GOSSIP: IDENTIFYING CENTRAL INDIVIDUALS IN A SOCIAL NETWORK

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

Abstract. Is it possible, simply by asking a few members of a community, to identify individuals who are best placed to diffuse information? A simple model of diffusion shows how boundedly rational individuals can, just by tracking gossip about people, identify those who are most central in a network according to "*diffusion centrality*" (a measure of network centrality which nests existing ones, and predicts the extent to which piece of information seeded to a network member diffuses in finite time). Using rich network data from 35 Indian villages, we find that respondents accurately nominate those who are diffusion central – not just traditional leaders or those with many friends. In a subsequent randomized field experiment in 213 villages, we track the diffusion of a piece of information initially given to a small number of "seeds" in each community. Seeds who are nominated by others lead to a near tripling of the spread of information relative to randomly chosen seeds. Diffusion centrality accounts for some, but not all, of the extra diffusion from these nominated seeds compared to other seeds (including those with high social status) in our experiment.

JEL Classification Codes: D85, D13, L14, O12, Z13

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

---

Date: This Version: February 14, 2016.

Michael Dickstein, Ben Golub, John Moore, and participants at various seminars/conferences provided helpful comments. Financial support from the NSF under grants SES-1156182 and 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 are thank Shobha Dundi, Devika Lakhote, Tithee Mukhopadhyay, and Gowri Nagraj for excellent research assistance.

[†]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 Dalí

Knowing who is influential, or central, in a community is important to anyone who wants to influence the choices made by community members. In particular, the extent to which a piece of information diffuses among a population often depends on how central the initially informed are within the network (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)). Policymakers, businesses and other organizations can thus benefit from targeting the right individuals for spreading valuable information.

However, learning who is central in a social network can be costly. For policymakers, collecting detailed network data is costly, and easy "fixes" (such as asking the traditional leaders, or geographically central households) may not identify people who are actually very central (see Beaman, BenYishay, Magruder, and Mobarak (2014); Banerjee, Chandrasekhar, Duflo, and Jackson (2013)). Even for members of the community, knowledge of the network structure beyond their immediate friends is far from automatic. In fact, individuals within a network tend to have little perspective on its structure, as found in important early research by Friedkin (1983) and Krackhardt (1987), among others (see Krackhardt (2014) for background and references). However, we can still ask whether, despite not knowing the network structure, people can make reliable guesses about who is central to the network and who, more specifically, is particularly well placed to diffuse information through the network. In this paper, we answer these questions both theoretically and empirically – finding positive answers in each case.

First, we develop a simple model, building 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 even *without knowing anything about the structure of the network*. Our model is about a process we call "gossip", where nodes generate pieces of information that are then stochastically passed from neighbor to neighbor, along with the identity of the node where it started. We assume only that individuals who hear the gossip are able to keep count of the

number of times each person in the network is mentioned as a source.[1] We show that for any listener in the network, the relative ranking under this count converges over time to the correct ranking of every node's propensity to send information to the rest of the network. The specific measure of a node's ability to send information that we use is given by its "diffusion centrality," introduced in Banerjee et al. (2013), which answers the question of how widely information from a given node diffuses in a given number of time periods and for a given random per-period transmission probability.[2]

In short, by listening and keeping count of how often they hear *about* someone, individuals learn the correct ranking of community members from the point of view of how effectively they can serve as a source of information to the rest of the community.

Second, we use a unique dataset to assess whether this holds empirically. 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 combine their answers (which we call their "nominations") with detailed network data that include maps of a variety of interactions in each of the 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 are significantly correlated with diffusion centrality even after conditioning on these characteristics. Finally, a LASSO regularization technique (Tibshirani (1996); Belloni and Chernozhukov (2009); Belloni et al. (2014b,a)) picks out diffusion centrality as the only relevant variable to predict the number of nominations, out of five possible measures of network positions (diffusion centrality, degree, eigenvector centrality, traditional leadership status and geographic centrality).

Thus, our model shows that it is possible for individuals to learn who are the most central people in their network, and our empirical work suggests that they do so. This data, of course, could still be consistent with other models of how people choose individuals to nominate, and we explore some of these alternatives in our analysis, showing that the nominees' centrality is significant in determining nominations well beyond other geographic and sociological attributes of the nominees. The correlation

[1]We use the term "gossip" to refer to the spreading of information about particular people. Our diffusion process is focused on basic information that is not subject to the biases or manipulations that might accompany some "rumors" (e.g., see Bloch, Demange, and Kranton (2014)).

[2]This measure of centrality nests three of the most prominent 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). For intermediate numbers of periods, diffusion centrality takes on a wide range of other values.

of nominations and centrality may have nothing to do with diffusion, and so our next step is to test whether the nominees are, indeed, good diffusers.

Third, to test this prediction, we conduct an additional large randomized field experiment in 213 different villages to test whether informing nominated individuals leads to wider diffusion of information than informing either randomly selected individuals or village elders. Our experiment consists of three pieces. In 71 villages we seed a piece information in 3 to 5 randomly selected households (the number of seeds to be reached was randomly selected). In 71 other villages we seed information in 3 to 5 village households who have status as "elders" in the village – leaders with a degree of authority in the community, who command great respect. Finally, in the remaining 71 villages, we seed information in 3 to 5 individuals nominated by others as being well suited to spread information ("gossip nominees"). The piece of information we want people to learn is very simple: anyone who calls a particular number will be entered in a raffle for a free cell phone and other cash prizes. The chance to win a prize is independent of the number of people who enter the raffle, ensuring that the information is non-rivalrous. The call itself is free. We then measure the extent of diffusion using the number of independent entries to the raffle. We get almost three times as many entries when we seed information with gossip nominees as compared to seeding with village elders or with random villagers (elders outperform random villagers). Furthermore, in a subsample of 69 out of the 71 villages where we did random seeding, we collect full network information. We find that the random seeds that happen to have high diffusion centrality and/or to be gossips lead to more diffusion of information. However, even controlling for diffusion centrality of a seed, gossip (nominated) seeds still provide greater information diffusion, suggesting that people do even better than we can at choosing initial seeds based on only diffusion centrality. This may be because they are incorporating other information, such as who is trustworthy or who is most charismatic or talkative, which may not be picked up in the pure network data. Or it may be because our measures of centrality are noisy and villagers are even more accurate at finding central individuals than we are.

To our knowledge, this is the first paper to demonstrate that members of communities are able, easily and accurately, to nominate people in the community who are good at diffusing information, and that these nominees are highly central in a network sense. It is also the first to describe a simple process by which people can

learn things about of their broader network that they have no direct access to.[3] Our results have important practical consequences, since policy makers and businesses are often looking for the best way to spread a given piece information, and asking people to identify the best person to spread the information is cheaper and easier than collecting data on the entire network.

The remainder of the paper is organized as follows. Section 2 develops the model of diffusion. In Section 3 we relate the notion of diffusion centrality to network gossip. Section 4.1 describes the setting and the data used in the empirical analysis. We examine whether individuals nominate central nodes in Section 4.2. Section 5 describes the experiment and the results. Section 6 concludes.

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 possibly directed[4] and weighted network, which has an adjacency matrix $\mathbf{g} \in \{0,1\}^{n\times n}$. Unless otherwise stated, we take the network $\mathbf{g}$ to be fixed and let $v^{(1)}$ be its first (right-hand) eigenvector, corresponding to the largest eigenvalue λ_1.[5] The first eigenvector is nonnegative and real-valued by the Perron-Frobenius Theorem.

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.[6]

2.2. **Diffusion Centrality.** In Banerjee, Chandrasekhar, Duflo, and Jackson (2013), we defined a notion of centrality called *diffusion centrality*, based on random information flow through a network according to the following process, which is a variant of the standard diffusion process that underlies many models of contagion.[7]

A piece of information is initiated at node i and then broadcast outwards from that node. In each period, with probability $q \in (0,1]$, independently across neighbors and

[3]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 test knowledge about peoples' own connections.

[4]When defining $\mathbf{g}$ in the directed case, the ij-th entry should indicate that i can tell something to j. In some networks, this may not be reciprocal.

[5]$v^{(1)}$ is such that $\mathbf{g}v^{(1)} = \lambda_1 v^{(1)}$ where λ_1 is the largest eigenvalue of $\mathbf{g}$ in magnitude.

[6]More generally, everything we say applies to the components of the network.

[7]See Jackson and Yariv (2011) for background and references. A continuous time version of diffusion centrality appears in Lawyer (2014).

history, each informed node informs each of its neighbors of the piece of information and the identity of its original source. The process operates for T periods, where T is a positive integer.

There are many reasons to allow T to be finite. For instance, a new piece of information may only be "news" for a limited time. After while boredom sets in or some other news arrives and the topic of conversation changes. By allowing for a variety of T's, diffusion centrality admits important finite-horizon cases, as well as more extreme cases where agents discuss a topic indefinitely.[8]

Diffusion centrality measures how extensively the information spreads as a function of the initial node. In particular, let

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

be the "hearing matrix." The ij-th entry of $\mathbf{H}$, $H(\mathbf{g}; q, T)_{ij}$, is the expected number of times, in the first T periods, that j hears about a piece of information originating from i. Diffusion centrality is then defined by

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

So, $DC(\mathbf{g}; q, T)_i$ is the expected total number of times that some piece of information that originates from i is heard by any of the members of the society during a T-period time interval.[9] Banerjee et al. (2013) showed that diffusion centrality of the initially informed member of a community was a statistically significant predictor of the spread of information – in that case, about a microfinance program.

Note that this measure allows people to hear the information multiple times from the same person and count those times as distinct reports, 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

[8]Of course this is an approximation and, moreover, different people may have different incentives to pass news, or time horizons over which they do so. The current model and definition already moves beyond the literature, but richer extensions would be easy to study.

[9] We note two useful normalizations. One is to compare it to what would happen if $q = 1$ 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 $DC(\mathbf{g}; q, T)$ 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 $DC(\mathbf{g}; q, T)$ by $\sum_i DC_i(\mathbf{g}; q, T)$. This normalization tracks how relatively diffusive one node is compared to the average diffusiveness in its society.

that tracks the expected number of informed nodes and avoids double-counting, our expression is *much* easier to calculate. Second, for many parameter values, the two measures are roughly proportional to each other. Third, 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. Finally, in a world in which multiple chances to hear the same thing lead to a greater probability of information retention, this count might be a better predictor of actual learning.[10]

2.3. **Properties of Diffusion Centrality.** It is useful to first remind the reader of diffusion centrality's relationship relative to other prominent measures of centrality in the literature, though a reader impatient to see our main results is welcome to bypass this sub-section and return to it at a later stage. As we state in Banerjee et al. (2013), as T is varied, diffusion centrality nests three of the most prominent and widely used centrality measures: degree centrality, eigenvector centrality, and Katz-Bonacich centrality. [11] It thus provides a foundation for these measures and spans between them.

