USING GOSSIPS TO SPREAD INFORMATION: THEORY AND EVIDENCE FROM TWO RANDOMIZED CONTROLLED TRIALS

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

Abstract. Is it possible to identify individuals who are highly central in a community without gathering any network information, simply by asking a few people to tell us whom we seed a information with if we want it to spread widely in the community? If we use people's nominees as seeds for a diffusion process, will it be successful? In a first "proof of concept" RCT run in 213 villages in Karnataka, India, information about opportunities to get a free cell phone or money diffused more widely when it was initially given to people nominated by others than when it was given to randomly selected people, or village elders. In a second large scale policy RCT in 517 villages in Haryana, India, the monthly number of vaccinations increased by 22% in randomly selected villages in which nominees were given information about about upcoming monthly vaccination camps and asked to spread it, than when randomly selected villagers were given the same information. After presenting the results from these RCTs, we show via a simple model how members of a community can just by tracking gossip about others, identify highly central individuals in their network. Asking villagers in rural Indian villages to name good seeds for diffusion, we find that they accurately nominate those who are central according to a measure tailored for diffusion – not just those with many friends or in powerful positions.

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

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

Date: This Version: March 12, 2018.
This paper supersedes an earlier paper "Gossip: Identifying Central Individuals in a Social Network." Financial support from the NSF under grants SES-1156182, SES-1155302, and SES-1629446, and from the AFOSR and DARPA under grant FA9550-12-1-0411, and from ARO MURI under award No. W911NF-12-1-0509 is gratefully acknowledged. We thank Shobha Dundi, Devika Lakhote, Francine Loza, Tithee Mukhopadhyay, Gowri Nagraj, and Paul-Armand Veillon for outstanding research assistance. We also thank Michael Dickstein, Ben Golub, John Moore, and participants at various seminars/conferences for helpful comments. Social Science Registry AEARCTR-0001770 and approved by MIT IRB COUHES # 1010004040.

[†]Department of Economics, MIT; NBER; J-PAL.
[‡]Department of Economics, Stanford University; NBER; J-PAL.
[§]Department of Economics, MIT; NBER; J-PAL.
[⋆]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

Policymakers and businesses often rely on key informants to diffuse new information to a community. The message is seeded to a number of people with the hope that it will diffuse via word-of-mouth. Even when there are alternatives available (e.g., broadcasting), seeding is still a commonly used technology.[1] For example, microcredit organizations use seeding to diffuse knowledge about their product, and agricultural extension agents try to identify leading farmers within each community (Bindlish and Evenson, 1997; Banerjee, Chandrasekhar, Duflo, and Jackson, 2013; Beaman, BenYishay, Magruder, and Mobarak, 2014). Such seeding is not restricted to developing economies: Gmail was first diffused by invitations to leading bloggers and then via sequences of invitations that people could pass to their friends, and seeding of apps and other goods to central individuals in viral marketing campaigns is common (e.g., see Aral, Muchnik, and Sundararajan (2013); Hinz, Skiera, Barrot, and Becker (2011a)).[2]

The central question of this paper is how to find the best (most effective) people to seed for a diffusion process. A body of work suggests that if the goal is to diffuse information by word-of-mouth then the optimal seeds are those who have central positions in the social network.[3] Moreover as shown in Banerjee, Chandrasekhar,

[1] The comparison between seeding and other methods is not the object of this paper, but seeding has many advantages. It is cheap, a peer may have a easier time getting someone's attention, and can also answer questions, etc.

[2] Beyond diffusion applications, there are many other reasons and contexts for wanting to identify highly central people. For instance, one may want to identify "key players" to influence behaviors with peer effects (e.g., see Ballester, Calvó-Armengol, and Zenou (2006)). These examples include peer effects in schooling, networks of crime and delinquent behavior, among other things. Similarly, the work of Paluck, Shepherd, and Aronow (2016) shows that "social referents" who are exposed to an intervention that encouraged taking public stances against conflict at school was particularly effective.

