Changes in Social Network Structure in Response to Exposure to Formal Credit Markets

Banerjee, Breza, Chandrasekhar, Duflo, Jackson and Kinnan

Online Appendix

Appendix B. Adding More General Dependencies to Our Model

We describe a variation on the subgraph formation model of Chandrasekhar and Jackson (2018).

Let G be some set of potential subgraphs on n nodes. For instance, instead of just a list of all possible links, it could also include triangles, or various other cliques, stars, and so forth.

We abuse notation and let $i \in g$ for some $g \in G$ denote that i is one of the nodes that has links in g. Let $v_i(g)$ denote the utility of i if g forms. The total utility that i obtains is the sum over all subgraphs that i is part of - so rather than just a network, the resulting object is a multigraph.

We let m_g denote a relative frequency adjustment for the type of subgraph in question, as some may be more or less likely to form as a function of the efforts.

The probability that some g forms if it is not present is then

$$m_g \times_{i \in g} e_i(1 - F(v_i(g)))$$

which is the product of the socialization efforts and the probability that each i involved in g finds it valuable to form g.

The probability that a subgraph is maintained if it is already present is[1]

$$\times_{i \in g} e_i.$$

Let $\mathrm{E}^+[v_i(g)]$ denote the expected utility that i gets from subgraph g conditional on finding it worthwhile to form, and $\mathcal{G}^t$ denote the set of subgraphs present at the beginning of time t. Then, the expected utility that i gets from effort e_i is

$$V_i(e_i) = u_{\theta_i} e_{\theta_i} - \frac{1}{2} c_{\theta_i} e_{\theta_i}^2 + \sum_{g \in \mathcal{G}^t : i \in g} \mathrm{E}^+[v_i(g)] \times_{j \in g} e_j + \sum_{g \notin \mathcal{G}^t : i \in g} \mathrm{E}^+[v_i(g)] m_g \times_{j \in g} e_j (1 - F(v_j(g))).$$

We say that a society is weakly connected in expectation if for all agents i and j, there exists a path between i and j if all $g \in G$ are formed. This is satisfied if at least one network that could conceivably form is path-connected, and is obviously satisfied if any two agents could be connected directly.

Proposition **2.** *Let $u_\theta > 0, c_\theta > 0$ for all θ, and the society be weakly connected in expectation. For sufficiently large $c_\theta > 0$'s, there is a unique equilibrium and it is stable and interior ($0 < e_i < 1$ for all θ). In addition, if $\mathrm{E}^+[v_i(g)] > 0, m_g > 0$ for each g, and $E^+[v_i(g)]$ is decreased for some g (holding all other parameters constant), then e_i decreases for all i, and $\mathrm{P}[g \in \mathcal{G}^t]$ decreases for all g.*[2]

[1] One could adjust the relative impact of effort for maintaining a subgraph to be some other function than simply the product, depending on the context.

[2] The proof of Proposition 2 is easy to construct as an extension of that of Proposition 1, and thus omitted.

From Proposition 2, it follows that, analogous to the dyads case above, all dyads and triads decrease in probability of forming if the value of at least one dyad or triad drops to some type. In particular, specializing the general model to the case of dyads and triads with types $\{H, L\}$, and corresponding values

$$v_{\theta\theta'}, v_{\theta\theta'\theta''}, (\theta, \theta, \theta') \in \{H, L\}^3,$$

we get the following corollary.

Corollary 1. *Let* $u_\theta > 0, c_\theta > 0$ *for all* θ*. For sufficiently large* $c_\theta > 0$*'s, if* $\mathrm{E}^+[v_{H,\theta'}]$ *or* $\mathrm{E}^+[v_{H,\theta',\theta''}]$ *decreases for some* $(\theta, \theta') \in \{H, L\}^2$*, and no values increase, then the formation of dyads and triads of all types decrease.*

Again, in cases in which the value of connections from Ls to Hs, in either dyads or triads, is high enough, the drop in efforts by Hs can lead to a drop in L efforts that is large enough to lead to a drop in LLL triangles that is as large or larger than the drop in HHH triangles.

Appendix C. Random Forest Model Description

We use a random forest algorithm (implemented in R) to classify our respondents into two types: those that have a high probability of taking up microfinance loans (H) and those that have a low probability (L), when offered.

C.1. **Algorithm Inputs.**

Input Data:

- N = Set of respondents from all villages,
- N_{mf} = Set of respondents from microfinance villages,
- Y_i = Loan take-up binary outcome for each $i \in N_{mf}$,
- X_i = Set of predictor variables for each $i \in N_{mf}$.

Algorithm Parameters:

- T = Set of trees to grow,
- p = Total number of predictors,
- m = Number of predictors selected at each split,
- c = Cut-off: a vector of length 2 (the winning class for an observation is the one with the maximum ratio of proportion of votes to cut-off),
- t = Fraction of sample to be used as training dataset.

C.2. **Basic Algorithm.**

Step 1: Randomly select (with replacement) training data S and testing data S' from N_{mf}. The size of S will be $t \cdot n(N_{mf})$ and the size of S' will be $(1-t) \cdot n(N_{mf})$.

Step 2: For each tree $t \in T$,

- Randomly select (without replacement) a sample of size $n(S)$ from S.
- At each node n of the tree t, randomly select (with replacement) a set of predictors of size m from p.
- At each node, construct a split based on a rule which uses Gini's Diversity Index (gdi) to determine the split.
- For every tree t, each $i \in N_{mf}$ will be assigned a classification $\hat{Y}_{it} \in \{0,1\}$.

Step 3: After classifying each $i \in N_{mf}$, for each tree t, the final classification can be computed as follows,

$$\hat{Y}_i = 1\left\{\frac{1}{n(S)}\sum_{t=1}^{n(T)}\hat{Y}_{it} > c[2]\right\}$$

and therefore $\theta_i = \hat{Y}_i \cdot H + (1-\hat{Y}_i) \cdot L$.

C.3. **Our Parameter Choices.**

- T: We use 1500 trees.
- p: We use 13 predictors for Karnataka and 19 predictors for Hyderabad. The choice of predictors is explained in subsection C.4.

- m: We use the basic R `randomForest` parameter which is equal to $\sqrt{p}$ for classification.
- c: We use the vector $(0.85, 0.15)$ for Karnataka panel and $(0.73, 0.27)$ for Hyderabad panel, chosen by cross-validation.
- t: We use 0.7 of our sample to train the data.

C.4. **Selection of predictors.** Where possible, in both settings, we select predictor variables that are likely correlated with microfinance eligibility, awareness and take-up.

C.4.1. *Karnataka predictors.* In the Karnataka sample, we have detailed information on how the MFI marketed its product to potential borrowers, along with the eligibility rules. The first five predictors capture key components of eligibility and awareness. The network measures are included to pick up the likelihood of households hearing about the product. We also include an additional set of household characteristics associated with wealth.

- dummy for being a BSS leader, who are the people that the MFI would approach when entering a village (the BSS definition of leader was defined by occupation, e.g., teachers; self-help group leaders; shopkeepers, so we can identify them similarly in MF and non-MF villages),
- dummy for whether the household has a female of eligible age (18-57) for a microfinance loan, which is a requirement for the household to be able to participate,
- the average closeness (mean of inverse of network distance) to leaders, which is relevant because those who are closer to leaders are more likely to hear of microfinance (Banerjee et al., 2013),
- the average closeness (mean of inverse distance) to same-caste leaders, because interactions within-caste are more likely and therefore should influence the likelihood of being informed,
- the share of same-caste leaders in the village, as above.
- GMOBC= a dummy for whether the household consists of general caste or other backward caste, so the omitted categories are scheduled caste and scheduled tribes (SCST); general and OBC are considered upper caste,
- household size,
- number of rooms,
- number of beds,
- dummy for access to electricity,
- dummy for access to latrine,
- dummy for RCC roof (considered a superior type of roof),
- dummy for thatched roof.

Table 2 shows a comparison of the H respondents and the L respondents along key dimensions. We see that H households are much more likely to be SCST, have smaller houses in terms of room count, are much less likely to have a latrine in the household, and are much less likely to have an RCC roof, all of which suggests that they tend to be poorer. Finally, we see that H households and L households have comparable degree (H types have 1.94 more friends on a base of 8.97), but the composition exhibits considerable homophily: H types have a lower number of links to L types and a higher number of links to H types. But H households are more eigenvector central in the network.

C.4.2. *Hyderabad predictors.* In Hyderabad, we have less precise information about the lending strategy followed by the local MFI. To be eligible, households needed to have a prime-aged woman, a key variable that we include as a predictor. We also include several neighborhood and household level variables that may proxy for credit supply, credit demand, demographic composition, and proxies for wealth.

- Total outstanding debt in area (baseline)
- Area population (baseline)
- Total number of businesses in area (baseline)
- Area mean monthly per-capita expenditure (baseline)
- Area literacy rate (HH heads only, baseline)
- Area literacy rate (all adults, baseline)
- Dummy for household operating any business(es) prior to 2006, when Spandana opened branches in treatment areas (endline 1)
- Adult equivalent household size (endline 1)
- Adults (16 and older) in household (endline 1)
- Children (15 and younger) in household (endline 1)
- Dummy for male household head (endline 1)
- Age of head of household (endline 1)
- Head of household with no education (endline 1)
- Prime-aged (18-45) women in household (endline 1)
- Any child 13-18 in household (endline 1)
- Dummy for literate spouse of household head (endline 1)
- Dummy for spouse who works for a wage (endline 1)
- Dummy for household owns land in Hyderabad (endline 1)
- Dummy for household owns land in native village (endline 1)

Note: All baseline variables were collected in 2005 and are area-level averages (Baseline households were not systematically resurveyed at endline). Endline 1 variables were collected in 2007-08 and are at the household level. The endline variables were selected to only include pre-determined household characteristics. For more information on the baseline and endline surveys, see Banerjee et al. (2015).

