GOSSIP: IDENTIFYING CENTRAL INDIVIDUALS IN A SOCIAL NETWORK

ABHIJIT BANERJEE[†], ARUN G. CHANDRASEKHAR[‡], ESTHER DUFLO[§], AND MATTHEW O. JACKSON[⋆]

ABSTRACT. We examine individuals' abilities to identify the highly central people in their social networks, where centrality is defined by *diffusion centrality* (Banerjee, Chandrasekhar, Duflo, and Jackson, 2013), which characterizes a node's influence in spreading information. We first show that diffusion centrality nests standard centrality measures – degree, eigenvector and Katz-Bonacich centrality – as extreme special cases. Next, we show that boundedly rational individuals can, simply by tracking sources of gossip, identify who is central in their social network in the specific sense of having high diffusion centrality. Finally, we examine whether the model's predictions are consistent with data in which we ask people in each of 35 villages whom would be the most effective point from which to initiate a diffusion process. We find that individuals accurately nominate central individuals in the diffusion centrality sense. Additionally, the nominated individuals are more central in the network than 'village leaders' as well as those who are most central in a GPS sense. This suggests that individuals can rank others according to their centrality in the networks even without knowing the network, and that eliciting network centrality of others simply by asking individuals may be an inexpensive research and policy tool.

JEL CLASSIFICATION CODES: D85, D13, L14, O12, Z13

KEYWORDS: Centrality, Gossip, Networks, Diffusion, Influence, Social Learning

Date: This Version: June 2014.

Ben Golub and participants at BREAD 2014 provided helpful comments. Financial support from the NSF under grant SES-1155302, from the AFOSR and DARPA under grant FA9550-12-1-0411, and from ARO MURI under award No. W911NF-12-1-0509 is gratefully acknowledged. We thank Francisco Muñoz for excellent research assistance. We are grateful to CMF at IFMR, Tanay Balantrapu, Gowri Nagraj, and Manaswini Rao.

[†]Department of Economics, MIT.

[‡]Department of Economics, Stanford University.

[§]Department of Economics, MIT.

[⋆]Department of Economics, Stanford University; Santa Fe Institute; and CIFAR.

1. Introduction

> *"The secret of my influence has always been that it remained secret."* – Salvador Dali

Knowing who is influential, or central, can be very important for members of a community as well as for businesses and policymakers.[1] For instance, the extent to which information diffuses among a population may depend on how central the initially informed are within the network. Policymakers and organizations can benefit from targeting the right individuals in efforts to effectively spread valuable information.

However, learning who is central in a social network has the potential to be quite difficult for several reasons. For policymakers, collecting detailed network data is costly. Even for members of the network, knowing the structure of the network beyond their immediate friends, and knowing how central other individuals are is far from automatic. Do people know how central other people are?

In this paper, we make two contributions.

First, we develop a simple model – based on our previous work (Banerjee, Chandrasekhar, Duflo, and Jackson, 2013) – to show that individuals in a network should be able to identify central individuals within their community without knowing the structure of the entire network. In our model, gossip randomly originates at various nodes: "Arun got a job", "Matt is teaching a new course", etc. That information is then randomly passed from neighbor to neighbor. Individuals track how frequently they hear gossip about various others simply by counting. We show that for any listener in the network, the relative ranking under this count converges over time to the correct ranking of all other nodes' abilities to send information. The specific sense in which we rank a node's ability to send information is given by the "diffusion centrality", introduced in Banerjee et al. (2013). It answers the question as to how widely information from a given node diffuses in a given number of time periods and

[1] For some discussion, see Katz and Lazarsfeld (1955); Rogers (1995); Kempe, Kleinberg, and Tardos (2003, 2005); Borgatti (2005); Ballester, Calvó-Armengol, and Zenou (2006); Banerjee, Chandrasekhar, Duflo, and Jackson (2013).

for a given random transmission probability? Additionally, we relate diffusion centrality to other standard measures of centrality in the literature, proving that it nests three of the most prominent centrality measures: degree centrality at one extreme if there is just one time period of communication, and eigenvector centrality and Katz-Bonacich centrality at the other extreme if there are unlimited periods of communication. Diffusion centrality takes on a wide range of other values for intermediate numbers of periods. Taken together, we show that by listening and doing simple addition, individuals can come to learn the correct ranking of the centralities of the people in their broader community.

Second, we use a unique dataset to study the implications of our theoretical predictions. We asked every adult in each of 35 villages to name the person in their village best-suited to initiate the spread of information. We couple their answers (which we call their 'nominations') with detailed network data that include maps of a variety of interactions in each of these 35 villages. We show that individuals nominate highly diffusion/eigenvector central people (on average at the 71st percentile of centrality). We also show that the nominations are not simply based on the nominee's leadership status or geographic position in the village, but is significantly correlated with diffusion centrality even after conditioning on these characteristics. Among these characteristics, nomination status is the best predictor of diffusion centrality of a household. This suggests that people understand our questions and are doing more than simply naming 'leaders' or the most geographically central individuals. Taken together, our results show that individuals are able to identify central individuals in their networks. Our model suggests why this might be the case. This is important because it is much cheaper to ask this question than to map out entire networks.

This is the first paper that we are aware of that shows how people can (easily) learn some things about their broader networks, and demonstrates that they are able to identify highly central people in their community.[2]

[2]There are some papers (e.g., Milgram (1967) and Dodds et al. (2003)) that have checked people's abilities to use knowledge of their friends' connections to efficiently route messages to reach distant people, but those rely on more local knowledge.

2. A Model of Network Communication

We consider the following model.

2.1. A Network of Individuals.

A society of n individuals are connected via a network, which has a (possibly directed) adjacency matrix $\mathbf{g} \in \{0,1\}^{n \times n}$.[3] Unless otherwise stated, we take the network $\mathbf{g}$ to be fixed and let $\mathbf{v}^{(1)}$ be its first eigenvector (which is nonnegative and real-valued by the Perron-Frobenius Theorem), corresponding to the largest eigenvalue λ_1.

Throughout, we assume that the network is (strongly) connected in that there exists a (directed) path from every node to every other node, so that information originating at any node could potentially make its way eventually to any other node.[4]

2.2. Diffusion Centrality.

Banerjee et al. (2013) defined a notion of centrality called *diffusion centrality*, based on random information flow through a network according to the following process, which is a variation of a standard diffusion process that underlies many contagion models.[5]

A piece of information is initiated at node i and then is broadcast outwards from that node. In each period each informed node informs each of its neighbors with a probability $p \in (0,1]$, independently across neighbors and history. The process operates for T periods, where T is a positive integer. Diffusion centrality measures how extensively the information spreads as a function of the initial node. In particular, let

$$\mathbf{H}(\mathbf{g}; p, T) := \sum_{t=1}^{T} (p\mathbf{g})^t \,,$$

be the 'hearing matrix'. The ij-th entry of $\mathbf{H}$, $H(\mathbf{g}; p, T)_{ij}$, is the expected number of times j hears about a piece of information originating from i in the first T periods. Diffusion centrality is then defined by

$$DC(\mathbf{g}; p, T) := \mathbf{H}(\mathbf{g}; p, T) \cdot \mathbf{1} = \left(\sum_{t=1}^{T} (p\mathbf{g})^t \right) \cdot \mathbf{1}.$$

[3]As will become clear, an easy extension is to allow $\mathbf{g}$ to be a weighted matrix.
[4]More generally, everything we say applies to the components of the network.
[5]See Jackson and Yariv (2011) for background and references.