In particular, it is straightforward to see that (i) diffusion centrality is proportional to degree centrality at the extreme at which $T = 1$, and (ii) if $q < 1/\lambda_1$, then diffusion centrality coincides with Katz-Bonacich centrality if we set $T = \infty$. Also, when $q > 1/\lambda_1$ diffusion centrality approaches eigenvector centrality as T approaches ∞.[12] For completeness, a proof of the last claim and a formal statement of these results appears in the appendix.

Between these extremes, diffusion centrality measures how diffusion process operates for some limited number of periods. As shown in Banerjee et al. (2013), the behavior in the intermediate ranges can be more relevant for certain diffusion phenomena than either extreme.

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

[11]Let $d(\mathbf{g})$ denote degree centrality: $d_i(\mathbf{g}) = \sum_j g_{ij}$. Eigenvector centrality corresponds to $v^{(1)}(\mathbf{g})$: the first eigenvector of $\mathbf{g}$. Also, let $KB(\mathbf{g}, q)$ denote Katz-Bonacich centrality – defined for $q < 1/\lambda_1$ by $KB(\mathbf{g}, q) := \left(\sum_{t=1}^{\infty} (q\mathbf{g})^t\right) \cdot \mathbf{1}$.

[12]It is useful to note that the difference between the extremes of Katz-Bonacich centrality and eigenvector centrality depends on whether q is sufficiently small so that limited diffusion takes place even in the limit of large T, or whether q is sufficiently large so that the knowledge saturates the network and then it is only relative amounts of saturation that are being measured. Saturation occurs when the entries of $\left(\sum_{t=1}^{\infty} (q\mathbf{g})^t\right) \cdot \mathbf{1}$ 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.

Exactly how should one choose the "right" q and T? Clearly this must be context dependent and should be treated as an empirical question. In some settings, people interact or communicate all the time and so q will be high, while in others their contact may be more limited, corresponding to a lower q. Likewise, there may be some things that are long lasting in terms of discussion or diffusion, corresponding to a high T, while others quickly subside, leading to a low q. Thus, the answer will be determined by the specifics of the context and setting of the application.

Despite the fact that the "right" answer is context dependent, it is useful to identify critical levels of q, T that differentiate the varying regimes of behavior of diffusion centrality. Our earlier results relating diffusion centrality to other standard measures at its extremes do not tell us at what levels of q, T we see fundamental changes in the measure's behavior. Here we provide some theoretical results on diffusion centrality that show that diffusion centrality behaves fundamentally differently depending on whether q is above or below $1/\lambda_1$ (the inverse of the first eigenvalue of $\mathbf{g}$), and whether T is smaller or bigger than the diameter of the graph. We use these to suggest that the threshold case of $q = 1/\mathrm{E}[\lambda_1]$ and $T = \mathrm{E}[Diam(\mathbf{g})]$ provides a natural benchmark value for these parameters.

The intuition behind these thresholds is as follows. Whether q is above or below $1/\lambda_1$ determines whether the sum in diffusion centrality converges or diverges – as we know from spectral theory that the first eigenvalue of a matrix governs its expansion properties. The role of T being above or below the diameter is also very intuitive. In many classes of large random graphs, the average distance between most nodes is actually almost the same as the diameter, something first discovered by Erdos and Renyi. Thus, if T is below the diameter, news from any typical node will not have a long enough time to reach most other nodes. In contrast, once T hits the diameter, then that permits news from any typical node to reach most others. If one moves beyond T, then many of the walks counted by $\mathbf{g}^T$ begin to have "echoes" in them: they visit the same node twice. For instance, news passing from node 1 to node 2 to node 3 then back to node 2 and then to node 4, etc. Once most paths have echoes in them, the measure begins to act differently, and that eventually converges to the ergodic distribution, and essentially the first eigenvector (provided q is large enough to get saturation).

Here we report a theorem and corollary that formalize some of these intuitive statements. To do this we consider a sequence of Erdos-Renyi networks, as those provide for clear limiting properties.[13]

Let $\mathbf{g}(n,p)$ denote an Erdos-Renyi random network drawn on n nodes, with each link having independent probability p. In the following, as is standard, p (and T) are functions of n, but we omit that notation to keep the expressions uncluttered. We also allow for self-links for ease of calculations. We consider a sequence of random graphs of size n and as is standard in the literature, consider what happens as $n \to \infty$.

THEOREM **1.** *If $T = o(pn)$, then $\frac{\mathrm{E}[DC(\mathbf{g}(n,p);q,T)]}{npq\frac{1-(npq)^T}{1-npq}} \to 1$.* [14]

Theorem 1 provides an expression for how we expect diffusion centrality to behave in large graphs. Provided that T grows at a rate that is not overly fast[15], then we expect diffusion centrality of a typical node to converge to $npq\frac{1-(npq)^T}{1-npq}$.

Theorem 1 thus provides us with the tool to see when a diffusion that begins at a typical node is expected to reach other nodes or not, and leads to the following corollary.

COROLLARY **1.** *Consider a sequence of Erdos-Renyi random networks $\mathbf{g}(n,p)$ for which $\frac{1-\varepsilon}{\sqrt{n}} \geq p \geq (1+\varepsilon)\frac{\log(n)}{n}$ for some $\varepsilon > 0$*[16] *and any corresponding $T = o(pn)$.*

(1) *A threshold q:*
   (a) *If $q = o(1/\mathrm{E}[\lambda_1])$, then $\mathrm{E}\left[DC\left(\mathbf{g}(n,p);q,T\right)\right] \to 0$.*
   (b) *If $1/\mathrm{E}[\lambda_1] = o(q)$, then $\mathrm{E}\left[DC\left(\mathbf{g}(n,p);q,T\right)\right] \to \infty$.*[17]

(2) *A threshold T:*[18]
   (a) *If $T < (1-\varepsilon)\mathrm{E}[Diam(\mathbf{g}(n,p))]$ for some $\varepsilon > 0$, then $\frac{\mathrm{E}[DC(\mathbf{g}(n,p);q,T)]}{n} \to 0$.*
   (b) *If $T \geq \mathrm{E}[Diam(\mathbf{g}(n,p))]$ and $q > 1/(\mathrm{E}[\lambda_1])^{1-\varepsilon}$ for some $\varepsilon > 0$, then $\frac{\mathrm{E}[DC(\mathbf{g}(n,p);q,T)]}{n} = \Omega(1)$.*

[13]These properties will extend to other classes of random graph models by standard arguments (e.g., see Jackson (2008a)), but a general exploration of such models takes us beyond our scope here.

[14]To remind the reader, $f(n) = o(h(n))$ for functions f, h if $f(n)/h(n) \to 0$, and $f(n) = \Omega(h(n))$ if there exists $k > 0$ for which $f(n) \geq kh(n)$ for all large enough n.

[15]But note that it can be a rate that can tend to infinity and is already very permissive as it is far beyond the growth of the diameter of the network. Here T can grow up to pn, which will generally be larger than $\log(n)$, while diameter is proportional to $\log(n)/\log(pn)$.

[16]This ensures that the network is connected almost surely as n grows, but not so dense that the diameter shrinks to be trivial. See Bollobas (2001).

[17]Note that $\mathrm{E}[\lambda_1] = np$.

[18]Again, note that $T = o(pn)$ is satisfied whenever $T = o(\log(n))$, and thus is easily satisfied given that diameter is proportional to $\log(n)/\log(pn)$.

Putting these results together, we know that by setting $q = 1/\mathrm{E}[\lambda_1]$ and $T = \mathrm{E}[Diam(\mathbf{g})]$ we are at the point at which diffusion is just expected to reach a non-trivial number of others from a typical node, but by moving either of the parameters above or below this level, we would expect in a large network either to reach almost nobody or saturate the network. This makes $DC\left(\mathbf{g}; 1/\mathrm{E}[\lambda_1], \mathrm{E}[Diam(\mathbf{g}(n,p))]\right)$ a nice benchmark centrality, which is what we use throughout the empirical sections. At these values it will differ from both degree and eigenvector centrality (and Katz-Bonacich centrality). Of course, fitting q, T from the data can provide for more accurate measures as it will be tailored to the context and setting, but if one wishes to use a benchmark measure that does not have free parameters, then these are the parameter values that most clearly distinguish this measure from standard measures.

3. Relating Diffusion Centrality to Network Gossip

We now investigate whether and how individuals living in network $\mathbf{g}$ can 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 stochastic information diffusion process but from a *receiver's* perspective. Over time, each individual hears information that originates from different sources in the network, and in turn passes that information on with some probability. The society discusses each of these pieces of information for T periods. The key point is that there are many such topics of conversation, originating from all of the different individuals in the society, with each topic being passed along for T periods.

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." Arun may also have told Ben that he thinks house prices will go up and Ben could have told Esther that "Arun thinks that house prices will go up". In this model Esther keeps track of the cumulative number of times that bits of information that originated from Arun reach her and compares it with the same number for information that originated from other people. What is crucial, therefore, is that the news involves the name of the node of origin – in this case "Arun" – and not what the information is about. The first piece of news originating from Arun could be about something he has done ("bought a car") but the second could just be an opinion ("Arun thinks house

prices will go up"). Esther keeps track of the fact that she has heard two different pieces of information originating from Arun.

Recall that

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

is such that the ij-th entry, $H(\mathbf{g}; q, T)_{ij}$, 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 $\mathbf{H}$,

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

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 $NG(\mathbf{g}; q, T)_{ij}$ is twice as high as $NG(\mathbf{g}; q, T)_{kj}$ 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 relatively how often a given node hears information from (or about) each of the other nodes.

To end this sub-section two remarks are in order. First, one could allow passing probabilities to differ by information type and pairs of nodes.[19] 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, the current simple version captures basic dynamics and relates naturally to centrality measures.

Second, we could allow nodes to differ in how frequently they generate new information which is then transmitted to its neighbors. Provided this transmission rate is positively related to nodes' centralities, the results that we present below still hold (and, in fact, the speed of convergence would be increased). If the rate of generation of information about nodes is negatively correlated with their position, then our results below would be attenuated.

[19]We can generalize our setup replacing q with a matrix $\mathbf{Q}$. Now define

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

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

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

THEOREM **2.** *For any* $(\mathbf{g}; q, T)$, $\sum_j \text{cov}(DC(\mathbf{g}; q, T), NG(\mathbf{g}; q, T)_j) = \text{var}(DC)$. *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, they are really two quite different objects. Diffusion centrality is a gauge of a node's ability to send information, while the network gossip measure tracks 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 for small T it is possible that some nodes have not even heard about other nodes, and moreover they might be biased towards their local neighborhoods.[20]

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.

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

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. Hence, in the limit, they assess the rankings according to diffusion/eigenvector centrality correctly. The result implies that even with very little computational ability beyond remembering counts and adding to them, agents can

[20] One might conjecture that more central nodes 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.

come to learn arbitrarily accurately complex measures of the centrality of everyone in the network, including those with whom they do not associate.

