Neighborhood composition and *Jati* Homophily: Evidence from Rural India

Pritha Dev, Hari K. Nagarajan, Abhishek Tripathy

Economics Area, Indian Institute of Management Ahmedabad

PRELIMINARY DRAFT

Abstract

This paper disentangles how physical proximity impacts network formation and homophily. We use the Rural Economic and Demographic Survey (REDS) conducted in 2006 which has data on social networks in 242 Indian villages. We have rich data on the *jati* affiliation as well as location of households such that we can identify consequent households situated on a street. We begin our analysis with a network formation model between households from two *jati*'s who first choose location and then links. The model results are empirically tested using three methodologies. We first run a dyadic regression to estimate the impact of *jati*, distance and time of entry into the village. Here we find a greater probability of same *jati* links, lower link probability as distance increases and as time gap between the entry of the two households increases. Next, we consider clusters of same *jati* households and we find that as the household moves closer to the periphery of the cluster, its homophily decreases. Finally, we estimate the causal impact of physical proximity of same *jati* households on *jati*-based homophily by instrumenting physical proximity by the variables related to the age of the household in the village. We find that *jati*-based homophily increases as the number of immediate neighbors from the same *jati* increases.

Keywords: Homophily, Spatial Distance, *Jati*, Instrumental Variables

1 Introduction

This paper explores forces behind household-level link formation and *jati*-based homophily in rural India. We focus on understanding how having immediate neighbors from the same *jati* impacts *jati*-based homphily. Homophily is the tendency of people to link with people who are alike (Burt, 1991; Lazarsfeld, Merton, et al., 1954; McPherson, Smith-Lovin, & Cook, 2001; McPherson, Smith-Lovin, & Rawlings, 2021). It has been shown to have wide ranging impact on behaviour and outcomes by having an impact on how information spreads over the network (Golub & Jackson, 2012; Halberstam & Knight, 2016) and because individuals like to conform to their networks (Bramoullé, Djebbari, & Fortin, 2009; Calvó-Armengol, Patacchini, & Zenou, 2009). We study village networks which are an important source of risk sharing especially in developing countries as highlighted by Alatas, Banerjee, Chandrasekhar, Hanna, and Olken (2016); Ambrus, Mobius, and Szeidl (2014); Bloch, Genicot, and Ray (2008); Jackson, Rodriguez-Barraquer, and Tan (2012); Mobius and Rosenblat (2016) and hence understanding the factors behind their formation is important.

We use data from the 2006 round of the Rural Economic and Demographic Survey (REDS) which allows us to observe physical proximity and also the nature of the social network. We are able to capture proximity exploiting the data gathering technique of REDS. The households were surveyed in order of their presence of the street and the household listing numbers can be used to approximate the physical proximity of the households. We can perfectly identify the immediate left and right neighbors via the household listing numbers. The network data was collected by asking each household to name three households it would reach out to for help with food items or money, in the event of an emergency. We measure homophily of the household by the proportion of same *jati* direct connections (made or received) as a proportion of total connections.

We first present a simple model of network formation where households from two *jati*s arrive in a predetermined order and choose location within a village. In the period of entry, households also choose whom to form links with where the link cost is increasing with physical distance and there is an additional cost of linking with the other jati. We find that in equilibrium households choose location next to other households who entered around the same time. In particular, whenever possible, households choose location next to members of their own *jati* leading to the formation of same *jati* clusters that we call runs. Links are formed with those closest to the household and hence heterophilous links are formed by households living close to households from the other *jati*. The model also predicts that homophily decreases as we move from the center towards the periphery of the run. Finally, the model shows that *jati* size matters and smaller jatis are more likely to be physically disperse and exhibit lower homophily.

In our empirical investigation we present three different ways in which location impacts links formation. We begin with a dyadic regression which explores the impact of increasing distance between a pair on the link formation probability. Here we find that link formation probability reduces with distance and that same *jati* links are more likely at all distances. We also find that the link probability is higher if households enter at around the same time.

Next, we consider the impact of location on homophily by focusing on runs of same *jati* households. In particular, we focus on households who are peripheral in the run and see if homophily reduces as households become more peripheral. We find that this is indeed the case.

Finally, we establish a causal relationship between neighborhood composition and homophily. We measure neighborhood composition as the number of immediate neighbors from the same *jati*. We use an instrumental variable regression since Heine et al. (2021) and Blumenstock and Fratamico (2013) show that homophily in social networks drives the choice of neighborhoods. Using the results from the model, we propose the use of an instrumental variable regression where we instrument the neighborhood by variables related to the age of the household in the village. Following the theoretical model, our novel instrument takes into account the *jati*-composition of the neighborhood at the time of the household's appearance in the village. We include both the relative size of the *jati* as well as its squared term, to capture possible non-linearities. In all the regressions, we also use various control variables as proposed by the literature on homophily in social networks. We find a positive and significant impact of more same *jati* neighbors on homophily.

Our key contribution is to show that spatial proximity with same-*jati* households is a key determinant of *jati*-based homophily in informal social networks in rural India. First, we show that links are made with other households which live in close proxmity to each other. Second, we find that households which are on the periphery of a *jati*-based run and hence have more diverse neighbours are less homophilous than those who are further removed from the periphery. Third, we find that households that have more neighbors of the same *jati* are significantly more homophilous.

Our paper adds to the literature on the relationship between physical distance and homophily. Previous literature linking physical distance and homophily includes Schelling (1971) who show how even a small preference for similar neighbors leads to residential neighborhoods becoming segregated. (Henry, 2011) extend Schelling (1971) to a network setting to show how a small preference for linking to own type leads to a segregated networks. Cutler, Glaeser, and Vigdor (1999); De Martí and Zenou (2008); Patacchini and Zenou (2016) also show how a homophily leads to residential segregation.

There exists a rich theoretical literature which tries to understand homophily in social networks. Wimmer and Lewis (2010), Yoshuke (2019) incorporate homophily in standard models of random network formation. Kets and Sandroni (2019) show how homophily is a result of a coordination game where agents from the same group receive correlated impulses which allow them to better coordinate. Currarini, Jackson, and Pin (2009) show why group size matters in homophily for using dynamic network formation model where agents have bias for meeting their own type. Bramoullé, Currarini, Jackson, Pin, and Rogers (2012) show the emergence of homophily in a dynamic network formation model where meetings are both random and based on network based search which allows for bias for own type. Jackson, Nei, Snowberg, and Yariv (2022) highlight the presence of assortativity of homophily where highly homophilous individuals tend to link with similar individuals

Many papers involving applied work have documented the existence of homophily in networks. Dyadic regressions to understand link formation have clearly shown the importance of homophily. For example, Fafchamps and Gubert (2007) provide a framework for the estimation of the probability of link formation between nodes as a function of the absolute difference in their respective characteristics. The literature offers various measures of homophily and segregation over social networks (Bojanowski & Corten, 2014). As indicated by the theoretical literature, there is a differential impact of the meeting process and homophilic preferences on observed homophily which empirical work has sought to disentangle (Chetty et al., 2022; Currarini et al., 2009; Kossinets & Watts, 2009; Mayer & Puller, 2008; Mosleh, Martel, Eckles, & Rand, 2021). In particular, Mayer and Puller (2008) show via dyadic regressions that common friends are an important driver of friendship formation.

The impact of distance on network formation has also been well documented. The phenomenon of distance decay or relationships being more likely to formed amongst the physically close has been empirically observed in various studies on telecommunication networks (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008; Krings, Calabrese, Ratti, & Blondel, 2009; Xu, Santi, & Ratti, 2022). Social media platforms and other online communities continue to exhibit connections driven by physical distance (Backstrom, Sun, & Marlow, 2010; Goldenberg & Levy, 2009; Liben-Nowell, Novak, Kumar, Raghavan, & Tomkins, 2005). In seminal papers Abu-Ghazzeh (1999); Festinger, Schachter, and Back (1950); Holahan, CJ, BL, MA, and RE (1978) have highlighted the impact of geographic proximity on interaction patterns. For social networks, Nahemow and Lawton (1975), Sacerdote (2001) and Small and Adler (2019) show the importance of physical distance on network formation.

Fafchamps and Gubert (2007) show the role of distance has been seen in dyadic regressions where distance is imputed by whether the two households live in the same cluster and their distance from the closest road. To understand the causal impact of distance on the formation of friendships, Marmaros and Sacerdote (2006) exploit the random assignment into rooms and dorms of students. They show that friendships are most likely to be formed with other students who are live close by and that distance in freshman year continues to impact friendships in later years.

The layout of the paper is as follows. The following Section 2 describes the data in detail and provides summary statistics as well as stylized facts emerging from the data. Section 3 describes the network formation model and its results. Section 4 presents the first empirical result which is the dyadic regression. Section 5 presents the results from a data sub-sample comprising of households which form the periphery of runs to show that the impact of location along a run on homophily. This is followed by Section 6 which includes all households and shows the causal relationship between neighborhood composition and homophily. Section 7 concludes.

2 Data and Summary Statistics

We use data from the 2006 round of the Rural Economic and Demographic Survey (REDS). The 2006 round covers 115,428 households belonging to 242 villages of 17 states in India. Data was captured using three modules: Listing, Household and Village. The Listing module is essentially a census of all the households (115,428). The Listing module contains data on social networks, *jati* of the household and location of the immediate neighbors of a household. In addition, the module also captures detailed information on the age of the household in the village. We also use data from the village module which captures village-level information in terms of the remoteness of the village from important locations and types of establishments in the vicinity, land-use practices, provision of public goods at the street level as well as details on elections for and proceedings in the local *Gram Panchayats*.

2.1 Social Networks

Our dependent variable uses information on social networks from the Listing module. Network data is collected using the following questions: a) *If you wish to borrow Rs. 1000 to meet a family emergency, Identify three households from this village in order of approachability.* and b) *If you wish to borrow simple food items such as chilies, spices, vegetables etc. Identify three households from this village in order of approachability.*. This method is consistent with established practice by following the "name generator" approach wherein the respondents ('ego') are asked to recall and mention other respondents ('alters') with whom they maintain social ties in a given context (Marsden, 2005; McCallister & Fischer, 1978).

We label answers to (a) as representative of the 'Money' networks that the household is a part of, while (b) allows us to observe 'Food' networks. These questions were asked of all the 115,428 households in the sample. In other words, we observe all the nodes and up to three links formed per node for each of the two networks.

The social network data was cleaned to ensure the household numbers listed as links were indeed part of the village. Additionally, some households listed themselves as links or listed the same other household as their link three times. Such cases were removed from the datasets. We find the following reporting patterns in the cleaned data: 92.98% (94.46%) of households report the first link in the Money (Food) network, 52% (66.93%) percent of the households report a second linked household and 20.21% (31.05%) percent of the households report a third link. Related to the reporting pattern, we find a density of 0.003 (0.004) in Money (Food) networks on average - a measure of the average number of links made per village as a proportion of total possible links.[1]

[1] Density is defined as the ratio of the total number of Observed Links with the total number of Possible Links. The low density is commonly seen is social networks and ours is additionally lower since number of links elicited from each household was restricted to a maximum of three.

2.2 *Jati*

The data on each household's *jati* identity was collected to ensure accurate identification. Enumerators were strictly instructed to not accept data on *varna* or *gotra*, both frequently misunderstood as proxies for *jatis*. The raw data on *jati* had to be cleaned to correct for errors of spelling and to ensure the same *jati* was correctly identified across the data set. Additionally, even though enumerators were explained the difference between *varna*, the primary *jati* was sometimes misidentified as the *varna*. In such cases, we used information present on sub-*jati* to correctly identify the *jati*.

After the cleaning process we document 2422 *jatis* present in the dataset overall or 10 *jatis* per village. The average number of households in a village is 1080 and 226 is the average size of each *jati*. We measure fractionalization by *jati* in each village and across all villages find an average fractionalization of 0.73, indicating that the average *jati* diversity is high. Consistent with Munshi and Rosenzweig (2015), we find that 45% of the links are within *jati* on average across all villages while the remaining is across *jatis*.

2.3 Neighborhood

Indian villages often do not have clearly marked streets. Instead of streets, the village might be organized by areas (colloquially called *para, tola, vada* amongst others) wherein each area is being serviced by meandering streets and sub-streets. The data collection was done such that enumerators were instructed to identify streets as those areas which had a clear entry and exit point. More importantly, within each street, houses were numbered consecutively in order of appearance. Further, houses on one side of the street are numbered first followed by the other side. House numbers thus allow us to recognize the left and right neighbors of a house. Therefore, if a household appears with ID n, then we know that it shares a boundary wall on the side with households numbered $n-1$ and $n+1$. Even when enumerators moved from one street to the other, they began with the closest household to the last household on the previous street.