So, $DC(\mathbf{g}; p, T)_i$ is expected total number of times that a piece of information that originates from i is heard by any of the members of the society during a T-period time interval.[6] Banerjee et al. (2013) showed that diffusion centrality was a statistically significant predictor of the spread of information about a microfinance program.

Note that this measure allows people to hear the information multiple times from the same person and counts those times, so that it is possible for an entry of DC to be more than n. There are several advantages to defining it in this manner. First, although it is possible via simulations to calculate a measure that tracks the expected number of informed nodes and avoid double-counting, this expression is *much* easier to calculate and for many parameter values the measures are roughly proportional to each other. Second, this version of the measure relates nicely to other standard measures of centrality in the literature, while a measure that adjusts for multiple hearing does not. Third, in a world in which multiple hearings lead to a greater probability of information retention, this count might be a better one.[7]

2.3. **Diffusion Centrality's Relation to Other Centrality Measures.** It is useful to situate diffusion centrality relative to other prominent measures of centrality in the literature.

Let $\mathbf{d}(\mathbf{g})$ denote degree centrality (so $d_i(\mathbf{g}) = \sum_j g_{ij}$). Recall that $\mathbf{v}^{(1)}(\mathbf{g})$ is the eigenvector centrality: the nonnegative vector such that $\mathbf{g}\mathbf{v}^{(1)} = \lambda_1 \mathbf{v}^{(1)}$ where λ_1 is the largest eigenvalue of $\mathbf{g}$ in magnitude.

[6] We note two useful normalizations. One is to compare it to what would happen if $p = 1$ and $\mathbf{g}$ were the complete network $\mathbf{g}^c$, which produces the maximum possible entry for each ij subject to any T. Thus, each entry of $DC(\mathbf{g}; p, T)$ could be divided through by the corresponding entry of $DC(\mathbf{g}^c; 1, T)$. This produces a measure for which every entry lies between 0 and 1, where 1 corresponds to the maximum possible numbers of expected paths possible in T periods with full probability weight and full connectedness. Another normalization is to compare a given node to the total level for all nodes; that is, to divide all entries of $DC(\mathbf{g}; p, T)$ by $\sum_i DC_i(\mathbf{g}; p, T)$. This normalization tracks how relatively diffusive one node is compared to the average diffusiveness in its society.

[7] One could also further complicate the measure by allowing for the forgetting of information, but with three parameters the measure would begin to become unwieldy.

Let $KB(\mathbf{g},p)$ denote Katz-Bonacich centrality - defined for $p < 1/\lambda_1$ by:[8]

$$KB(p,\mathbf{g}) := \left(\sum_{t=1}^{\infty} (p\mathbf{g})^t\right) \cdot \mathbf{1}.$$

It is direct to see that (i) diffusion centrality is proportional to degree centrality at the extreme at which $T = 1$, and (ii) if $p < 1/\lambda_1$, then diffusion centrality coincides with Katz-Bonacich centrality if we set $T = \infty$. We now show that when $p > 1/\lambda_1$ diffusion centrality approaches eigenvector centrality as T approaches ∞, which then completes the picture of the relationship between diffusion centrality and extreme centrality measures.

The difference between the extremes of Katz-Bonacich centrality and eigenvector centrality depends on whether p is sufficiently small so that limited diffusion takes place even in the limit of large T, or whether p is sufficiently large so that the knowledge saturates the network and then it is only relative amounts of saturation that are being measured.[9]

Theorem **1.**

(1) *Diffusion centrality is proportional to degree when $T = 1$:*

$$DC(\mathbf{g};p,1) = pd(\mathbf{g}).$$

(2) *If $p \geq 1/\lambda_1$, then as $T \to \infty$ diffusion centrality approximates eigenvector centrality:*

$$\lim_{T\to\infty} \frac{1}{\sum_{t=1}^{T} (p\lambda_1)^t} DC(\mathbf{g};p,T) = \mathbf{v}^{(1)}.$$

(3) *For $T = \infty$ and $p < 1/\lambda_1$, diffusion centrality is Katz-Bonacich centrality:*

$$DC(\mathbf{g};p,\infty) = KB(\mathbf{g},p); \quad p < 1/\lambda_1.$$

All proofs appear in the Appendix.

[8]See (2.7) in Jackson (2008) for additional discussion and background. This was a measure first discussed by Katz, and corresponds to Bonacich's definition when both of Bonacich's parameters are set to p.

[9]Saturation occurs when the entries of $\left(\sum_{t=1}^{\infty} (p\mathbf{g})^t\right) \cdot \mathbf{1}$ diverge (note that in a (strongly) connected network, if one entry diverges, then all entries diverge). Nonetheless, the limit vector is still proportional to a well defined limit vector: the first eigenvector.

The result shows that as T is varied, diffusion centralities nests three of the most prominent and used centrality measures: degree centrality, eigenvector centrality, and Katz-Bonacich centrality. It thus provides a foundation for these measures and spans between them.[10] Between these extremes, diffusion centrality measures how diffusion process operates for some limited number of periods. Importantly, as in Banerjee et al. (2013), the behavior in the intermediate ranges can be more relevant for certain diffusion phenomena.

3. Relating Diffusion Centrality to Network Gossip

We now investigate whether and how individuals living in $\mathbf{g}$ end up with knowledge of others' positions in the network that correlates with diffusion centrality without knowing anything about the network structure.

3.1. **A Gossip Process.** Diffusion centrality considers diffusion from the *sender's* perspective. Let us now consider the same diffusion process but from a *receiver's* perspective. Over time, news about any given individual arrives randomly and that person is 'gossiped' about by her neighbors. In particular, her neighbors transmit this news to each of their neighbors, and so forth just as in the diffusion process described above. In particular, in each of a discrete number $T > 0$ periods that piece of information is diffused through the network. For instance, Arun may tell Matt that he has a new car. Matt then may tell Abhijit that "Arun has a new car," and then Abhijit may tell Esther that "Arun has a new car". In each period, each individual that has some information tells tells it to each of her friends with a probability $p > 0$, independently across neighbors. What is crucial is that the news involves the name of the node of origin – in this case "Arun".

[10]We also remark on the comparison to another measure: the influence vector that appears in the DeGroot learning model (see, e.g., Golub and Jackson (2010)). That metric captures how influential a node is in a process of social learning. It corresponds to the (left-hand) unit eigenvector of a stochasticized matrix of interactions rather than a raw adjacency matrix. While it might be tempting to use that metric here as well, we note that it is the right conceptual object to use in a process of *repeated averaging* through which individuals update opinions based on averages of their neighbors' opinions. It is thus conceptually different from the diffusion process that we study. One can define a variant of diffusion centrality that works for finite iterations of DeGroot updating.

Recall that

$$\mathbf{H}(\mathbf{g};p,T) := \sum_{t=1}^{T} (p\mathbf{g})^t ,$$

is such that the ij-th entry, $H(\mathbf{g};p,T)_{ij}$, is the expected number of times j hears a piece of information originating from i.

We define the *network gossip heard* by node j to be the j-th column of $\mathbf{H}$,

$$NG(\mathbf{g};p,T)_j = H(\mathbf{g};p,T)_{\cdot j}.$$

Thus, NG_j lists the expected number of times a node j will hear a given piece of news as a function of the node of origin of the information. So, if $NG(\mathbf{g};p,T)_{ij}$ is twice as high as $NG(\mathbf{g};p,T)_{kj}$ then j is expected to hear news twice as often that originated at node i compared to node k, presuming equal rates of news originating at i and k.

Note the different perspectives of DC and NG: diffusion centrality tracks how well information spreads from a given node, while network gossip tracks how relatively often a given node hears information from (or about) each of the other nodes.