[3] There is a voluminous literature on the importance of opinion leaders and key individuals in diffusing products and information. This ranges from the early sociology literature (e.g, classic studies by Simmel (1908); Katz and Lazarsfeld (1955); Coleman, Katz, and Menzel (1966)), to the vast literature on diffusion of innovations (e.g., Rogers (1995); Centola (2010, 2011); Jackson and Yariv (2011)), to theoretical research on what are appropriate measures of centrality in a network (e.g., Bonacich (1987); Borgatti (2005, 2006); Ballester, Calvó-Armengol, and Zenou (2006); Valente, Coronges, Lakon, and Costenbader (2008); Valente (2010, 2012); Lim, Ozdaglar, and Teytelboym (2015); Bloch and Tebaldi (2016)), to a literature on identifying central and influential individuals in marketing (e.g., Krackhardt (1996); Iyengar, den Bulte, and Valente (2010); Hinz, Skiera, Barrot, and Becker (2011b); Katona, Zubcsek, and Sarvary (2011)), to the computational issues of identifying

Duflo, and Jackson (2013) and Beaman, BenYishay, Magruder, and Mobarak (2014), even though many measures of centrality are correlated, successful diffusion requires seeding information via people who are central according to specific measures. A practical challenge is that the relevant centrality measures are based on extensive network information, which can be costly and time consuming to collect in many settings.

In this paper, we thus ask the following question. How can one easily and cheaply identify highly central individuals without gathering network data? As shown in the research mentioned above, superficially obvious proxies for individuals who are central in the network sense – such as targeting people with leadership or special status, or who are geographically central, or even those with many friends – can fail when it comes to diffusing information. So, how can one find highly central individuals without network data and in ways that are more effective than relying on such proxies? We explore a direct technique that turns out to be remarkably effective: simply asking a few individuals in the community who would be the best individuals for spreading information.

Surprisingly, this is not a solution that had been recommended in theory or, to our knowledge, tried in practice by any organization in the field. This is perhaps because there is ample reason to doubt that such a technique would work. Previous studies have shown that people's knowledge about the networks in which they are embedded is surprisingly lacking. In fact, individuals within a network tend to have little perspective on its structure, as found in important research by Friedkin (1983) and Krackhardt (1987), among others.[4] Indeed, in data collected in the same villages in Karnataka as we used for part of this study, Breza, Chandrasekhar, and Tahbaz-Salehi (2017) show that individuals have very limited knowledge of the network. 47% of randomly selected individuals are unable to offer a guess about whether two others in their village share a link and being one step further from the pair corresponds to a 10pp increase in the probability of misassessing link status. There is considerable uncertainty over network structure by those living in the network.

This raises the question of whether and how, despite not knowing the structure of the network in which they are embedded, people know who is central and well-placed to diffuse information through the network.

multiple individuals for seeding (e.g., (Kempe, Kleinberg, and Tardos, 2003, 2005)), to a method of finding influential individuals via the friendship paradox (e.g., see Feld (1991); Krackhardt (1996); Kim, Hwong, Staff, Hughes, OâĂŹMalley, Fowler, and Christakis (2015); Jackson (2016).

[4] See Krackhardt (2014) for background and references.

In this paper, we examine people's ability to identify highly central individuals and effective seeds for a diffusion process. We make three main contributions.

Our first contribution is empirical. Via two different RCTs we show that, in practice, it is possible to cheaply identify influential seeds by asking community members.

The first randomized controlled trial was run in 213 villages in Karnataka. We asked villagers who would be a good diffuser of information. In 71 of those villages, we then used those nominations to seed information about a (non-rival) raffle for cell phone and cash prizes. We compare how well these nominated seeds do compared to another 71 villages in which we selected seeds who villagers reckon to have high social status (village elders), and yet another 71 villages in which we selected the seeds randomly.

Specifically, in each village, we seeded a piece information in 3 to 5 households. In the 71 "random seeding" the seed households were randomly selected. In the 71 "social status" villages, they had status as "elders" in the village – leaders with a degree of authority in the community, who command respect. In the remaining 71 villages, the seeds were those nominated by others as being well suited to spread information ("gossip nominees"). The piece of information that we spread was simple: anyone who gives a free calls a particular phone number will have a chance to win a free cell phone, and if they do not win the phone, they are guaranteed to win some cash. The chances to win cash and phones are independent of the number of people who respond, ensuring that the information was non-rivalrous and everyone was informed of that fact. We then measured the extent of diffusion using the number of independent entrants.

We received on average 8.1 phone calls in villages with random seedings, 6.9 phone calls in villages with village elder seedings, and 11.7 in villages with gossip seedings. Thus, a policymakers would accelerate diffusion by identifying "gossip" seeds in this easy way. Moreover, since we also track whether at least one gossip was hit in the "random seeding" villages, we can instrument "hitting at least one gossip" with the gossip treatment. This gives us a similar result: seeding at least one gossip seed yields an extra 7.4 calls, or nearly double the base rate.