More sophisticated strategies where individuals try to infer network topology, could accelerate learning. Nonetheless, learning is possible even in an environment where individuals do not know the structure of the network and do not tag anything but the source of the information.

The restriction to $q \geq 1/\lambda_1$ is important. When q tends to 0, individuals hear about others in the network with vanishing frequency, and as a result, the network distance between people can influence who they think is the most important.

4. Evidence: who are the gossips?

4.1. **Data Collection.** To investigate whether individuals' nomination of who is best at diffusing is related to the nominee's centrality, we use a unique network data that we gathered from villages in rural Karnataka (India). The data consists of a complete description of the network combined with "gossip" information for 35 villages.

To collect the network data (described in detail in Banerjee, Chandrasekhar, Duflo, and Jackson (2013), and publicly available at http://economics.mit.edu/faculty/eduflo/social), we asked adults to name those with whom they interact in the course of daily activities.[21] We have data concerning 12 types of interactions for a given survey respondent: (1) whose houses he or she visits, (2) who visits his or her house, (3) his or her relatives in the village, (4) non-relatives who socialize with him or her, (5) who gives him or her medical help, (6) from whom he or she borrows money, (7) to whom he or she lends money, (8) from whom he or she borrows material goods (e.g., kerosene, rice), (9) to whom he or she lends material goods, (10) from whom he or she gets important advice, (11) to whom he or she gives advice, (12) with whom he or she goes to pray (e.g., at a temple, church or mosque). Using these data, we construct one network for each village, at the household level where a link exists between households if any member of either household is linked to any other member of the other household in at least one of the 12 ways. Individuals can communicate if they interact in any of the 12 ways, so this is the network of potential communications, and using this network avoids the selection bias associated with data-mining to find the most predictive subnetworks. The resulting objects are undirected, unweighted networks at the household level.

[21]We have network data from 89.14% of the 16,476 households based on interviews with 65% of all adult individuals aged 18 to 55. This is a new wave of data relative to our original microfinance study.

After this data was collected, to collect gossip data, we asked the adults the following two additional questions:

(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?*

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

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

Only half of the households were willing to name a good "gossip". This is in itself intriguing. Perhaps people are unwilling to offer an opinion when they are unsure of the answer.[22] Conditional on naming someone, however, there is substantial concordance of opinion. Only 4% of households were nominated in response to the Event question (and 5% for the Loan question) with a cross-village standard deviation of 2%. Conditional on being nominated, the median household was nominated nine times.[23] This is perhaps a first indication that the answers may be meaningful, since if people are good at identifying central individuals we would expect their nominations to coincide.

We label as "leaders", shopkeepers, teachers and leaders of self-help groups – 13% of households fall into this category. This was how the microfinance organization Bharatha Swamukti Samsthe (BSS) defined leaders, who were identified as people to be seeded with information about their product. BSS's theory was that such social leaders were a priori likely to be important in the social learning process and thereby would contribute to more diffusion of microfinance.[24]

4.2. **Do individuals nominate central nodes?** Our theoretical results suggest that people can learn others' diffusion or eigenvector centralities simply by tracking

[22]See Alatas et al. (2014) for a model that builds on this idea.
[23]We work at the household level, in keeping with Banerjee et al. (2013) who used households as network nodes; a household receives a nomination if any of its members are nominated.
[24]In our earlier work, Banerjee et al. (2013), we show that there is considerable variation in the centrality of these "leaders" in a network sense, and that this variation predicts the eventual take up of microfinance.

news that they hear through the network, and therefore should name central individuals when asked whom to use as a "seed" for diffusion. In this section, we examine whether this is the case.

4.2.1. *Data description.* Figure 1 shows that people who are nominated as gossips, as well as people who are considered by the microfinance institution to be good "seeds" for their product (the "leaders") are significantly more central than randomly picked households. Moreover, gossip nominees are more central than the leaders. Indeed, 47% of households that are both nominated and have a leader are in the top decile of the eigenvector centrality distribution. Furthermore, 23% of the households that are nominated but are not leaders are in the top decile of the centrality distribution, while only 16% of households that are not nominated but have a leader, are in the top decile of the eigenvector centrality distribution. Finally, only 7% of households that are not nominated and contain no leader are in the top decile of the eigenvector centrality distribution.[25]

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 with whom they are directly connected. Panel A of Figure 2 shows that less than 20% of individuals nominate someone within their direct neighborhood, compared to about 10% of nodes within this category. At the same time, over 27% of nominations come from a network distance of at least three or more (40% of nodes are in these category). Therefore, while respondents tend to nominate people who are closer to them than the average person in the village, they are also quite likely to nominate someone who is far away. Moreover, it is important to note that highly central individuals are generally closer to people than the typical household (since they have many friends – the famous "friendship paradox"), so it does make sense that people tend to nominate individuals who are closer to them. Taken together, this suggests that information about centrality does indeed travel through the network.

Furthermore, Panel B of Figure 2 shows 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 comparably accurate

[25]The difference between the 23% of households that are in the top decile given that they are nominated but not leaders and the 16% in the top decile given that they are not nominated but are leaders is significant with a p-value of 0.00 under a Welch test.

information about central individuals in the community who are near or far from them.

4.2.2. *Regression Analysis.* Motivated by this evidence, we present a more systematic analysis of the correlates of nominations, using a discrete choice framework for the decision to nominate someone.

Our theory suggests that if people choose whom to nominate based on who they hear about most frequently, then diffusion centrality should be a leading predictor of nominations. While the aforementioned results are consistent with this prediction, there are several plausible alternative interpretations which do not rely on the information mechanism we outline in the model. For example, individuals may nominate the person with the most friends, and people with many friends tend to be more diffusion central than those with fewer friends (i.e., diffusion centrality with $T = 1$ and $T > 1$ can be positively correlated). Alternatively, it may be that people simply nominate the "leaders" within their village, or people who are central geographically, and these also correlate with diffusion/eigenvector centrality. There are indeed 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. Previous research has also shown that geographic proximity increases the probability of link formation (Fafchamps and Gubert, 2007; Ambrus et al., 2014; Chandrasekhar and Lewis, 2014) and therefore one might expect geographic data to be a useful predictor of centrality. In addition to leadership data we have detailed GPS coordinates for every household in each village. We include these in our analysis below as controls.[26]

To deal with this concern we show that our diffusion centrality measure does not simply pick out degree centrality, geography or traditional leadership. We recognize that the correlations below do not constitute proof that the causal mechanism is indeed gossip, but they do rule out these obvious confounding factors.

[26]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 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.

To operationalize our analysis we use $DC\left(1/\mathrm{E}[\lambda_1], \mathrm{E}[Diam(\mathbf{g}(n,p))]\right)$ as our measure of diffusion centrality, as discussed in Section 2.3. This is what we mean when we refer to diffusion centrality.

We estimate a discrete choice model of the decision to nominate an individual. Note that we have large choice sets as there are $n-1$ possible nominees and n nominators per village network. We model agent i as receiving utility $u_i(j)$ for nominating individual j:

$$u_i(j) = \alpha + \beta' x_j + \gamma' z_j + \mu_v + \epsilon_{ijv},$$

where x_j is a vector of network centralities for j (eigenvector centrality, diffusion centrality and degree centrality), z_j is a vector of demographic characteristics (e.g., leadership status, geographic position and caste controls), μ_v is a village fixed effect, and ϵ_{ijv} is a Type-I extreme value distributed disturbance. For convenience given the large choice sets, we estimate the conditional logit model by an equivalent Poisson regression, where the outcome is the expected number of times an alternative is selected (Palmgren, 1981; Baker, 1994; Lang, 1996; Guimaraes et al., 2003). This is presented in Panel A of Table 2. For comparison, Panel B presents the corresponding OLS results.

We begin with a number of bivariate regressions in Table 2. First we show that diffusion centrality is a significant driver of an individual nominating another (column 1). A one standard deviation increase in diffusion centrality is associated with a 0.607 log-point increase in the number of others nominating a household (statistically significant at the 1% level). Columns 2 to 5 repeats the exercise with two other network statistics (degree and eigenvector centrality), with the "leader" dummy, and with an indicator for geographic centrality. All of these variables, except for geographic centrality, predict nomination, and the coefficients are similar in magnitude.

The different network centrality measures are all correlated. To investigate whether diffusion centrality remains a predictor of gossip nomination after controlling for the other measures, we start by introducing them one by one as controls in column 1 to 4 in Table 3. Degree and eigenvector centrality are insignificant, and do not affect the coefficient of diffusion centrality. The leader dummy continues to be significant, but the coefficient of diffusion centrality remains strong and significant. The geographic centrality variable now has a negative coefficient, and does not affect coefficient of the diffusion centrality variable.

Our results provide suggestive evidence that a key driver of the nomination decision involves diffusion centrality with $T > 1$. The point estimates point towards the diffusion centrality as the most robust predicting factor.

To confirm this pattern, in the last column of Table 3, we introduce all the variable together and perform a LASSO analysis which "picks" out the variables that are strongly associated with the nomination variable. Specifically, we use the post-LASSO procedure of Belloni and Chernozhukov (2009). It is a two-step procedure. In the first step, standard LASSO is used to select the support: which variables matter in predicting our outcome, the number of nominations. In the second step, standard OLS (or Poisson, for Panel A) is run on the support selected in the first stage.[27,28]

Interestingly, LASSO picks out only one variable: diffusion centrality. The post-LASSO coefficient and standard error thus exactly replicate the OLS of using just diffusion centrality. This confirms that diffusion centrality is the key predictor of gossip nomination.

5. Experiment: Do gossip nominees spread information widely?

We have shown that individuals nominate central people in the network. Prior research demonstrated that providing information to more central individuals leads to greater diffusion (Banerjee et al., 2013; Beaman et al., 2014). Therefore, a natural question is whether using our gossip nomination protocol picks out those individuals that lead to faster diffusion of information compared to other obvious ways of choosing the seeds This is the most relevant policy implication of our theory.

5.1. **Information Diffusion and Gossip Seeding.**

We want to compare seeding information to gossips (nominees) to two benchmarks: (1) a set of village elders and (2) randomly selected households. Seeding information among random households provides the most relevant benchmark because it allows us to study how information circulates starting from random (non gossip) households. Seeding information with village elders provides an interesting benchmark because they are traditionally respected as social and political leaders. They are generally easy to identify, and it

[27]The problem with the returned coefficients from LASSO in the first step is that it shrinks the coefficients towards zero. Belloni and Chernozhukov (2009), Belloni et al. (2014b) and Belloni et al. (2014a) show that running out usual OLS on the support selected in the first stage in a second step will recover consistent estimates for the parameters of interest.