We make two remarks about the process. First, one could allow passing probabilities to differ by information type and pairs of nodes.[11] Indeed, in Banerjee et al. (2013) we allowed different nodes to pass information with different probabilities, and in Banerjee et al. (2014) we allow the probability of communication to depend on the listener's network position. Although one can enrich the model in many ways to capture specifics of information passing, this simple version captures basic dynamics and relates naturally to centrality measures. Second, this process keeps track of information as if fountains of information come from each node with equal rates and then listeners track the information that they hear. It could be that information springs in manners that are correlated with nodes' attributes. Provided that information springs

[11]We can generalize our setup replacing p with a matrix $\mathbf{P}$. Now define

$$\mathbf{H}(\mathbf{g};\mathbf{P},T) := \left(\sum_{t=1}^{T} (\mathbf{P} \circ \mathbf{g})^t \right).$$

Here $\mathbf{P}$ can have entries P_{ij} which allow the transmission probabilities to vary by pair. Note that P_{ij} can depend on characteristics of those involve and encode strategic behavior based on the economics being modeled.

at rates that are positively related to nodes' centralities, the results that we present below go through.

3.2. **Identifying Central Individuals.** With this measure of gossip in hand, we show how individuals in a society can estimate who is central by simply counting how often they hear gossip about others. We first show that, on average, individuals' rankings of others according to how much gossip they hear about the others, given by NG_j, are positively correlated with diffusion centrality.

THEOREM **2.** *For any* $(\mathbf{g}; p, T)$, $\sum_j \mathrm{cov}(DC(\mathbf{g}; p, T), NG(\mathbf{g}; p, T)_j) = \mathrm{var}(DC)$. *Thus, in any network with differences in diffusion centrality among individuals, the average covariance between diffusion centrality and network gossip is positive.*

It is important to emphasize that although both measures, NG_i and DC_i, are based on the same sort of information process, these are really two different measures. Diffusion centrality is a measure of a node's ability to send information, while the network gossip measure is tracking the reception of information by different nodes. Indeed, the reason that Theorem 2 is only stated for the sum rather than any particular individual j's network gossip measure is that after a (small) number of periods it is possible that some nodes have not even heard about other nodes, and moreover they might be biased towards their local neighborhoods.[12]

Next, we show that if individuals exchange gossip over extended periods of time, every individual in the network is eventually able to *perfectly* rank others' centralities.

[12]One might conjecture that more central listeners would be better "listeners": for instance having more accurate rankings than less central listeners after a small number of periods. Although this might happen in some networks, and for many comparisons, it is not guaranteed. None of the centrality measures considered here ensure that a given node, even the most central node, is positioned in a way to "listen" uniformly better than all other less central nodes. Typically, even a most central node might be farther than some less central node from some other important nodes. This can lead a less central node to hear some things before even the most central node, and thus to have a clearer ranking of at least some of the network before the most central node. Thus, for small T, the $\sum$ is important in Theorem 2.

THEOREM **3.** *If $p \geq 1/\lambda_1$, then as $T \to \infty$ every individual j's ranking of others under $NG(\mathbf{g}; p, T)_j$ will be according to the ranking of diffusion centrality, $DC(\mathbf{g}; p, T)$, and hence according to eigenvector centrality, $\mathbf{v}^{(1)}$.*

The intuition is that individuals hear (exponentially) more often about those who are more diffusion/eigenvector central, as the number of rounds of communication tends to infinity. As such, in the limit, they assess the rankings according to diffusion/eigenvector centrality correctly. The result implies that with very little computational ability other than remembering counts and adding, agents come to learn arbitrarily accurately complex measures of others' centralities.

More sophisticated strategies in which individuals try to infer network topology, could accelerate learning. Nonetheless, this result holds even in a minimal environment wherein individuals do not know the structure of the network and do not tag anything but the topic of conversation ("Arun has a new car").

The restriction to $p \geq 1/\lambda_1$ is important. For example, as p tends to 0, then individuals hear with vanishing frequency about others in the network, and network distance between people can matter in determining whom whom they think is the most important.

4. The Data

To investigate the theory presented above, we examine new data, coupled with detailed network data gathered from villages in rural Karnataka (India).

The network data consist of network information from each of the 75 villages. Adult members of households were surveyed.[13] We have data concerning 12 types of interactions: (1) whose houses one visits, (2) who visit one's house, (3) relatives in the village, (4) non-relatives who socialize with the respondent, (5) who gives the respondent medical advice, (6) from whom the respondent borrows money, (7) to whom the respondent lends money, (8) from whom the respondent borrows material goods (e.g., kerosene, rice), (9) to whom the respondent lends material goods, (10) from whom the respondent gets advice,

[13]We have network data from 89.14 percent of the 16,476 households based on interviews with 65 percent of all adult individuals aged 18-55.

(11) to whom the respondent gives advice, (12) with whom the respondent goes to pray (e.g., at a temple, church or mosque). Using these data, we construct 75 networks, one for each village, at the household level where a link exists between households if any member of any household is linked to any other member of any other household in at least one of the twelve ways. The resulting objects are undirected, unweighted networks at the household level.

We then asked the adults in each of 35 villages the following two additional questions:

(Loan) *If we want to spread information about a new loan product to everyone in your village, to whom do you suggest we speak?*

(Event) *If we want to spread information to everyone in the village about tickets to a music event, drama, or fair that we would like to organize in your village, to whom should we speak?*

Table 1. Summary Statistics

	mean	sd
households per village	196	61.70
household degree	17.72	(9.81)
clustering in a household's neighborhood	0.29	(0.16)
avg distance between nodes in a village	2.37	(0.33)
fraction in the ingiant	0.98	(0.01)
is a 'leader'	0.13	(0.34)
nominated someone for loan	0.48	(0.16)
nominated someone for event	0.38	(0.16)
was nominated for loan	0.05	(0.03)
was nominated for event	0.04	(0.02)
number of nominations received for loan	0.45	(3.91)
number of nominations received for event	0.34	(3.28)

Notes: for the variables nominated someone for loan (event) and was nominated for loan (event) we present the cross-village standard deviation.

Table 1 provides summary statistics of our data. The networks are sparse: the average number of households is 196 with a standard deviation of 61.7, while the average degree is 17.7 with a standard deviation of 9.8.

We see that only 5 percent of households were named in response to the Loan question (and 4 percent for Event) with a cross-village standard deviation of 3 percent. Thus there is a substantial concordance in who is named as a good initiator of diffusion within a village. In fact, conditional on being nominated, a household was nominated on average 9 times.[14]

The fact that less than half of the households were willing to name someone, is itself intriguing. Perhaps people are unwilling to offer an opinion when they are unsure of the answer[15] or possibly were afraid of offending someone given that they were asked to name just one person.

13 percent of households are "leaders": self-help group leaders, shopkeepers, teachers, etc. We use this definition as it was used by the microfinance organization BSS in our earlier work Banerjee et al. (2013) as their strategy for identifying people from which to initiate diffusion. BSS' approached such social leaders because they were a priori likely to be important in the social learning process and thereby would contribute to more diffusion of microfinance. This sort of approach is common in outreach programs and agricultural extension schemes.

5. Empirical Analysis

5.1. **How central are the nominees?**

We begin by looking at the relative network position of a given individual's nominee. To begin, we operationalize the theory with eigenvector centrality, which is a limit as $T \to \infty$ of diffusion centrality for large enough p. We return to compare what happens with other p and T, below.