While this RCT is a useful proof of concept, and has the advantage of being clearly focused on a pure information diffusion process, the information that was circulated is not particularly important. This is potentially concerning for two reasons. First, the application itself is not of direct policy interest. Second, targeting gossips might have been successful because the information was anodyne. Perhaps the elders and the randomly selected seeds would have more aggressively circulated a more relevant piece of information.

To find out whether the success of the technique carries over to a policy relevant setting, we conducted out a second large scale policy relevant RCT, in the context of a unique collaboration with the Government of Haryana (India) on their immunization program. Immunization is an important policy priority in Haryana, because it is remarkably low. This projects takes place in seven low performing district where full immunization rates were around 40% or less at baseline. We worked some of the villages that were part of the sample of a randomized controlled trial of the impact of incentives in immunization. In those villages, all immunizations delivered in monthly camps were tracked via a tablet-based e-Health application. Prior to the launch of the incentive programs and the tablets, we identified 517 villages for a "seed" intervention. Those villages were randomly assigned to four groups. In the first group ("gossip"), 17 randomly selected households were surveyed and asked to identify who would be good diffusers of information; in the second group ("trust") we asked 17 randomly selected households who people in the village tend to trust; in the third one, we asked who is both good at diffusing information *and* trusted. In the fourth group, no nominations were elicited. We then visited the six most nominated individuals in each village (or the head of six randomly selected households in the fourth group) and asked them to become the program's ambassadors. Throughout the year, they receive regular SMS reminding them to spread information about immunization.[5] We have administrative data on immunization (from the tablets) for about one year after launch of the program.

The results of this RCT are consistent with those of the first study. In the average monthly camp with random seeds, 17 children attended and received at least one shot. In village with gossip seeds, the number was 21, or 22% higher. We find a significant increase for all types of vaccines. For example, the monthly number of children immunized for measles, the most deadly disease and one where immunization rates are particularly low, increased from 3.66 in villages with random seeds to 4.6 in villages with gossip seeds. The other seedings are in between: neither statistically different from random seeding (for most vaccines), nor statistically different from gossip seeds.

Thus, these both RCTs carried out in very different context illustrate that villagers identify people that effectively spread information.

[5] 79% of people contacted agreed to participate, and every village had at least one seed.

Our second contribution is theoretical. We answer the question of how it could be that people name highly diffusion central individuals (more on this later) without knowing anything about their networks. Because we are interested in diffusion, the feature of the network that we hope people would have implicit knowledge about is a notion of centrality which relates to iterative expansion properties of the social network, which we have defined as "diffusion centrality". Needless to say, this is a complicated concept, and so superficially it may seem implausible that people could estimate it, especially since it is a function of an object (the network) that they do not know well. Our main theoretical result shows that there is a very simple argument for why even very naive agents, simply by counting how often they hear pieces of gossip, would have accurate estimates of others' diffusion centralities. This result demonstrates what is special about this notion of centrality. In particular, we model a process that we call "gossip" in which nodes generate pieces of information that are stochastically passed from neighbor to neighbor, along with the identity of the node from which the information emanated. We assume only that individuals who hear the gossip are able to keep count of the number of times that each person in the network is mentioned as a source.[6] 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 centrality.[7] Even without any knowledge of the network, gossip is information of which individuals can easily be aware.

It is worth underscoring this is just a possibility result. There are other ways in which people can learn who is central. The theory suggests one reason for believing our empirical results–but the empirical results are not a test of the theory to the exclusion of other possible explanations.

Our final contribution is to check that people are actually nominating people who are high in diffusion centrality, consistent with our model's prediction. In 33 villages in which we had previously collected detailed network data (not part of the RCTs), we collected new data on who villagers think would be good at spreading information. We then find that, indeed, individuals nominate highly diffusion central people. Nominees consistently rank in the top quartile of diffusion centrality, and many rank in the top decile. We also show that the nominations are not simply based on the nominee's

[6] 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)).

[7] The specific definition of centrality we use here is diffusion centrality (Banerjee et al., 2013) but a similar result holds for eigenvector centrality, as is shown in the Appendix.

leadership status, degree, or geographic position in the village, but are significantly correlated with diffusion centrality even after controlling for these characteristics.