Exploiting this feature of the data, we can proxy the distance between the two households by the difference between their IDs. For runs of same *jati* households we can identify the households at the periphery. Finally, we can correctly identify the two adjacent households for each household. We then measure how many of these immediate neighbors are from the same *jati* as the household. For the one of our regressions, the main independent variable is Nbd_Jati which takes the value $\{0, 1, 2\}$ depending on how many of the two adjacent households have the same *jati* as the focal household. Thus Nbd_Jati takes value 0 if none of the immediate neighbors are of the same *jati* (12% of the sample), it takes value 1 if one of the two immediate neighbors is of the same *jati* (28%) and it takes value 2 if both the two immediate neighbors are of the same *jati* (59%). The state with the highest average value of Nbd_Jati is West Bengal, where households have 1.78 same-*jati* neighbors on an average, while Kerala has the lowest average value at 1.07.

2.4 Stylized Facts

[Figure 1 about here.]

In this section we provide some basic summary statistics related to the social networks we observe and some basic socio-demographic statistics. Figure 1 shows the network graph for two villages where the nodes represent households and the colors represent *jati*. We see varying levels of homophily and also number of *jatis*. Table 1 contains data on socio-demographics at the household and village level while Table 2 contains some basic network statistics at the household and village-levels. The average household in the REDS sample comprises of 5 members, is 46% likely to own land (conditionally owns 1.15 acres of land), has an in-degree of 1.48 in food and 1.88 in money networks. The households, on an average, arrived in the village around the year 1804, suggesting that they are approximately 200 years old.

Villages in the REDS sample are 600 years old on an average and comprise of 1080 households spanning across 15 distinct *jatis*. Fractionalization of 0.73 suggests that the *jati* diversity in the average (and indeed, the median) village remains high. We also note that 52% (35%) of the streets contain public taps (drinking water wells), indicating inter-street variation in the availability of public spaces that facilitate interactions, and therefore spatial inter-mixing.

From Table 2, we note that Food (Money) networks comprise of 59 (83) distinct non-singleton components in an average village. Such components, on an average, span 15 *jatis*, 6 streets and 3 jati-street combinations in the case of food networks, suggesting that a majority of the members in a component are more likely to belong to the same *jati* and street. Similar observations hold for money networks. We note an average clustering coefficient of 0.227 (0.140) at the *jati*-street level for Food (Money) networks[2]. This implies that 22.7% of food network triads observed among households belonging to the same *jati* and street are closed and therefore represent a complete triad. Finally, we note that the diameter - the largest distance between any two nodes in a network graph/ sub-graph is 18 for Food and 21 for Money networks. We also find that restricting the sub-graphs to the same *jati*-street combination yields significantly lower diameter values on an average. Taken together, we note that network formation is chiefly influenced by spatial and *jati*-proximity among the households.

In addition, we note the following features of network formation among the households. First, we note a positive relationship between a household's age in the village and the average age of its immediate neighbors. Figure 2 shows this relationship by plotting household age quintiles against the average age of neighboring households. We note, for example, that households belonging to the highest quintile (that is, households that are 498 years old in their respective villages, on an average) tend to have neighbors that have a mean tenure of 472 years. This is suggestive of a clustering of similarly-aged households in the village that plays a key role in determining the spatial distribution

[2] Clustering measures the ratio of connected triplets among all the possible triplets that have a common node. For example, if Household a is linked to b and c, then the Clustering coefficient for a is the probability that b and c are also linked to each other (Jackson et al., 2008). That is, it is the probability of finding 'closed' triads from the set of all possible triads in an observed network.

of households observed by the survey.

[Figure 2 about here.]

Second, we define a 'run' as a set contiguously located households, all of which belong to the same *jati*. By definition, therefore, households appearing on the periphery of a run (that is, it is situated at the end of a contiguous set of households) will have exactly one neighbor of the same-*jati*. Figure 3 shows that households on the periphery also exhibit significantly lower homophily across both types of networks.

[Figure 3 about here.]

Third, we calculate *jati*-level averages of homophily and neighborhood composition to examine a relationship between the two variables. Figure 4 plots the results for Food and Money networks. The patterns suggest a positive relationship between neighborhood composition and homophily.

[Figure 4 about here.]

Fourth, we note that *jatis* that form a smaller share of a particular village's population are more diverse in terms of outcomes related to spatial distribution as measured by the average number of same-*jati* neighbors for a household in the *jati*. This relationship has been plotted in Figure 5. Smaller *jati*s can either have on average very few neighbours from the same *jati* or many. In other words, households from smaller *jati*s are either isolated from other members of the *jati* or they all live in close proximity. The pattern clearly suggests a positive correlation between the two variables. Note that part of the correlation is mechanical since as the *jati* becomes large, neighbours are bound to be from the same *jati*. The graph shows that the spatial distribution of households is determined by the relative size of their *jati* within the village.

[Figure 5 about here.]

Finally, Figure 6 shows that the relationship between demographic dominance of a *jati* and its neighborhood composition translates into greater homophily exhibited by households. We note a positive correlation among the two variables. As in the case above, note that part of this correlation is mechanical. Again, it is interesting to note that homophily is more variable for smaller *jati*s.

[Figure 6 about here.]

[Table 1 about here.]

[Table 2 about here.]

3 Network Formation Model with Location Choice

In this section, we consider a simple model of network formation to explain the stylized facts observed in the section above. In this model, the players are the households in a village. There are n households and we consider a network formation model where all links are made within the village households. Each household has two decisions to take - the first regards choice of location and the second one regards the links they choose to form with other households. The only source of heterogeneity within the households is that they belong to one of two jatis J_1, J_2. We assume that the payoffs are derived only from the network formation game where the costs of link formation are driven by the jati identity of the two linked households and the distance between them determined by their location choice.

The game is played over n rounds. Each round, a household is arrives at the village. For simplicity, we assume that household h arrives in period h. Household i's jati is denoted by $j_i \in \{J_1, J_2\}$ such that jati J_1 forms a fraction p of the population. Wlog we assume that $p \leq 1/2$. We also assume that the order of arrival of households is known.

The new household first chooses their location from those remaining. We assume there are as more locations than households and the village locations are represented along a line. Location is indexed by $d \in \{0, m\}$ where $m \geq n$. If households i, j choose locations d_i, d_j then the distance between the locations is given by $d_{ij} = |d_i - d_j|$.

Each entering household also announces which links they want to form with the set of households currently in the village. If the link is accepted, both households bear the cost. The links proposed by household i is denoted by $g_i = \{g_{ik}\}_{k \neq i}$ where $g_{ik} = 1$ indicates that i forms a link with k and $g_{ik} = 0$ indicates that i does not form a link with k. The cost of link formation depends on the distance between the two households as well as their jati with lower costs for same jati links. In the period of their entry, households derive value from the links they form but in the future periods they derive value from the links they made and receive. The model does not allow for older households to update their link offers given the new entrant household except for accepting or rejecting links offers by this household. We assume that houeseholds live on forever after entry. The household i's realised utility is then:

$$u_i(g) = \sum_{t \geq i} \delta^{t-i} \left(\sum_{k<i} g_{ik} - \sum_{k<i} g_{ik} c(d_{ik}, j_i, j_k) + \sum_{l>i}^{l \leq t} g_{li} - \sum_{l>i}^{l \leq t} g_{li} c(d_{li}, j_l, j_i) \right) \tag{1}$$

The household discounts the future at the rate δ. At the period of entry, the household makes links with households who entered before ($k < i$) and in the future receives links from households who enter later ($l > i$).

For an entering household, it makes links today and estimates the links it will receive in the future and its expected utility is as follows:

$$E(u_i(g))|_{t=i} = \sum_{t \geq i} \delta^{t-i} \left(\sum_{k<i} g_{ik} - \sum_{k<i} g_{ik} c(d_{ik}, j_i, j_k) + E\left(\sum_{l>i}^{l \leq t} g_{li} - \sum_{l>i}^{l \leq t} g_{li} c(d_{li}, j_l, j_i) \right) \right) \tag{2}$$

Note that at the period of entry, the household can only estimate how many links it recieves while it knows how many links it makes. They also know where locations are currently available and they can estimate how households will settle.

A1 We assume that cost function takes the following form:

$$c(d_{ik}, j_i, j_k) = \kappa * d_{ki} + \gamma * I_{j_i \neq j_k}$$

Note that κ is the incremental cost of linking with a household based on the physical distance while γ is the additonal cost of linking to someone from a different jati. Note that the lowest cost of link is with a person of the same jati living right next a household and this is given by κ.

Given the structure of the utility, each link is evaluated individually and thus the only factors determining link formation are distance and jati-affiliation. There will be maximum distance such that same jati link are formed only if distance is below that maximum and another such maximum will exist for the other jati.

We define a **run** as the set of households from the same jati located consequetively. The run is taken to be the set such that no superset is also a run.

We evaluate the subgame perfect equilibrium outcomes in two cases: one where households are impatient ($\delta \to 0$) and the other where households are patient ($\delta \to 1$). The proofs for both of the following propositions is in the Appendix. Here, we provide the intuition behind the proof. In both cases, note that the last household to enter will prefer to locate close to a household of the same *jati*. Additionally, in case links can be formed, entering households will find available locations next to households who arrived in the recent past and choose from amongst these locations. Thus, households who arrived at around the same time will locate next to each other and also form links with each other. Also note that if the cost of links are such that either no links can be formed, only homophilous links can be formed or all links can be formed. In case no links can be formed, households will choose locations at random from those available since their utility from all locations is zero.

In the case of impatient households and $\delta \to 0$, note that the households while making their location choice only care about the links that they form and not about the links they might receive in the future. In this case, if a household enters at a time when it cannot form any links, it is indifferent with respect to choice of location. Next, consider the case where only homophilous links are possible. Here, the first entrant of each *jati* can form no links and chooses a location at random. Households which enter later will always choose a location closest to the largest run of their own jati. Multiple runs of each jati can emerge if the random choice of location by the first entering households of the two *jati*s are such that one of the *jati*s cannot be fully accomodated next to the first entering household. Note that this location pattern may give rise to multiple components within a *jati* but each component contains members of a single *jati*. Next, if all links are possible, only the first entering household can form no links and chooses a location at random. Following households can always form a homophilous or heterophilous link and they choose a location next to the existing households. Households will still prefer to locate next to own *jati* households whenever feasible since those links are cheaper and the location pattern will exhibit runs.

Proposition 1 *The subgame perfect equilibrium network and spatial structure when households are impatient and $\delta \to 0$ is as follows:*

- *If only homophilous links are possible i.e. $\kappa \leq 1$ but $\kappa + \gamma > 1$ then the resulting network has components where each component has members from the same jati and no mixed-jati components exist. Additionally, the the two jatis will be spatially segregated with there being few runs of each jati.*

- *If heterophilous links are possible i.e. $\kappa + \gamma \leq 1$ then the network is connected. Multiple runs of the jatis are likely to emerge.*

In the case of patient households and $\delta \to 1$, households do care about the links they will receive in the future. When only homophilous links are possible, the first entering household of each jati will thus choose location such that there are enough empty locations left to allow the entire jati to settle next to each other. The resulting network is segregated with each jati forming a component. If heterophilous links are possible, then the resulting network is connected. In this case also, households from the same *jati* prefer to locate next to each other and form homophilous links which are cheaper. The exception to this is if the size of *jati* is very small and there are large gaps in the entry of members. In this case, the first entrant has to choose between choosing a location far from the other *jati* forgoing current heterophilous links for homophilous links in the future. If the wait time is long and the additional cost of heterophilous links is low, the agent might choose to locate next to the other *jati* and forgo the possibility of locating and links with own *jati* household.

Proposition 2 *The subgame perfect equilibrium network and spatial structure when households are patient and $\delta \to 1$ is as follows:*

- *If only homophilous links are possible i.e. $\kappa \leq 1$ but $\kappa + \gamma > 1$ then the resulting network has components where each component has members from the same jati and no mixed-jati components exist. Additionally, the two jatis are spatially segregated with there always being exactly one run in each jati.*

- *If heterophilous links are possible i.e. $\kappa + \gamma \leq 1$ then the network is connected. The spatial structure and exact network topology depend on the size of the jatis and order of arrival. If one of the jatis is very small and arrival time of its members is disperse, households from the smaller jati are located in between runs of the larger jati and form heterophilous links. When jatis are of equal size, the spatial structure will be a single runs of each jati with heterophilous links between closely located households from different jatis.*

The two propositions above give the following insights related to network and structural patterns which align with the stylised facts presented earlier and can be rigourously tested with data:

- Households choose location close to others who have arrived at around the same time in the village. Links are also formed with those who arrived at around the same time.

