of interest

[1]To whom correspondence should be addressed. E-mail: venelin.v.martinov@gmail.com

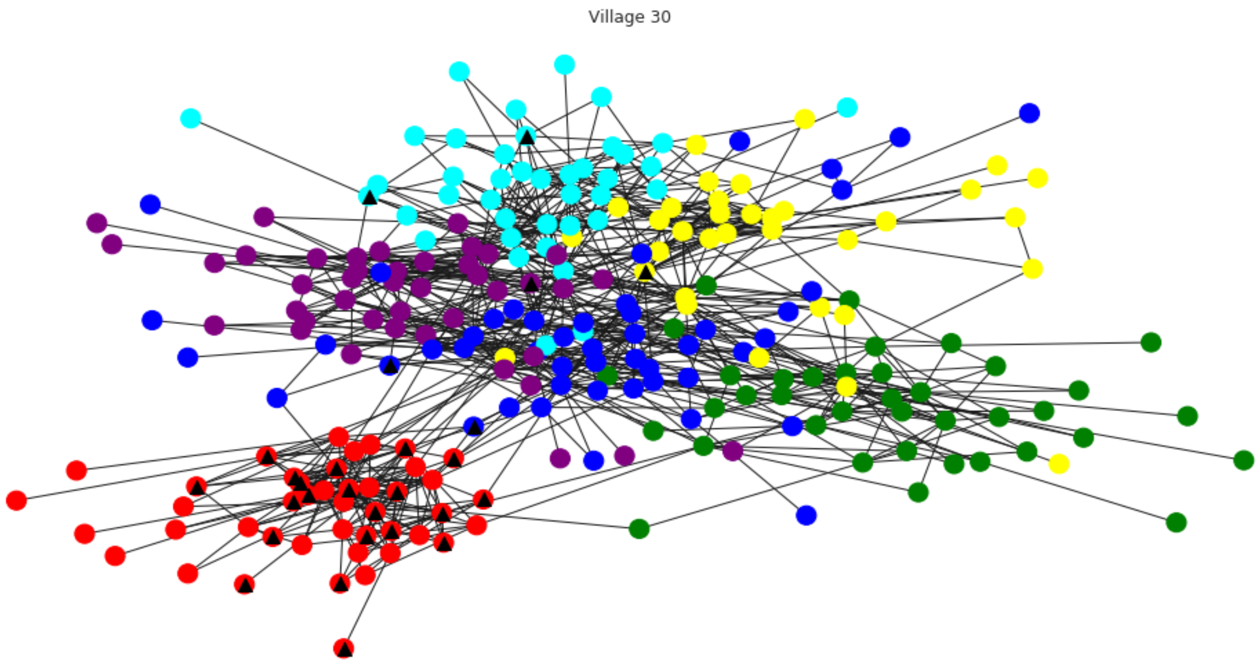

Fig. 1. A visual representation of one of the village social networks. The colours of the nodes correspond to communities identified by the Louvain algorithm, while the nodes marked with a triangle are ones which adopted the scheme. The network exhibits clear community-dependent adoption ratio, with the red community containing almost all adopting nodes.

Robustness of the Adopter Community. The robustness of the adopter community is a measure of how good is the partition of the network into adopters and non-adopters. Here we are using the Louvain algorithm, which is a greedy algorithm optimizing modularity, as defined in (5). We feed it the partition of the network into adopters and non-adopters and define O_{obs} as the maximum overlap of a resulting community identified by the algorithm with one of the original ones. We then repeat the process with a random rewiring of the network, preserving the degrees and adoption status of nodes and define O_{rand} as the maximum overlap of a resulting community with one of the original ones. The robustness score is then defined as $\frac{O_{obs}}{O_{rand}}$. This normalization is necessary to account for any inherent properties of the network (e.g. an unusual degree distribution). Due to the somewhat random results of the Louvain algorithm the whole process is repeated multiple times and a mean is taken as the result.