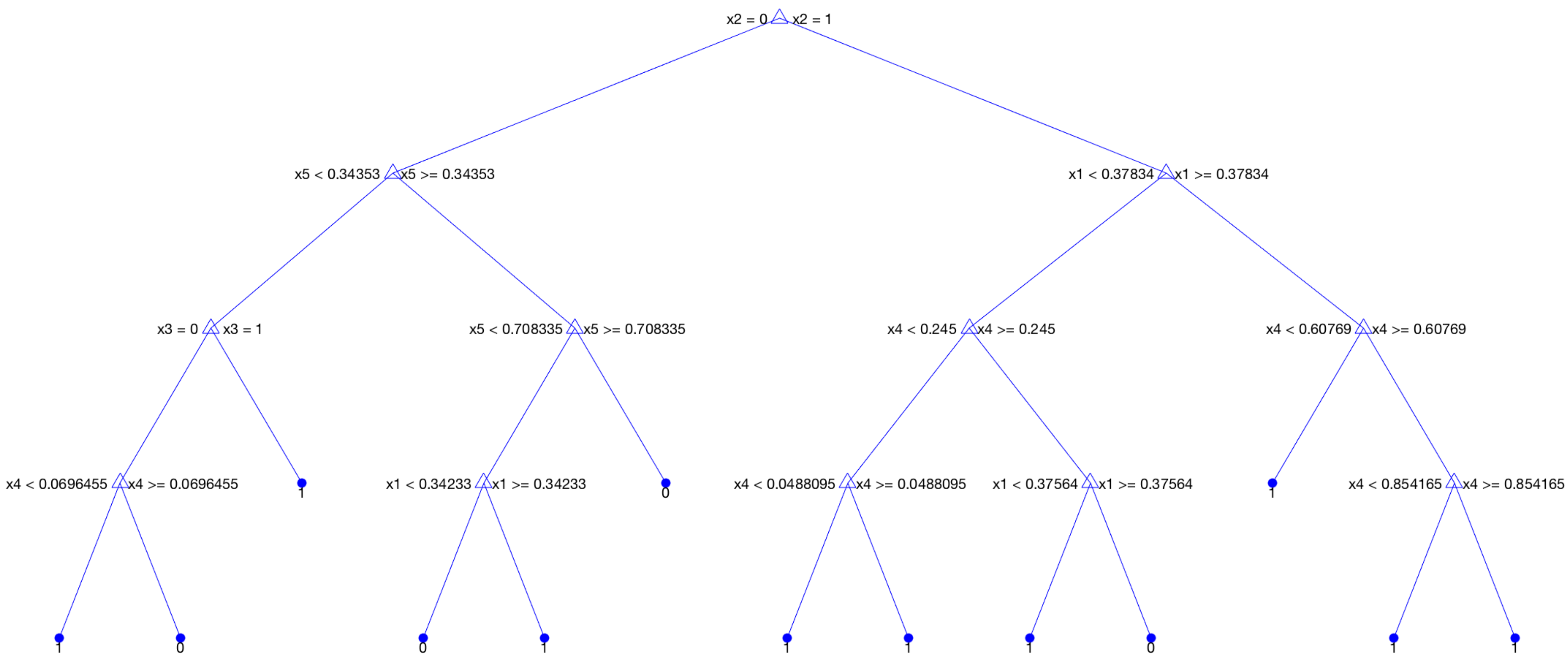


Figure C.1. This presents an example of a decision tree. For the sake of simplicity, we limit the maximum number of splits to 12. The actual procedure has a considerably more complex tree. Here x_1 is the average closeness to leaders, x_2 is whether the household is eligible by having a female of eligible age, x_3 is whether the household is a leader, x_4 is the share of same-caste leaders in the village, and x_5 is the closeness to same-caste leaders.

Table C.1. Confusion Matrices for H and L classification, Karnataka

		Predicted L	Predicted H	Total
Observed	L	1469	898	2367
	H	204	308	512
	Total	1673	1206	$N = 2879$

Notes: This table presents the confusion matrix for the validation sample for Karnataka. The following metrics on this confusion matrix capture classification quality: DOR = 2.47, F1 = 0.359, MCC = 0.172.

Table C.2. Confusion Matrices for H and L classification, Hyderabad

		Predicted L	Predicted H	Total
Observed	L	661	174	835
	H	129	105	234
	Total	790	279	$N = 1069$

Notes: This table presents the confusion matrix for the validation sample for Hyderabad. The following metrics on this confusion matrix capture classification quality: DOR = 3.09, F1 = 0.409, MCC = 0.226.

C.5. **Random Forest Classifier quality metrics and comparison with Logistic Classifier.** Here we compare the performance of the random forest and logistic classifiers. Appendix Tables C.1 and C.2 present the confusion matrices for the random forest classifiers in Karnataka and Hyderabad, respectively. Appendix Tables G.6 and G.12 present the confusion matrices for the logistic classifiers in both samples.

The confusion matrices present the fractions of true positives, true negatives, false positives and false negatives. Ideally, the true positives and negatives would be 100% each and the rate of false positives and negatives would be 0%. In each of the table notes, we also present several commonly-used diagnostic measurements for assessing the quality of classification. These include Matthews correlation coefficient (the preferred diagnostic measure), the F1 score (TP: True Positive, FP: False Positive, TN: True Negative, FN: False Negative), and the the diagnostic odds ratio:

- Matthews correlation coefficient: $MCC = \frac{TP.TN - FP.FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$
- F1 score: $F1 = 2\frac{PPV.TPR}{PPV+TPR}$ where $PPV = \frac{TP}{TP+FP}$ and $TPR = \frac{TP}{TP+FN}$
- Diagnostic odds ratio: $DOR = \frac{TP.TN}{FP.FN}$

Both random forest and logistic classifiers have a cut-off parameter that was chosen by 3-fold cross-validation to maximize the Matthews correlation coefficient. We next compare each classification metric across random forest and logistic. Overall, random forest outperforms logit. The difference is particularly stark in the Hyderabad sample.

C.5.1. *Karnataka Sample.* We compare the quality of the classification for Karnataka:

(1) Confusion Matrix
  (a) True negative rate (TNR):
    - Random forest: 62%
    - Logit: 37%
  (b) True positive rate (TPR):
    - Random forest: 60%
    - Logit: 84%
  (c) Positive predicted value (PPV, probability of true positive vs all positive):
    - Random forest: 26%
    - Logit: 22%
  (d) Negative predicted value (NPV, probability of true negative vs all negative):
    - Random forest: 88%
    - Logit: 91%
(2) Metrics for quality of classification are:
  (a) Matthews correlation coefficient (from [-1,1]): the preferred diagnostic measure
    - Random forest: 0.172
    - Logit: 0.164

(b) F1 score (from $[0, 1]$)
   - Random forest: 0.359
   - Logit: 0.351

(c) Diagnostic odds ratio (positive odds ratio / negative odds ratio, from $[0, \infty)$):
   - Random forest: 2.47
   - Logit: 2.94

C.5.2. *Hyderabad Sample.* We compare the quality of the classification for Hyderabad:

(1) Confusion Matrix
  (a) True negative rate:
   - Random forest: 79%
   - Logit: 78%

  (b) True positive rate:
   - Random forest: 45%
   - Logit: 35%

  (c) Positive predicted value (probability of true positive vs all positive):
   - Random forest: 38%
   - Logit: 31%

  (d) Negative predicted value (probability of true negative vs all negative):
   - Random forest: 84%
   - Logit: 81%

(2) Metrics for quality of classification are:
  (a) Matthews correlation coefficient (from [-1,1]): the preferred diagnostic measure
   - Random forest: 0.226
   - Logit: 0.120

  (b) F1 score (from $[0, 1]$)
   - Random forest: 0.409
   - Logit: 0.325

  (c) Diagnostic odds ratio (positive odds ratio / negative odds ratio, from $[0, \infty)$):
   - Random forest: 3.09
   - Logit: 1.87

Appendix D. Balance

Table D.1. Covariate balance

Karnataka Wave 1 Villages

				Treatment - control			Number of Households Adjusted Treatment - control		
	Obs	Control Mean	Control SD	Coeff.	p-value	5% limit	Coeff.	p-value	5% limit
Number of Households	75	165.812	48.945	57.397	0.000	24.354			
Average Density	75	0.119	0.042	-0.020	0.011	0.015	0.005	0.467	0.013
Average Clustering	75	0.334	0.074	-0.041	0.009	0.030	-0.018	0.284	0.032
Average Closeness	75	0.379	0.046	-0.013	0.183	0.019	0.009	0.320	0.019
Harm mean distance to leaders	7511	0.474	0.079	-0.021	0.049	0.020	0.007	0.481	0.020
WC harm mean dist	7511	0.495	0.186	-0.020	0.127	0.028	-0.004	0.761	0.028

Notes: This table reports Treatment-Control balance on additional variables used as predictors to classify households as H (high MF propensity) vs. L (low MF propensity). 5% limit shows the size of the difference between treatment and control that would be powered to detect at the 5% level, holding the standard error fixed. Unit of observation: household (except for area variables). p-values of differences reflect standard errors clustered at the area level.