- Runs of consecutive households from the same jati arise since in many cases, households choose locations close to others from the same jati who entered the village at around the same time. The households who are peripheral in these runs are more likely to form heterophilous links in contrast to those who are more central.
- The composition of the immediate neighborhood of a household impacts the overall level of homophily of a household.
- The size of the jati will have an impact on the level of homophily in households of that jati.

4 Household Link Formation Decisions

We begin our empirical analysis by examining the factors that determine the probability of the existence of a link between two households of the same village. We are particularly interested in the impact of distance between households, their *jati* identity and the time of their entry on link formation. We measure the distance as the difference between their household IDs and it is important to note this only allows us to correctly identify close by neighbors. We do note that we do not observe the full network since households were asked to report upto three households they link to.

We estimate the following logistic regression specification:

$$\begin{aligned} Link_{ij} &= \gamma_0 + \sum_{d=1}^{5} \gamma_{1d} Dist_{ij,d} + \gamma_2 SameJati_{ij} \\ +\gamma_3 \Big| X_i - X_j \Big| &+ +\gamma_4 (X_i - X_j) + VillageFE + \epsilon_{ij} \end{aligned} \tag{3}$$

where the dependent variable $Link_{ij}$ equals 1 if households i and j are linked in the network under consideration which is either the Money (M) or the Food (F) networks. Our main independent variables are $Dist_{ij,d}$ which are indicator variables taking value

1 if the distance between the households is d and 0 otherwise. Our other independent variable of interest is $SameJati_{ij}$ which is an indicator variable taking value 1 if the two households belong to the same jati and 0 otherwise. Our control variables are included in $|X_i - X_j|$ which represents the absolute difference in the characteristics of the two households. An important control measures the difference in arrival times to the village. Other differences are measured over a variety of characteristics such as household size, income, position in the village land-ownership distribution, education levels and age of the household head. Finally, we control for unobservable factors influencing the structure of networks (arising out of dyad formation) in the form of a vector of Village dummies. For example, extant social norms and vulnerability to covariate and idiosyncratic risks might significantly impact the likelihood of dyad formation between households. Such effects can be considered to be subsumed through the usage of village dummies. We use standard errors corrected for cross-observation correlation across error terms, consistent with Cameron, Gelbach, and Miller (2011) and Cameron and Miller (2015).

We now present results from the estimation of dyadic regressions. Given the high number of dyads possible in the entire dataset, we only show results for the most populous state Uttar Pradesh in order to provide a general interpretation of the various features that determine network formation.[3] Table 3 shows results for food networks, wherein we observe from Col. (1) that the probability of link formation declines monotonously with the distance between households.

[Table 3 about here.]

Table 3 shows that older households are less likely to form links with newer households (and vice versa) in the case of both food and money networks, indicating that *jati*-based homophily might be linked to the age of the household in the village. We also find that a higher difference in household positions on the village-level land ownership

[3] Results for other states can be made available upon request.

distribution also reduces the probability of link formation. In other words, households who own more land are less likely to form links with those that do not (and vice versa).

Using the previous results, we present a graphical representation of how the probability of link formation declines with an increase in distance between households. Results shown graphically in Fig. 7 for food and money networks respectively suggest that the probability of link formation between two households is close to zero (yet statistically significant) if they are located at a distance of 5 units from each other.[4]

[Figure 7 about here.]

We next investigate if *jati*-compatibility leads to heterogeneity in the declining dependence on distance as observed previously. Results from segregating the data by jati have been presented in Table 4. We notice again that link formation between older and newer households is less likely, irrespective of their *jati* compatibility. However, the reduction in probability is significantly lower when households belong to the same *jati*.

Figure 8 shows results in graphical form, by plotting the marginal effects of distance between households of a dyad on the probability of dyad formation. In both panels, we note that while there is consistency in the declining dependence on distance (as was seen in Figures 4 and 5), the probability figures are albeit statistically higher for same *jati* households. This implies that when two households are situated at a distance of, say, 3 units the probability of link formation for both network types is higher when they belong to the same *jatis*.[5]

[Table 4 about here.]

[Figure 8 about here.]

We find that the probability of link formation decreases with the distance between households in a non-linear fashion. This relationship is robust to the inclusion of other

[4] This trend is robust to the restriction of distance to 7, 9 and 11 units.

[5] We also note that the probability of link formation between same-*jati* households is always significantly higher for food networks than for money networks. A possible reason for this is the deeper interaction of food items with potential taboos and restrictions due to norms that govern the identity of the household from where food could be borrowed.

controls as well as village dummies. Households of the same *jati* are more likely to form a link - a preliminary evidence of *jati*-based homophily.

5 Behaviour of Households along a *Run*

In this section, we estimate a homophily regression where we restrict attentions to households that lie on the periphery of runs. Given the methodology used by the enumerators for covering households in the listing module, we are able to accurately determine the order in which households are located on a particular street in the village. The model highlights that households that lie on the periphery of a run might behave differently as compared to households that are less peripheral.

Consider 10 consecutive households on in a village. For simplicity, let these households belong to 2 different *jatis* A and B. Let the arrangement of households be the following: $A_1, A_2, A_3, A_4, A_5, A_6, B_1, B_2, A_7, B_3$. We are interested in uncovering heterogeneity in link formation behaviour and homophily for the A_5 and A_6 households.

In order to uncover such heterogeneous effects, we select data in the following manner. We first retain all data pertaining to run sizes of at least 9 – that is, all households that are part of a run that contains less than 9 are eliminated.[6] This enables us to identify the 5 peripheral households in any remaining run and we narrow our data to include only such households to see if their homophily levels differ. If link formation in our sample is consistent with theory, we should be able to see a monotonic relationship between how peripheral a household is and its homophily.

Our dependent variable is $Homophily_{hvjs}$ which measures the proportion of total links of the household h with other households of the village v belonging to the same *jati* j with respect to either the food (F) or money (M) network. Our regression

[6] Our results and qualitative conclusions are robust to the choice of other thresholds as well, such as 5, 7, 11, 15. Results have been provided in the Appendix.

specification takes the following form:

$$Homophily_{hjv} = \beta_0 + \sum_d \beta_{1d} RunPosition^d_{hjv} + X_H\Gamma + X_J\Omega + X_V\Theta + \epsilon_{hjv} \qquad (4)$$

where d takes values from 1 to 5 and $RunPosition^d_{hjv}$ is a discrete variable that takes the value 1 for a household taking the d-th position on the periphery of a run and 0 otherwise. If $RunPosition^1_{hjv} = 1$ then the household is the peripheral household while if $RunPosition^2_{hjv} = 1$ then the household is once removed from the periphery. After estimating β_{1d}, we run pair-wise t-tests to infer the statistical difference between the effect sizes for households in across different positions. Our control variables in the regression are X_H and X_V which are vectors of various household and village-level controls, respectively.

[Table 5 about here.]

The results presented in Table 5 suggest that as the household move away from the periphery of a run, they become more homophilous as compared to those closer to the periphery.

6 Homophily by *Jati*

We first outline the econometric specification for identifying the impact of the *jati* composition of a household's immediate neighborhood on *jati*-homophily displayed by them in both money and food-related networks. In general, the following regression provides us with a starting point in investigating the relationship between neighborhood composition and *jati*-homophily:

$$Homophily_{hjv} = \beta_0 + \beta_1 Nbd\_Jati_{hjv} + X_H\Gamma + X_J\Omega + X_V\Theta + \epsilon_{hjv} \qquad (5)$$

As before, the dependent variable $Homophily_{hvjs}$ measures the proportion of total links of the household h with other households of the village v belonging to the same *jati* j

with respect to either the food (F) or money (M) network. The main dependent variable is Nbd_Jati_{hvj} which measures the number of immediate neighbors of household h that belong to the same *jati* j as the household. The other variables in the regression are X_H and X_V which are vectors of various household and village-level controls, respectively. We include the level and squared terms of the population share of the household's *jati* in the village as a control as indicated by the model. Some of the other important variables included in these vectors are fractionalization index for the village [7]. In addition, we also include average village income calculated by netting out the income of the household under consideration at the village level (net-of-own income), the number of agricultural equipment owned, the household's position in the within-village land ownership and income distribution (measured in terms of their z-scores) as well as standard demographic variables such as the size of the household, dependency ratio, a dummy for whether the household is male-headed and education level of the head of the household.

6.1 Identification

Our coefficient of interest, β_1, will provide an unbiased point estimate of the impact of neighborhood *jati* composition on the proportion of *jati*-based homophily in links formed and recieved by the household, provided the *jati*-composition of neighboring households is orthogonal to other factors that might determine homophilous behaviour. However, a key threat to our identification comes from the non-random distribution of households across the village space as seen in our model where households tend to form homogeneous neighborhoods keeping in mind the links they plan to form.

To address such endogeneity concerns, we employ an Instrumental Variables (IV)-based regression framework by identifying instruments that are not only orthogonal to the other determinants of neighborhood formation (*exogeneity*), but also significantly

[7] Fractionalization is calculated as $1 - \sum_{j=1}^{J} s_j^2$ where s_j is *jati* j's share of the village population. Fractionalization is also expressed as $1 - HHI$ where HHI is the Herfindahl-Hirschmann Index - a measure of concentration.

predict it independently (*relevance*). Finally, our choice of instruments must be such that they close the 'back-door' effect on the outcome variable (homophily), which implies that the causal route runs exclusively through neighborhood formation (*exclusion restriction*).

Using the intuition from our model, we use as instruments various variables which capture the age at which the focal household and other households entered the village. The REDS dataset captures information on the various historical aspects of village formation. We are able to identify when a village was formed and when a particular household appeared in the village. Since we have already noted that the structure of neighborhoods observed in the survey is a result of variations in such factors, defining instruments based on historical data provide us with exogenous variation in the *jati*-composition of neighborhoods in such a way that the variation is orthogonal to observed household characteristics.

We now turn to an explanation of the definition of the instruments. Given accurate data on the age of a particular household, we are able to back out the year of their arrival in the village. This enables us to identify the number of same-*jati* households living on the same street as the household in question *at the time of its arrival in the village*. As pointed out in the model, the number of such households reduces the cost for the newly arrived household to choose the same neighborhood. Such an action, over time, produces neighborhoods that tend to spatially cluster along the *jati* dimension (Bharathi, Malghan, & Rahman, 2023; Munshi & Rosenzweig, 2015).

Therefore, we define the first instrument as the log of the number of same-*jati* households living on the same street at the time of the arrival of the household in question. In addition, we recognize the role of the age of the village in determining the extent of spatial inter-mixing of neighborhoods. We expect both instruments to *positively* influence the composition of a neighborhood in the sense that a greater number of extant same-*jati* households and an older village encourage the formation of *jati*-based neighborhoods.

The IV-2SLS specification might still not lead to an alleviation of concerns over *jati*-level characteristics - both observable or otherwise - potentially biasing homophilous behavior, *independent* of neighborhood composition. For example, Dev, Mberu, and Pongou (2016) suggest that although "*...[ethnic] groups are exogenous, adherence to group values is endogenous...*" (p. 652-53), which in our context implies the role of specific *jati* characteristics significantly determining link formation decisions. In other words, specific *jatis* might invest more time in forming and maintaining links owing to a set of *jati*-specific characteristics, most of which are un-observable. This could lead to spurious point estimates of our main coefficient of interest, β_1, owing to it capturing the joint effect of neighborhood composition and such un-observable factors. Therefore, we account for such variation by including dummy variables for each *jati* in the above regression[8]. Further, in order to account for the effect of various legal and constitutional aspects of the categorization of specific *jatis* into caste categories, we also include state dummies in the above regression. These dummies subsume within themselves all the observable and un-observable characteristics that might influence the structure of the network, thereby allowing us to isolate the true impact of neighborhood composition on homophilous behavior. Therefore, our IV-2SLS specification can be shown in the following form below:

$$Nbd\_Jati_{hjv} = \pi_0 + \pi_1 SameJatiHH_{hjv} + \pi_2 AgeVillage_v + \epsilon_{1,hjv} \tag{6}$$

$$Homophily^k_{hjvs} = \beta_0 + \beta_1 \widehat{Nbd\_Jati}_{hjvs} + X_H\Gamma + X_J\Omega + X_V\Theta + \alpha_s + \gamma_j + \epsilon_{2,hjvs} \tag{7}$$

where state-level unobserved heterogeneity is controlled for by including α_s. Similarly, we control for *jati*-level heterogeneity by including γ_j. All other variables carry the same meaning as in Equation 5. Having defined the exogenous instruments as described

[8] The alloment of unique codes to each *jati* in the sample was a non-trivial exercise. We allot a unique code for each *jati* after cleaning and standardizing the names as they appear in the dataset. The codes have been allotted at the state-level. For example, Yadavs in Uttar Pradesh receive a code that is distinct from those received by Yadavs in Madhya Pradesh. We do this to capture state-specific *jati* characteristics. This, in turn enables us to control for unobservable jati-state specific characteristics that might bias homopilous behaviour by providing varying conditions for same-*jati* neighborhood formation.

above, Equation 6 enables us to isolate the variation in neighborhood composition as a result of the variation in *jati* composition of streets prior to the migration of a household. $\widehat{\beta_1}$ then represents the impact of predicted neighborhood composition (observed in 2006) on households' homophilous behavior in Money and Food networks, providing a causal effect.