Measures of Community Adoption Rates. Lastly, we directly measure some of the properties of the individual adoption rates of the communities in the networks. Using the Louvain algorithm again, we split each network into communities and calculate the adoption rate within each community. Then for each network we measure the variance in its communities' adoption, averaged over multiple trials. We also measure the maximum adoption rate of a community for each network and normalize it by the maximum adoption of a community of the null model of each network, defined as before using the configuration model. The result is again averaged over multiple runs, due to the randomness in the community algorithm. We will refer to the normalized rate as the maximum adoption rate.

Figure 2 presents the mixing matrix for one of the networks, showing a well-defined community structure and Table 1 gives the community adoption rates for the same village. The table clearly show a significantly higher maximum adoption rate for one of the communities and a big variance in the community adoption rates. Moreover the clusters found by the Louvain algorithm seem relatively even, so questions of disproportionate communities seem to not be relevant (more in the Discussion).

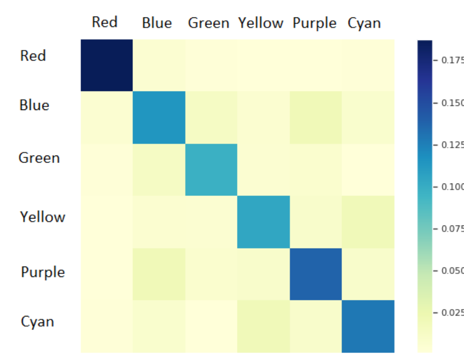

Fig. 2. A heat map of the mixing matrix of the communities in the same village as Fig. 1. The gradient on the right shows correspondence between colours and numbers. The strong colours in the diagonal entries and the weak ones in the off-diagonal entries show well-defined communities.

Simulating Diffusion and Comparing Injection Point Choices

After showing the presence and effect of community structures on the diffusion process, we move on to reconstructing the simulation model in (2) and test different initial injection node strategies and their effects on the spread of microfinance

Table 1. Comparison of the characteristics of the communities in the same village as in Fig. 1. The different communities have a comparable number of leaders, but very different numbers of adopters. A possible explanation, as identified in the original paper, is the difference in the average eigenvector centrality of the leaders.

Community	Size	Leaders	Leader Centrality	Adoption
1. Red	44	7	0.16	0.48
2. Blue	48	7	0.03	0.04
3. Green	38	7	0.01	0.00
4. Yellow	34	5	0.04	0.03
5. Purple	42	12	0.02	0.02
6. Cyan	40	6	0.04	0.05

adoption.

Simulation Model. The model we use for simulating microfinance adoption on the village networks is the information model used in (2). A probability of adoption is defined for each node, based on a logistic regression between the household covariates and the adoption status. An information transmission probability is defined, which is different for adopters and non-adopters, but is independent of whether or not the neighbours of the node adopted. The model proceeds as follows:

- Initially a number of nodes are informed of the microfinance program and each one chooses whether to participate or not. Participation choices are final.
- After that the newly informed nodes spread information to their neighbours with some probability, depending only on their adoption status.
- Each of their neighbours that was informed chooses whether to participate, they inform their neighbours and the process repeats.

The process is repeated for a number of steps and the adoption rate for the network is calculated. The performance of the initial injection points is the average adoption rate over multiple simulations of the process. The information model with no endorsement effects was chosen, based on the finding of the authors of the original paper that the adoption choices of neighbours do not significantly influence the choice of a node. The model naturally incorporates the homophily-driven behaviour of similar nodes by dividing the information spreading from the adoption process.

Initial Injection Strategies. We test two different centrality-based strategies for choosing the initially informed nodes and compare their performance to the performance of the original leaders informed in the program, and a null strategy which does not consider community structures. For each strategy we consider a local and global version, where in the local version we only calculate the centralities of the nodes inside the subgraph defined by the community. In the global versions we calculate centralities of the nodes as usual. The two strategies considered are one based on eigenvector centrality and one based on degree centrality. In each one, we choose the same number of leaders in each network as were originally chosen. The difference is that we distribute them between the communities proportionally to the size of each community. In each case, we choose the most central nodes from the communities.