Table D.2. Endline network summary statistics

	Non-Microfinance villages
Panel A: Karnataka Wave 2 Data	
Average Degree (Mean)	17.46
Average Degree (Std. Dev.)	4.34
Average Clustering (Mean)	0.32
Average Clustering (Std. Dev.)	0.06
Average Closeness (Mean)	0.48
Average Closeness (Std. Dev.)	0.05
Number of Households (Mean)	175.84
Number of Households (Std. Dev.)	53.49
Panel B: Hyderabad Data	
Average Degree (Mean)	5.949
Average Degree (Std. Dev.)	0.833
Average Clustering (Mean)	0.060
Average Clustering (Std. Dev.)	0.036
Average Closeness (Mean)	0.004
Average Closeness (Std. Dev.)	0.013
Number of Households (Mean)	200.738
Number of Households (Std. Dev.)	101.377

Appendix E. Alternative Models: Existing Models in the Literature

In this section we describe several alternative models, emphasizing why they, in their basic forms, cannot generate the patterns in our data. We also describe an extension of our model that includes direct payoff externalities across links. While richer environments with components from these alternative models may be able to rationalize our empirical findings, we believe our model offers a new, parsimonious, empirically plausible, and valuable theoretical perspective by considering *global* externalities in network formation. In particular, our model places a central focus on the fact that a change, such as the introduction of a new formal financial product, can have effects even for those who do not adopt it and who may be arbitrarily distant from those who do adopt, by affecting the equilibrium incentives to engage in link formation.

We study four specific alternatives and work through each model using the setup of Section 4.6. What follows, of course, is not an exhaustive list, but is representative of the types of models that would be natural candidates for this application.

The first two involve exogenous random matching and mutual consent. These are analogous to the type of models studied by Watts (2001); Jackson and Watts (2002); Christakis, Fowler, Imbens, and Kalyanaraman (2010); Mele (2017), albeit presented in a simplified manner for clarity of argument.

First, Section E.1, presents the case when links are historically given but may break as a result of a shock, such as the introduction of microfinance. New links are however slow to form, and, in the short run, the dominant effect of shock is that links break (in the longer run new links presumably form). This is as in Jackson et al. (2012). The second model takes on the opposite extreme case where links get renewed every period from scratch. So in section E.2, we imagine an exogenous set of unlinked individuals who form new links, with random matching opportunities and mutual consent for link formation.

The third model, presented in Section E.3, returns to the case where links are easy to break but slow to form, but focuses on triads rather than pairs. This introduces the idea of support–that the presence of one link may help sustain other links involving some of the same set of people (Jackson et al., 2012).

Despite their very different perspectives, these three models all point to similar conclusions: that the number of HL links should go down in microfinance villages, while the number of LL links should stay the same or, if it does decline, should decline less than mixed link types. Further LLL triads should decline less than LLH or LHH.

The fourth model, presented in Section E.4, returns to the setup where networks essentially re-form every period, but now introduces "directed search". With directed search, agents are free to choose which other types of agents they want to link with. In such a model, we find that while HL links should decline in microfinance villages, LL links should go up. This fits with the main strand of the network formation literature (e.g., Jackson and Wolinsky (1996); Dutta and Mutuswami (1997); Bala and Goyal (2000); Currarini and Morelli (2000); Jackson and Van den

Nouweland (2005); Bloch, Genicot, and Ray (2008); Herings, Mauleon, and Vannetelbosch (2009); Jackson, Rodriguez-Barraquer, and Tan (2012); Boucher (2015)).

Our conclusion is that these four approaches, based on either exogenous or directed search, with mutual consent and perhaps support requirements for forming links, cannot, at least in their basic versions, generate patterns consistent with the data.

E.1. **The impact on pre-existing links.** The first model takes the view that villagers are in a pre-existing network, and while links are easy to break, forming new links can be very slow and is thus not on the same time-scale. We start from a setting where we take these network connections as given before the arrival of microcredit. Where microcredit arrives, people have the choice of continuing or breaking off those relationships, and breaking is unilateral (consistent with mutual consent models). In control villages we assume that nothing changes.

Let us write that the payoff to node i of type θ_i of being linked to j of type θ_j is given by

$$\alpha_{\theta_i}\beta_{\theta_j}r + \beta_{\theta_i}\alpha_{\theta_j}b - \epsilon_{ij},$$

where G is the CDF of ϵ, a mean-zero random variable, so as before the expected value is

$$v_{\theta\theta'} = \alpha_\theta\beta_{\theta'}r + \beta_\theta\alpha_{\theta'}b.$$

Recall that α_θ is the probability of having money to lend and β_θ is the probability of needing to borrow. So we may imagine that, due to microfinance entry, α_H declines with high frequency repayments (or may increase in the world of relending, which appears to be empirically less common) and β_H increases where the microcredit loans allow the borrower to overcome a non-convexity (or may decline if microcredit loans are substitutes for informal loans). The parameters for Ls, who do not borrow from microfinance, remain the same. As a consequence, we imagine that v_{HH} ought to decline and v_{LL} ought to not change. Whether v_{LH} and v_{HL} is a more delicate matter and in general ambiguous.

With this in mind, what is the effect on the number of relationships of each type: HH, LH, and LL? Clearly the number of HH relations goes down and the number of LL relationships should be unchanged. The number of HL relationships however depends on both the willingness of the H to partner with an L, which has gone down and the willingness of an L to partner with an H, which might have gone up. The number of LH pairs in MF villages is given by

$$G(v_{HL} + \Delta v_{HL}) \cdot G(v_{LH} + \Delta v_{LH})$$

compared to

$$G(v_{HL}) \cdot G(v_{LH})$$

in non-MF villages. For relatively small changes in the value of the relationships the difference in the number of HL pairs can be written as

$$G'(v_{HL})\Delta v_{HL} + G'(v_{LH})\Delta v_{LH} = G'(v_{HL})[\Delta v_{HL} + \Delta v_{LH}]$$

$$= (\alpha_H \Delta\beta_H + \beta_H \Delta\alpha_H)(r + b) < 0.$$

The last inequality follows from the fact that if relending is small relative to the change in appetite for borrowing (as is the case in the empirical literature), then $\Delta_{HH} < 0$, which is the same condition as above.

Therefore the number of HL relations must also fall. Only the number LL relationships do not go down when MF arrives.

CLAIM **1.** *Starting with a given set of links, the introduction of microfinance should*

(1) *reduce HH links,*
(2) *reduce LH links,*
(3) *leave LL links unchanged,*
(4) *and the total number of links should decline and be less than in non-microfinance villages.*

E.2. **Introducing link formation.** We now turn to a model at the other extreme: there is no persistence in links whatsoever, so we can consider the formation of new links from an unmatched population.

As before the pairs are formed if both parties want the link, which happens with probability $G(v_{\theta\tilde{\theta}}) \cdot G(v_{\tilde{\theta}\theta})$ for a $\theta\tilde{\theta}$ link. From above, the fraction of new HH and LH links should go down in microfinance villages but that of new LL links should remain the same.

CLAIM **2.** *If new links are formed by randomly matching, the introduction of microfinance should*

(1) *reduce new HH links,*
(2) *reduce new LH links,*
(3) *leave new LL links unchanged,*
(4) *and the total number of new links should be less than in non-microfinance villages.*

E.3. **A model with supported links.** Our third model again takes the perspective that links are easy to break but slow to form, but in this case we focus on the value of a link being supported in the sense of Jackson et al. (2012). Jackson, Rodriguez-Barraquer, and Tan (2012) introduce the notion of support, which correlates the presence of links based on incentives to exchange favors (including lending to each other).The idea is that two households in isolation may not have enough bilateral interaction to be able to sustain cooperation with each other, but if they both also have relationships with some other households in common, then the relationships can all "support" each other: if someone fails to cooperate with one of their friends then beyond losing that relationship, they also lose relationships with all the other friends that they had in common with the friend with whom they did not cooperate. Fear of losing all of those relationships if they misbehave provides added incentives to maintain cooperation.[3] This leads relationships to be correlated: forming them in supported combinations provides stronger incentives, and then both their presence and disappearance ends up being correlated.

This model builds a natural connection between what happens to the Hs (who are directly affected by microcredit) and what happens to Ls. An LL link can break because it is no longer supported by an H. However, for reasons that will become clear, it cannot explain the patterns we observe in the data.

E.3.1. *Payoffs.* We start with a set of HH, LH, and LL links. However some of these links also support each other in the sense that some are part of HHH, LHH, LLH, or LLL triangles. We assume that no one has more than two links to keep the problem manageable. We assume that the payoff to i from the links between i (a type θ) and j (a type $\widetilde{\theta}$) that is supported by k (a type θ') is given by

$$W_{ijk}(\theta, \widetilde{\theta}|\theta') = v_{\theta\widetilde{\theta}} + \max\{\varepsilon_{ij}, \varepsilon_{ik}\}$$

where $v_{\theta\widetilde{\theta}}$ is defined as in Section 4, and ε_{ij} and ε_{ik} are drawn, as before, i.i.d. from a distribution G.

This formulation makes sense in a world where there is no crowd-out in borrowing or lending – when an agent is in the borrowing state he gets twice the benefit b if he can borrow from two sources and when he is in the lending state he gets twice the benefit r if he can lend to two people.

When the relation is not supported, i.e., there is either just one pair or there is a potential triad but not all 3 pairs are connected, the payoff from it is, as before

$$W_{ij}(\theta, \widetilde{\theta}|\emptyset) = v_{\theta\widetilde{\theta}} + \varepsilon_{ij}$$

where the ε_{ij} is drawn, as before, i.i.d. from a distribution G.

E.3.2. *Analysis of the model.* The decision to be made is simple: whether to stay linked. However starting from a trilateral relationship, there are potentially multiple equilibria: i might leave because she expects k to leave and vice versa. To reduce the number of cases, assume that the equilibrium selection rule is always to choose the triad equilibrium if it existed in the pre-period and is still an equilibrium. In other words, each participant of triad only checks whether they want to stay in the relationship if the other two members of the triad were to stay. If the triad is no longer an equilibrium, then each pair in the erstwhile triad independently decides whether or not to stay together as a pair (and clearly at least one will not), and the equilibrium is unique.