[28]To our knowledge, the post-LASSO procedure has not been developed for nonlinear models, so we only conduct selection using OLS.

could be, for instance, that information spreads widely only if it has the backing of someone who can influence opinion, not just convey information.

We conducted an experiment in 213 new villages in Karnataka that were not involved in the microfinance diffusion project, and where we had not worked before. In every village, we attempted to contact k households and inform them about a promotion run by our partner, a cellphone sales firm, that gave them a chance to enroll in a (non-rival) raffle to win a new mobile phone or a cash prize.[29] These individuals were encouraged to inform others in their community about the promotion. If an individual in the village heard about the promotion, she could give us a call to take part[30] Our primary outcome data is thus the number of calls from unique households that we received[31]. In half of the villages, we set $k = 3$ and in half of the villages we set $k = 5$.

We randomly divided the sample of 213 into three arms of 71 villages, where the k seeds were selected as follows.

T1. Village Elders: k households were chosen from a list of village elders obtained one week prior, by interviewing up to 15 households in the village (selected randomly via circular random sampling via the right-hand rule method that is commonly used in surveying).[32]

T2. Random: k households were chosen uniformly at random, also using the right-hand rule method and going to every n/k households.

T3. Gossip: k households were chosen from a list of gossip nominees obtained one week prior, by interviewing up to 15 households in the village (selected randomly via circular random sampling).

[29]The promotion was as follows. If an individual gave us a call, we would enter them in a non-rival raffle. We would then come to the village a week later and give every entered individual the opportunity to win a cellphone. The subject rolls a pair of dice and if she rolls a 12 she wins the phone. If she rolled less than a 12, lesser (but still substantial for these villagers), cash prizes of 25 to 275 rupees were given out. These terms were explained to the seeds and to the callers when they called to enter the raffle.

[30]In fact, we used what are known as "missed calls", a Indian institution. It is a call that is not picked up (and thus not charged) but serves as a ping, which we then call back so that the villagers are not charged for the call.

[31]Seed calls are included but there are seed number fixed effects.

[32]Circular sampling is a standard survey methodology where the enumerator starts at the end of a village and, using a right-hand rule, spirals throughout the entire village, when enumerating households.

The main outcome variable we are interested in is the number of calls we received. This represents the number of people who heard about the promotion and wanted to participate.

Given that the seeding does not exclude gossips in random villages, in some random villages gossip nominees were included in our seeding set by chance. We reclassify these villages as gossip seeding, so the random benchmark should be interpreted, as intended, as random conditional on none of the seed households being gossip nominees. Gossip seeding should be interpreted as "at least one seed is a gossip nominee". The reclassification is valid because the selection of households under this treatment is random, and hence the re-classification is random.

Subsequently, we collected full network data in 69 of the 71 villages in group T2 (two villages were no longer accessible to our surveyors). This data is used in Tables 6 and 7.

Figure 3 presents the results graphically. The distribution of calls in the gossip villages clearly stochastically dominates that of the leader and random graphs. Moreover, the incidence of a diffusive event, where a large number of calls is received, is rare when we seed information randomly or with village elders – but we do see such events when we seed information with gossip nominees.

Table 4 presents the regression analysis. Columns (1) and (2) do not include any controls. Columns (3) and (4) control for a potentially endogenous variable, "non broadcast communication": in one village the seed (a gossip nominee) made a flyer and printed many photocopies to advertise the promotion. The number of calls received in this village (106 calls) is much larger than the median of 5 in the gossip villages and 3 in the entire sample. The 95th percentile is 39 in the full sample and 47 in the gossip sample. The results are similar, showing that this village does not drive this result.

In columns (1) and (3) the outcome variable is the number of calls received. In the village elder treatment, we receive 5.5 phone calls on average, whereas in the random treatment we receive 3.9 calls on average. When the seeds are gossip nominees, we receive 13.8 calls on average. In columns (2) and (4) we normalize by the number of seeds. In the random seeding treatment we receive just over 1 calls per seed, which increases to about 1.4 call per seed in the village elder treatment (though the difference is not significant). In contrast, under gossip seeding we have 3.54 calls per seed, more than 3 times the ratio. This is our key experimental result: gossip nominees are indeed extremely good starting points to diffuse a piece of information.

5.2. **Mechanism: does gossip seed diffusion capture diffusion centrality?** We have seen in the first section of the paper that villagers seem to nominate individuals who are diffusion central. In our previous work (Banerjee, Chandrasekhar, Duflo, and Jackson, 2013), we showed that diffusion central seeds are associated with faster diffusion. To what extent is the faster diffusion of information in the experiment mediated by the diffusion centrality of the gossip seeds, and to what extent does it reflect the villagers' ability to capture other dimensions of individuals that would make them good at diffusing information?

To get at this issue, a few weeks after the experiment, we collected complete network data sample in 69 villages where seeds were randomly selected (2 of the 71 village were no longer accessible). In these villages, by chance, some seeds happen to be gossips and/or elders. We create a measure of centrality that exactly parallels the gossip dummy and elder dummy by forming a dummy for "high diffusion centrality". A household has "high diffusion centrality" if its diffusion centrality is greater than one standard deviation above the mean. As reported in Table 5, with this measure 24 villages have exactly one high diffusion centrality seed and 14 have more than one. For comparison, 23 villages have exactly one gossip seed, and 8 have more than one.

In Panel A of Table 6, we run Poisson regressions where the dependent variable is the number of calls we received, without control variables. As before, Panel B controls for the "broadcast diffusion" (the village where a flyer was made by one of the seeds and photocopies were distributed throughout the village, and where we received 106 calls while the next highest number of calls in the random sample was 58). This is an endogenous control variable, but an important one (since the diffusion does not take place via the network learning model). This village's seeds happened to include two gossips, but none with high diffusion centrality. Table 7 performs the same regressions, but using an OLS specification, and the results are qualitatively very similar.

Column 1 of Table 6 shows that more calls are received when enough seeds have high diffusion centrality, although the significance of the results is strongly influenced by the single "broadcast village". Without a control for this, in the Poisson specification the coefficient of at least one highly diffusion central seed is -0.0677, at least 2 is 0.59, and 3 or more 0.335. With control for broadcast diffusion, the coefficient of at least one high diffusion centrality seed is 0.348 (not significant), at least 2 is 1.008 (p value = 0.006), and 3 or more is 0.677 (p value = 0.056). In Table 7 we see that we receive on average 6.9 (9.8) more calls if we reached at least two high diffusion

centrality seeds without (with) control for broadcast diffusion. Although controlling for the broadcast diffusion does not produce results that are statistically different from those without controls, the point estimates are larger and the standard errors smaller because the broadcast village is a huge outlier in terms of number of calls, and the seed was a gossip who was not central. With this caveat, this result confirms the non-experimental results from our previous work that diffusion centrality captures the potential of seeds to diffuse information in a network.

Column 2 of both tables shows the impact of hitting the gossip seeds. For instance, in Table 7 we see with one gossip seed reached, the coefficient is positive but not significantly different from 0, but with 2, it is large and significant: we received 12 more calls if we reached two gossip seeds than if we reach none (24 if we do not control for the broadcast village!). Column 3 shows no effect of reaching elders. Columns 4 introduces gossip and high centrality dummies together. The results are somewhat noisy since the variables are collinear (as we saw before), but the key result is that the point estimate of the coefficient on there being "two gossips" remains large and significant in all specifications (OLS, Poisson, with and without control for broadcast).

Taken together, the results show that diffusion centrality captures part of the impact of gossip, but not all of it. An obvious example of something that is not captured by centrality is the fact that the gossip in the broadcast village had imaginative ways of conveying information (e.g., by making a poster), and people knew that, which is why they nominated this person in the first place. Note, however, that the impact of gossip seeds remains strong after controlling for the centrality of seeds, even in villages where there was no broadcasting. This suggests that there are other features of gossip nominees that make them good at diffusing information, and villagers must know about these traits.

To further explore this, in column 7, we implement a LASSO procedure to check whether the value of gossip seeds in explaining calls is robust to the inclusion of other controls. Following Belloni et al. (2014b), we perform a double-post LASSO to select the optimal set of control variables in a regression of calls on gossip nominations. Our basic regression relates y (number of calls) to Z (gossip variables) and some set of control variables X (diffusion centrality, elder dummies, etc.). We first perform a LASSO of Z on X to select the variables in X that are relevant in explaining Z (step 1). We then perform a LASSO of y on X, to select the control variable that are important in explaining y (step 2). Finally we perform a regression of y on Z and the union of variables that were selected in step 1 and in step 2. Belloni et al. (2014b)

shows that this method has the following desirable properties: it allows for imperfect model selection and allows for robust estimation and inference even if the underlying assumptions about the sparsity of the data generating process does not exactly hold. In our case, no control variable is selected in step 1 or in step 2 (regardless of whether we control for broadcast), so we end up regressing number of calls on the number of gossip seed dummies. This means that none of the other variables at our disposal sufficiently explain variation in the number of calls or gossip nomination itself so as to be selected as a control when looking at the partial correlation of gossip nomination with number of calls. The exercise confirms that having seeds that are gossips is strongly correlated with number of calls: having two gossip seeds corresponds to about 24 more calls received.

To summarize, this subsection presents some evidence in favor of the mechanism we emphasize in the model, but also highlights the fact that the gossip variable is actually a richer and more important proxy of information diffusion than the model based measure. Gossip seeds tend to be highly central, and information does spread faster from highly central seeds. This accounts for some part of the reason why information diffuses rapidly from gossip nominated seeds. At the same time it is also clear that the model does not capture the entire reason why gossip seeds are best to diffuse information: even controlling for their diffusion centrality, gossip seeds still lead to faster diffusion. Furthermore, we find that being nominated as a gossip is the only factor that predicts diffusion. There are clearly other factors that predict whether a seed will be good at diffusing information beyond their centrality (altruism, interest in the information, etc.) and villagers seem to be good at capturing those factors.

6. Conclusion

Our model illustrates that it should be easy for even very myopic and non-Bayesian agents, simply by counting, to have an idea of 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 (the most central among those they know), but actually name ones that are *globally* central within the village. Moreover, in a specially designed experiment, we find that nominated individuals are indeed extremely effective at diffusing information. This suggests that individuals may use simple protocols to learn valuable things about the complex systems within which they are embedded.

While our model focuses on the network-based mechanics of communication, in practice, considerations beyond 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 seems that villagers take such characteristics into account and thus nominate individuals who are even more successful at diffusing information than the most central individual in the network.

Our findings have important policy implications, since such nomination data are easily collected and therefore can be used in a variety of contexts, either on their own or combined with other easily collected data, to identify who would be a good seed for information diffusion. Thus, using this sort of protocol may be a cost-effective way to improve diffusion and outreach.

Beyond these applications, the work presented here opens a rich agenda for further research, as one can explore which other aspects of agents' social environments can be learned in simple ways. For example, can individuals also identify individuals who are trusted and trusted by others? A piece of information about a raffle is probably innocuous enough to be transmitted by a "gossip", but what about advice on immunization, for example?