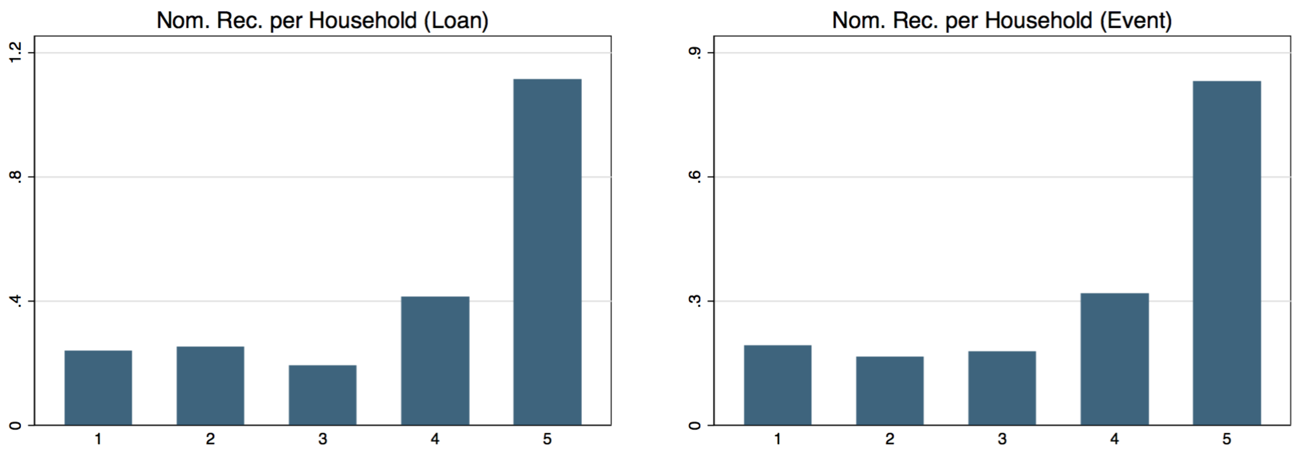
The average eigenvector centrality percentile for leaders is 0.62, consistent with the idea that leaders are more central than a typical villager. However,

[14]We work at the household level, in following with Banerjee et al. (2013) in which households were used, and so a household receives a nomination if any of its members are nominated.

[15]See Alatas et al. (2014) for a model that builds on this idea.

the average eigenvector centrality percentile for those nominated for loan is 0.71 (and 0.7 for event), above the leader level.[16]

Figure 1 shows that households with higher levels of eigenvector centrality receive more nominations on average than households of lower centrality, by a factor of about four when comparing the highest quintile to the lowest.



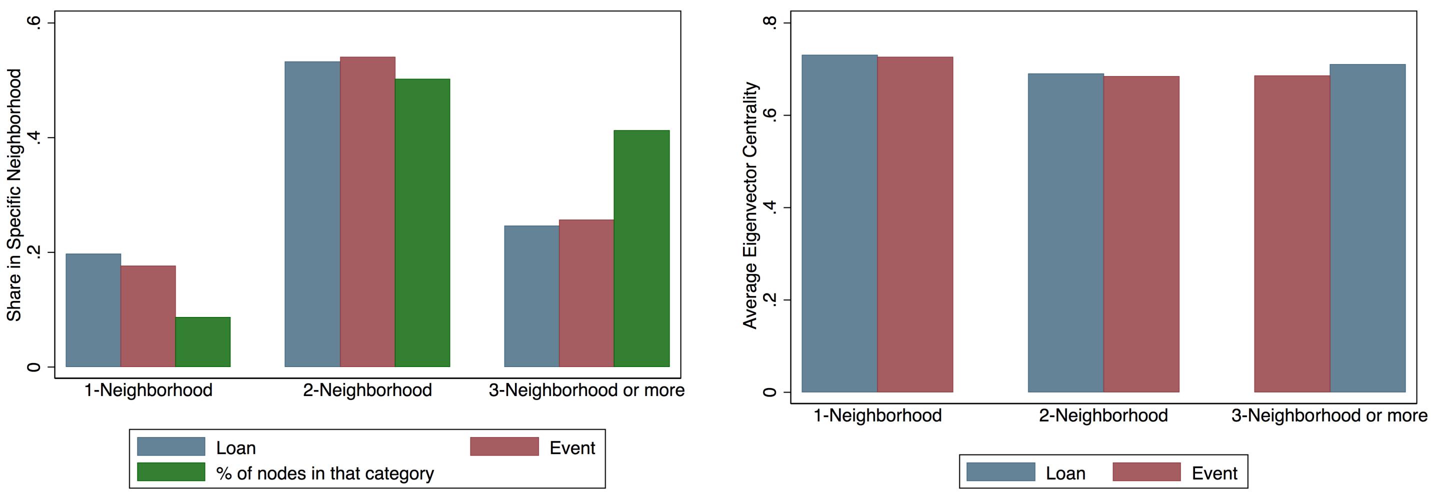
(A) Loan Question (B) Event Question

FIGURE 1. The average number of nominations per household vs the quintile of the households' eigenvector centralities.

Figure 2 presents the distribution of nominations as a function of the network distance from a given household. If information did not travel well through the social network, we might imagine that individuals would only nominate households to whom they are directly connected. Panel A of Figure 2 shows that fewer than 20 percent of individuals nominate someone within their direct neighborhood. At the same time, over 27 percent of nominations come from a network distance of at least three or more. Taken together, this suggests that information about centrality does travel through the network.

From Panel A of Figure 2, we also see that when compared to the relative frequency that individuals appear in households' first, second, and third or more degree neighborhoods (the third columns), people's nominations are

[16]Table 2 explores the difference thoroughly.



(A) Share of nominees in specified neighborhood

(B) Average eigenvector centrality percentile of nominees in specified neighborhood

FIGURE 2. Distribution of centralities of nominees

biased towards being closer than the typical household. However, they are still generally naming people outside of their immediate neighborhoods and sometimes quite far. Moreover, it is important to note that highly central individuals are generally closer to people than the typical household, so we should expect such a bias.

It is plausible that individuals may not have good information about who is central in parts of the network that are far from them. However, in Panel B of Figure 2 we see that the average eigenvector centrality percentile of those named at distance 1 is the same as at distance 2 or distance 3 or more. This suggests that individuals have reasonable and equivalently accurate information about central individuals in the community who are immediate neighbors or at greater distance from them.

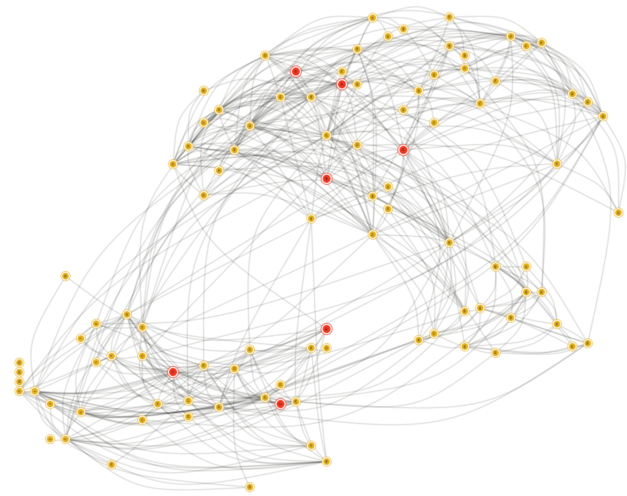
5.2. **Predicting Centrality with Nominations.** Our theoretical results suggest that people can learn others' diffusion or eigenvector centralities simply by tracking news they hear through the network. Thus, they should be able to name central individuals when asked whom to use as a "seed" for diffusion. Ultimately, such direct questions, which can easily be added to standard survey modules, could be used as a method of identifying central individuals without obtaining detailed network data. Thus we examine how well a household's nomination status correspond to its centrality. Furthermore, is being nominated a more valuable predictor than other typical data that a researcher or policymaker may have at her disposal? A policymaker may easily have access to other household characteristics such as leadership status as well as geographic position.