[3]This setup also nests risk sharing, which can be seen as another form of favor exchange in which i gives a transfer to j if j is hit by a negative shock. The third friend, k, can be valued for two (non mutually exclusive) reasons: because k will punish i if i reneges on her expected transfer to j and/or because k can make a transfer to i which can then be shared in turn with j.

Clearly some of the Hs who are in a triad and have access to microfinance will want to break at least one link since both v_{HH} and v_{HL} decline. Once this is taken as given, the value of each remaining relationship goes down at least weakly, and in some fraction of cases those relationships will also break up because they were sustained by the higher ε associated with the triad. The only triads that will be unaffected are the LLL triads. All other types of triads will break up more in MF villages than in non-MF villages. It is also easy to see that LHH triads are more likely to break up than LLH triads with microfinance, simply because the LH links are the vulnerable points.

This model can explain why lots of pre-existing LL links break up in MF villages. The argument would be that most of these links were part of a triad with an H and that the H has less incentive to continue in the triad. It does however suggest that fewer LL links should break up than LH links, since under this theory LL links only break up because an LH link that sustained that LL relationship broke up.

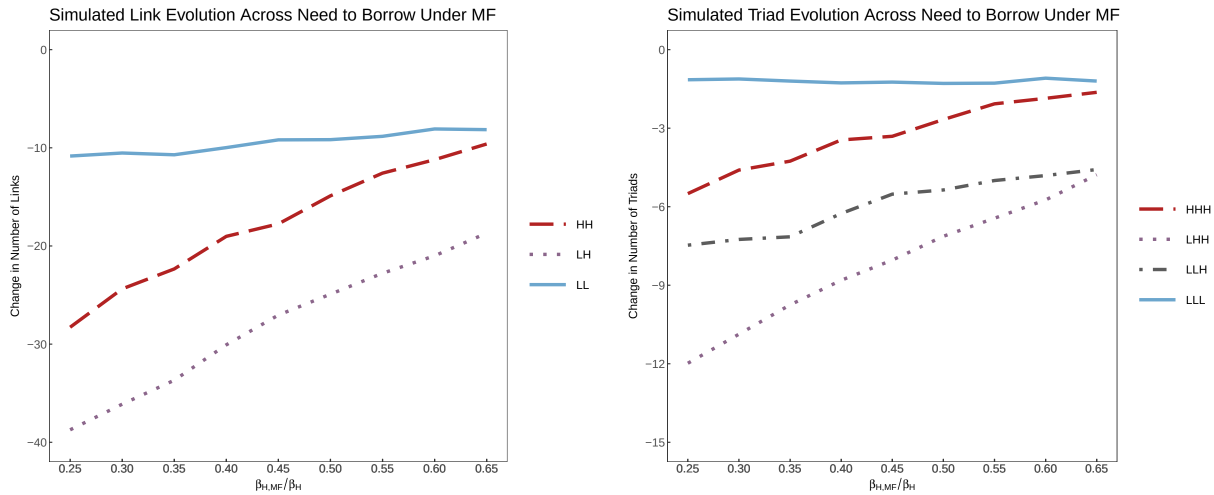
Claim **3.** *In the model with supported links, when microfinance is introduced,*

(1) *LLs decline but LHs should decline by more,*
(2) *LHHs are more likely to decline than LLHs, which are more likely to decline than LLLs.*

E.3.3. *Simulation.* To make this transparent, we present a simulation exercise. We look at networks of size $n = 300$. We set the payoff parameters $r = 0.1$, $b = 1$, and $\alpha_H = \alpha_L = \beta_H = \beta_L = 1/3$. We set $\alpha'_H = 1.45\alpha_H$ and vary the needing to borrow probability under microfinance, $\beta'_H \in \{0.25, 0.3, \ldots, 0.65\}$, for the simulations. Under these parameters we have v_{HH}, v_{HL}, v_{LH}, and v_{LL} satisfying the assumptions maintained throughout this paper, described in Section 4. We let $G(\varepsilon) = \mathcal{N}(0, 1/100)$ and let half the population be H and the other half be L.

We repeat 100 simulations of the following procedure. We seed the graph by connecting collections of mutually exclusive sets of three nodes at random. We then draw ε_{ij} and compute an equilibrium network under no-MF payoffs and an equilibrium network under MF payoffs, holding fixed the seed and the shocks as above. Specifically, any triangle that exists initially and for which it is still an equilibrium under the shocks and payoff parameters to maintain are maintained. If not, then constituent links are checked. A resulting equilibrium graph holding fixed seeds and shocks can be computed for each simulation drawn under both non-MF and MF payoffs.

Figure E.1 presents the results. We plot the change in the number of links (and the change in the number of triangles) comparing MF networks to non-MF networks. We see that MF networks uniformly lead to a decline in every link and triangle type. Furthermore, the gap between the models declines the closer β'_H is to β_H. Nonetheless, what is striking is that LL links drop much less than its counterparts HH and LH, as do LLL triangles compared to HHH, LLH and LHH.



(A) Evolution of HH, LH, and LL

(B) Evolution of HHH, LHH, LLH, and LLL

FIGURE E.1. Supported Links Model

E.3.4. *Summary so far.* The models discussed so far, with or without the idea of support, all point to the same conclusion: the number of LH and HH links should go down faster in MF villages than the number of LL links. Moreover LLL triads should be least affected. This is inconsistent with our empirical findings, suggesting that, at least in their basic versions, these models cannot be the (sole) data generating process in our settings.

There is however one additional factor that the previous models ignore. As Feigenberg et al. (2013) show, empirically, microfinance itself may promote connections between group members, who will, by definition, tend to be Hs. This could lead to offsetting effects on HH links and HHH triads, making net predictions ambiguous. We next consider a model of directed search which accommodates this possibility.

E.4. **A model of directed search.** Let us take the set up of the model where networks essentially re-form every period, but now introduce directed search. Instead of matching randomly, we now assume that each agent can select the population within which they will match. Once they observe who they are matched to, which happens randomly within the group, they get to decide whether they will actually form a link. Link formation is unilateral. There are three possible populations: HH (i.e., just Hs), LL (i.e., just Ls), and LH (i.e., mixed, with the fractions endogenously determined). Within the HH and LL groups everyone will get matched (assuming even numbers). Within the HL group the outcomes depends on the fraction of the two types, but we assume that the maximum possible number of matches are formed.

In this model there are spillovers from the decisions of the Hs on the decisions of the Ls. If Hs decide to stop matching with the Ls, then Ls might be forced to change their matching habits. However for reasons that will become clear, this model does not deliver the desired patterns.

In non-MF villages we have assumed that the payoffs for Hs and Ls are identical and therefore there are many possible equilibria. However, in all equilibria the shares of H and L types in the LH group must be the same.

In MF villages, observe that

$$\Delta v_{HL} - \Delta v_{HH} = \alpha_H \Delta\beta_H b + \beta_H \Delta\alpha_H r - (\alpha_H \Delta\beta_H + \beta_H \Delta\alpha_H)(r+b) = -\alpha_H \beta_H \left[\frac{\Delta\beta_H}{\beta_H} r + \frac{\Delta\alpha_H}{\alpha_H} b \right].$$

This leaves us with two possibilities. Either $\frac{\Delta\beta_H}{\beta_H} r + \frac{\Delta\alpha_H}{\alpha_H} b > 0$ or not. Assume the expression is positive. Since we started from a situation where $v_{HL} = v_{HH} = v_{LL}$, the condition implies that in MF villages $v_{HH} > v_{HL}$. Therefore all Hs will chose the HH option. Paradoxically the same condition also tells us that $\Delta v_{LH} > 0$, so in MF villages $v_{LL} < v_{LH}$. In other words, an L will prefer to be matched with an H. However, the probability of being matched with an H is zero for an L, since all Hs will choose the HH option. Therefore all Ls will choose the LL option.

Or second, $\frac{\Delta\beta_H}{\beta_H} r + \frac{\Delta\alpha_H}{\alpha_H} b < 0$. In this case Hs will want to match with Ls but not the other way around. Therefore once again we will see full homophily. The fraction of both the HH and LL populations will go up and that of HL will go down in both cases. However in both cases the value of HH links has gone down ($\Delta v_{HH} < 0$), while that of LL links is unchanged. Therefore the fraction of HH links actually formed may go up or down. The fraction of LL links should however go up and therefore on aggregate, the LH population turns into HHs and LLs in MF villages. Randomly formed LL pairs out of this population have the same probability of turning into an actual link as randomly formed LH pairs, but randomly formed HH pairs have lower chance of turning into an actual link. The total number of realized links should therefore be lower in MF villages.

This example is extreme but it captures a robust intuition. If microfinance makes Ls want to pair with Hs rather than with Ls, it also makes Hs want to pair with Hs, and vice versa, which is why there are no LH pairs in MF villages.

Claim **4.** *If new links are formed by directed matching, the introduction of microfinance should*

(1) *either reduce or increase new HH links,*
(2) *reduce new LH links,*
(3) *increase new LL links,*
(4) *and the total number of new links should be less than in non-microfinance villages.*

We can see from the result that the predictions of directed search are inconsistent with the data, namely because the effect on LLs should be positive rather than negative, whereas the number of LH links would go down.