Next we use these data to examine another implication of our model of gossip and knowledge about the diffusion of others. Under the model, if the process only runs for finite time, agents can have different rankings others' centralities. We show that, conditional on individual fixed effects and even controlling for the diffusion centrality of the agent being assessed, a villager is more likely to nominate an agent who is of higher rank in the respondent's personal/subjective finite-time ranking under the model.

Finally, to test whether the increase in diffusion from gossip nominees is in fact fully accounted for by their diffusion centrality, we went back to the villages with random seeding in the cell phone RCT, and collected full network data. Consistent with network theory, we find that information diffuses more extensively when we hit at least one seed with high diffusion centrality. However, when we include both gossip nomination and diffusion centrality of the seeds in the regression, the coefficient of gossip centrality does not decline much (although it becomes less precise). This suggests that diffusion centrality does not explain all of the extra diffusion from gossip nominees. People's nominations may incorporate additional attributes, such as who is listened to in the village, or who is most charismatic or talkative, etc., which goes beyond a nominee's centrality. Alternatively, it may be that our measure of the network and diffusion centrality are noisy, and villagers are even more accurate at finding central individuals than we are.

To summarize, we suggest a process by which, by listening and keeping count of how often they hear *about* someone, individuals learn the correct ranking of community members in terms of how effectively they can spread information. And, we show that, in practice, individuals nominated by others are indeed effective seeds of information.

The remainder of the paper is organized as follows. In Section 2, we describe the two RCTs and results. Section 3 develops our model of diffusion and presents the theoretical results relating network gossip to diffusion centrality. Section 4 describes the the data used in the empirical analysis of how diffusion central nominees are, and presents the analysis of the relation between being nominated and being central. Section 5 concludes.

2. Experiments: Do gossip nominees spread information widely?

In two "proof of concept" RCTs, we show that when a simple piece of information is given to people who are nominated by their fellow villagers as being good information spreaders, it diffuses more widely than when it is given to people with high social status or to random people. The first experiment concerned information about an opportunity to get free cash or a cell phone, while the second experiment concerned information about a vaccination program.

2.1. **Study 1: The cell phone and cash raffle RCT.** We conducted an RCT in 213 villages in Karnataka (India) to investigate if people who are nominated by others as being good "gossips" (good seeds for circulating information) are actually more effective than other people at transmitting a simple piece of information.

We compare seeding of information with gossips (nominees) to two benchmarks: (1) village elders and (2) randomly selected households. Seeding information among random households is obviously a natural benchmark. Seeding information with village elders provides an interesting alternative because they are traditionally respected as social and political leaders and one might presume that they would be effective seeds. They have the advantage of being easy to identify, and it could be, for instance, that information spreads widely only if it has the backing of someone who can influence opinion, not just convey information.

In every village, we attempted to contact a number k (detailed below) of households and inform them about a promotion run by our partner, a cellphone sales firm. The promotion gave villagers a non-rivalrous chance to win a new mobile phone or a cash prize. Most villagers in this area of India already have a cell phone or access to one, but the phone was new, of decent quality, and unlocked and could be resold. It is common for people in India to frequently change handsets and to buy and sell used ones. Thus, the cell phone can be taken to be worth to villagers roughly its cash value (Rs. 3,000). All the other prizes were direct cash prizes.

The promotion worked as follows. Anyone who wanted to participate could give us a "missed call" (a call that we registered, but did not answer, and which was thus free).[8] In public, a few weeks later, the registered phone numbers were randomly awarded cash prizes ranging from Rs. 50 to Rs. 275, or a free cell phone. Which prize any given entrant was awarded was determined by the roll of two dice (the total of the two dice times 25 rupees, unless they rolled a 12, in which case they got a

[8] This is a common technique in this region.

cell phone), regardless of the number of participants, ensuring that the awarding of all prizes was fully non-rivalrous and there was no strategic incentive to withhold information about the promotion.

In each treatment, the seeded individuals were encouraged to inform others in their community about the promotion. In half of the villages, we set $k = 3$, and in half of the villages we set $k = 5$. This was done because we were not sure how many seeds were needed to avoid the extremes of the process dying out or diffusing to everybody. In practice, we find that there is no significant difference between 3 and 5 seeds on our outcome variable (the number of calls received).