Figure 3 shows a village network with nominees and village leaders highlighted. Note that this is not a geographic representation of nodes, but rather, a representation using a simple algorithm to best visually represent the subcommunity structure of the network. We see in the figure that 'leaders' include central households but also some peripheral households, and that nominees appear to be more central (which we confirm with regressions below). Additionally, nodes that are both leaders and nominated are *highly* central.

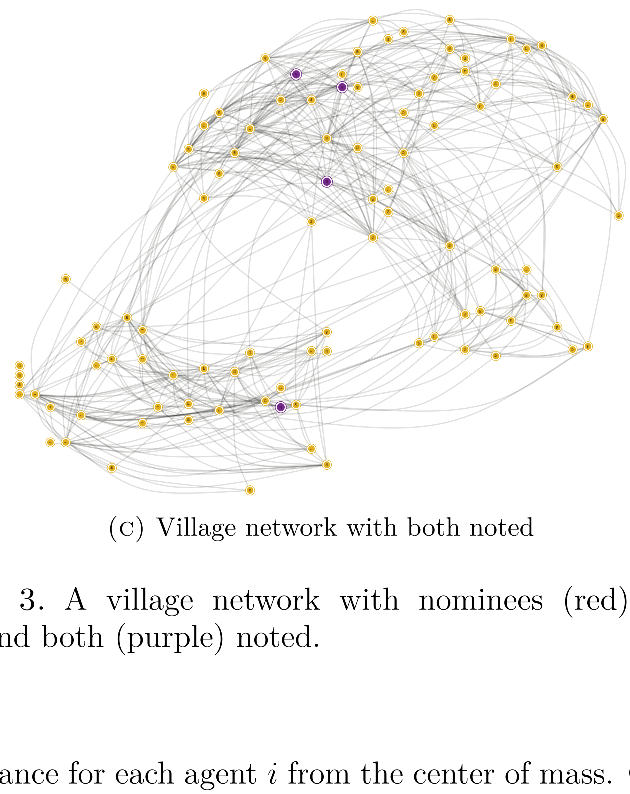
There are a priori reasons to think that leadership status and geography may be good predictors of network centrality, since, as noted in Banerjee et al. (2013), the microfinance organization selected 'leaders' precisely because they believed these people would be informationally central. The microfinance organization clearly had an interest in maximal diffusion and therefore this is the sort of logic a policymaker might use. Second, as noted in Table 1, leaders are more central than the average person in the village network. Therefore, we have data-based reasons to think that using leadership status alone may be a reasonable prediction strategy.

Similarly, geographic data is an alternative characteristic that a policymaker might use. We have detailed GPS coordinate data for every household in each village. Previous research has shown that geographic proximity strongly increases the probability of link formation (Fafchamps and Gubert, 2007; Ambrus et al., 2012; Chandrasekhar and Lewis, 2014) and therefore there are a priori grounds to think that geographic data may be a useful predictor of centrality. To operationalize geographic centrality, we use two measures. The first uses the center of mass. We compute the center of mass and then compute the

(A) Village network with leaders noted



(B) Village network with nominees noted



(C) Village network with both noted

FIGURE 3. A village network with nominees (red), leaders (blue) and both (purple) noted.

geographic distance for each agent i from the center of mass. Centrality is the inverse of this distance, which we normalize by the standard deviation of this measure by village. The second uses the geographic data to construct an adjacency matrix. We denote the ij entry of this matrix to be $\frac{1}{d(i,j)}$ where $d(\cdot,\cdot)$

is the geographic distance. Given this weighted graph, we compute the eigenvector centrality measure associated with this network. Results are robust to either definition.

In sum, both geographic and leadership data are important to consider because they are the sorts of covariates that policymakers or researchers may readily be able to collect. Therefore we ask among these candidate predictors of centrality – nomination status, leadership status, and geographic status – which is the best predictor of a household's centrality in the network?

Including these covariates is also useful for a secondary reason. It can help test whether the mechanism is different from what we suggest. For instance, perhaps people simply nominate 'leaders' within their village, or people who are central geographically, and these correlate with diffusion/eigenvector centrality.

In Table 2 we present regressions of the eigenvector centrality of a household on whether a household was nominated, leadership status, and geographic centrality of the household.

TABLE 2. How nominations, leadership status and geographic centrality predict network centrality

	(1)	(2)	(3)	(4)	(5)	(6)
	Eig. Cen.	Eig. Cen.	Eig. Cen.	Eig. Cen.	$DC(0.2, 3)$	$DC(0.2, 3)$
Nominated (Event)	0.769***		0.678***		0.687***	
	(0.073)		(0.081)		(0.082)	
Nominated (Loan)		0.604***		0.538***		0.554***
		(0.062)		(0.061)		(0.063)
Leader			0.380***	0.382***	0.404***	0.405***
			(0.045)	(0.046)	(0.041)	(0.041)
Geographic Centrality			0.013	0.012	0.000	-0.001
			(0.032)	(0.032)	(0.029)	(0.028)
Observations	6,466	6,466	5,733	5,733	5,733	5,733
R-squared	0.101	0.096	0.115	0.110	0.080	0.076
Village FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Eigenvector centrality, diffusion centrality and geographic centrality are standard deviation normalized at the village level. Clustered standard errors at the village level in parenthesis. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

In column (1) we see that being nominated for an event corresponds to a 0.77 standard deviation increase in eigenvector centrality. Column (2) shows that this is robust to looking at nomination for a loan, though the point estimate is somewhat smaller. Columns 3 and 4 show these results are robust to the inclusion of a leadership status dummy variable as well as geographic centrality. Here we see that being nominated is the best predictor of eigenvector centrality. Specifically, the increase in centrality corresponding to being nominated for an event is 79% more than the increase in centrality associated with being a leader in the village. Moreover, being nominated for an event question (or a loan question) is associated with a greater increase in centrality than being a leader and the difference is statistically significant ($p = 0.005$ and $p = 0.07$ for events and nominations, respectively). Further, geographic centrality, conditional on nominations and leadership, is not a significant predictor of the centrality of the individual. This is robust to either definition of geographic centrality. Finally, in columns 5 and 6 we show that our results are robust to looking at other values of diffusion centrality. In particular, we look at $DC(0.2, 3)$ where $p = 0.2$ is motivated from previous work (Banerjee et al., 2013)[17] and $T = 3$ is motivated below.

While Table 2 treats increases in centrality throughout the distribution symmetrically, note that we are interested in identifying (and villagers' knowledge of) *maximally* central individuals. As such, we do a related exercise in Appendix B, Table B.1. We show that being nominated corresponds to a 20pp increase in the probability that a node is in the top decile of the centrality distribution. Meanwhile being a leader corresponds to only a 9pp increase in the probability that the node is in the top decile of the centrality distribution. This suggests that nomination status is at least twice as good in terms of generating a prediction of a *highly central* household – which is the object that a policymaker would want.