Table 2. The table shows the coefficients in the linear regressions between each pair of metrics.

Variable 1	Variable 2	Slope	p-value
Assortativity	Robustness	0.95	10^{-7}
Assortativity	Variance	4.64	10^{-8}
Assortativity	Maximum Adoption	3.26	10^{-6}
Robustness	Variance	2.94	10^{-5}
Robustness	Maximum Adoption	1.97	10^{-4}
Variance	Maximum Adoption	0.59	10^{-8}

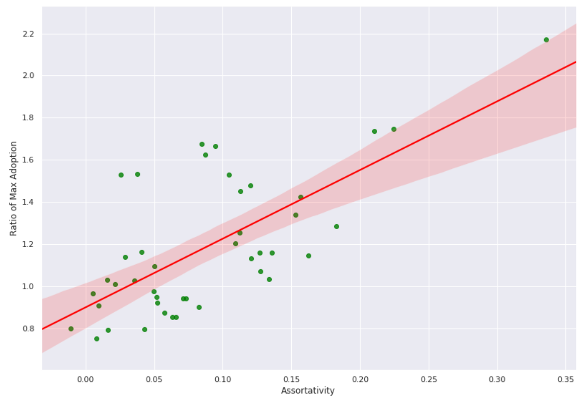

Fig. 3. Linear regression between the assortativity coefficients of the village networks and their maximum community adoption ratio.

The reason for deploying the eigenvector strategy is that in (2), the authors found a strong dependence of the performance of the program in a given village on the eigenvector centrality of the leaders in the network.

Results

Here we present our findings on the different metrics for the relevance of the community structures in the networks and on the strategies for choosing initial injection nodes.

A. Agreement Between the Different Metrics. We find a relatively good agreement between each pair. To show this we calculate the slope and p-values of a linear regression between each pair, after scaling each to zero mean and variance one. Table 2 presents the results of the regressions and 3 shows a plot of one of the regressions. The high agreement between the different metrics signifies a good definition of relevant community structure for the purposes of the program diffusion.

B. Performance of the Strategies. After testing the strategies for choosing initially informed nodes with the simulation model defined above, we find a consistent (around $5\% - 10\%$) improvement of the proposed strategies over the original leader choices. We present the normalized results for the eigenvector centrality strategies in Fig. 4, along with the performance of the null strategy. We find similar results for the degree centrality strategies.

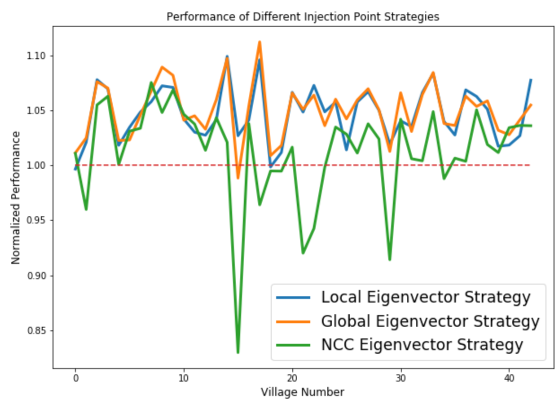

Fig. 4. This figure shows the performance of the different eigenvector-based strategies in each village, normalized by the performance of the strategy, where the initial injection points are the originally identified leaders. Both the local and global strategy choose a number of initial nodes from each community proportional to the size of the community. The local eigenvector strategy, in blue, chooses the nodes with highest eigenvector centrality in the subgraph consisting only of the nodes in the community. The global eigenvector strategy, in orange, chooses injection points based on the eigenvector centrality of the whole graph. The NCC (no community considerations) strategy, in green, chooses the top eigenvector centrality nodes, without any community considerations. The red line is the performance of the original leaders. The number of initially informed nodes varies between villages, but does not vary between strategies. The strategies which take into account the community structures consistently outperform the ones that do not. Note villages here are zero-indexed.