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," *NBER Working Paper*. 22

Ambrus, A., M. Mobius, and A. Szeidl (2014): "Consumption Risk-Sharing in Social Networks," *American Economic Review*, 104, 149–82. 4.2.2

Baker, S. G. (1994): "The multinomial-Poisson transformation," *The Statistician*, 495–504. 4.2.2

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. 1, 2.2, 2.3, 3.1, 4.1, 23, 24, 4.2.2, 5, 5.2, A.1

Beaman, L., A. BenYishay, J. Magruder, and A. M. Mobarak (2014): "Can Network Theory based Targeting Increase Technology Adoption?" . 1, 5

Belloni, A. and V. Chernozhukov (2009): "Least squares after model selection in high-dimensional sparse models," *MIT Department of Economics Working Paper*. 1, 4.2.2, 27

Belloni, A., V. Chernozhukov, and C. Hansen (2014a): "High-dimensional methods and inference on structural and treatment effects," *The Journal of Economic Perspectives*, 29–50. 1, 27

——— (2014b): "Inference on treatment effects after selection among high-dimensional controls," *The Review of Economic Studies*, 81, 608–650. 1, 27, 5.2

Benzi, M. and C. Klymko (2014): "A matrix analysis of different centrality measures," *arXiv:1312.6722v3*. A.1

Bloch, F., G. Demange, and R. Kranton (2014): "Rumors and Social Networks," *Paris School of Economics, Working paper 2014 - 15*. 1

Bollobas, B. (2001): *Random Graphs*, Cambridge University Press. 16

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. 4.2.2

Dodds, P. S., R. Muhamad, and D. J. Watts (2003): "An Experimental Study of Search in Global Social Networks," *Science*, 301, 827–829. 3

Fafchamps, M. and F. Gubert (2007): "The formation of risk sharing networks," *Journal of Development Economics*, 83, 326–350. 4.2.2

Friedkin, N. E. (1983): "Horizons of Observability and Limits of Informal Control in Organizations," *Social Forces*, 61:1, 54–77. 1

Golub, B. and M. Jackson (2010): "Naive Learning in Social Networks and the Wisdom of Crowds," *American Economic Journal: Microeconomics*, 2, 112–149. A.1

Guimaraes, P., O. Figueirdo, and D. Woodward (2003): "A tractable approach to the firm location decision problem," *Review of Economics and Statistics*, 85, 201–204. 4.2.2

Jackson, M. (2008a): "Average Distance, Diameter, and Clustering in Social Networks with Homophily," in *the Proceedings of the Workshop in Internet and Network Economics (WINE 2008), Lecture Notes in Computer Science, also: arXiv:0810.2603v1*, ed. by C. Papadimitriou and S. Zhang, Springer-Verlag, Berlin

Heidelberg. 13

——— (2008b): *Social and Economic Networks*, Princeton: Princeton University Press. 33

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.* 7

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

Krackhardt, D. (1987): "Cognitive social structures," *Social Networks*, 9, 109–134. 1

——— (2014): "A Preliminary Look at Accuracy in Egonets," *Contemporary Perspectives on Organizational Social Networks, Research in the Sociology of Organizations*, 40, 277–293. 1

Lang, J. B. (1996): "On the comparison of multinomial and Poisson log-linear models," *Journal of the Royal Statistical Society. Series B (Methodological)*, 253–266. 4.2.2

Lawyer, G. (2014): "Understanding the spreading power of all nodes in a network: a continuous-time perspective," *arXiv:1405.6707v2*. 7

Milgram, S. (1967): "The small world problem," *Psychology Today*. 3

Palmgren, J. (1981): "The Fisher information matrix for log linear models arguing conditionally on observed explanatory variables," *Biometrika*, 563–566. 4.2.2

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

Tibshirani, R. (1996): "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288. 1

Figures

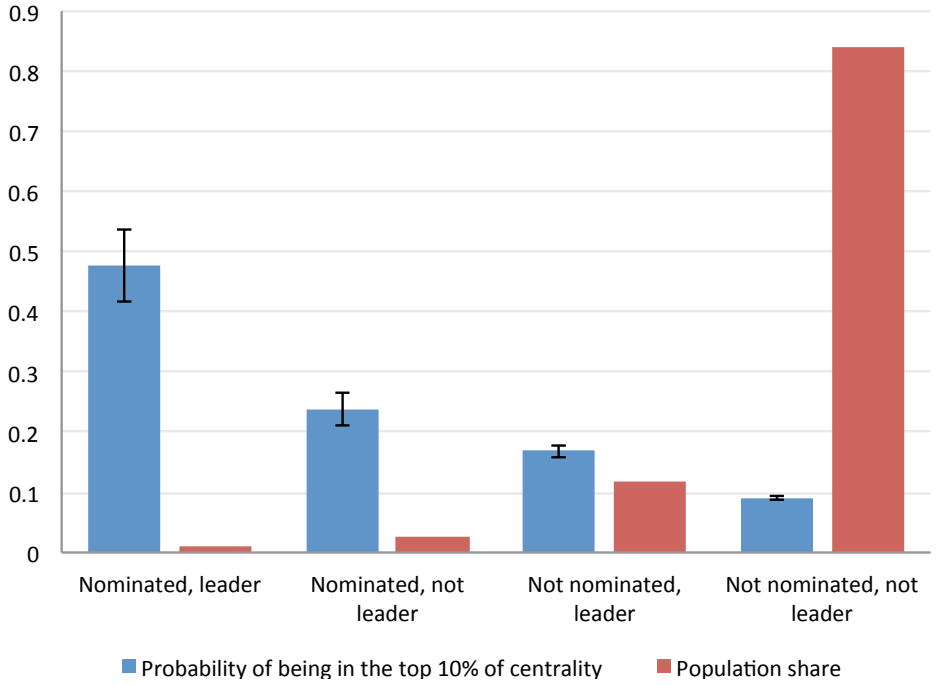
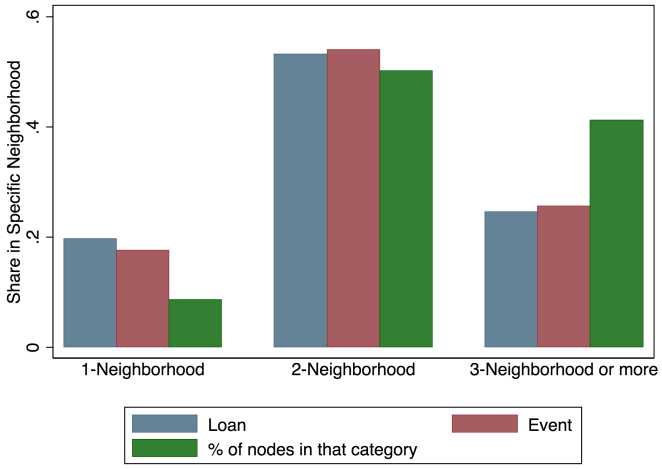


Figure 1. This figure uses the Phase 1 dataset. The probability that a randomly chosen node with a given classification (whether or not it is nominated under the event question and whether or not it has a village leader) is in the top decile of the eigenvector centrality distribution. 95% confidence intervals are displayed.



(A) Share of nominees in specified neighborhood

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

FIGURE 2. Distribution of centralities of nominees with the Phase 1 dataset.

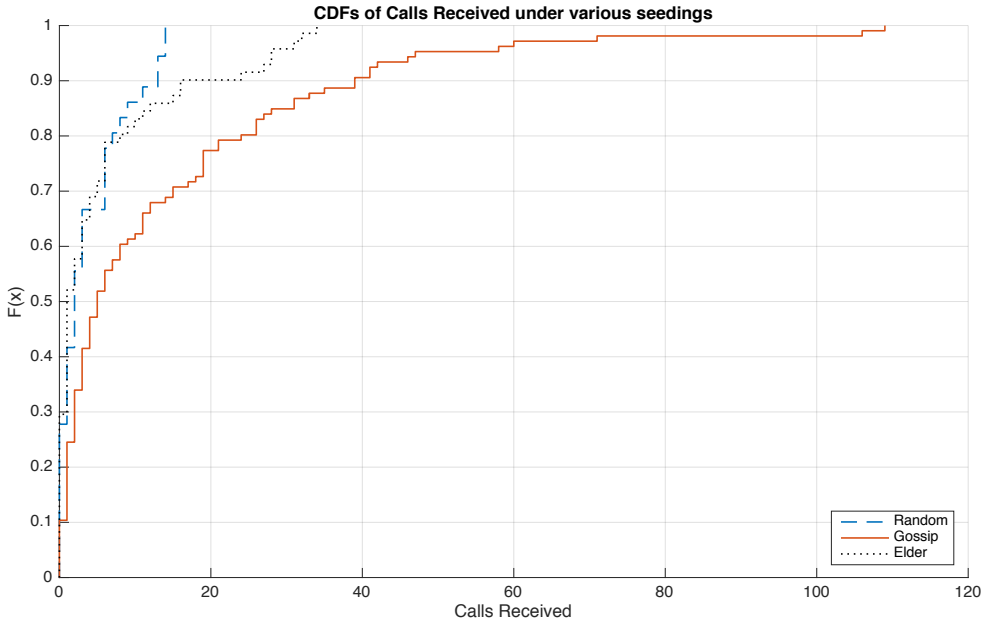


FIGURE 3. Distribution of calls received by treatment in the Phase 2 experiment.

Tables

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 giant component	0.98	(0.01)
is a "leader"	0.13	(0.34)
nominated someone for event	0.38	(0.16)
nominated someone for loan	0.48	(0.16)
was nominated for event	0.04	(0.02)
was nominated for loan	0.05	(0.03)
number of nominations received for loan	0.45	(3.91)
number of nominations received for event	0.34	(3.28)

Notes: This table presents summary statistics from the Phase 1 dataset: 35 villages of the Banerjee et al. (2013) networks dataset where nomination data was originally collected in 2011/2012. For the variables "nominated someone for loan (event)" and "was nominated for loan (event)" we present the cross-village standard deviation.