Finally, we investigate which diffusion centrality measure best determines the nomination decision. At its core this is a discrete choice: an individual

[17]This comes from the average of the transmission probability parameters estimated in Banerjee et al. (2013): $(0.35 + 0.05)/2$.

chooses to nominate an individual as a function of possibly demographic characteristics, network position, and unobservables. Our theory suggests that a leading factor should be network position as captured by diffusion centrality. A plausible alternative – beyond correlated unobservables – may be that individuals have some idea of the number of friends an individual has. There are two challenges in estimating such discrete choices in our data. First, the two extremes of diffusion centrality – degree and eigenvector – are correlated (0.91 when looking at the measures standardized by the standard deviation within villages). This, of course, depends on the network topology and happens to be true of our data. Second, a simple multinomial logistic approach will not suffice because we have $n-1$ possible nominees and n observations of nominators and an asymptotically growing choice set at the same rate as the number of observations presents statistical problems. Thus, we directly examine which centrality measure predicts nominations via regressions of the number of nominations on centrality and other node characteristics (leadership status, geography, using village fixed effects).

Table 3. Which centralities explain the number of nominations

	(1) Event	(2) Loan	(3) Event	(4) Loan	(5) Event	(6) Loan	(7) Event	(8) Loan	(9) Event	(10) Loan
Degree Centrality	0.130	0.179	0.099	0.171	0.098	0.149	-0.031	-0.003		
	(0.105)	(0.107)	(0.104)	(0.115)	(0.119)	(0.134)	(0.205)	(0.234)		
Eigenvector Centrality	0.169*	0.218**	0.213**	0.242**	0.208*	0.257*			0.056	0.058
	(0.100)	(0.101)	(0.102)	(0.111)	(0.116)	(0.129)			(0.164)	(0.179)
$DC(0.2,3)$							0.332	0.404*	0.247	0.344*
							(0.203)	(0.228)	(0.173)	(0.186)
Leader					0.291	0.470*	0.288	0.468*	0.289	0.469*
					(0.178)	(0.232)	(0.178)	(0.232)	(0.179)	(0.233)
Geographic Centrality					-0.014	-0.096*	-0.014	-0.096*	-0.014	-0.096*
					(0.082)	(0.054)	(0.082)	(0.055)	(0.081)	(0.054)
Observations	6,466	6,466	6,466	6,466	5,733	5,733	5,733	5,733	5,733	5,733
R-squared	0.008	0.010	0.010	0.011	0.012	0.013	0.012	0.013	0.012	0.013
Village FE	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Outcome variable is the number of nominations for either the event or loan question. Degree centrality, eigenvector centrality and $DC(0.2,3)$ are normalized by within-village standard deviation. Clustered standard errors at the village level in parenthesis. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

In Table 3, we find that eigenvector centrality is a significant driver of the number of nominations and degree cannot be separated from zero (Table 3). A one standard deviation in eigenvector centrality is associated with a 0.208

increase in the number of nominations under the event question (statistically significant at the 10% level) as compared to a 0.098 increase when looking at degree (not statistically significant, Column 5). However, note that parameter estimates are noisy and it cannot be rejected that coefficients on degree and eigenvector centralities are statistically the same, despite the vast difference in the point estimate.

Looking at Columns 7-10, we add $DC(p, T)$ to the mix. We set $p = 0.2$ as discussed above and we pick $T = 3$ to contrast with degree and eigenvector centrality. An a priori reason to set $T = 3$ is that the average distance between nodes in our network is 2.7 and thus 3 represents the distance where information travels before it begins to echo backwards, whereas large T levels allow for the recurrent information to matter more. We find that $DC(0.2, 3)$ is a crucial driver of the number of nominations as compared to either degree or eigenvector centrality. A one standard deviation increase is associated with a 0.3 to 0.4 increase in the number of nominations as compared to a near-zero effect for eigenvector and degree. The effects are statistically significant at the 10% level for loans and have p-values of 0.11 and 0.15 for events. However, again given the amount of noise, we are unable to reject that degree centrality and $DC(0.2, 3)$ or eigenvector centrality and $DC(0.2, 3)$ have the same effects on the outcome variables of interest, despite the enormous differences in point estimates. Finally, we note that leadership status also matters – as would be expected – in explaining the number of nominations received by a household.

Taken together, this provides only suggestive evidence that a key driver of the nomination decision involves a diffusion centrality metric with $T > 1$. However, given the high degree of correlation of these metrics in our sample, it is unsurprising that we are unable to cleanly separate the measures. Regardless, it is still clear that individuals are able to accurately name highly central individuals, well beyond using other status variables.

6. Concluding Remarks

Our model illustrates that it should be easy for even very myopic and non-Bayesian agents, simply by counting, to have an idea as to who is central in their community (according to fairly complex definitions). Motivated by this, we asked villagers to identify central individuals in their village. They do not simply name locally central individuals, but actually name ones that are globally central within the village. This suggests that individuals may use simple protocols to learn valuable things about the complex systems within which they are embedded. This suggests a rich agenda for further research, as one can explore which other aspects of agents' social environments can be learned in simple ways.

It is also worth commenting that our work focuses on the network-based mechanics of communication. In practice, more than simple network position may determine who the 'best' person is to spread information, as other characteristics may affect the quality and impact of communication. It might actually be that the people nominate individuals who are even *better* than the most central individual in the network. This is another important issue for future research.

References

Alatas, V., A. Banerjee, A. G. Chandrasekhar, R. Hanna, and B. A. Olken (2014): "Network structure and the aggregation of information: Theory and evidence from Indonesia," . 15

Ambrus, A., M. Mobius, and A. Szeidl (2012): "Consumption risk-sharing in social networks," *NBER Working Paper*. 5.2

Ballester, C., A. Calvó-Armengol, and Y. Zenou (2006): "Who's who in networks, wanted: the key player," *Econometrica*, 74, 1403–1417. 1

Banerjee, A., E. Breza, A. Chandrasekhar, E. Duflo, and M. Jackson (2014): "Come play with me: information diffusion about rival goods," *mimeo: MIT, Columbia, and Stanford*. 3.1

Banerjee, A., A. Chandrasekhar, E. Duflo, and M. Jackson (2013): "Diffusion of Microfinance," *Science*, 341, DOI: 10.1126/science.1236498, July 26 2013. (document), 1, 1, 2.2, 2.3, 3.1, 4, 14, 5.2, 5.2, 17

Borgatti, S. P. (2005): "Centrality and network flow," *Social Networks*, 27, 55 – 71. 1

Chandrasekhar, A. and R. Lewis (2014): "Econometrics of sampled networks," Stanford working paper. 5.2

Dodds, P. S., R. Muhamad, and D. J. Watts (2003): *Science*, 301, 827–829. 2

Fafchamps, M. and F. Gubert (2007): "The formation of risk sharing networks," *Journal of Development Economics*, 83, 326–350. 5.2

Golub, B. and M. Jackson (2010): "Naive Learning in Social Networks and the Wisdom of Crowds," *American Economic Journal: Microeconomics*, 2, 112–149. 10

Jackson, M. (2008): *Social and economic networks*, Princeton: Princeton University Press. 8

Jackson, M. and L. Yariv (2011): "Diffusion, strategic interaction, and social structure," *Handbook of Social Economics, San Diego: North Holland, edited by Benhabib, J. and Bisin, A. and Jackson, M.O.* 5

Katz, E. and P. Lazarsfeld (1955): *Personal influence: The part played by people in the flow of mass communication*, Free Press, Glencoe, IL. 1

Kempe, D., J. Kleinberg, and E. Tardos (2003): "Maximizing the Spread of Influence through a Social Network," *Proc. 9th Intl. Conf. on Knowledge Discovery and Data Mining*, 137 – 146. 1

——— (2005): "Influential Nodes in a Diffusion Model for Social Networks," *In Proc. 32nd Intl. Colloq. on Automata, Languages and Programming*, 1127 – 1138. 1

Milgram, S. (1967): "The small world problem," *Psychology today.* 2

Rogers, E. (1995): *Diffusion of Innovations*, Free Press. 1

Appendix A. Proofs

The following lemma is used in proofs of the theorems.