6.2 Homophily Regressions Results

We begin by discussing the results from our main IV-2SLS specification. The results have been shown in Table 6 below. Cols. 1-3 contains coefficient estimates for food networks followed by Cols. 4-6 for money networks. Col. 1/4 show estimates for the OLS specification, Col. 2/5 shows first-stage estimates while Col. 3/6 shows the final estimates for the second-stage regression of the IV-2SLS system.

[Table 6 about here.]

We note from Col. 2 that the exogenous instruments return significant positive coefficients in the first stage. That is, for a 1% increase in the number of same-*jati* households that lived on the same street as the arriving household, the number of same-*jati* neighbors increases by 0.138 (over a base of 1.48). Similarly, a 1% increase in the age of the village also increases the number of same-*jati* neighbors of a household, indicating conformance with existing results that suggest an 'ossification' of neighborhoods along the *jati* dimension with the passage of time. The observations hold for Money networks as well with minor changes due to changes in the observations.

Our interpretations depend, crucially, on the instruments satisfying certain statistical requirements. In support of the relevance criterion, we note that the first-stage F-statistic is greater than 10 for estimations related to both food and monetary networks, thereby allowing us to reject the null of weak identification of endogenous neighborhood preferences by the exogenous instruments. With a Hansen statistic of 10.76 (7.40) for Food (Money) networks and associated p-values of 0.215 (0.494), we fail to reject the null of no over-identification of the first-stage outcome variable, in effect strengthening the statistical evidence in defense of our choice of exogenous instruments.

Cols. 3 and 6 show results from the second stage regression for food and money networks, respectively. We note that, after endogenizing neighborhood composition with the help of the exogenous instruments, we obtain a Localized Average Treatment Affect (LATE) of 0.256 (0.260) for food (money) networks, both of which are significant. This implies that an additional same-*jati* neighbor leads to the household maintaining 25.6% (26%) more links in the same *jati* in money (food) networks. Based on this evidence, it is clear that neighborhood composition plays a crucial role in determining homophilous network formation by rural households.

It is important to point out that these estimates retain significance even after accounting for unobservable characteristics in the form of *jati* and state dummies, thereby enabling us to highlight the important role of neighborhood composition in determining *jati*-based homophily in social networks in rural India, *independent of such characteristics*. Further, we find that the share of the *jati* in the village population also plays a significant role in determining homophily. In particular, for a 1 percentage point in the population share, the household will maintain 57% (59%) additional links of the same-*jati* in food (money) networks. We however fail to find evidence of non-linearity in the dependence of homophily on the *jati* population fraction, owing to insignificant coefficients on the squared term.

6.3 Robustness Checks

We undertake a series of robustness checks in order to check for the sensitivity of our main causal estimates due to changes in the definitions of key variables and variations in the methods of estimation.

Changing the neighborhood

We first consider a different definition of same-*jati* neighborhood by considering the proportion of same-*jati* households living in the same street as the focal household. This allows us to expand the definition of a household's neighborhood to include same-*jati* households on the street. In other words, we change the composition of neighbors from being located in the 'immediate' vicinity to being located on the same street. We address endogeneity concerns by estimating an IV-2SLS system of equations using the same set of instruments as earlier. Results from the estimation have been presented in Table 7. We notice that a higher number of same-*jati* households on the same street leads to greater homophily in both types of networks. Our key finding of homophily being significantly determined by spatial neighborhood composition is therefore robust to varying definitions of the 'neighborhood' itself.

[Table 7 about here.]

Neighborhoods and Split households

We next carry out a modification of the main homophily equation by accounting for neighboring same-*jati* households belong to the same parent household. We use household listing IDs from the 1999 round of the REDS to identify whether two neighboring households observed in 2006 were split off the same parent household in 1999. Typically, households split along patrilineal lines and therefore, the split-off household belongs to the same *jati* as the parent household.

Therefore, we modify our variable of interest *nbd_jati* by netting out those households that belong to the same *jati*, are neighbors and share the same parent household as per the 1999 survey. Our new variable *nbd_jati_mod* is defined after re-classifying 3590 (out of a total of 100,244) households.

For example, consider a set of 10 households: $A_1^1, A_2^2, A_3^3, A_4^3, B_5^4, A_6^5, A_7^6, A_8^6, A_9^6, A_{10}^7$. The subscripts denote the serial number of the household (from 1 to 10) in the current survey round and the superscript denotes the serial number of the household in the previous survey round. As before, households belong to two possible *jatis* - A and B. Note that household A_3^3 has two neighbors of the same *jati*, one of which, however, belonged to the same parent household (3). Hence, for A_3^3, we redefine *nbd_jati* by netting out A_4^3, implying that *nbd_jati* reduces from 2 to 1. Using this approach, *nbd_jati* for A_8^6 will reduce from 2 to 0. Such a re-classification allows us to obtain a definition of same-*jati* neighborhoods net of those who belonged to the same parent household and hence, by construction, belong to the same *jati*. This addresses issues related to the potential overstating of *nbd_jati* used in our main specification (Eq. 2). Table 8 summarizes the number of households that were re-classified using this approach.

[Table 8 about here.]

IV-2SLS estimates presented in Table 9 show that our main causal estimates remain stable to the re-classification of the endogenous variable. In particular, after controlling for unobservables at the state and *jati* levels, we find that neighborhood composition matters in determining homophilous behavior across both food and money networks.

[Table 9 about here.]

Control Function approach

As noted earlier, our key endogenous variable *nbd_jati* can take three values - 0, 1 and 2. Hence, estimating the first stage as an OLS might lead to an assumption of the true error structure, leading to biased estimates across both the stages. We therefore, opt for the control function approach wherein we estimate the first stage equation separately from the second stage by using Poisson regressions. We obtain the residuals of the endogenous variable and plug it into the second stage equation.

Results have been shown in Table 10. We only show the second stage results for the sake of brevity. In particular, we note that neighborhood composition still significantly determines homophilous behavior.

[Table 10 about here.]

7 Discussion and Conclusion

This paper highlights the importance of neighbourhoods in determining the structure of networks maintained by households. It showed a game theoretic model of network formation where households from different *jatis* enter a village to first choose location and then form links with other households. We then tested for the results of the model using three different empirical techniques allowing for locations to be interpreted in three different ways, in the process finding that our main findings are robust to such changes. In the dyadic regression, we consider location by measuring the distance between households in a dyad and find that link formation probability falls with distance. In the run regression, we considered location with reference to the periphery of the run and find that moving away from the periphery increases homophily. Finally, we ran a household level regression with homophily as the dependent variable and the *jati* identity of the immediate neighbours as the main independent variable. We addressed causality by instrumenting the endogenous composition of neighborhoods with exogenous variables related to the age of the household in a village. Again, we found that location as measured by immediate neighbours matters in explaining homophily with more same *jati* neighbours leading to more homophily.

References

Abu-Ghazzeh, T. M. (1999). Housing layout, social interaction, and the place of contact in abu-nuseir, jordan. *Journal of environmental psychology*, *19*(1), 41–73.

Alatas, V., Banerjee, A., Chandrasekhar, A. G., Hanna, R., & Olken, B. A. (2016). Network structure and the aggregation of information: Theory and evidence from indonesia. *American Economic Review*, *106*(7), 1663–1704.

Ambrus, A., Mobius, M., & Szeidl, A. (2014). Consumption risk-sharing in social networks. *American Economic Review*, *104*(1), 149–82.

Backstrom, L., Sun, E., & Marlow, C. (2010). Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on world wide web* (pp. 61–70).

Bharathi, N., Malghan, D., & Rahman, A. (2023). Ethnic diversity and economic development with spatial segregation. *Economics Letters*, *222*, 110951.

Bloch, F., Genicot, G., & Ray, D. (2008). Informal insurance in social networks. *Journal of Economic Theory*, *143*(1), 36–58.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *2008*(10), P10008.

Blumenstock, J., & Fratamico, L. (2013). Social and spatial ethnic segregation: a framework for analyzing segregation with large-scale spatial network data. In *Proceedings of the 4th annual symposium on computing for development* (pp. 1–10).

Bojanowski, M., & Corten, R. (2014). Measuring segregation in social networks. *Social networks*, *39*, 14–32.

Bramoullé, Y., Currarini, S., Jackson, M. O., Pin, P., & Rogers, B. W. (2012). Homophily and long-run integration in social networks. *Journal of Economic Theory*, *147*(5), 1754–1786.

Bramoullé, Y., Djebbari, H., & Fortin, B. (2009). Identification of peer effects through social networks. *Journal of econometrics*, *150*(1), 41–55.

Burt, R. S. (1991). Measuring age as a structural concept. *Social Networks*, *13*(1), 1–34.

Calvó-Armengol, A., Patacchini, E., & Zenou, Y. (2009). Peer effects and social networks in education. *The review of economic studies*, *76*(4), 1239–1267.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, *29*(2), 238–249.

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of human resources*, *50*(2), 317–372.

Chetty, R., Jackson, M. O., Kuchler, T., Stroebel, J., Hendren, N., Fluegge, R. B., ... others (2022). Social capital ii: determinants of economic connectedness. *Nature*, *608*(7921), 122–134.

Currarini, S., Jackson, M. O., & Pin, P. (2009). An economic model of friendship: Homophily, minorities, and segregation. *Econometrica*, *77*(4), 1003–1045.

Cutler, D. M., Glaeser, E. L., & Vigdor, J. L. (1999). The rise and decline of the american ghetto. *Journal of political economy*, *107*(3), 455–506.

De Martí, J., & Zenou, Y. (2008). *Friendship formation, oppositional identity, and segregation* (Tech. Rep.). Citeseer.

Dev, P., Mberu, B. U., & Pongou, R. (2016). Ethnic inequality: Theory and evidence from formal education in nigeria. *Economic Development and Cultural Change*, *64*(4), 603–660.

Fafchamps, M., & Gubert, F. (2007). The formation of risk sharing networks. *Journal of development Economics*, *83*(2), 326–350.

Festinger, L., Schachter, S., & Back, K. (1950). Social pressures in informal groups; a study of human factors in housing.

Goldenberg, J., & Levy, M. (2009). Distance is not dead: Social interaction and geographical distance in the internet era. *arXiv preprint arXiv:0906.3202*.

Golub, B., & Jackson, M. O. (2012). How homophily affects the speed of learning and best-response dynamics. *The Quarterly Journal of Economics*, *127*(3), 1287–1338.

Halberstam, Y., & Knight, B. (2016). Homophily, group size, and the diffusion of political information in social networks: Evidence from twitter. *Journal of public economics*, *143*, 73–88.

Heine, C., Marquez, C., Santi, P., Sundberg, M., Nordfors, M., & Ratti, C. (2021). Analysis of mobility homophily in stockholm based on social network data. *PloS one*, *16*(3), e0247996.

Henry, A. D. (2011). Ideology, power, and the structure of policy networks. *Policy Studies Journal*, *39*(3), 361–383.

Holahan, C., CJ, H., BL, W., MA, B., & RE, C. (1978). Social satisfaction and friendship formation as a function of floor level in high-rise student housing.

Jackson, M. O., Nei, S., Snowberg, E., & Yariv, L. (2022). The evolution of networks and homophily. *Available at SSRN*.

Jackson, M. O., et al. (2008). *Social and economic networks* (Vol. 3). Princeton university press Princeton.

Jackson, M. O., Rodriguez-Barraquer, T., & Tan, X. (2012). Social capital and social quilts: Network patterns of favor exchange. *American Economic Review*, *102*(5), 1857–97.

Kets, W., & Sandroni, A. (2019). A belief-based theory of homophily. *Games and Economic Behavior*, *115*, 410–435.