TABLE 2. Factors predicting nominations

Panel A: Poisson Regression	(1)	(2)	(3)	(4)	(5)
Diffusion Centrality	0.607***				
	(0.085)				
Degree Centrality		0.460***			
		(0.078)			
Eigenvector Centrality			0.605***		
			(0.094)		
Leader				0.868***	
				(0.288)	
Geographic Centrality					-0.082
					(0.136)
Observations	6,466	6,466	6,466	5,733	6,466
Panel B: OLS	(1)	(2)	(3)	(4)	(5)
Diffusion Centrality	0.285***				
	(0.060)				
Degree Centrality		0.250***			
		(0.061)			
Eigenvector Centrality			0.283***		
			(0.064)		
Leader				0.422**	
				(0.172)	
Geographic Centrality					-0.025
					(0.038)
Observations	6,466	6,466	6,466	5,733	6,466

Notes: This table uses data from the Phase 1 dataset. Panel A reports estimates of Poisson regressions where the outcome variable is the expected number of nominations under the event question. Panel B reports the same using OLS. Results are robust to including caste fixed effects and village fixed effects, available upon request. Degree centrality, eigenvector centrality and diffusion centrality, $DC\left(\mathbf{g}; 1/E[\lambda_1], E[Diam(\mathbf{g}(n,p))]\right)$, are normalized by their standard deviations. Standard errors (clustered at the village level) are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

TABLE 3. Factors predicting nominations

Panel A: Poisson Regression	(1)	(2)	(3)	(4)	(5)
Diffusion Centrality	0.642***	0.354**	0.553***	0.606***	0.607***
	(0.127)	(0.176)	(0.098)	(0.085)	(0.085)
Degree Centrality	-0.039				
	(0.101)				
Eigenvector Centrality		0.283			
		(0.186)			
Leader			0.541*		
			(0.305)		
Geographic Centrality				-0.082	
				(0.142)	
Observations	6,466	6,466	5,733	6,466	6,466
Post-LASSO					✓
Panel B: OLS	(1)	(2)	(3)	(4)	(5)
Diffusion Centrality	0.303***	0.161*	0.278***	0.285***	0.285***
	(0.091)	(0.087)	(0.069)	(0.060)	(0.060)
Degree Centrality	-0.020				
	(0.066)				
Eigenvector Centrality		0.138			
		(0.095)			
Leader			0.297		
			(0.175)		
Geographic Centrality				-0.026	
				(0.039)	
Observations	6,466	6,466	5,733	6,466	6,466
Post-LASSO					✓

Notes: This table uses data from the Phase 1 dataset. Panel A reports estimates of Poisson regressions where the outcome variable is the expected number of nominations under the event question. Panel B reports the same using OLS. Results are robust to including caste fixed effects and village fixed effects, available upon request. Degree centrality, eigenvector centrality and diffusion centrality, $DC\left(\mathbf{g}; 1/E[\lambda_1], E[Diam(\mathbf{g}(n,p))]\right)$, are normalized their standard deviations. Column (5) uses a post-LASSO procedure where in the first stage LASSO is implemented to select regressors and in the second stage the regression in question is run on those regressors. Omitted terms indicate they were not selected in the first stage. Standard errors (clustered at the village level) are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

TABLE 4. Calls received by treatment

	(1) Calls Received	(2) Calls Received	(3) Calls Received Seeds	(4) Calls Received Seeds
Gossip HHs Informed	9.875***	8.997***	2.498***	2.196***
	(2.075)	(1.898)	(0.568)	(0.485)
Elders Informed	1.632	1.632	0.443	0.443
	(1.284)	(1.287)	(0.371)	(0.372)
Constant (Random Non-Gossip)	3.889***	3.889***	1.044***	1.044***
	(0.743)	(0.745)	(0.213)	(0.214)
Observations	213	213	213	213
R-squared	0.081	0.246	0.068	0.333
Broadcast Control		✓		✓

Notes: This table uses data from the Phase 2 experimental dataset. Columns (1) and (2) use the number of calls received as the outcome variable. Columns (3) and (4) normalize the number of calls received by the number of seeds, 3 or 5, which is randomly assigned. Columns (2) and (4) control for broadcast. Robust standard errors are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

TABLE 5. Number of seeds by type

	None	Exactly 1	Exactly 2	3 or more
High Diffusion Centrality	30	24	12	2
Gossip Nomination	37	23	8	1
Elder	53	10	5	0

Notes: This table presents summary statistics from the 69 random villages in the Phase 2 experiment, where we collected network data. We were unable to collect network data from 2 villages due to issues of conflict and we omit one village where a seed made fliers. A seed is considered to have high diffusion centrality if it is at least one standard deviation above the mean in terms of centrality.

TABLE 6. Calls received by seed traits, Poisson Regression

Panel A: No control for broadcast diffusion	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1 High DC Seed	-0.0677			-0.302	-0.0477	-0.247	
	(0.433)			(0.437)	(0.431)	(0.377)	
2 High DC Seeds	0.590			-0.0329	0.645	0.190	
	(0.437)			(0.652)	(0.415)	(0.528)	
At least 3 High DC Seeds	0.335			-0.571	0.671	0.431	
	(0.374)			(0.618)	(0.498)	(0.479)	
1 Gossip Nominations		0.536		0.547		0.719*	0.524
		(0.383)		(0.455)		(0.433)	(0.374)
2 Gossip Nominations		1.721***		1.754**		2.054***	1.593***
		(0.514)		(0.705)		(0.587)	(0.439)
At least 3 Gossip Nominations		-0.0732		0.0648		0.139	-0
		(0.378)		(0.489)		(0.456)	(0.240)
1 Elder			0.225		-0.0587	-0.393	
			(0.385)		(0.342)	(0.374)	
2 Elders			-0.591		-0.828*	-1.794***	
			(0.518)		(0.486)	(0.653)	
Observations	69	69	69	69	69	69	69
Double Post-LASSO							✓
Control for broadcast diffusion							
Panel B: Control for broadcast diffusion	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1 High DC Seed	0.348			0.237	0.365	0.226	
	(0.336)			(0.343)	(0.335)	(0.349)	
2 High DC Seeds	1.008***			0.750*	1.021***	0.777*	
	(0.348)			(0.419)	(0.346)	(0.422)	
At least 3 High DC Seeds	0.677**			0.252	0.970**	0.757	
	(0.329)			(0.370)	(0.465)	(0.499)	
1 Gossip Nominations		0.397		0.234		0.354	0.524
		(0.370)		(0.378)		(0.393)	(0.374)
2 Gossip Nominations		1.108***		0.788**		1.087***	1.593***
		(0.369)		(0.401)		(0.389)	(0.439)
At least 3 Gossip Nominations		-0.299		-0.371		-0.311	-0
		(0.330)		(0.326)		(0.334)	(0.240)
1 Elder			0.378		0.0228	-0.172	
			(0.393)		(0.345)	(0.338)	
2 Elders			-0.408		-0.683	-1.227**	
			(0.485)		(0.484)	(0.546)	
Observations	69	69	69	69	69	69	69
Double Post-LASSO							✓
Control for broadcast diffusion	✓	✓	✓	✓	✓	✓	✓

Notes: This table presents data from the 69 random villages in the Phase 2 experiment, where we collected network data. The table presents Poisson regressions of number of calls received by characteristics of the set of seeds. High DC refers to a seed being above the mean by one standard deviation of the centrality distribution. Columns (1)-(6) control for number of seeds, village size and the interaction. Panel A does not control for the broadcast diffusion village whereas Panel B does. Column (7) performs a double post-LASSO procedure, to select optimal controls in a regression of number of calls on gossip variables, where the set of controls are diffusion centrality dummies, elder dummies, number of seeds, village size, and the interaction. No controls are selected by the procedure. Robust standard errors are used.

TABLE 7. Calls received by seed traits, OLS

Panel A: No control for broadcast diffusion	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1 High DC Seed	-0.775			-2.728	-0.731	-2.876	
	(4.067)			(4.406)	(4.114)	(4.439)	
2 High DC Seeds	6.861			0.990	7.227	1.782	
	(5.562)			(8.087)	(5.389)	(7.656)	
At least 3 High DC Seeds	3.173			-8.125	6.142	-2.805	
	(3.551)			(10.86)	(4.729)	(7.310)	
1 Gossip Nominations		4.229		4.209		6.492	4.130
		(3.657)		(4.589)		(5.292)	(3.312)
2 Gossip Nominations		24.42*		24.90*		29.50*	23.50**
		(12.38)		(14.71)		(15.55)	(11.17)
At least 3 Gossip Nominations		-0.758		0.580		0.915	-0
		(3.781)		(4.694)		(4.654)	(1.473)
1 Elder			2.432		-0.314	-4.320	
			(4.874)		(4.524)	(5.526)	
2 Elders			-4.110		-6.131*	-16.14*	
			(3.283)		(3.595)	(8.778)	
Observations	69	69	69	69	69	69	69
Double Post-LASSO							✓
Control for broadcast diffusion							
Panel B: Control for broadcast diffusion	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1 High DC Seed	2.190			1.313	2.180	1.051	
	(2.659)			(2.680)	(2.718)	(2.767)	
2 High DC Seeds	9.843**			7.372	9.836**	7.265	
	(4.693)			(5.403)	(4.676)	(5.478)	
At least 3 High DC Seeds	5.377**			0.906	7.709*	3.765	
	(2.672)			(3.712)	(4.094)	(3.782)	
1 Gossip Nominations		2.929		1.517		2.704	4.130
		(3.401)		(3.788)		(4.013)	(3.312)
2 Gossip Nominations		12.12**		9.352		12.31**	23.50**
		(4.847)		(5.630)		(4.910)	(11.17)
At least 3 Gossip Nominations		-2.959		-3.184		-2.792	-0
		(3.194)		(3.363)		(3.391)	(1.473)
1 Elder			3.945		0.568	-1.274	
			(4.836)		(4.458)	(4.337)	
2 Elders			-2.567		-4.698	-9.080**	
			(2.821)		(3.609)	(4.505)	
Observations	69	69	69	69	69	69	69
Double Post-LASSO							✓
Control for broadcast diffusion	✓	✓	✓	✓	✓	✓	✓

Notes: This table presents data from the 69 random villages in the Phase 2 experiment, where we collected network data. The table presents Poisson regressions of number of calls received by characteristics of the set of seeds. High DC refers to a seed being above the mean by one standard deviation of the centrality distribution. Columns (1)-(6) control for number of seeds, village size and the interaction. Panel A does not control for the broadcast diffusion village whereas Panel B does. Column (7) performs a double post-LASSO procedure, to select optimal controls in a regression of number of calls on gossip variables, where the set of controls are diffusion centrality dummies, elder dummies, number of seeds, village size, and the interaction. No controls are selected by the procedure. Robust standard errors are used.

Appendix A. Proofs

A.1. **Relation of Diffusion Centrality to Other Measures.**

We prove all of the statements for the case of weighted and directed networks.

Let $v^{(L,k)}$ indicate k-th left-hand side eigenvector of $\mathbf{g}$ and similarly let $v^{(R,k)}$ indicate $\mathbf{g}$'s k-th right-hand side eigenvector. In the case of undirected networks, $v^{(L,k)} = v^{(R,k)}$. In the case of directed networks, eigenvector $v^{(1)}$ in the main body corresponds to $v^{(R,1)}$.