Lemma 1. *Consider a positive and diagonalizable* $\mathbf{g}$*. Then* $\mathbf{g}$ *has a unique largest eigenvalue. Moreover, letting* $\widetilde{\mathbf{g}} = \mathbf{g}/\lambda_1$*,*

$$[\widetilde{g}^T]_{ij} \to_T v_i^{(1)} v_j^{(1)}.$$

This means that the columns of $\widetilde{g}^T$ *are converging to the first eigenvector* $\mathbf{v}^{(1)}$ *times a constant scale for each column,* $v_j^{(1)}$*.*

Proof of Lemma 1. The uniqueness of the largest eigenvalue follows from the Perron-Frobenius Theorem. By diagonalizability we can write $\mathbf{g}$ as

$$\mathbf{g} = \mathbf{V}\Lambda\mathbf{V}^{-1}$$

where Λ is the matrix with eigenvectors on the diagonal (ordered from 1 to n in order of magnitude) and $\mathbf{V}$ is the matrix with rows equal to the left hand eigenvectors.[18] By normalizing $\widetilde{\mathbf{g}} = \mathbf{g}/\lambda_1$ we have $\widetilde{\mathbf{g}} = \mathbf{V}\widetilde{\Lambda}\mathbf{V}^{-1}$. Note that $\widetilde{\Lambda} = \text{diag}\left\{1, \widetilde{\lambda}_2, ..., \widetilde{\lambda}_n\right\}$. Since the largest eigenvalue is unique, it also follows that $0 \leq \left|\widetilde{\lambda}_k\right| < 1$ for $k > 1$.

It helps to write $\mathbf{V} = \left[v_i^{(j)}\right]_{i,j}$. Then

$$\widetilde{\mathbf{g}}^T = \mathbf{V}\widetilde{\Lambda}^T\mathbf{V}^{-1} = \left[v_j^{(k)}\widetilde{\lambda}_k^T\right]_{j,k} V' = \left[\sum_k v_i^{(k)} v_j^{(k)} \widetilde{\lambda}_k^T\right]_{i,j}.$$

Then $\widetilde{\mathbf{g}}^T$ is given by a matrix with jth column entries and ith row entries:

$$\sum_k v_i^{(k)} \widetilde{\lambda}_k^T v_j^{(k)} = v_i^{(1)} v_j^{(1)} + \sum_{k \geq 2}^{n} v_i^{(k)} \widetilde{\lambda}_k^T v_j^{(k)} \to_{T\to\infty} v_i^{(1)} v_j^{(1)},$$

as claimed. ■

Proof of Theorem 1. We show the second statement as the others follow directly.

First, note that in any neighborhood of any nonnegative matrix $\mathbf{g}$, there exists a positive and diagonalizable matrix $\mathbf{g}'$. Next, consider any nonnegative

[18] Note that $\mathbf{V}^{-1}$ is not only $\mathbf{V}$'s inverse, but is also the matrix with columns equal to the right hand eigenvectors.

$\mathbf{g}$. If the statement holds for any arbitrarily close positive and diagonalizable $\mathbf{g}'$, then since $\frac{DC(\mathbf{g};p,T)}{\sum_{t=1}^{T}(p\lambda_1)^t}$ is a continuous function (in a neighborhood of a non-negative and strongly connected $\mathbf{g}$) as is the first eigenvector, the statement also holds at $\mathbf{g}$. Thus, it is enough to prove the result for a positive and diagonalizable $\mathbf{g}$, as in what follows.

In fact, we show the following strengthening of the statement for a positive and diagonalizable $\mathbf{g}$.

- The 'tail terms' approach eigenvector centrality:
$$(\mathbf{g}/\lambda_1)^t \cdot \mathbf{1} \to \mathbf{v}^{(1)} \text{ as } t \to \infty,$$
where $\mathbf{v}^{(1)}$ is the eigenvector corresponding to the largest eigenvalue.
- Diffusion centrality with $p > \lambda_1^{-1}$ is ratio-consistent for eigenvector centrality:
$$\lim_{T\to\infty} \frac{DC\left(\mathbf{g};p,T\right)}{\sum_{t=1}^{T}\left(p\lambda_1\right)^t} = \lim_{T\to\infty} \frac{DC\left(g;p,T\right)}{\frac{p\lambda_1-(p\lambda_1)^{T+1}}{1-(p\lambda_1)}} = \mathbf{v}^{(1)},$$
and, also with $p = \lambda_1^{-1}$,
$$\lim_{T\to\infty} \frac{1}{T} DC\left(\mathbf{g};\lambda_1^{-1},T\right) = \mathbf{v}^{(1)}.$$

By diagonalizability we can write

$$\mathbf{g} = \mathbf{V}\Lambda\mathbf{V}^{-1}$$

and by normalizing $\widetilde{\mathbf{g}} = \mathbf{g}/\lambda_1$ it follows that $\widetilde{\mathbf{g}} = \mathbf{V}\widetilde{\Lambda}\mathbf{V}^{-1}$. Normalize the eigenvectors in ℓ_1, $\sum_j v_j^{(1)} = 1$, so the columns sum to 1. Then, applying Lemma 1, it follows that $\widetilde{\mathbf{g}}^T \cdot \mathbf{1}$ is such that

$$\widetilde{\mathbf{g}}^T \cdot \mathbf{1} \to \mathbf{v}^{(1)}.$$

This completes the proof of the first statement.

Next we turn to the sum $\sum_{t=1}^{\infty} \widetilde{\mathbf{g}}^t \cdot \mathbf{1}$ and show the second statement for the case where $p = 1/\lambda_1$. It is sufficient to show

$$\lim_{T\to\infty} \left\| \frac{DC\left(\mathbf{g};\lambda_1^{-1},T\right)}{T} - \mathbf{v}^{(1)} \right\| = 0.$$

This follows from the fact that

$$\left|\frac{DC_i\left(\mathbf{g};\lambda_1^{-1},T\right)}{T}-v_i^{(1)}\right| = \left|\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{n}\sum_{k\geq 2}^{n} v_i^{(k)}v_j^{(k)}\widetilde{\lambda}_k^t\right| \leq \frac{1}{T}\sum_{t=1}^{T}\sum_{k\geq 2}^{n} 1\cdot\underbrace{\left|\sum_{j=1}^{n} v_j^{(k)}\right|}_{\leq 1}\cdot\left|\widetilde{\lambda}_k^t\right|$$

$$\leq \frac{n}{T}\sum_{t=1}^{T}\left|\widetilde{\lambda}_2^t\right| = \frac{n}{T}\frac{\left|\widetilde{\lambda}_2\right|}{1-\left|\widetilde{\lambda}_2\right|}\left(1-\left|\widetilde{\lambda}_2\right|^T\right)\to 0.$$

Since the length of the vector (which is n) is unchanging in T, pointwise convergence implies convergence in norm, proving the result.