Appendix F. Hyderabad Network Elicitation

F.1. **Survey Questions.**

F.1.1. *Direct Link Elicitation.* We first ask the following set of network questions

(1) *Financial* relationships
   (a) If your gas cylinder, kerosene or any other cooking fuel runs out while cooking and you don't have it readily available at home, who would you go to in this neighborhood to borrow some and who would come to you in a similar situation?
   (b) If you need 50 or 100 Rupees because you're falling short for some payment, who in this neighborhood would you borrow this money from and who from this basti would come to you in a similar situation?
   (c) If you had visitors and needed some milk or sugar to make tea but the shop is closed, who in this neighborhood would you borrow it from and who would come to you in a similar situation?
(2) If you needed advice on financial matters, for example, opening a savings account, buying gold, taking a loan, buying insurance, making investments, etc. who in this neighborhood would you go to and who would come to you for similar advice?
(3) *Information* relationships (non-finance)
   (a) If you needed advice on which school/college to put your children in, who in this neighborhood would you go to and who would come to you for similar advice?
   (b) If you had to move to another house in this neighborhood, who would you ask for help to find a house and who would come to you for help to find a house?
   (c) If your child or another member of your family falls sick, who in this neighborhood would you go to for advice and who would come to you for similar advice?
(4) *Social* relationships
   (a) Who would come or send their children to your house to watch television and whose house would you or your children go to for the same purpose?

While these questions resemble those in a full network elicitation, there are several key differences. First, we only interview a subsample of the neighborhood. Second, we do not have a census enumeration of the full neighborhood, so consequently, third, we do not attempt to match survey responses to form an adjacency matrix.

F.1.2. *ARD Questions.* We collected Aggregated Relational Data (ARD) using the following questions:

How many other households do you know in your neighborhood ...

(1) where a woman has ever given birth to twins?
(2) where there is a permanent government employee?
(3) where there are 5 or more children?

(4) where anye child has studied past 10th standard?
(5) where any adult has had typhoid, malaria, or cholera in the past six months?
(6) where any adult has been arrested by the police?
(7) where at least one woman has had a second marriage?
(8) where at least one man currently has more than one wife?
(9) where at least one member has migrated abroad for work?

Each respondent was also asked whether her household posessed each of these traits.

F.2. **ARD Algorithm.** We adapt the ARD algorithm from Breza et al. (2020b). Here we provide an overview of the method. Suppose that a researcher is interested in studying networks in a set of distinct communities. A network with n households is given by $\mathbf{g}$, which is a collection of links ij where $g_{ij} = 1$ if and only if households i and j are linked and $g_{ij} = 0$ otherwise. Our goal is to estimate characteristics of g, such as the probability that arbitrary pairs or triples of nodes are linked.

I. **Conduct ARD survey:** Sample a share ψ (e.g., 30 percent) of households. Have each enumerate a list of their network links. Note that this gives a direct estimate of the respondent's degree. Ask 5-9 ARD questions, such as

"How many households among your network list do you know where any adult has had typhoid, malaria, or cholera in the past six months?"

The ARD response for a household i is

$$y_{ik} = \sum_j g_{ij} \cdot \mathbf{1}\{j \text{ has had one of those diseases in past 6 mo.}\}$$

where trait k denotes the disease question. This adds up all links that have had the diseases over the last six months. Ask whether the respondent household themselves has each ARD trait k as well to generate population estimates for the prevalence of each trait.

II. **Estimate network formation model with ARD:** Use the information from the ARD survey and the trait prevalences to estimate the parameters of a network formation model. In this model, the probability that two households i and j are linked depends on household fixed effects (ν_i) and distance in a latent space (latent locations z_i) with

$$\mathrm{P}(g_{ij} = 1 | \nu_i, \nu_j, \zeta, z_i, z_j) \propto \exp(\nu_i + \nu_j + \zeta \cdot \text{distance}(z_i, z_j)).$$

We use a Bayesian framework to estimate the model parameters and latent space locations. We assume a latent space on $\mathcal{S}^3$, the surface of a sphere. Priors for latent positions of each individual and trait group follow a von Mises-Fisher distribution:

$$\begin{aligned} z_i | \upsilon_z, \eta_z &\sim \mathcal{M}(\upsilon_z, 0) \\ z_{j \in G_k} | \upsilon_k, \eta_k &\sim \mathcal{M}(\upsilon_k, \eta_k) \end{aligned}$$

The centers $\{\upsilon_z, \upsilon_k\}_{z,k}$ and concentration parameters $\{\eta_k\}_k$ of the prior distributions also need to be estimated.

For respondent i and subpopulation G_k, recall that we observe:

$$y_{ik} = \sum_{j \in G_k} g_{ij}.$$

Conditional on latent positions, $(z_i, z_{j\in G_k})$, and household fixed effects $\{\nu_i\}_i$, y_{ik} approximately follows a Poisson distribution when the number of individuals in the subpopulation n_k is large. The Poisson parameter is:

$$\lambda_{ik} = \sum_{j \in G_k} \mathrm{P}(g_{ij} = 1 | \nu_i, \nu_{j\in G_k}, \zeta, z_i, z_{j \in G_k})$$

The expected ARD response by i for category k can be expressed as

$$\lambda_{ik} = \mathrm{E}[y_{ik}] = d_i b_k \left(\frac{C_{p+1}(\zeta) C_{p+1}(\eta_k)}{C_{p+1}(0) C_{p+1} \sqrt{\zeta^2 + \eta_k^2 + 2\zeta\eta_k \cos(\theta_{(z_i, \upsilon_k)})}} \right),$$

where d_i is the respondent expected degree, $b_k = n_k/n$ the share of nodes in group k, $C_{p+1}(\cdot)$ is the normalizing constant of the von Mises-Fisher distribution, and $\theta_{(z_i,\upsilon_k)}$ is the angle between the two vectors. Note that the resulting likelihood relies entirely on ARD:

$$y_{ik} | d_i, b_k, \zeta, \eta_k, \theta_{(z_i,\upsilon_k)} \sim \text{Poisson}\left(\lambda_{ik}\right).$$

Finally, to estimate the household fixed effects $\{\nu_i, \nu_j\}$, note that

$$d_i = n \exp(\nu_i) \mathrm{E}[\exp(\nu_j)] \left(\frac{C_{p+1}(0)}{C_{p+1}(\zeta)} \right).$$

Letting $\boldsymbol{\theta}$ be a shorthand for all parameters, we can estimate the posterior

$$\boldsymbol{\theta} | y_{ik} \propto \prod_{k=1}^{K} \prod_{i=1}^{n} \exp(-\lambda_{ik}) \lambda_{ik}^{y_{ik}} \prod_{i=1}^{n} \text{Normal}(\log(d_i) | \mu_d, \sigma_d^2)$$
$$\times \prod_{k=1}^{K} \text{Normal}(\log(b_k) | \mu_b, \sigma_b^2) \prod_{k=1}^{K} \text{Normal}(\log(\eta_k) | \mu_{\eta_k}, \sigma_{\eta_k}^2) \text{Gamma}(\zeta | \gamma_\zeta, \psi_\zeta).$$

Given this posterior distribution, the probability of any network $\mathbf{g}$ being drawn is fully computed.

III. **Compute network statistics of interest:** Use the estimated probability model (using ζ, fixed effects ν_i and latent locations z_i) to compute $\mathrm{E}[S(\mathbf{g})|\mathbf{Y}]$, where $S(\mathbf{g})$ is, for example, the probability of a link between any arbitrary pair (or triple) of nodes. The data and replication code is freely available (Breza et al., 2020a).

Appendix G. Alternate Logistic Classification

G.1. Karnataka Exhibits.

Table G.1. Characteristics of H versus L, Karnataka

Panel A: Karnataka - Demographics and Amenities variables

	(1) GMOBC	(2) Latrine	(3) Num. Rooms	(4) Num. Beds	(5) Thatched Roof	(6) RCC Roof
H	−0.275	−0.358	−0.732	−0.787	0.018	−0.187
	(0.014)	(0.015)	(0.039)	(0.063)	(0.003)	(0.012)
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Depvar Mean	0.7	0.261	2.36	0.84	0.0235	0.117
Observations	14,904	14,904	14,904	14,904	14,904	14,904

Panel B: Karnataka - Network variables

	(1) Degree	(2) Links to L	(3) Links to H	(4) Eig. Cent.
H	2.998	−0.222	2.962	0.017
	(0.178)	(0.138)	(0.099)	(0.001)
	[0.000]	[0.000]	[0.025]	[0.000]
Depvar Mean	8.97	2	5.7	0.0524
Observations	14,904	14,904	14,904	14,904

Notes: Standard errors (clustered at the village level) are reported in parentheses. p-values are reported in brackets. Panels A and B pertains to Karnataka, based on Wave 1 data only. GMOBC = A dummy for whether the household consists of general caste or backwards caste, so the omitted categories are scheduled caste and scheduled tribes. General and OBC are considered upper caste.

Table G.2. Link Evolution, Karnataka

	(1) Linked Post-MF	(2) Linked Post-MF	(3) Linked Post-MF	(4) Linked Post-MF
Microfinance	−0.071	−0.082	−0.025	−0.023
	(0.023)	(0.021)	(0.010)	(0.008)
	[0.003]	[0.000]	[0.010]	[0.005]
Microfinance × LH	0.021	0.021	0.002	0.004
	(0.020)	(0.019)	(0.005)	(0.004)
	[0.315]	[0.283]	[0.631]	[0.335]
Microfinance × HH	0.022	0.021	0.005	0.008
	(0.025)	(0.021)	(0.007)	(0.006)
	[0.364]	[0.333]	[0.419]	[0.150]
LH	−0.068	−0.054	−0.012	−0.016
	(0.017)	(0.017)	(0.004)	(0.003)
	[0.000]	[0.002]	[0.006]	[0.000]
HH	−0.043	−0.024	−0.001	−0.012
	(0.019)	(0.019)	(0.006)	(0.005)
	[0.025]	[0.204]	[0.825]	[0.022]
Observations	57,376	57,376	846,561	846,561
Linked Pre-MF	Yes	Yes	No	No
Controls		✓		✓
Depvar Mean	0.441	0.441	0.0636	0.0636
LL, Non-MF Mean	0.523	0.523	0.0848	0.0848
MF + MF x $LH = 0$ p-val	0.017	0.004	0.001	0.002
MF + MF x $HH = 0$ p-val	0.006	0.001	0.005	0.017
MF + LH x MF = MF + HH x MF p-val	0.911	0.998	0.423	0.225

Notes: Classification of H type is based on logistic regression. Standard errors (clustered at the village level) are reported in parentheses. p-values are reported in brackets. Controls are selected by double post lasso among centrality controls (vector of flexible controls for centrality of both nodes), household characteristics (caste, a number of wealth proxies including number of rooms, number of beds, electrification, latrine presence, and roofing material) and all variables that are used in the random forest classification.