Kossinets, G., & Watts, D. J. (2009). Origins of homophily in an evolving social network. *American journal of sociology*, *115*(2), 405–450.

Krings, G., Calabrese, F., Ratti, C., & Blondel, V. D. (2009). Scaling behaviors in the communication network between cities. In *2009 international conference on computational science and engineering* (Vol. 4, pp. 936–939).

Lazarsfeld, P. F., Merton, R. K., et al. (1954). Friendship as a social process: A substantive and methodological analysis. *Freedom and control in modern society*, *18*(1), 18–66.

Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., & Tomkins, A. (2005). Geographic routing in social networks. *Proceedings of the National Academy of Sciences*, *102*(33), 11623–11628.

Marmaros, D., & Sacerdote, B. (2006). How do friendships form? *The Quarterly Journal of Economics*, *121*(1), 79–119.

Marsden, P. V. (2005). Recent developments in network measurement. *Models and methods in social network analysis*, *8*, 30.

Mayer, A., & Puller, S. L. (2008). The old boy (and girl) network: Social network formation on university campuses. *Journal of public economics*, *92*(1-2), 329–347.

McCallister, L., & Fischer, C. S. (1978). A procedure for surveying personal networks. *Sociological Methods & Research*, *7*(2), 131–148.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, 415–444.

McPherson, M., Smith-Lovin, L., & Rawlings, C. (2021). The enormous flock of homophily researchers: Assessing and promoting a research agenda. *Personal Networks: Classic Readings and New Directions in Egocentric Analysis*, *51*, 459.

Mobius, M., & Rosenblat, T. (2016). Informal transfers in social networks.

Mosleh, M., Martel, C., Eckles, D., & Rand, D. G. (2021). Shared partisanship dramatically increases social tie formation in a twitter field experiment. *Proceedings of the National Academy of Sciences*, *118*(7), e2022761118.

Munshi, K., & Rosenzweig, M. (2015). *Insiders and outsiders: local ethnic politics and public goods provision* (Tech. Rep.). National Bureau of Economic Research.

Nahemow, L., & Lawton, M. P. (1975). Similarity and propinquity in friendship formation. *Journal of Personality and Social Psychology*, *32*(2), 205.

Patacchini, E., & Zenou, Y. (2016). Social networks and parental behavior in the intergenerational transmission of religion. *Quantitative Economics*, *7*(3), 969–

995.

Sacerdote, B. (2001). Peer effects with random assignment: Results for dartmouth roommates. *The Quarterly journal of economics*, *116*(2), 681–704.

Schelling, T. C. (1971). Dynamic models of segregation. *Journal of mathematical sociology*, *1*(2), 143–186.

Small, M. L., & Adler, L. (2019). The role of space in the formation of social ties. *Annual Review of Sociology*, *45*, 111–132.

Wimmer, A., & Lewis, K. (2010). Beyond and below racial homophily: Erg models of a friendship network documented on facebook. *American journal of sociology*, *116*(2), 583–642.

Xu, Y., Santi, P., & Ratti, C. (2022). Beyond distance decay: Discover homophily in spatially embedded social networks. *Annals of the American Association of Geographers*, *112*(2), 505–521.

Table 1: Summary Statistics

Variables	**Mean**	**CV=$\frac{\sigma}{\mu}$**	**Median**
Household level statistics			
Household Size	5.21	0.54	5
nbd_jati	1.48	0.47	2
In-degree (Food)	1.88	1.62	1
In-degree (Money)	1.62	2.20	0
Household Income (Rs.)	52,741.23	1.69	32,500
Mean education level	4.71	.305	4.43
Household owns land (=1)	0.46	1.08	-
Land Holding by household (acres)	1.15	15.97	0
Hindu (=1)	0.84	0.43	-
Year of Arrival in village	1804	0.07	1805
Village-level statistics			
Age of the Village (Years)	603	0.615	500
Distance to nearest Railway Station	22.71	1.05	14
Distance to nearest Pucca Road	0.60	3.27	0
# HHs in village	1,080.83	1.33	610
# *Jatis* in village	15.68	0.57	14
# HHs per *Jati*	226.37	1.45	123
Fractionalization	0.73	0.22	0.78
% HHs with Landline	0.15	1.39	-
% HHs with Mobile	0.14	1.17	-
% HHs with Electricity	0.64	0.81	-
% of Streets with Public Taps	0.52	0.96	-
% of Streets with Drinking Water Wells	0.35	1.35	-
Income Gini Index	0.45	0.19	0.45
Run Size	28.46	1.86	9

Table 2: Networks statistics - Components

Food Networks	**Mean**	**CV=$\frac{\sigma}{\mu}$**	**Median**
# Components in the village network	59.68	2.18	12
# Singleton Components in the village	214.79	0.89	133
# HHs per component	674.64	1.55	348
Distinct components spanned by a *jati*	15.31	2.05	4
Distinct components spanned by a street	6.95	2.21	3
Distinct components spanned by a *jati*+street	3.65	1.70	2
Clustering coefficient: Street-level	0.204	0.675	-
Clustering coefficient: *Jati*-level	0.202	0.853	-
Clustering coefficient: *Jati*+Street-level	0.227	0.898	-
Diameter: Overall	18.306	0.414	-
Diameter: Street-level	7.785	0.726	-
Diameter: *Jati*-level	7.509	0.755	-
Diameter: *Jati*+Street-level	4.089	1.018	-
Money Networks			
# Components in the village network	83.80	2.01	21
# Singleton Components in the village	239.29	0.91	182
# HHs per component	643.09	1.66	320
Distinct components spanned by a *jati*	20.83	2.04	6
Distinct components spanned by a street	9.81	2.19	4
Distinct components spanned by a *jati*+street	4.93	1.81	2
Clustering coefficient: Street-level	0.129	0.905	-
Clustering coefficient: *Jati*-level	0.121	1.112	-
Clustering coefficient: *Jati*+Street-level	0.140	1.203	-
Diameter: Overall	21.991	0.414	-
Diameter: Street-level	9.484	0.882	-
Diameter: *Jati*-level	8.861	0.886	-
Diameter: *Jati*+Street-level	5.078	1.293	-

Table 3: Dyadic regressions (State: Uttar Pradesh)

VARIABLES	(1) Link (Food) = 1	(2) Link (Money) =1
Distance = 2	-0.402***	-0.568***
	(0.0106)	(0.0267)
Distance = 3	-0.790***	-1.015***
	(0.0213)	(0.0331)
Distance = 4	-1.148***	-1.266***
	(0.0300)	(0.0411)
Distance = 5	-3.504***	-3.310***
	(0.0365)	(0.0456)
Same Jati (=1)	0.711***	0.711***
	(0.0334)	(0.0371)
Diff (HH age in village)	-0.00173***	-0.00216***
	(0.000194)	(0.000238)
Same Occupation (=1)	-0.0475*	-0.0910***
	(0.0247)	(0.0317)
Diff (HH Size)	0.00135	0.00789
	(0.00367)	(0.00557)
Diff (Age of Head)	-0.00157**	-0.00344***
	(0.000801)	(0.000993)
Diff (Head education)	-0.00656***	0.00582**
	(0.00212)	(0.00293)
Diff (Dependency ratio)	0.00497	-0.0106
	(0.00610)	(0.00686)
Diff (z score or land ownership)	-0.0498***	0.0337
	(0.0112)	(0.0252)
Ln Diff (HH Income)	0.0123*	0.0650***
	(0.00673)	(0.0112)
Constant	-2.910***	-4.248***
	(0.106)	(0.166)
Observations	5,318,946	5,318,946

Col. 1 shows results for the probability of dyad formation in food networks, Col. 2 shows similar results for money networks. The variable *Distance* measures the difference between household IDs and is constructed as a multinomial. *Same_jati* is a dummy variable that equals 1 if both households in a dyad belong to the same *jati*. *Same_Occupation* measures if the household heads of a dyad are in the same occupation. Difference variables measure the absolute differences between household characteristics such as household size, age and the education of the head, age of the household in the village, dependency ratio (proportion of household members who are below 18 years or above 60 years). The relative difference between the position in the land ownership and income distribution of households in the dyad has also been captured. Standard errors have been corrected using the method proposed by Fafchamps and Gubert (2007).

Table 4: Dyadic regressions (State: Uttar Pradesh)

	(1)	(2)	(3)	(4)
	DV: Link_Food = 1		DV: Link_Money = 1	
VARIABLES	Same Jati = 0	Same Jati = 1	Same Jati = 0	Same Jati = 1
Distance = 2	-0.453***	-0.388***	-0.530***	-0.554***
	(0.0740)	(0.0142)	(0.104)	(0.0326)
Distance = 3	-0.823***	-0.772***	-0.856***	-1.016***
	(0.0762)	(0.0255)	(0.117)	(0.0405)
Distance = 4	-1.259***	-1.104***	-1.119***	-1.259***
	(0.0939)	(0.0355)	(0.113)	(0.0492)
Distance = 5	-3.462***	-3.439***	-3.082***	-3.314***
	(0.0731)	(0.0408)	(0.0999)	(0.0512)
Diff (HH Age in village)	-0.000914***	-0.00326***	-0.00161***	-0.00295***
	(0.000225)	(0.000401)	(0.000274)	(0.000439)
Same Occupation (=1)	-0.000512	-0.0512*	-0.106**	-0.0511
	(0.0389)	(0.0288)	(0.0491)	(0.0353)
Diff (HH Size)	0.00389	-0.00155	0.0101	0.00545
	(0.00493)	(0.00450)	(0.00652)	(0.00713)
Diff (Head Age)	-0.00147	-0.00203**	-0.00403***	-0.00267**
	(0.00112)	(0.00101)	(0.00134)	(0.00124)
Diff (Head Education)	-0.0141***	0.00208	0.00962**	0.00158
	(0.00309)	(0.00287)	(0.00414)	(0.00349)
iff (Dep. Ratio)	0.00644	0.00378	-0.0124	-0.00696
	(0.00804)	(0.00742)	(0.00915)	(0.00892)
Diff (z score in land distribution)	-0.0418***	-0.0443***	0.0622**	0.00329
	(0.0162)	(0.0138)	(0.0287)	(0.0211)
Ln Diff (Household Income)	0.00341	0.0208**	0.0454***	0.0838***
	(0.0101)	(0.00815)	(0.0157)	(0.0133)
Constant	-2.816***	-2.287***	-4.341***	-3.642***
	(0.160)	(0.116)	(0.229)	(0.182)
Village Dummies	Yes	Yes	Yes	Yes
Observations	4,062,890	1,256,056	4,062,890	1,256,056

Cols. 1 and 2 show the differential impact of $same - jati$ on the probability of dyad formation in food networks. Cols 3-4 show similar results for money networks. The multinomial variable $Distance$ measures the difference between household IDs. Distances have been capped at 5. $Same\_Occupation$ measures if the household heads of a dyad are in the same occupation. Difference variables measure the absolute differences between household characteristics such as household size, age and the education of the head, age of the household in the village, dependency ratio (proportion of household members who are below 18 years or above 60 years). The relative between the position in the land ownership and income distribution of households in the dyad has also been captured. All specifications control for unobservables at the village-level. Standard errors have been corrected using the method proposed by Fafchamps and Gubert (2007).