Let $d(\mathbf{g})$ denote degree centrality (so $d_i(\mathbf{g}) = \sum_j g_{ij}$). Eigenvector centrality corresponds to $v^{(1)}(\mathbf{g})$: the first eigenvector of $\mathbf{g}$. Also, let $KB(\mathbf{g}, q)$ denote Katz-Bonacich centrality – defined for $q < 1/\lambda_1$ by:[33]

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

It is direct to see that (i) diffusion centrality is proportional to degree centrality at the extreme at which $T = 1$, and (ii) if $q < 1/\lambda_1$, then diffusion centrality coincides with Katz-Bonacich centrality if we set $T = \infty$. We now show that when $q > 1/\lambda_1$ 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 q is sufficiently small so that limited diffusion takes place even in the limit of large T, or whether q is sufficiently large so that the knowledge saturates the network and then it is only relative amounts of saturation that are being measured.[34]

Theorem **A.1.**

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

$$DC(\mathbf{g}; q, 1) = qd(\mathbf{g}).$$

[33] See (2.7) in Jackson (2008b) 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 q.

[34] Saturation occurs when the entries of $\left(\sum_{t=1}^{\infty} (q\mathbf{g})^t\right) \cdot \mathbf{1}$ 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.

(2) *If* $q \geq 1/\lambda_1$ *and* $\mathbf{g}$ *is aperiodic, then as* $T \to \infty$ *diffusion centrality approximates eigenvector centrality:*

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

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

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

This is a result we mention in Banerjee, Chandrasekhar, Duflo, and Jackson (2013). An independent formalization appears in Benzi and Klymko (2014).

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. Nonetheless, one can also define a variant of diffusion centrality that works for finite iterations of DeGroot updating.

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

First, consider any irreducible and aperiodic nonnegative (and hence primitive) $\mathbf{g}$. If the statement holds for any arbitrarily close positive and diagonalizable $\mathbf{g}'$ (which are dense in a nonnegative neighborhood of $\mathbf{g}$), then since $\frac{DC(\mathbf{g};q,T)}{\sum_{t=1}^{T}(q\lambda_1)^t}$ is a continuous function (in a neighborhood of a primitive $\mathbf{g}$, which has a simple first eigenvalue) as is the first eigenvector, then the statement also holds at $\mathbf{g}$.[35] Thus, it is enough to prove the result for a positive and diagonalizable $\mathbf{g}$.

We show the following for a positive and diagonalizable $\mathbf{g}$:

- If $q > \lambda_1^{-1}$, then

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

- If $q = \lambda_1^{-1}$, then

[35] As is shown below, $\frac{DC(\mathbf{g};q,T)}{\sum_{t=1}^{T}(q\lambda_1)^t}$ has a well-defined limit, and so this holds also for the limit.

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

Let $\widetilde{\mathbf{g}}=\mathbf{g}/\lambda_1$, and normalize the eigenvectors to lie in ℓ_1, so that the entries in each column of $\mathbf{V}^{-1}$ and each row of $\mathbf{V}$ sum to 1.

Let us show the statement for the case where $q=1/\lambda_1$. It is sufficient to show

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

First, note that given the diagonalizable matrix, straightforward calculations show that

$$DC_i\left(\mathbf{g};\lambda_1^{-1},T\right)=\sum_j\sum_{t=1}^{T}\sum_k v_i^{(R,k)}v_j^{(L,k)}\widetilde{\lambda}_k^t.$$

Thus,

$$\begin{aligned}\left|\frac{DC_i\left(\mathbf{g};\lambda_1^{-1},T\right)}{T}-v_i^{(R,1)}\right| &= \left|\frac{\sum_j\sum_{t=1}^{T}\sum_{k=1}^{n}v_i^{(R,k)}v_j^{(L,k)}\widetilde{\lambda}_k^t}{T}-v_i^{(R,1)}\right|= \\ &= \left|\frac{1}{T}\sum_j\sum_{t=1}^{T}\sum_{k=2}^{n}v_i^{(R,k)}v_j^{(L,k)}\widetilde{\lambda}_k^t\right|\leq\frac{1}{T}\sum_{t=1}^{T}\sum_{k=2}^{n}1\cdot\underbrace{\left|\sum_{j=1}^{n}v_j^{(L,k)}\right|}_{\leq 1}\cdot\left|\widetilde{\lambda}_k^t\right| \\ &\leq \frac{n}{T}\sum_{t=1}^{T}\left|\widetilde{\lambda}_2^t\right|=\frac{n}{T}\frac{\left|\widetilde{\lambda}_2\right|}{1-\left|\widetilde{\lambda}_2\right|}\left(1-\left|\widetilde{\lambda}_2\right|^T\right)\to 0.\end{aligned}$$

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 $q>1/\lambda_1$. It follows that the eigenvalues of $q\mathbf{g}$ are $\widetilde{\Lambda}=\text{diag}\left\{\widetilde{\lambda}_1,...,\widetilde{\lambda}_n\right\}$ with $q\lambda_k=\widetilde{\lambda}_k$. We show

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

By similar derivations as above,

$$\begin{aligned}
\left|\frac{DC_i\left(\mathbf{g};\lambda_1^{-1},T\right)}{\sum_{t=1}^T \widetilde{\lambda}_1^t} - v_i^{(R,1)}\right| &= \left|\frac{\sum_j \sum_{t=1}^T \sum_{k=1}^n v_i^{(R,k)} v_j^{(L,k)} \widetilde{\lambda}_k^t}{\sum_{t=1}^T \widetilde{\lambda}_1^t} - v_i^{(R,1)}\right| \\
&= \left|\frac{\sum_j \sum_{t=1}^T \sum_{k=2}^n v_i^{(R,k)} v_j^{(L,k)} \widetilde{\lambda}_k^t}{\sum_{t=1}^T \widetilde{\lambda}_1^t} + \frac{\sum_j \sum_{t=1}^T v_i^{(R,1)} v_j^{(L,1)} \widetilde{\lambda}_1^t}{\sum_{t=1}^T \widetilde{\lambda}_1^t} - v_i^{(R,1)}\right| \\
&= \left|\frac{\sum_j \sum_{t=1}^T \sum_{k=2}^n v_i^{(R,k)} v_j^{(L,k)} \widetilde{\lambda}_k^t}{\sum_{t=1}^T \widetilde{\lambda}_1^t} + \frac{\sum_{t=1}^T v_i^{(R,1)} \widetilde{\lambda}_1^t}{\sum_{t=1}^T \widetilde{\lambda}_1^t} - v_i^{(R,1)}\right| \\
&= \left|\frac{1}{\sum_{t=1}^T \widetilde{\lambda}_1^t} \sum_j \sum_{t=1}^T \sum_{k=2}^n v_i^{(R,k)} v_j^{(L,k)} \widetilde{\lambda}_k^t\right| \\
&\leq \frac{1}{\sum_{t=1}^T \widetilde{\lambda}_1^t} \sum_{t=1}^T \sum_{k=2}^n 1 \cdot \left|\sum_{j=1}^n v_j^{(L,k)}\right| \cdot \left|\widetilde{\lambda}_k^t\right| \\
&\leq \frac{n}{\sum_{t=1}^T \widetilde{\lambda}_1^t} \sum_{t=1}^T \left|\widetilde{\lambda}_2^t\right|.
\end{aligned}$$

Note that this last expression converges to 0 since $\widetilde{\lambda}_1 > 1$, and $\widetilde{\lambda}_1 > \widetilde{\lambda}_2$.[36] which completes the argument. ■

A.2. **Other Proofs.**

Proof of Theorem 1 .

$$\begin{aligned}
\mathrm{E}\left[DC\left(\mathbf{g}(n,p);q,T\right)\right]_i &= \left[\sum_1^T \mathrm{E}\left[q^t \mathbf{g}(n,p)^t\right] \cdot \mathbf{1}\right]_i \\
&= \sum_1^T q^t n \mathrm{E}\left[\mathbf{g}(n,p)^t\right]_{ij},
\end{aligned}$$

where the last equality comes from the fact that $\mathrm{E}\left[\mathbf{g}(n,p)^t\right]_{ij} = \mathrm{E}\left[\mathbf{g}(n,p)^t\right]_{ik}$ for all i,j,k in an Erdos-Renyi random graph.

Next, note that

$$\mathrm{E}\left[\mathbf{g}(n,p)^t\right]_{ij} = \mathrm{E}\left[\sum_{k_1,k_2,\ldots,k_{t-1}\in\{1,\ldots,n\}^{t-1}} g_{ik_1} g_{k_1k_2} \cdots g_{k_{t-1}j}\right]$$

[36]Note that it is important that $q \geq 1/\lambda_1$ for this claim, since if $q < 1/\lambda_1$, then $q\lambda_1 < 1$. In that case, observe that

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

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

If all the indexed $g_{..}$'s were distinct, the right hand side of this equation would simply be $n^{t-1}p^t$. However, some terms repeat, in which case, since they are bernoulli random variables, the expression would be even less for some terms. Thus, it follows directly that

$$\mathrm{E}\left[\mathbf{g}(n,p)^t\right]_{ij} \geq n^{t-1}p^t$$

and so

$$\begin{aligned}\mathrm{E}\left[DC\left(\mathbf{g}(n,p);q,T\right)\right]_i &= \sum_1^T q^t n \mathrm{E}\left[\mathbf{g}(n,p)^t\right]_{ij} \\ &\geq \sum_1^T q^t n^t p^t = npq\frac{1-(npq)^T}{1-npq}\end{aligned}$$

Note also, that

$$\mathrm{E}\left[\sum_{k_1,k_2,\ldots,k_t\in\{1,\ldots,n\}^t} g_{ik_1}g_{k_1k_2}\cdots g_{k_{t-1}j}\right] \leq n^{t-1}p^t + tn^{t-2}p^{t-1} + t^2n^{t-3}p^{t-2} + \ldots + t^t.$$

This last inequality is a very loose upper bound simply by loosely upper-bounding how many $g_{..}$'s could conceivably repeat, and then putting in the expression that would ensue if they did repeat. Despite how loose the bound is, it suffices for our purposes.

Given that $t \leq T < pn$, it follows that

$$\begin{aligned}\mathrm{E}\left[\sum_{k_1,k_2,\ldots,k_t\in\{1,\ldots,n\}^t} g_{ik_1}g_{k_1k_2}\cdots g_{k_{t-1}j}\right] &\leq n^{t-1}p^t\left(1+\frac{t}{pn}+\left(\frac{t}{pn}\right)^2\ldots+\left(\frac{t}{pn}\right)^t\right) \\ &= n^{t-1}p^t\left(\frac{1-\left(\frac{t}{pn}\right)^t}{1-\left(\frac{t}{pn}\right)}\right).\end{aligned}$$

Thus,

$$\mathrm{E}\left[\mathbf{g}(n,p)^t\right]_{ij} \leq n^{t-1}p^t\frac{1}{1-\frac{T}{pn}}.$$

Since $T << pn$ it follows that (here $o(1)$ is with respect to n):

$$\begin{aligned}\mathrm{E}\left[DC\left(\mathbf{g}(n,p);q,T\right)\right]_i &= \sum_1^T q^t n \mathrm{E}\left[\mathbf{g}(n,p)^t\right]_{ij} \\ &\leq \sum_1^T q^t n^t p^t(1+o(1)) = npq\frac{1-(npq)^T}{1-npq}(1+o(1)).\end{aligned}$$

The theorem follows directly. ■

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

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

Additionally,

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

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 Corollary 1 . To see (1) first note that $x\frac{1-x^T}{1-x} \to 0$ if $x \to 0$, and that $x\frac{1-x^T}{1-x} \to x\frac{x^T}{x} \to \infty$ if $x \to \infty$. Replacing x with npq and then applying Theorem 1 yields the result under (a) and (b), respectively.