Discussion

The methods and results in this work seek to explore how relevant community structure affects the diffusion process on a network. The high agreement between the different metrics provided indicates a good definition of community structure and the results on choosing initially informed nodes takes advantage of the findings.

We now move on to outlining possible extensions of the analysis done in this work and raise some of the questions prompted from our findings. Finding reasons for the clustering observed in the villages and characterising the sensitivity of the results to subsampling would be a natural continuation of the results. Exploring how the effectiveness of the initial choice strategy relates to network characteristics is another possible extension of the work.

The results from (2) also raise the question whether a better strategy for choosing initially informed nodes might be available, based on choosing important nodes who are likely to adopt the program. Since adopters are more likely to inform their neighbours, a combination of community considerations, importance in the network and likelihood of adopting might yield better diffusion in the network.

Usage of the Louvain algorithm also gives rise to several issues related to the division of the networks into communities. The first one is on the stability of the communities found by the algorithm between different iterations of it. Motivation for this question is that some of the networks had very low robustness of their adopter community, while maintaining a disproportionately large maximum community adoption rate. In this work, we have relied on the fact that averaging over multiple runs should give relevant results, but this does not preclude the possibility that two completely different near-optimal modularity partitions are very near each other in the resulting space of divisions.

The question of community stability also prompts an analysis of the results on the local eigenvector strategy deployed for choosing the initially informed nodes. Eigenvectors are quite sensitive to changes in the underlying network, so any differences in the clusters might cause big changes in the results of the strategy. We tried counteracting this in the work by averaging the results over multiple runs of the methods.

Another issue, as mentioned in the work is the relative sizes of communities that each graph was split into. Although the Louvain algorithm has a "size preference" for the communities it identifies (6), it is not very strict so it is possible for the methods here to identify communities of very different sizes and skew the results. It is not what we observed in our work, but a more rigorous analysis of the question might be useful.

There are slight differences in the number of communities into which the Louvain algorithm divides a given network, but the metrics tracked in the work are agnostic to that number. A possible continuation of the work might be to see how the sizes of communities affects the results presented by varying the time parameter in the clustering algorithm.

A word of warning is also in order. The methods presented here are mostly theoretical and based on limited information on the actual households in the villages. These ideas might be useful to help choices of real-world actions, but great care should be taken, as with all correlation models. As an example, including information on the households' religious affiliation in the logistic regression yields very different results for the different religions present in the area, which is not a desired outcome.

Materials and Methods

The covariates used for the logistic regression in the calculation of the individual adoption probabilities were, as in the original paper: the number of rooms in each house, the number of beds, the presence and source of electricity and the presence of a latrine. A generalized linear model with Binomial distribution and a logit link function was used for the regression.

Throughout the whole work, whenever any community structure was calculated, only the largest connected component of the network was considered (a fairly standard practice). Since none of the networks in the dataset have many disconnected nodes, this does not significantly affect the results.

Each time the Louvain clustering algorithm was used it was repeated 100 times to counteract the randomness in its results. Each of the averages taken was performed over 1000 trials, so that averages of results of the Louvain algorithm were actually taken over 100000 runs (including the results of the initial node choosing strategies).

1. 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.
2. Banerjee A, Chandrasekhar AG, Duflo E, Jackson MO (2012) The diffusion of microfinance, (National Bureau of Economic Research), Working Paper 17743.
3. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008:10008.
4. Newman ME (2003) Mixing patterns in networks. 67:026126.
5. Newman MEJ (2006) Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* 103(23):8577–8582.
6. Fortunato S, Barthélemy M (2007) Resolution limit in community detection. *Proceedings of the National Academy of Sciences* 104(1):36–41.