Table G.3. Link Evolution for Info and Financial Links, Karnataka

	(1) Financial Linked Post-MF	(2) Financial Linked Post-MF	(3) Info Linked Post-MF	(4) Info Linked Post-MF
Microfinance	−0.060	−0.013	−0.061	−0.016
	(0.027)	(0.007)	(0.026)	(0.007)
	[0.026]	[0.056]	[0.019]	[0.022]
Microfinance × LH	0.010	0.001	0.025	0.003
	(0.026)	(0.004)	(0.023)	(0.004)
	[0.690]	[0.800]	[0.270]	[0.499]
Microfinance × HH	0.013	0.002	0.017	0.003
	(0.028)	(0.005)	(0.027)	(0.005)
	[0.651]	[0.722]	[0.514]	[0.549]
LH	−0.067	−0.008	−0.069	−0.011
	(0.020)	(0.004)	(0.019)	(0.004)
	[0.001]	[0.042]	[0.000]	[0.003]
HH	−0.049	0.0001	−0.045	−0.003
	(0.021)	(0.004)	(0.022)	(0.004)
	[0.020]	[0.991]	[0.038]	[0.490]
Observations	27,072	876,865	37,044	866,893
Linked Pre-MF	Yes	No	Yes	No
Depvar Mean	0.333	0.0341	0.326	0.0377
LL, Non-MF Mean	0.415	0.0458	0.403	0.0534
MF + MF x LH= 0 p-val	0.024	0.003	0.055	0.002
MF + MF x HH= 0 p-val	0.012	0.008	0.009	0.003
MF + LHxMF = MF + HHxMF p-val	0.897	0.759	0.634	0.934

Notes: Classification of H type is based on logistic regression. Standard errors (clustered at the village level) are reported in parentheses. p-values are reported in brackets. Columns 1-2 restrict to financial links, while columns 3-4 restrict to non-financial links. Columns 1 and 3 consider links that existed in Wave 1, while columns 2 and 4 consider pairs of nodes that were not linked in Wave 1.

TABLE G.4. Triples Evolution, Karnataka

	(1) Full triangle linked Post-MF	(2) Full triangle linked Post-MF	(3) Any link in triangle survived Post-MF	(4) Any link in triangle survived Post-MF
Microfinance	−0.142	−0.131	−0.087	−0.080
	(0.041)	(0.031)	(0.041)	(0.030)
	[0.001]	[0.000]	[0.033]	[0.009]
Microfinance × LLH	0.072	0.054	0.027	0.016
	(0.032)	(0.027)	(0.030)	(0.024)
	[0.026]	[0.044]	[0.378]	[0.508]
Microfinance × LHH	0.105	0.078	0.036	0.017
	(0.041)	(0.033)	(0.042)	(0.032)
	[0.010]	[0.018]	[0.391]	[0.593]
Microfinance × HHH	0.107	0.069	0.044	0.017
	(0.042)	(0.033)	(0.044)	(0.032)
	[0.012]	[0.036]	[0.327]	[0.603]
LLH	−0.086	−0.064	−0.031	−0.011
	(0.021)	(0.022)	(0.016)	(0.016)
	[0.000]	[0.005]	[0.052]	[0.482]
LHH	−0.110	−0.071	−0.043	−0.004
	(0.026)	(0.028)	(0.024)	(0.023)
	[0.000]	[0.012]	[0.068]	[0.878]
HHH	−0.100	−0.053	−0.018	0.033
	(0.029)	(0.030)	(0.023)	(0.024)
	[0.001]	[0.081]	[0.444]	[0.181]
Observations	53,233	53,233	53,233	53,233
Linked Pre-MF	Yes	Yes	Yes	Yes
Controls		✓		✓
Depvar Mean	0.197	0.197	0.808	0.808
LLL, Non-MF Mean	0.324	0.324	0.868	0.868
MF + MF x HHH = 0 p-val	0.147	0.004	0.019	0.000
MF + MF x LLH = 0 p-val	0.008	0.002	0.011	0.003
MF + MF x LHH = 0 p-val	0.142	0.032	0.015	0.002
MF + MF x HHH = MF + MF x LLH p-val	0.230	0.550	0.527	0.984
MF + MF x HHH = MF + MF x LHH p-val	0.936	0.627	0.661	0.973
MF + MF x LLH = MF + MF x LHH p-val	0.083	0.148	0.612	0.952

Notes: Classification of H type is based on logistic regression. Standard errors (clustered at the village level) are reported in parentheses. p-values are reported in brackets. Controls are selected by double post lasso among centrality controls (vector of flexible controls for centrality of both nodes), household characteristics (caste, a number of wealth proxies including number of rooms, number of beds, electrification, latrine presence, and roofing material) and all variables that are used in the random forest classification.

Table G.5. Borrowing patterns

	(1) MFI	(2) Friends	(3) SHG	(4) Moneylender	(5) Family
Microfinance × Post	263.311	−601.585	−620.619	−722.385	788.165
	(142.672)	(406.126)	(410.764)	(1,115.592)	(690.808)
	[0.065]	[0.139]	[0.131]	[0.518]	[0.254]
Microfinance × Post × H	1,032.967	244.268	−330.376	650.958	−956.507
	(204.985)	(258.923)	(379.616)	(1,136.715)	(878.447)
	[0.000]	[0.346]	[0.385]	[0.567]	[0.277]
Microfinance × H	41.647	−71.273	−123.645	419.473	771.010
	(53.205)	(57.566)	(169.093)	(621.046)	(725.419)
	[0.434]	[0.216]	[0.465]	[0.500]	[0.288]
Post × H	167.255	−554.681	553.549	−800.524	521.388
	(118.240)	(190.730)	(334.329)	(999.243)	(662.376)
	[0.158]	[0.004]	[0.098]	[0.424]	[0.432]
Observations	28,062	27,194	28,062	28,062	28,062
Depvar Mean	596.976	860.228	1863.324	2667.56	1656.881
L, Non-MF Mean	173.353	1235.576	1528.414	3183.32	2001.767
MF x Post x H + MF x Post =0 p-val	0.000	0.202	0.038	0.926	0.795

Notes: Classification of H type is based on logistic regression. This table presents the effect of microfinance access on the loan amounts borrowed from various sources. Outcomes are winsorized to the 1% level. All of the columns control for surveyed in wave 1 fixed effects. Here all specifications include demographic household and village controls (the same ones used in random forest classification of H vs L) subject to double-post LASSO. Standard errors (clustered at the village level) are reported in parentheses. p-values are reported in brackets. MFI: Microfinance Institution; SHG: Self-Help Group

Table G.6. Confusion Matrices for H and L classification, Karnataka

		Predicted L	Predicted H	Total
Observed	L	866	1501	2367
	H	84	428	512
	Total	950	1929	$N = 2879$

Notes: This table presents the confusion matrix for the validation sample for Karnataka. The following metrics on this confusion matrix capture classification quality: DOR = 2.94, F1 = 0.351, MCC = 0.164.

G.2. **Hyderabad Exhibits.**

TABLE G.7. Characteristics of H versus L, Hyderabad

Panel A: Hyderabad - Demographics and Amenities variables

	(1)	(2)	(3)	(4)	(5)
	GMOBC	Latrine	Num. Rooms	Thatched Roof	RCC Roof
H	0.024	0.072	0.193	−0.003	−0.027
	(0.025)	(0.019)	(0.081)	(0.007)	(0.018)
	[0.351]	[0.000]	[0.019]	[0.690]	[0.141]
Depvar Mean	0.429	0.578	2.314	0.025	0.882
Observations	4,520	4,483	4,516	4,516	4,508

Panel B: Hyderabad - Network Variables

	(1)	(2)	(3)	(4)
	Expected Degree	Expected Links to L	Expected Links to H	Expected Centrality
H	0.373	0.100	0.272	0.010
	(0.123)	(0.123)	(0.090)	(0.003)
	[0.003]	[0.418]	[0.003]	[0.001]
Depvar Mean	5.806	4.144	1.664	0.074
Observations	4,523	4,523	4,523	4,523

Notes: Classification of H type is based on logistic regression. Standard errors (clustered at the village level) are reported in parentheses. p-values are reported in brackets. GMOBC = A dummy for whether the household consists of general caste, otherwise back-wards caste, so the omitted categories are scheduled caste and scheduled tribes. General andOBC are considered upper caste. RCC is Reinforced Cement Concrete.

Table G.8. Link Evolution for Financial and Non Financial Links, Hyderabad

	(1) Financial Links	(2) Financial Links	(3) Non Financial Links	(4) Non Financial Links
Microfinance	−0.355	−0.347	−0.232	−0.192
	(0.127)	(0.131)	(0.120)	(0.116)
	[0.007]	[0.010]	[0.057]	[0.101]
Microfinance × H	0.475	0.483	0.692	0.694
	(0.173)	(0.164)	(0.182)	(0.178)
	[0.008]	[0.005]	[0.0003]	[0.0002]
H	0.138	−0.340	−0.085	−0.218
	(0.124)	(0.175)	(0.127)	(0.167)
	[0.269]	[0.055]	[0.502]	[0.195]
Observations	4,429	4,429	4,429	4,429
Double-Post LASSO	No	Yes	No	Yes
Depvar Mean	4.24	4.24	2.87	2.87
MF + MF × H = 0 p-val	0.586	0.505	0.024	0.008

Notes: Classification of H type is based on logistic regression. Standard errors (clustered at the village level) are reported in parentheses. p-values are reported in brackets. All columns include a full set of controls. Centrality controls are a vector of flexible controls (a polynomial) for centrality of both nodes. Household characteristics are caste and a number of wealth proxies including number of rooms, number of beds, electrification, latrine presence, and roofing material. Household predictor variables consist of all variables that are used in the random forest classification. In every case we include interactions of all of these network, demographic, and classification variables with microfinance.