To see (2), we consider the case in which $q > 1/(\mathrm{E}[\lambda_1])^{1-\varepsilon}$, and so in which $npq > (np)^{\varepsilon}$. This is the case under which (b) applies. This also implies the result in (a), since if the conclusion of (a) holds for such a q it will also hold for all lower q, given that DC is monotone in q.

Again, since $npq > 1$, it follows that if T is growing, then

$$\mathrm{E}\left[DC\left(\mathbf{g}(n,p); q, T\right)\right]_i \to npq\frac{1-(npq)^T}{1-npq} \to (npq)^T.$$

So, to have

$$\mathrm{E}\left[DC\left(\mathbf{g}(n,p); q, T\right)\right]_i \geq kn$$

for some $k > 0$, it is sufficient that $(npq)^T \geq kn$, or

$$T \geq \frac{\log(n) + \log(k)}{\log np + \log(q)} \to \frac{\log(n)}{\log np} \sim \mathrm{E}[Diam\left(\mathbf{g}(n,p)\right)],$$

where the last comparison is a property of Erdos-Renyi random networks given that $\frac{1-\varepsilon}{\sqrt{n}} \geq p \geq (1+\varepsilon)\frac{\log(n)}{n}$, and so this establishes (b). From the analogous calculation, if T is below $\frac{\log(n)}{\log np}$, then $\mathrm{E}\left[DC\left(\mathbf{g}(n,p); q, T\right)\right]_i \leq kn$ for any k, and so (a) follows. ■

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

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

Also, let $\tilde{\lambda}_k = q\lambda_k$. It then follows that we can write

$$\mathbf{H} = \sum_{t=1}^{T} (q\mathbf{g})^t = \sum_{t=1}^{T} \left(\sum_{k=1}^{n} v_i^{(R,k)} v_j^{(L,k)} \tilde{\lambda}_k^t \right).$$

By the ordering of the eigenvalues from largest to smallest in magnitude,

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

So, since the largest eigenvalue is unique, it follows that

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

Note that the last expression converges to 0 since $\tilde{\lambda}_1 > 1$, and $\tilde{\lambda}_1 > \tilde{\lambda}_2$. Thus,

$$\frac{\mathbf{H}_{\cdot,j}}{\sum_{t=1}^{T} \tilde{\lambda}_1^t} \to v^{(R,1)} v_j^{(L,1)}$$

for each j. This completes the proof since each column of $\mathbf{H}$ is proportional to $v^{(R,1)}$ in the limit, and thus has the correct ranking for large enough T.[37] Note that the ranking is up to ties, as the ranking of tied entries may vary arbitrarily along the sequence. That is, if $v_i^{(R,1)} = v_\ell^{(R,1)}$, then the ranking that j has over i and ℓ could vary arbitrarily with T, but their rankings will be correct relative to any other entries with higher or lower eigenvector centralities. ■

[37] The discussion in Footnote 36 clarifies why $q > 1/\lambda_1$ is required for the argument.

Appendix B. Alternative Outcome for Experiment

Table B.1. Showed up for payment, no controls for broadcast diffusion

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1 High DC Seed	0.0483			-0.200	0.0475	-0.142	
	(0.413)			(0.383)	(0.416)	(0.343)	
2 High DC Seeds	0.381			-0.0594	0.492	0.153	
	(0.427)			(0.464)	(0.429)	(0.443)	
At least 3 High DC Seeds	0.0560			-0.673	0.251	0.0404	
	(0.341)			(0.605)	(0.368)	(0.324)	
1 Gossip Nominations		0.197		0.215		0.401	0.0793
		(0.350)		(0.399)		(0.363)	(0.339)
2 Gossip Nominations		1.605***		1.664***		1.841***	1.383***
		(0.433)		(0.528)		(0.449)	(0.396)
At least 3 Gossip Nominations		0.0445		0.122		0.132	-0.147
		(0.362)		(0.420)		(0.374)	(0.229)
1 Elder			-0.0751		-0.320	-0.647**	
			(0.409)		(0.421)	(0.328)	
2 Elders			-0.544*		-0.539**	-1.119***	
			(0.316)		(0.254)	(0.243)	
Observations	48	48	48	48	48	48	48
Double Post-LASSO							✓
Control for broadcast diffusion							

Notes: This table presents data from the 48 of the 69 random villages in the Phase 2 experiment, where we collected network data and had payment data. The table presents Poisson regressions of number of respondents who showed up (to play their payment lottery) calls received by characteristics of the set of seeds. The sample includes the 48 villages for which we have this data. High DC refers to a seed being above the mean by one standard deviation of the centrality distribution. Columns (1)-(6) control for number of seeds, village size and the interaction. Column (7) performs a double post-LASSO procedure, to select optimal controls in a regression of number of calls on gossip variables, where the set of controls are diffusion centrality dummies, elder dummies, number of seeds, village size, and the interaction. No controls are selected by the procedure. Robust standard errors are used.

TABLE B.2. Showed up for payment, controlling for broadcast diffusion

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1 High DC Seed	0.435			0.280	0.436	0.258	
	(0.317)			(0.338)	(0.318)	(0.332)	
2 High DC Seeds	0.802**			0.573	0.869**	0.627	
	(0.338)			(0.372)	(0.355)	(0.388)	
At least 3 High DC Seeds	0.365			-0.0146	0.516	0.324	
	(0.298)			(0.428)	(0.342)	(0.329)	
1 Gossip Nominations		0.0779		-0.0246		0.134	0.0793
		(0.337)		(0.350)		(0.337)	(0.339)
2 Gossip Nominations		1.040***		0.823**		1.037***	1.383***
		(0.335)		(0.359)		(0.385)	(0.396)
At least 3 Gossip Nominations		-0.183		-0.292		-0.250	-0.147
		(0.324)		(0.317)		(0.311)	(0.229)
1 Elder			0.113		-0.206	-0.394	
			(0.413)		(0.421)	(0.301)	
2 Elders			-0.396		-0.398	-0.761***	
			(0.295)		(0.300)	(0.262)	
Observations	48	48	48	48	48	48	48
Double Post-LASSO							✓
Control for broadcast diffusion	✓	✓	✓	✓	✓	✓	✓

Notes: This table presents data from the 48 of the 69 random villages in the Phase 2 experiment, where we collected network data and had payment data. The table presents Poisson regressions of number of respondents who showed up (to play their payment lottery) calls received by characteristics of the set of seeds. All regressions control for a dummy for broadcast village. The sample includes the 48 villages for which we have this data. High DC refers to a seed being above the mean by one standard deviation of the centrality distribution. Columns (1)-(6) control for number of seeds, village size and the interaction. Column (7) performs a double post-LASSO procedure, to select optimal controls in a regression of number of calls on gossip variables, where the set of controls are diffusion centrality dummies, elder dummies, number of seeds, village size, and the interaction. No controls are selected by the procedure. Robust standard errors are used.

Appendix C. Loan Question

Table C.1. Factors predicting nominations

Panel A: Poisson Regression	(1)	(2)	(3)	(4)	(5)
Diffusion Centrality	0.625***				
	(0.075)				
Degree Centrality		0.490***			
		(0.067)			
Eigenvector Centrality			0.614***		
			(0.084)		
Leader				0.950***	
				(0.271)	
Geographic Centrality					-0.113
					(0.082)
Observations	6,466	6,466	6,466	5,733	6,466
Panel B: OLS	(1)	(2)	(3)	(4)	(5)
Diffusion Centrality	0.391***				
	(0.071)				
Degree Centrality		0.367***			
		(0.065)			
Eigenvector Centrality			0.378***		
			(0.074)		
Leader				0.629***	
				(0.229)	
Geographic Centrality					-0.045
					(0.029)
Observations	6,466	6,466	6,466	5,733	6,466

Notes: This table uses data from the Phase 1 dataset. Panel A reports estimates of Poisson regressions where the outcome variable is the expected number of nominations under the loan question. Panel B reports the same using OLS. Results are robust to including caste fixed effects and village fixed effects, available upon request. Degree centrality, eigenvector centrality and diffusion centrality, $DC\left(\mathbf{g}; 1/\mathrm{E}[\lambda_1], \mathrm{E}[Diam(\mathbf{g}(n,p))]\right)$, are normalized their standard deviations. Standard errors (clustered at the village level) are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

TABLE C.2. Factors predicting nominations

Panel A: Poisson Regression	(1)	(2)	(3)	(4)	(5)
Diffusion Centrality	0.560***	0.431***	0.565***	0.624***	0.560***
	(0.122)	(0.130)	(0.086)	(0.075)	(0.122)
Degree Centrality	0.070				0.070
	(0.086)				(0.086)
Eigenvector Centrality		0.219			
		(0.138)			
Leader			0.613**		
			(0.290)		
Geographic Centrality				-0.115	
				(0.089)	
Observations	6,466	6,466	5,733	6,466	6,466
Post-LASSO					✓
Panel B: OLS	(1)	(2)	(3)	(4)	(5)
Diffusion Centrality	0.310***	0.266***	0.383***	0.391***	0.310***
	(0.112)	(0.089)	(0.081)	(0.071)	(0.112)
Degree Centrality	0.091				0.091
	(0.079)				(0.079)
Eigenvector Centrality		0.138			
		(0.089)			
Leader			0.457*		
			(0.231)		
Geographic Centrality				-0.045	
				(0.030)	
Observations	6,466	6,466	5,733	6,466	6,466
Post-LASSO					✓

Notes: This table uses data from the Phase 1 dataset. Panel A reports estimates of Poisson regressions where the outcome variable is the expected number of nominations under the loan question. Panel B reports the same using OLS. Results are robust to including caste fixed effects and village fixed effects, available upon request. Degree centrality, eigenvector centrality and diffusion centrality, $DC\left(\mathbf{g}; 1/\mathrm{E}[\lambda_1], \mathrm{E}[Diam(\mathbf{g}(n,p))]\right)$, are normalized their standard deviations. Column (5) uses a post-LASSO procedure where in the first stage LASSO is implemented to select regressors and in the second stage the regression in question is run on those regressors. Omitted terms indicate they were not selected in the first stage. Standard errors (clustered at the village level) are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.