Table 5: Household-level Heterogeneity regressions

VARIABLES	(1) Homophily (Food)	(2) Homophily (Food)	(3) Homophily (Money)	(4) Homophily (Money)
Periphery=2	0.0563***	0.0569***	0.0373***	0.0387***
	(0.00581)	(0.00641)	(0.00640)	(0.00736)
Periphery=3	0.0899***	0.0965***	0.0672***	0.0790***
	(0.00870)	(0.0101)	(0.00781)	(0.00953)
Periphery=4	0.110***	0.118***	0.0891***	0.104***
	(0.00937)	(0.0112)	(0.00876)	(0.0111)
Periphery=5	0.121***	0.133***	0.0860***	0.0999***
	(0.0107)	(0.0131)	(0.00936)	(0.0120)
Run Size	0.000929***	0.00127***	0.00107***	0.00163***
	(0.000295)	(0.000424)	(0.000267)	(0.000341)
Periphery=2 * Run Size		-0.000111		-0.000219
		(0.000179)		(0.000187)
Periphery=3 * Run Size		-0.000522		-0.000924***
		(0.000317)		(0.000238)
Periphery=4 * Run Size		-0.000546*		-0.000986***
		(0.000316)		(0.000269)
Periphery=5 * Run Size		-0.000648*		-0.000877***
		(0.000347)		(0.000279)
Jati Population fraction	0.604***	0.601***	0.705***	0.701***
	(0.142)	(0.142)	(0.144)	(0.144)
(Jati Population fraction)^2	-0.0999	-0.0979	-0.112	-0.109
	(0.186)	(0.186)	(0.193)	(0.192)
Fractionalization	0.116	0.116	0.138	0.140
	(0.0844)	(0.0844)	(0.0875)	(0.0875)
Constant	0.208	0.203	0.685	0.677
	(0.499)	(0.499)	(0.544)	(0.543)
Controls	Yes	Yes	Yes	Yes
State Dummies	Yes	Yes	Yes	Yes
Jati Dummies	Yes	Yes	Yes	Yes
Observations	42,106	42,106	41,579	41,579
R-squared	0.261	0.261	0.256	0.256
Periphery[1] = Periphery [2]	***	***	***	***
Periphery[1] = Periphery [3]	***	***	***	***
Periphery[1] = Periphery [4]	***	***	***	***
Periphery[1] = Periphery [5]	***	***	***	***
Periphery[2] = Periphery [3]	***	***	***	***
Periphery[2] = Periphery [4]	***	***	***	***
Periphery[2] = Periphery [5]	***	***	***	***
Periphery[3] = Periphery [4]	***	***	***	***
Periphery[3] = Periphery [5]	***	***	***	***
Periphery[4] = Periphery [5]	***	***		

Cols. (1) and (2) show IV-2SLS estimation results for Eq. 7. The dependent variable is defined as the number of same-*jati* links maintained by a household in food (Cols. (1)-(2)) and money (Cols. (3)-(4)). Households belonging to runs with a length of at least 9 have been retained for this estimation. Periphery is a multinomial variable that takes values from 1 to 5. *Periphery*=1 is taken as the base category and signifies a household at the end of a same-*jati* run. *Periphery*=5 signifies a household in the middle of the run. All specifications include state and *jati* dummies. Standard errors have been clustered at the village level.

Table 6: Household-level Homophily regressions

Variables	Food Networks			Money Networks		
	(1)	(2)	(3)	(4)	(5)	(6)
	Prop. Links	nbd_jati	Prop. Links	Prop. Links	nbd_jati	Prop. Links
	OLS	First Stage	2nd Stage	OLS	First Stage	2nd Stage
nbd_jati	0.136***		0.256***	0.115***		0.260***
	(0.00835)		(0.0321)	(0.00804)		(0.0318)
Ln (# same jati-street HHs during arrival)		0.138***			0.139***	
		(0.0111)			(0.0112)	
Ln (Age of the village)		0.0531***			0.0535***	
		(0.0194)			(0.0198)	
Jati population share	0.811***	1.230***	0.574***	0.884***	1.241***	0.595***
	(0.143)	(0.135)	(0.141)	(0.143)	(0.136)	(0.142)
(Jati population share)^2	-0.131	-0.879***	0.0489	-0.121	-0.890***	0.0977
	(0.218)	(0.182)	(0.205)	(0.217)	(0.185)	(0.204)
Jati Fractionalization	0.133	-0.288***	0.172**	0.148*	-0.289***	0.196**
	(0.0909)	(0.105)	(0.0859)	(0.0862)	(0.105)	(0.0841)
Constant	-0.594	0.722	-0.730	-0.179	0.701	-0.340
	(0.654)	(0.466)	(0.648)	(0.701)	(0.469)	(0.700)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
State FE	Yes	Yes	Yes	Yes	Yes	Yes
Jati FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	91,060	91,060	91,060	89,806	89,806	89,806
R-squared	0.369		0.340	0.348		0.307
Kleibergen-Papp F Statistic (Weak Identification)		79.42			78.19	
Kleibergen-Papp LM Statistic (Underidentification)		55.04			55.65	
Hansen test statistic (Overidentification)		0.43			2.04	
Hansen p value		0.613			0.153	

Cols. (1) and (4) show results from the OLS estimation of the main homophily specification, for food and money networks, respectively. Cols. (2) and (3) show results for the first and second stage of the IV-2SLS specification for food networks, whereas Cols. (5) and (6) show corresponding results for the money networks. The endogenous variable, *nbd_jati*, is the number of same-*jati* households living either to the left or right of the household in question. This variable takes on possible values of 0, 1 or 2. The dependent variable in Cols. (1)-(3) is the proportion of same-*jati* links maintained by a household in food networks. A similar definition extends to the case of money networks, applicable to Cols. (4)-(6). Each specification includes State and *Jati* dummies. Standard errors are clustered at the Village level.

Table 7: Robustness: Different definition of endogenous variable

	(1)	(2)
VARIABLES	Homophily (Money)	Homophily (Food)
Ln (# same-*jati* HHs on street)	0.0701***	0.0681***
	(0.0101)	(0.0102)
Jati Population fraction	0.590***	0.573***
	(0.151)	(0.151)
(Jati Population fraction)^2	0.125	0.0715
	(0.215)	(0.217)
Fractionalization	0.0933	0.0684
	(0.0869)	(0.0948)
State Dummies	Yes	Yes
Jati Dummies	Yes	Yes
Observations	88,630	89,869

Cols. (1) and (2) show results from the IV-2SLS estimation of the modified homophily specification, for food and money networks, respectively. The endogenous variable, $Ln(\#same - jatiHHsonstreet)$, is the number of same-*jati* households living on the same street as the household in question. The dependent variable in Cols. (1) and (2) is the proportion of same-*jati* links maintained by a household in food networks. Each specification includes State and *Jati* dummies. Standard errors are clustered at the Village level.

Table 8: Re-classifying *nbd_jati*

nbd_jati (OLD)	nbd_jati (NEW)			
	0	1	2	Total
0	4192	67	0	4259
1	1012	10300	0	11312
2	0	2578	22852	25430
Total	5204	12945	22852	**41001**

$nbd\_jati(OLD)$ is the actual definition of the neighborhood composition of a household, as used in the original homophily equation. The variable $nbd\_jati(NEW)$ is the new definition - used for robustness checks - that is obtained after netting out households that split-off the same-*jati* parent household as a neighbor.

Table 9: Re-classifying *nbd_jati*

	(1)	(2)
VARIABLES	Food Homophily	Money Homophily
nbd_jati_new	0.265***	0.295***
	(0.0526)	(0.0523)
Jati Population Fraction	0.669***	0.712***
	(0.162)	(0.161)
(*Jati* Population Fraction)^2	-0.0542	-0.0356
	(0.222)	(0.215)
Fractionalization	0.159*	0.168*
	(0.0933)	(0.0955)
State Dummies	Yes	Yes
Jati Dummies	Yes	Yes
Observations	37,286	36,927

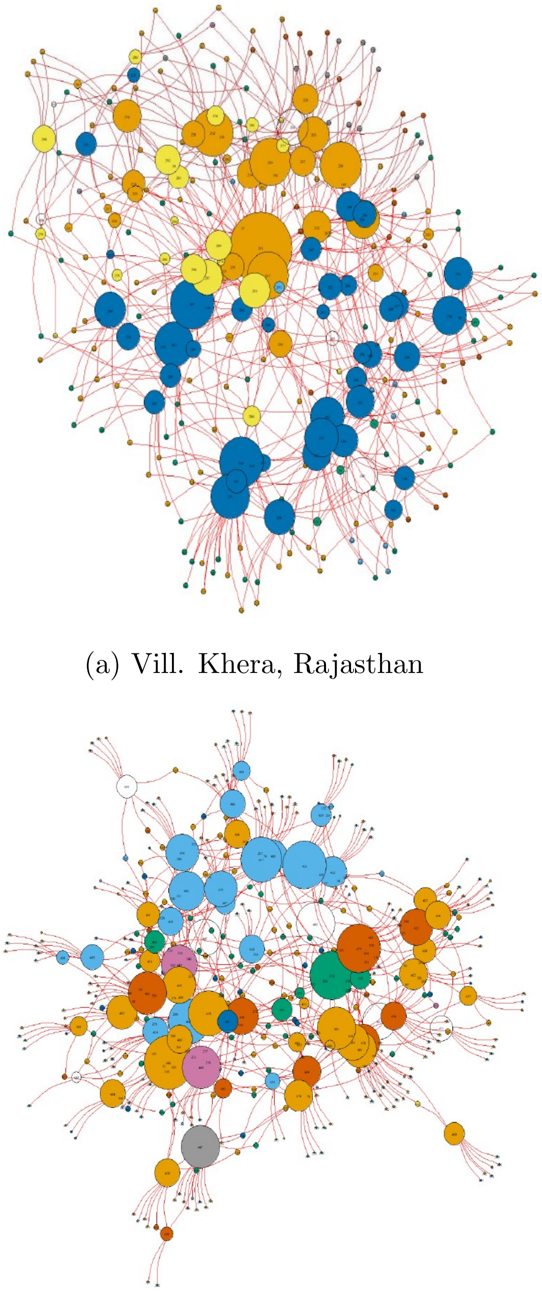
Cols. (1) and (2) show results from the IV-2SLS estimation of the modified homophily specification, for food and money networks, respectively. The endogenous variable, *nbd_jati_new*, is the number of same-*jati* households living in the immediate neighborhood of the household in question, after netting out those that have split from the same parent household in the 1999 rounds of the REDS. The dependent variable in Cols. (1) and (2) is the proportion of same-*jati* links maintained by a household in food networks. Each specification includes State and *Jati* dummies. Standard errors are clustered at the Village level.

Table 10: Robustness: Control Function approach

VARIABLES	(1) Homophily (Food)	(2) Homophily (Money)
$\widehat{nbd\_jati}$	0.125***	0.105***
	(0.00793)	(0.00781)
Jati Population Fraction	1.033***	1.068***
	(0.142)	(0.142)
(*Jati* Population Fraction)^2	-0.287	-0.248
	(0.219)	(0.216)
Fractionalization	0.0977	0.119
	(0.0962)	(0.0875)
Constant	-0.557	-0.132
	(0.677)	(0.720)
State Dummies	Yes	Yes
Jati Dummies	Yes	Yes
Observations	91,060	89,806

Cols. (1) and (2) show second stage results from the estimation of the base homophily specification using the Control Function approach, for food and money networks, respectively. The endogenous variable, $\widehat{nbd\_jati}$, is the predicted number of same-*jati* households living in the immediate neighborhood of the household in question, estimated using a Poisson first-stage regression. The dependent variable in Cols. (1) and (2) is the proportion of same-*jati* links maintained by a household in food networks. Each specification includes State and *Jati* dummies. Standard errors are clustered at the Village level.