Table G.9. Link Evolution, Hyderabad

	(1) Prob. Linked	(2) Prob. Linked
Microfinance	−0.004	−0.006
	(0.002)	(0.002)
	[0.147]	[0.012]
Microfinance x HH	−0.011	−0.008
	(0.006)	(0.003)
	[0.049]	[0.029]
Microfinance x LH	−0.003	−0.001
	(0.002)	(0.001)
	[0.161]	[0.308]
HH	0.016	0.015
	(0.005)	(0.004)
	[0.004]	[0.0005]
LH	0.005	0.004
	(0.002)	(0.001)
	[0.007]	[0.003]
Observations	141,996	141,996
Controls	No	Yes
Depvar Mean	0.0255	0.0255
LL, Non MF Mean	0.0257	0.0257
MF + MF x $HH = 0$ p-val	0.02	0.001
MF + MF x $LH = 0$ p-val	0.063	0.009
MF + MF x $HH =$ MF + MF x LH p-val	0.033	0.009

Notes: Classification of H type is based on logistic regression. Standard errors (clustered at the village level) are reported in parentheses. p-values are reported in brackets. The controls are selected by double post lasso among all the variables that are used for its random forest classification, and includes several household and village level characteristics.

TABLE G.10. Triples Evolution, Hyderabad

All variables x 1000	(1) Full Triangle Linked	(2) Full Triangle Linked
Microfinance	−0.007	−0.036
	(0.009)	(0.023)
	[0.434]	[0.109]
Microfinance × LLH	−0.012	−0.006
	(0.007)	(0.005)
	[0.106]	[0.298]
Microfinance × LHH	−0.051	−0.035
	(0.027)	(0.017)
	[0.056]	[0.044]
Microfinance × HHH	−0.160	−0.122
	(0.076)	(0.050)
	[0.035]	[0.015]
LLH	0.017	0.018
	(0.007)	(0.007)
	[0.010]	[0.017]
LHH	0.062	0.062
	(0.026)	(0.023)
	[0.018]	[0.007]
HHH	0.176	0.166
	(0.075)	(0.060)
	[0.019]	[0.006]
Observations	3,341,006	3,341,006
Controls	No	Yes
Depvar Mean	0.0353	0.0353
LLL, Non-MF Mean	0.0299	0.0299
MF + MF × $HHH = 0$ p-val	0.03	0.006
MF + MF × $LLH = 0$ p-val	0.164	0.103
MF + MF × $LHH = 0$ p-val	0.051	0.024
MF + MF × HHH = MF + MF × LLH p-val	0.037	0.015
MF + MF × HHH = MF + MF × LHH p-val	0.033	0.015
MF + MF × LLH = MF + MF × LHH p-val	0.058	0.032

Notes: Classification of H type is based on logistic regression. Standard errors (clustered at the village level) are reported in parentheses. p-values are reported in brackets. The controls are all the variables that are used for its random forest classification, and includes several household and village level characteristics.

Table G.11. Borrowing patterns, Hyderabad

	(1) MFI	(2) Friends	(3) SHG	(4) Moneylender	(5) Family
Microfinance	1,175.608	−0.860	−1,840.964	−1,740.473	125.632
	(260.751)	(846.228)	(848.555)	(1,560.136)	(662.533)
	[0.00002]	[1.000]	[0.033]	[0.268]	[0.850]
Microfinance × H	1,891.065	−265.623	−213.995	841.310	836.203
	(483.311)	(1,330.141)	(1,587.697)	(2,420.973)	(1,071.399)
	[0.0002]	[0.843]	[0.894]	[0.729]	[0.437]
H	−1,114.618	−1,546.096	267.459	−765.199	−298.855
	(437.060)	(1,373.934)	(1,392.183)	(2,102.069)	(833.856)
	[0.013]	[0.264]	[0.849]	[0.717]	[0.721]
Observations	6,811	6,863	6,863	6,863	6,863
Depvar Mean	3107.86	7895.05	6935.66	18805.06	2620.97
L, Non MF Mean	2079.1	7895.74	7020.19	19372.79	2634.7
MF + MF × H = 0 p-val	0	0.837	0.208	0.725	0.387

Notes: Classification of H type is based on logistic regression. These tables present the effect of microfinance access on the loan amounts borrowed from various sources. Outcomes are winsorized to the 2.5% level. Here all specifications include demographic household and village controls (the same ones used in random forest classification of H vs L) subject to double-post LASSO. Standard errors (clustered at the village level) are reported in parentheses. p-values are reported in brackets.

Table G.12. Confusion Matrices for H and L classification, Hyderabad

		Predicted L	Predicted H	Total
Observed	L	651	184	835
	H	153	81	234
	Total	804	265	$N = 1069$

Notes: Classification of H type is based on logistic regression. This table presents the confusion matrix for the validation sample for Hyderabad. The following metrics on this confusion matrix capture classification quality: DOR = 1.87, F1 = 0.325, MCC = 0.120.

Table G.13. Risk sharing, Hyderabad

	(1) Expenditures: Non-Food	(2) Expenditures: Total
Microfinance × Income	0.079	0.071
	(0.030)	(0.037)
	[0.009]	[0.056]
Microfinance × Income $\times H$	−0.076	−0.096
	(0.045)	(0.059)
	[0.097]	[0.107]
Household Income per capita	0.054	0.113
	(0.020)	(0.026)
	[0.011]	[0.000]
Household Income per capita $\times H$	0.025	0.041
	(0.028)	(0.043)
	[0.389]	[0.347]
Observations	10,502	10,593
Depvar Mean	1193	2040
L, Non-MF Depvar Mean	1184	2055
Income Mean	1440	1437
L, Non-MF Income Mean	1437	1434
Test: MF x Income + MF x Income x H = 0	0.931	0.618

Notes: Classification of H type is based on logistic regression. Income is total household, monthly per capita earnings from employment or business activities, excluding private and government transfers. Dependent variable is monthly per capita household expenditure. In col. 1, expenditure excludes food and in col. 2, we present total expenditure. Data is from the first (2007-08) and third (2012) waves of the Hyderabad survey. Regression includes controls for household fixed effects and wave-by-neighborhood-by-type fixed effects. Additional controls are selected by double post lasso from the set of variables used in the prediction exercise, interacted with type. Standard errors (clustered at the neighborhood level) are reported in parentheses. p-values are reported in brackets.

Appendix H. Hyderabad Consumption Smoothing Robustness

Table H.1. Microfinance Treatment Effects on Income, Hyderabad

	(1) Income
Microfinance	1.730
	(63.477)
	[0.979]
Microfinance $\times H$	125.279
	(85.296)
	[0.145]
Observations	11,768
Depvar Mean	1423
L, Non-MF Depvar Mean	1417
Test: MF + MF x H = 0	0.103

Notes: Income is total household, monthly per capita earnings from employment or business activities. Controls are selected by double post lasso from strata fixed effects and the set of variables used in the prediction exercise, interacted with type. Standard errors (clustered at the neighborhood level) are reported in parentheses. p-values are reported in brackets.

Appendix I. Model Simulation

Here we provide simulation evidence that, in our model, LL links can drop more than HL or HH links in response to an exogenous change in relationship value. We consider the model specialized to the case of two types, as in Section 4.6.

In our simulations, we maintain the following parameters:

- $n = 250$; population
- $\lambda_H = 0.2$; the share of high types, which approximates the empirical frequency
- $\lambda_L = (1 - \lambda_H)$; the share of low types
- $F(\cdot)$ is the uniform distribution on [0,1]
- values for links
  - $v_{HH} = 0.46$
  - $v_{LL} = 0.27$
  - In our simulations we vary v_{HL} and v_{LH} over the range

$$(v_{HL}, v_{LH}) \in [0.05, 0.5]^2$$

    to trace out relative changes in linking effort across types over the parameter space.
- base costs and benefits of socializing
  - $c = 4$; homogenous cost
  - $u = 0.3$; homogenous benefit

We show below the resulting stylized network is sparse and homophilic, much like the data. Nevertheless, in the homophilic network, we will find a non-trivial part of the parameter space where declines of effort are greater for Ls than Hs and where links decline more between two L types than any other configuration of types.

As shown in the paper, unique equilibrium efforts can be calculated as

$$e = (I - E)^{-1} u$$

with E being the $|\Theta| \times |\Theta|$ matrix with θ, θ' entries

$$\frac{1}{c_\theta} \mathrm{E}^+[v_{\theta\theta'}] n_{\theta\theta'} \left(1 - F(-v_{\theta\theta'})\right) \left(1 - F(-v_{\theta'\theta})\right).$$

Further, as shown in Section 4.6, we can directly calculate the expected degrees of nodes of each type, d_θ, as well as the link counts from θ to θ', denoted $d_{\theta\theta'}$.

We are interested in how e changes with the introduction of microcredit. Microfinance is assumed to affect valuation only through one parameter; it reduces v_{HL}:

$$v_{HL}^{\mathrm{mf}} = 0.75 v_{HL}^{\mathrm{no\ mf}}$$
$$v_{LH}^{\mathrm{mf}} = v_{LH}^{\mathrm{no\ mf}}$$
$$v_{LL}^{\mathrm{mf}} = v_{LL}^{\mathrm{no\ mf}}$$

$$v_{HH}^{\text{mf}} = v_{HHH}^{\text{no mf}}.$$

By varying the valuation of cross-type links, we can study how the equilibrium efforts change across the $(v_{HL}^{\text{no mf}}, v_{LH})$-plane. In what follows, v_{HL} denotes $v_{HL}^{\text{no mf}}$.