The final piece repeats the argument for $p > 1/\lambda_1$, but now uses the definition $\widetilde{\mathbf{g}} = (p\mathbf{g})$. Then we have $\widetilde{\Lambda} = \text{diag}\left\{\widetilde{\lambda}_1, ..., \widetilde{\lambda}_n\right\}$ with $p\lambda_k = \widetilde{\lambda}_k$. We show[19]

$$\lim_{T\to\infty}\left\|\frac{DC\left(\mathbf{g};p,T\right)}{\sum_{t=1}^{T}\left(p\lambda_1\right)^t}-\mathbf{v}^{(1)}\right\| = 0.$$

Again it is easy to see that

$$\left|\frac{DC_i\left(\mathbf{g};\lambda_1^{-1},T\right)}{\sum_{t=1}^{T}\widetilde{\lambda}_1^t}-v_i^{(1)}\right| = \left|\frac{1}{\sum_{t=1}^{T}\widetilde{\lambda}_1^t}\sum_{t=1}^{T}\sum_{j=1}^{n}\sum_{k\geq 2}^{n} v_i^{(k)}v_j^{(k)}\widetilde{\lambda}_k^t\right| \leq \frac{1}{\sum_{t=1}^{T}\widetilde{\lambda}_1^t}\sum_{t=1}^{T}\sum_{k\geq 2}^{n} 1\cdot\left|\sum_{j=1}^{n} v_j^{(k)}\right|\cdot\left|\widetilde{\lambda}_k^t\right|$$

$$\leq \frac{n}{\sum_{t=1}^{T}\widetilde{\lambda}_1^t}\sum_{t=1}^{T}\left|\widetilde{\lambda}_2^t\right|\to 0$$

which completes the argument. ■

Proof of Theorem 2. Recall that $\mathbf{H} = \sum_{t=1}^{T}\left(p\mathbf{g}\right)^t$ and $DC = \left(\sum_{t=1}^{T}\left(p\mathbf{g}\right)^t\right)\cdot\mathbf{1}$ and so

$$DC_i = \sum_j H_{ij}.$$

Additionally,

$$\text{cov}(DC, H_{\cdot,j}) = \sum_i\left(DC_i - \sum_k\frac{DC_k}{n}\right)\left(H_{ij}-\sum_k\frac{H_{kj}}{n}\right).$$

[19]Note that it is important that $p \geq 1/\lambda_1$ in the proof, since if $p < 1/\lambda_1$, then $p\lambda_1 < 1$. In this case, observe that

$$\frac{\sum_{t=1}^{T}|\tilde{\lambda}_2|^t}{\sum_{t=1}^{T}\tilde{\lambda}_1^t} = \frac{\tilde{\lambda}_2}{\tilde{\lambda}_1}\cdot\frac{1-\tilde{\lambda}_1}{1-\tilde{\lambda}_2}$$

by the properties of a geometric sum, which is of constant order. Thus, higher order terms ($\tilde{\lambda}_2$, etc.) persistently matter and are not dominated relative to $\sum_t^T\tilde{\lambda}_1^t$.

Thus

$$\sum_j \operatorname{cov}(DC, H_{\cdot,j}) = \sum_i \left(DC_i - \sum_k \frac{DC_k}{n} \right) \left(\sum_j H_{ij} - \sum_k \frac{\sum_j H_{kj}}{n} \right),$$

implying

$$\sum_j \operatorname{cov}(DC, H_{\cdot,j}) = \sum_i \left(DC_i - \sum_k \frac{DC_k}{n} \right) \left(DC_i - \sum_k \frac{DC_k}{n} \right) = \operatorname{var}(DC),$$

which completes the proof. ∎

Proof of Theorem 3. Again, we prove the result for a positive diagonalizable $\mathbf{g}$, noting that it then holds for any (nonnegative) $\mathbf{g}$.

Again, let $\mathbf{g}$ be written as

$$\mathbf{g} = \mathbf{V}\Lambda\mathbf{V}^{-1}$$

and let $\widetilde{\mathbf{g}} = \mathbf{g}/\lambda_1$. Then as in Lemma 1,

$$\widetilde{\mathbf{g}}^T = \mathbf{V}\widetilde{\Lambda}^T\mathbf{V}^{-1} = \left[v_j^{(k)} \widetilde{\lambda}_k^T \right]_{j,k} V' = \left[\sum_k v_i^{(k)} v_j^{(k)} \widetilde{\lambda}_k^T \right]_{i,j}.$$

Then $\widetilde{\mathbf{g}}$ is given by a matrix with jth column entries and ith row entries:

$$\sum_k v_i^{(k)} \widetilde{\lambda}_k^T v_j^{(k)} = v_i^{(1)} v_j^{(1)} + \sum_{k\geq 2}^{n} v_i^{(k)} \widetilde{\lambda}_k^T v_j^{(k)} \to_{T\to\infty} v_i^{(1)} v_j^{(1)},$$

and so the columns of $\widetilde{g}^T$ are approximated by the first eigenvector $\mathbf{v}^{(1)}$ times a constant scale for each column, $v_j^{(1)}$.

Recall that

$$\mathbf{H} = \sum_{t=1}^{T} (p\mathbf{g})^t .$$

Then, let $\tilde{\lambda}_k = p\lambda_k$ as a normalization. By the ordering of the eigenvalues,

$$\begin{aligned}\mathbf{H}_{\cdot,j} &= \sum_{t=1}^{T}\left[\mathbf{v}^{(1)}v_j^{(1)}\tilde{\lambda}_1^t + \mathbf{v}^{(2)}v_j^{(2)}\left|\tilde{\lambda}_2\right|^t + O\left(\left|\tilde{\lambda}_2\right|^t\right)\right] \\ &= \sum_{t=1}^{T}\left[\mathbf{v}^{(1)}v_j^{(1)}\tilde{\lambda}_1^t + O\left(\left|\tilde{\lambda}_2\right|^t\right)\right] \\ &= \mathbf{v}^{(1)}v_j^{(1)}\sum_{t=1}^{T}\tilde{\lambda}_1^t + O\left(\sum_{t=1}^{T}\left|\tilde{\lambda}_2\right|^t\right).\end{aligned}$$

So

$$\frac{\mathbf{H}_{\cdot,j}}{\sum_{t=1}^{T}\tilde{\lambda}_1^t} = \mathbf{v}^{(1)}v_j^{(1)} + O\left(\frac{\sum_{t=1}^{T}\left|\tilde{\lambda}_2\right|^t}{\sum_{t=1}^{T}\tilde{\lambda}_1^t}\right) = \mathbf{v}^{(1)}v_j^{(1)} + o\left(1\right).$$

This completes the proof since the ranking matrix based on **H** therefore has identical columns in the limit.[20] ■

[20]The discussion in Footnote 19 clarifies why $p > 1/\lambda_1$ is required for the argument.

Appendix B. Supplementary Material

Table B.1. How nominations, leadership status and geographic centrality predict being in the top decile of network centrality

	(1) Eig. Cen.	(2) Eig. Cen.	(3) Eig. Cen.	(4) Eig. Cen.	(5) $DC(0.2, 3)$	(6) $DC(0.2, 3)$
Nominated (Event)	0.205***		0.195***		0.195***	
	(0.026)		(0.028)		(0.026)	
Nominated (Loan)		0.150***		0.143***		0.147***
		(0.024)		(0.025)		(0.023)
Leader			0.088***	0.089***	0.091***	0.092***
			(0.014)	(0.014)	(0.014)	(0.014)
Geographic Centrality			0.007	0.006	0.007	0.007
			(0.008)	(0.008)	(0.008)	(0.008)
Observations	6,466	6,466	5,733	5,733	5,733	5,733
R-squared	0.017	0.011	0.028	0.022	0.029	0.024
Village FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Eigenvector centrality and diffusion centrality are dummies for the upper decile of the village distribution. gps centrality are standard deviation normalized at the village level. Clustered standard errors at the village level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.