Figure 1: Network Maps

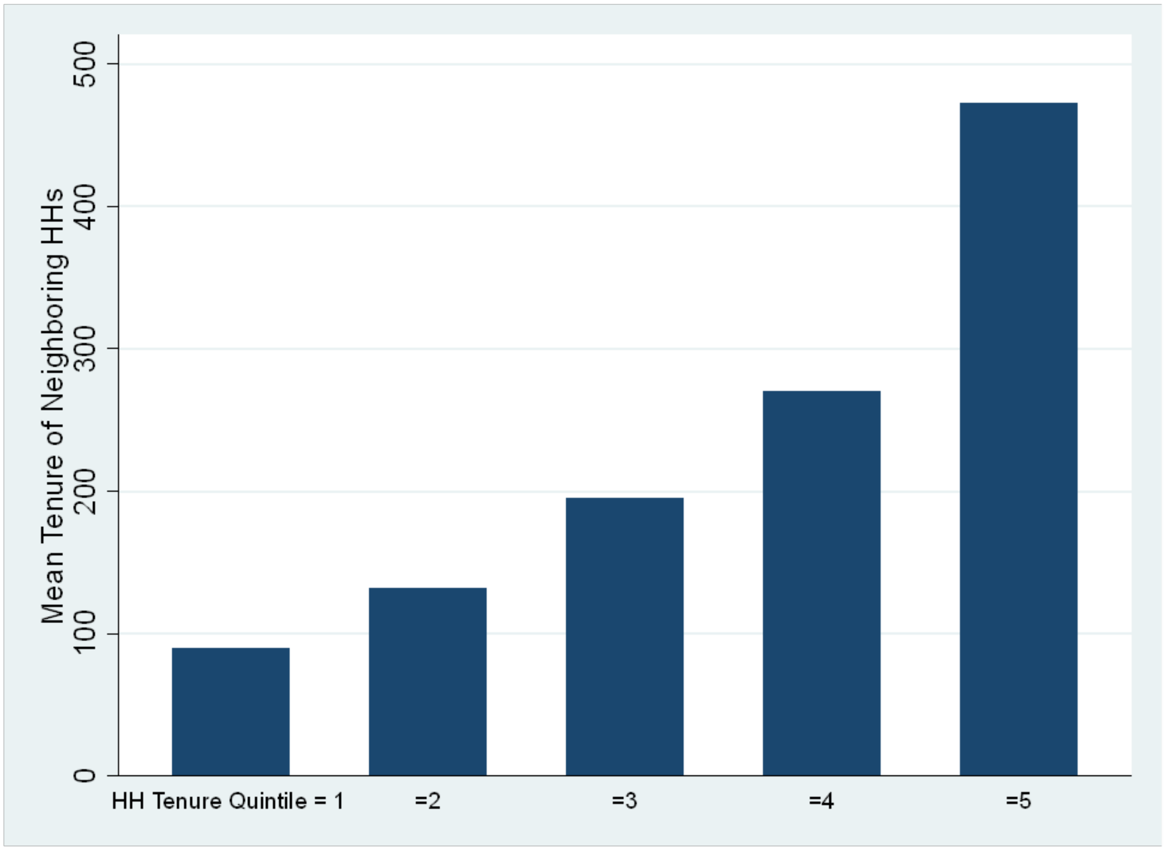


(a) Vill. Khera, Rajasthan

(b) Vill. Thirubuvanam, Tamil Nadu

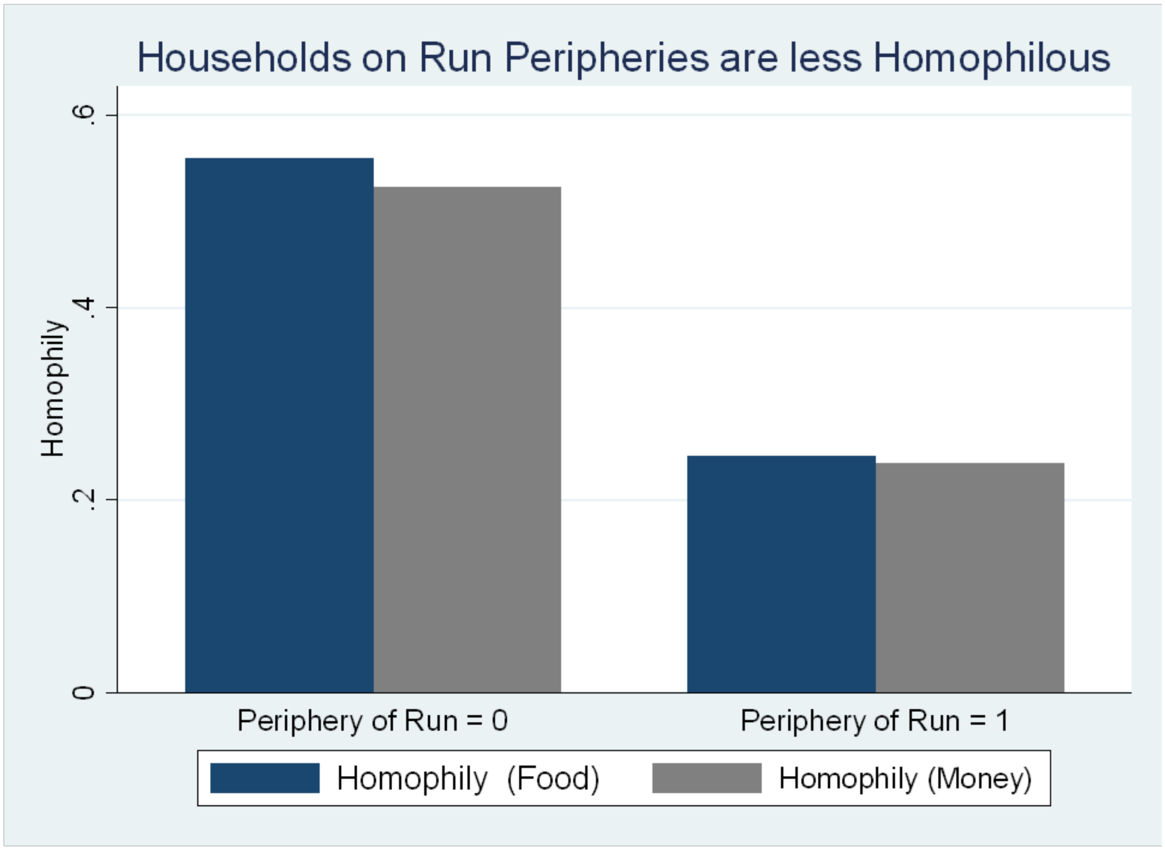
This figure presents network graphs for two REDS villages. For the sake of brevity, we have presented the graphs for Money Networks. Each node represents a household. Node size is weighted by household centrality, whereas each *jati* is represented by a unique color. Lines represent reported links between two nodes (households).

Figure 2: Households arriving at the same time tend to be spatially concentrated



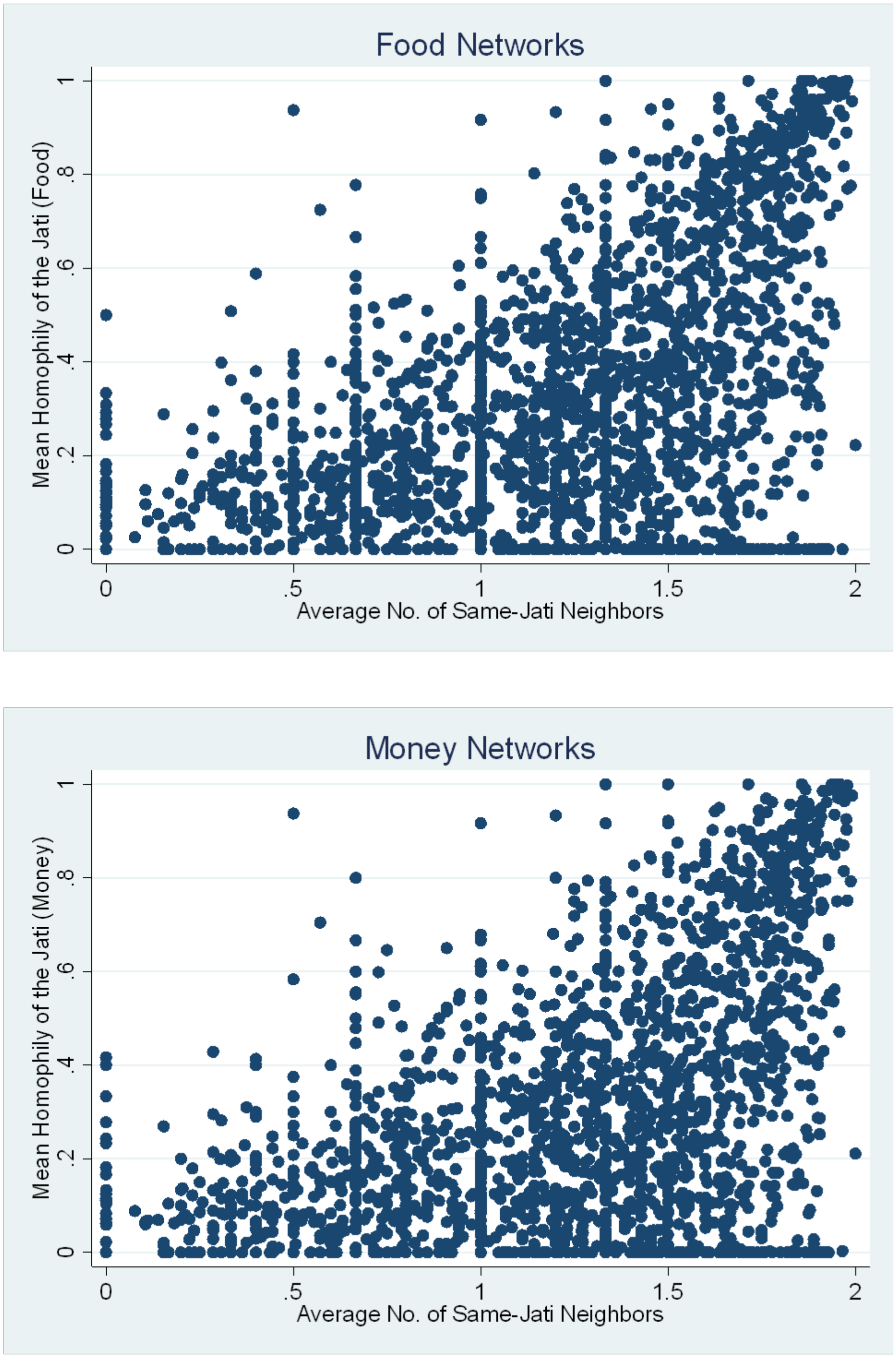
HH Tenure is the age of the household in a village. This variable has been classified into quintiles to facilitate easier comparison. The y-axis is the average tenure of the immediate left and right-hand side neighbors of a household.

Figure 3: Relationship between Household position in Same-*Jati* Run and Homophily



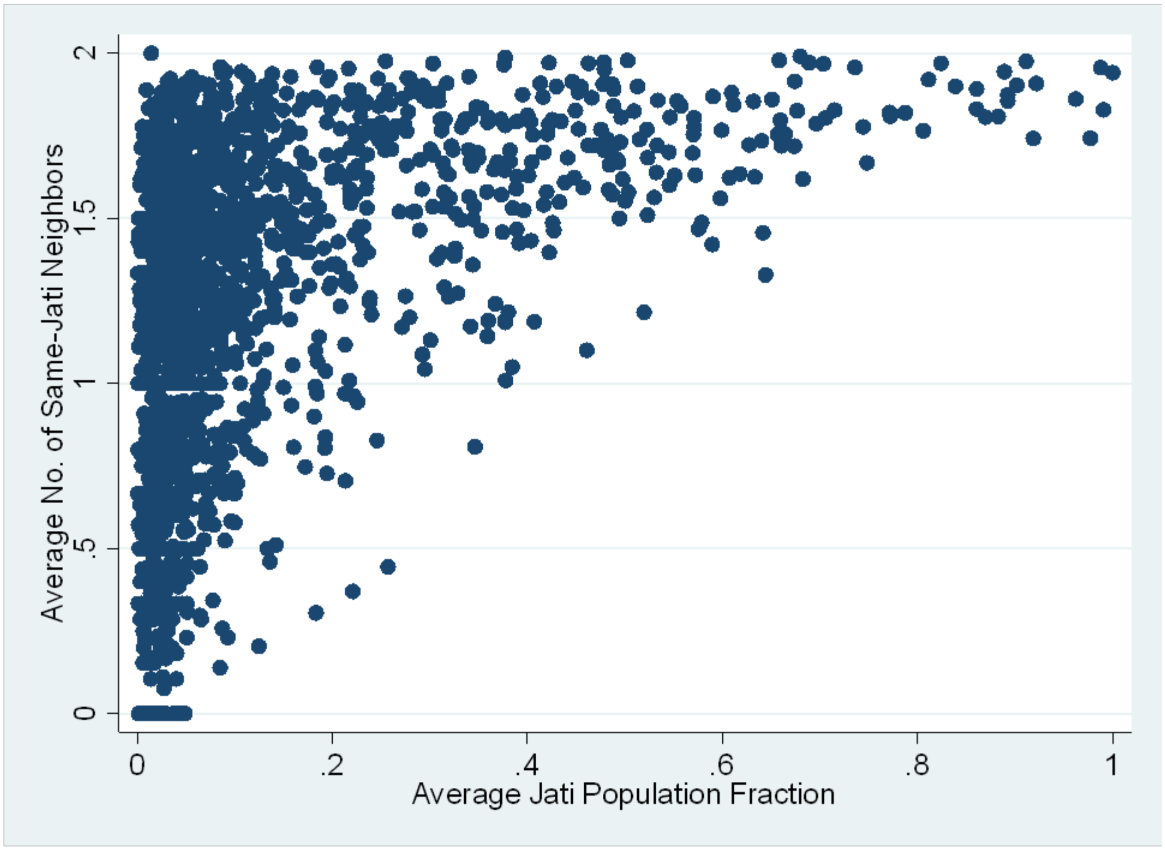
Homophily is defined as the proportion of links maintained by a household with other households that belong to the same *jati* as itself. A 'Run' is defined as a set contiguously located households, all of which belong to the same *jati*. Households situated at the end of such a 'run' are labelled as being on the 'periphery'. We define a dummy variable called *Run*$_P$*eriphery* which takes a value of 1 for households lying on a run periphery.