Specifically, we look the quantity

$$\delta(v_{HL}, v_{LH}) := \frac{e_H^{\text{mf}} - e_H^{\text{no mf}}}{e_L^{\text{mf}} - e_L^{\text{no mf}}}$$

which is a positive ratio, since both efforts decline due to microcredit as per Proposition 1.

Figure I.1 plots $\log \delta$ as a function of (v_{HL}, v_{LH}). Because δ is a ratio of changes in fractions, the logarithmic transformation makes visualization easier by dampening extremes for visualization. So, notice that if $\log \delta < 0$, then the decline in effort for H types is less than that of L types which is the pathology made possible under this model.

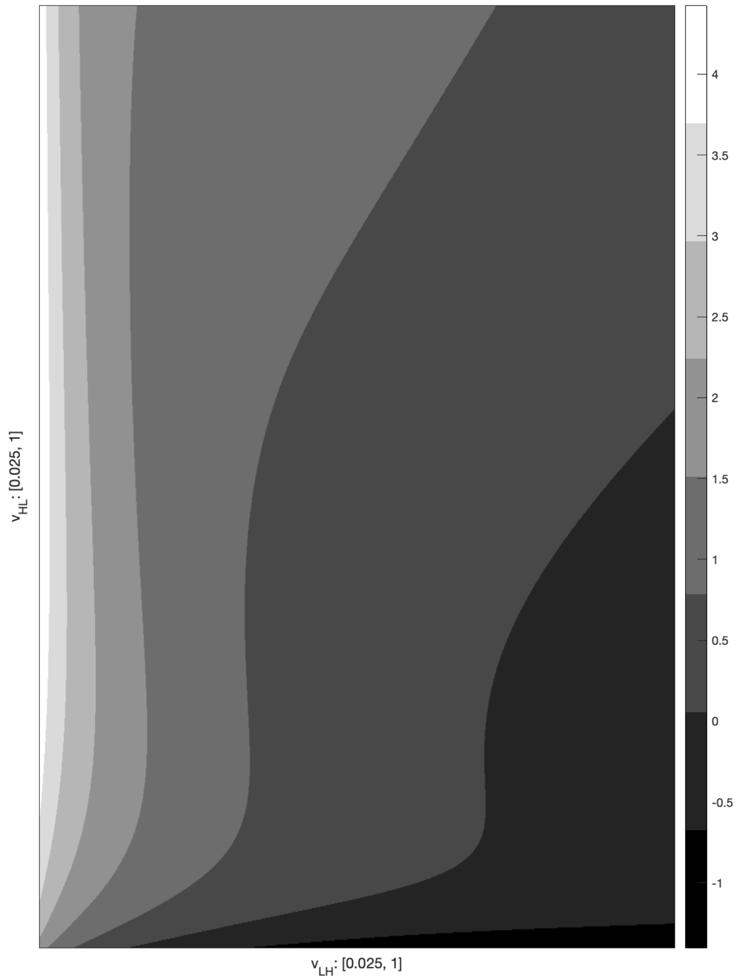


Figure I.1. Plot of $\log \delta$ as a function of (v_{HL}, v_{LH})

From the figure it is clear that effort declines more for Ls when $v_{LH} >> v_{HL}$ (bottom, right area of the parameter space). Moreover, for larger values of v_{LH}, we see larger declines in e_L than e_H (or $\log \delta < 0$), even for larger values of v_{HL}. Further, the intuition is that when Hs place lower value on Ls, but Ls still value Hs quite a bit, then microcredit greatly reduces mutual consent. As a consequence, much of the motivation on behalf of the Ls themselves globally reduces and it can result in greater declines in effort for Ls.

In Table I.1, we present the basic link patterns for an example in the parameter space, with $v_{HL} = 0.01$ and $v_{LH} = 0.8$. With the chosen parameters, the network is sparse (an average degree of roughly 9) but also exhibits considerable homophily. The vast majority of links are within type, rather than across type. This also points to the fact that just because Ls value Hs more does not mean that the network need to be heterophilic. It is useful to note that links require mutual consent, so even if Ls highly value Hs, the fact that Hs do not value Ls much can generate homophily. This is amplified by the fact that $\lambda_H = 0.2$, so, as there are few Hs to begin with, the resulting network would still be homophilic.

In this example we see a sparse network, with strong homophily, with uniform declines in link patterns of all types and efforts in linking when microfinance is introduced. Further, the declines for d_{LL} are the largest and also e_L declines more than e_H.

Table I.1. Link and Effort Patterns for an Example

	No MF	MF	Difference
d_H	9.614	9.287	-0.327
d_L	8.689	8.226	-0.463
e_H	0.898	0.896	-0.002
e_L	0.760	0.742	-0.018
d_{HH}	8.523	8.490	-0.033
d_{HL}	1.091	0.798	-0.293
d_{LH}	0.273	0.199	-0.073
d_{LL}	8.416	8.027	-0.389

Appendix J. Effects of microfinance: Example

Consider the special case where $\alpha_L = \alpha_H$, $\beta_L = \beta_H$ and $\alpha_H \Delta\beta_H + \beta_H \Delta\alpha_H = 0$. In this case $\Delta_{HH} = 0$.

Now suppose first that $\Delta\beta_H > 0$ and therefore $\Delta\alpha_H < 0$. In this case

$$0 < \Delta_{HL} = \alpha_H \Delta\beta_H b + \beta_H \Delta\alpha_H r \Leftrightarrow r\frac{\beta_H |\Delta\alpha_H|}{\alpha_H \Delta\beta_H} = r < b$$

and

$$0 < \Delta_{LH} = \alpha_H \Delta\beta_H r + \beta_H \Delta\alpha_H b \Leftrightarrow r\frac{\alpha_H \Delta\beta_H}{\beta_H |\Delta\alpha_H|} = r > b.$$

In the case where $\Delta\beta_H < 0$ and therefore $\Delta\alpha_H > 0$, these inequalities get reversed and we get

$$0 < \Delta_{HL} \Leftrightarrow r > b$$

and

$$0 < \Delta_{LH} \Leftrightarrow r < b.$$

In other words, in this special case, Δ_{HL} and Δ_{LH} move in opposite directions and which one goes up depends on which of r and b is bigger and whether or not $\Delta\beta_H > 0$.

Since $b > r$, in this special example we would expect Δ_{HL} to be positive and Δ_{LH} to be negative as long as $\Delta\beta_H > 0$ and the reverse otherwise. In other words, it is entirely possible for v_{HL} to go up, v_{LH} to go down and v_{HH} to be unchanged but it requires α_H to go down and β_H to go up.

References

Bala, V. and S. Goyal (2000): "A noncooperative model of network formation," *Econometrica*, 68, 1181–1229.

Banerjee, A., A. Chandrasekhar, E. Duflo, and M. O. Jackson (2013): "Diffusion of Microfinance," *Science*, 341, DOI: 10.1126/science.1236498, July 26 2013.

Banerjee, A., E. Duflo, R. Glennerster, and C. Kinnan (2015): "The miracle of microfinance? Evidence from a randomized evaluation," *American Economic Journal: Applied Economics*, 7, 22–53.

Bloch, F., G. Genicot, and D. Ray (2008): "Informal insurance in social networks," *Journal of Economic Theory*, 143, 36–58.

Boucher, V. (2015): "Structural homophily," *International Economic Review*, 56, 235–264.

Breza, E., A. G. Chandrasekhar, T. H. McCormick, and M. Pan (2020a): "Data and Code for: Using aggregate relational data to feasibly identify network structure without network data," .

——— (2020b): "Using aggregated relational data to feasibly identify network structure without network data," *American Economic Review*, 110, 2454–84.

Chandrasekhar, A. and M. O. Jackson (2018): "A Network Formation Model Based on Subgraphs," *SSRN Working Paper: http://ssrn.com/abstract=2660381*.

Christakis, N. A., J. H. Fowler, G. W. Imbens, and K. Kalyanaraman (2010): "An empirical model for strategic network formation," Tech. rep., National Bureau of Economic Research.

Currarini, S. and M. Morelli (2000): "Network formation with sequential demands," *Review of Economic Design*, 5, 229 – 250.

Dutta, B. and S. Mutuswami (1997): "Stable networks," *Journal of economic theory*, 76, 322–344.

Feigenberg, B., E. Field, and R. Pande (2013): "The economic returns to social interaction: Experimental evidence from microfinance," *Review of Economic Studies*, 80, 1459–1483.

Herings, P. J.-J., A. Mauleon, and V. Vannetelbosch (2009): "Farsightedly stable networks," *Games and Economic Behavior*, 67, 526–541.

Jackson, M. O., T. R. Rodriguez-Barraquer, and X. Tan (2012): "Social Capital and Social Quilts: Network Patterns of Favor Exchange," *American Economic Review*, 102, 1857–1897.

Jackson, M. O. and A. Van den Nouweland (2005): "Strongly stable networks," *Games and Economic Behavior*, 51, 420–444.

Jackson, M. O. and A. Watts (2002): "The evolution of social and economic networks," *Journal of Economic Theory*, 106(2), 265–295.

Jackson, M. O. and A. Wolinsky (1996): "A Strategic Model of Social and Economic Networks," *Journal of Economic Theory*, 71, 44–74.

Mele, A. (2017): "A structural model of Dense Network Formation," *Econometrica*, 85, 825–850.

Watts, A. (2001): "A dynamic model of network formation," *Games and Economic Behavior*, 34, 331–341.