Figure 4: Relationship between Neighborhood composition and Homophily



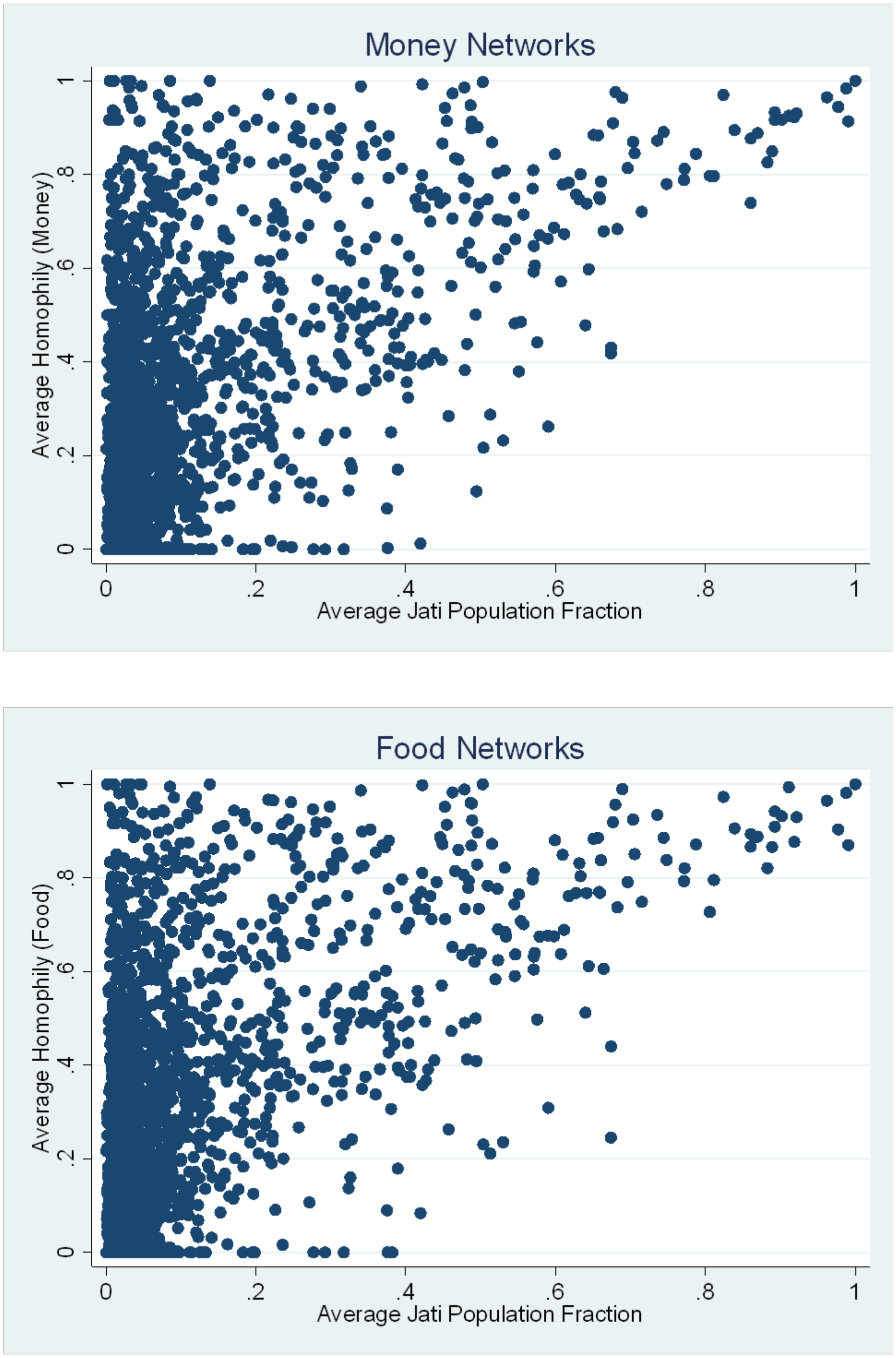
Each point on the graph represents a *jati*. Mean Homophily is defined as the average proportion of same-*jati* links maintained by households belonging to a particular *jati*. It is measure of average homophily calculated the *jati*-level. Average number of same-*jati* neighbors represents the *jati*-level average of neighborhood composition variable $Nbd\jmath ati$.

Figure 5: *Jatis* Size and Same-*Jati* Neighbors



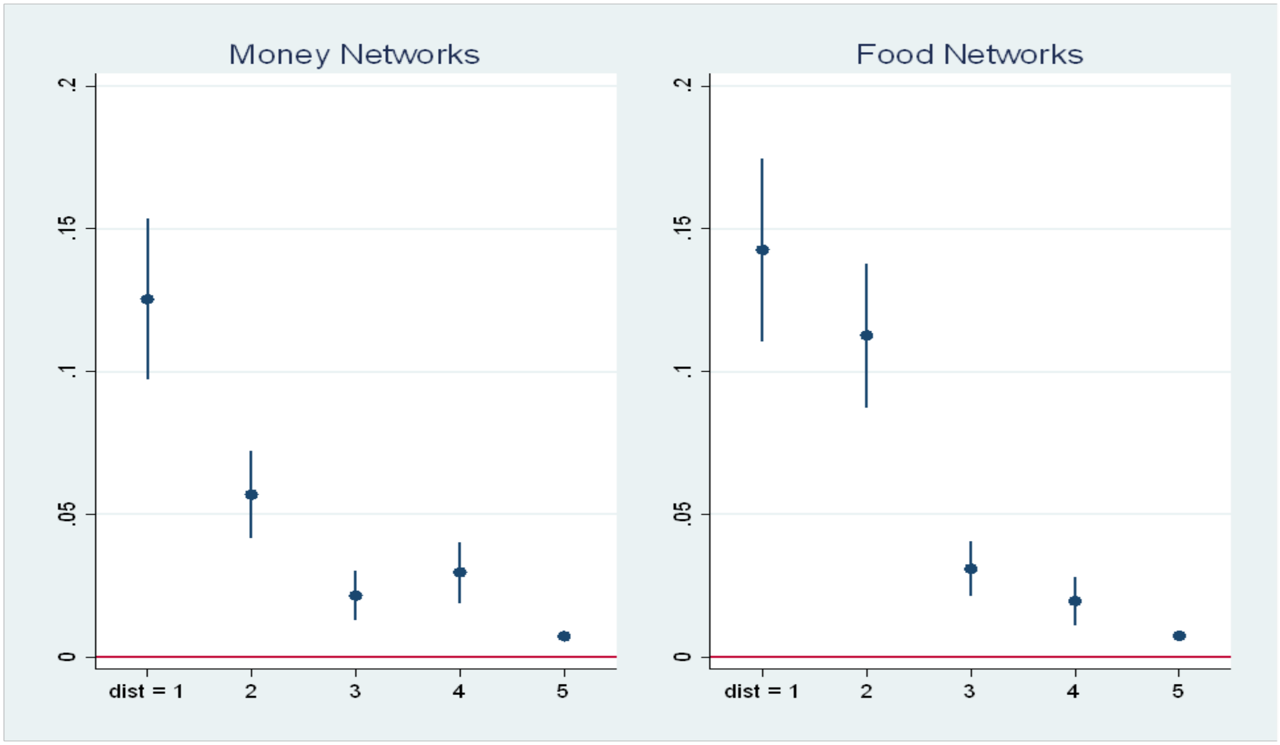
Each point on the graph represents a *jati*. The average number of same-*jati* neighbors represents the *jati*-level average of neighborhood composition variable Nbd_{Jati}. *Jati* Population Fraction is a measure of the share of village population represented by a *jati*. This figure is averaged across all villages, producing a variable that measures the average population fraction of a particular *jati* across all villages in the sample.

Figure 6: *Jatis* Size and Homophily



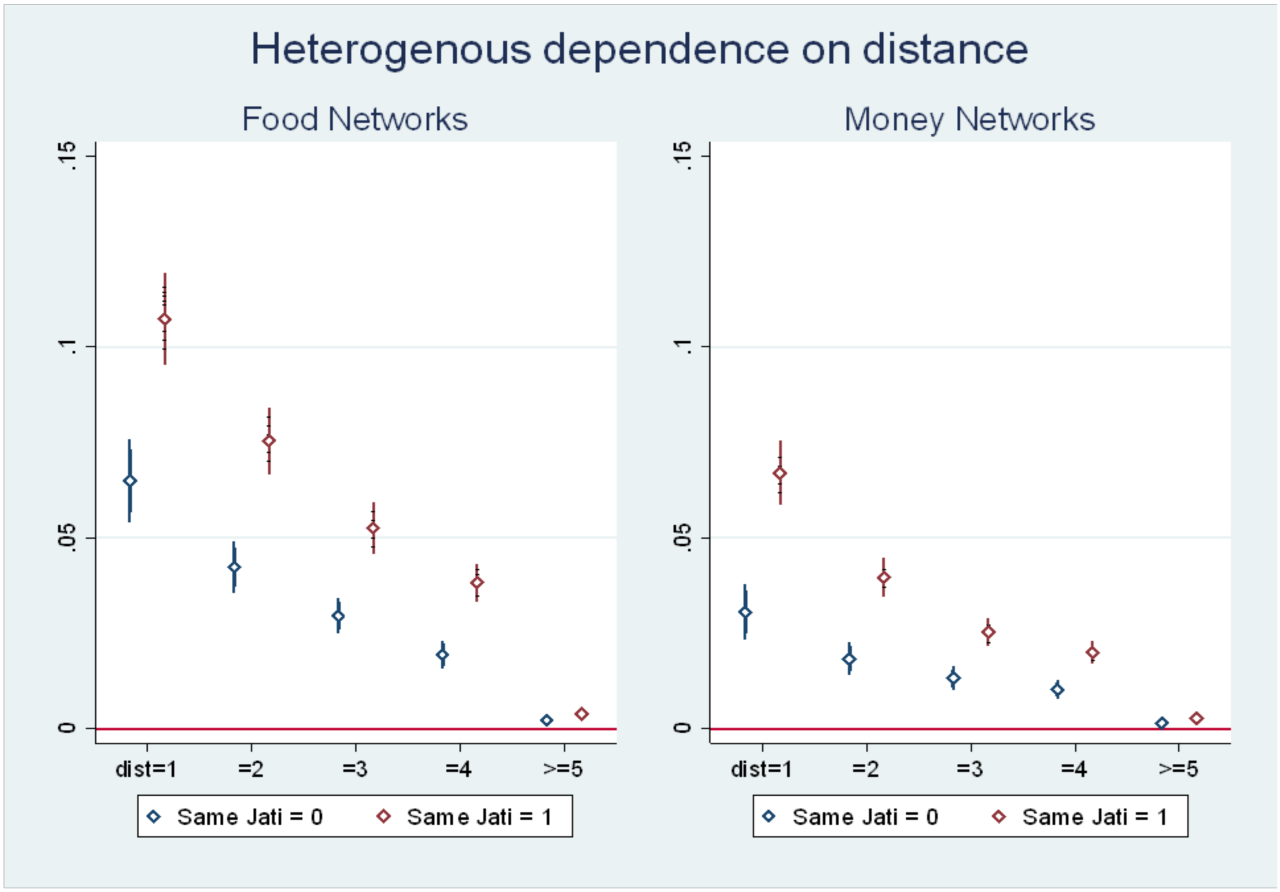
Each point on the graph represents a *jati*. Average Homophily is calculated at the level of each *jati* for Money and Food networks. *Jati* Population Fraction is a measure of the share of village population represented by a *jati*. This figure is averaged across all villages, producing a variable that measures the average population fraction of a particular *jati* across all villages in the sample.

Figure 7: Decreasing Dependence on Distance



The plot is obtained from the estimates of the Dyadic Regression shown in Equation 3. It shows the marginal impact of distance categories on the probability of link formation by households in the dyad. Distance is defined as categorical variables taking a maximum of five values. Distance = 1 represents a dyad with two neighboring households and Distance = 5 represents a dyad with two households that have 4 intervening households. 95% Confidence Intervals have been plotted along with the point estimates of the marginal effects.

Figure 8: Heterogenous dependence on distance



The plot is obtained from the estimates of the Dyadic Regression shown in Equation 4. It shows the heterogenous impact of *jati* identity on the marginal impact of distance categories on the probability of link formation by households in the dyad. The series represented in Red (Blue) is for households in a dyad that belong to (do not belong to) the same *jati*. Distance is defined as categorical variables taking a maximum of five values. Distance = 1 represents a dyad with two neighboring households and Distance = 5 represents a dyad with two households that have 4 intervening households. 95% Confidence Intervals have been plotted along with the point estimates of the marginal effects.