We randomly divided the sample of 213 villages into three sets of 71, where the k seeds were selected as follows. A few days before the experiment, we interviewed up to 15 households in every village (selected randomly via circular random sampling via the right-hand rule method) to identify "elders" and "gossips".[9] We asked the same questions in all villages to allow us to identify the sorts of seeds that were reached in each treatment.

The question that was asked for the 15 households to identify the gossip nominees was:[10]

> *"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 or a new loan product, to whom should we speak?"*

The notion of "village elder" is well understood in these villages: there are people who are recognized authorities, and believed to be influential. To elicit who was an elder, we asked the following question:

> *"Who is a well-respected village elder in your village?"*

In summary, there were three treatments groups:

T1. Random: k households were chosen uniformly at random, also using the right-hand rule method and going to every n/k households.

T2. Gossip: k households were chosen from the list of gossip nominees obtained one week prior.

[9] 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. This allows us to cover the entire geographic span of the village which is desirable in this application, particularly as castes are often segregated, which may lead to geographic segregation of the network. We want to make sure the nominations reflect the entire village.

[10] Our question in this RCT is an aggregation of two questions that we used in a prior survey that asked villagers to nominate gossips and we studied whether they were central in the network. That exercise is described in Section 4.1, where the disaggregated questions can be found.

T3. Elder: k households were chosen from the list of village elders obtained one week prior.

Note that this seeding does not address the challenging problem of choosing the optimal set of nodes for diffusion given their centralities. The solution typically will not be to simply pick the highest ranked nodes, since the positions of the seeds relative to each other in the network also matters. This results in a computationally challenging problem (in fact, an NP-complete one, see Kempe, Kleinberg, and Tardos (2003, 2005)). Here, we randomly selected seeds from the set of nominees, which if anything biases the test against the gossip treatment–we could have, for example, used the most highly nominated nodes from each caste group, which might have delivered combinations of highly central nodes that are well-spaced in the network.

The main outcome variable that we are interested in is the number of calls received. This represents the number of people who heard about the promotion and wanted to participate.[11] The mean number of calls in the sample is 9.35 (with standard deviation 15.64). The median number of villagers who participated is 3 across all villages. In 80.28% of villages, we received at least one call, and the 95th percentile is 39. It is debatable whether these are large or small numbers for a marketing campaign. Nonetheless, there is plenty of variation from village to village to allow us to identify the effect of the seeding on information diffusion.

We exclude one village, in which the number of calls was 106, from our analysis. In this village one of the seeds (who happened to be a gossip nominee in a random village) prepared posters to broadcast the information broadly. The diffusion in this villages does not have much to do with the network process we have in mind. We thus use data from 212 villages in all the regressions that follow. The results including this village are presented in Appendix E. They are qualitatively similar: the OLS of the impact of hitting at least one gossip is in fact larger and more precise when that village is included, while the Reduced form and IV estimates are similar but noisier.

Figure 1 presents the results graphically. The distribution of calls in the gossip villages clearly stochastically dominates that of the elder 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.

[11] The calls from the seeds are included in the main specification, and so we include seed number fixed effects.

We begin with the analysis of our RCT, which is the policymaker's experiment: what is the impact on diffusion of purposefully seeding gossips or elders, as compared to random villagers?

$$\begin{aligned}(2.1)\qquad y_j = \theta_0 + \theta_1 GossipTreatment_j + \theta_2 ElderTreatment_j + \\ \theta_3 NumberSeeds_j + \theta_4 NumberGossip_j + \theta_5 NumberElder_j + u_j,\end{aligned}$$

where y_j is the number of calls received from village j (or the number of calls per seed), $GossipTreatment_j$ is a dummy equal to 1 if seeds were assigned to be from the gossip list, $ElderTreatment_j$ is a dummy equal to 1 if seeds were assigned to be from the elder list, $NumberSeeds_j$ is the total number of seeds, 3 or 5, in the village, $NumberGossip_j$ is the total number of gossips nominated in the village, and $NumberElder_j$ is the total number of elders nominated in the village.

Table 1 presents the regression analysis. The results including the broadcast village are presented in Appendix E. Column 1 shows the reduced form (2.5). In control (random) villages, we received 8.077 calls, or an average of 1.967 per seeds. In gossip treatment villages, we received 3.65 more calls ($p = 0.19$) in total or 1.05 per seed ($p = 0.13$).

This exercise is of independent interest since it is the answer to the policy question of how much a policy makers would gain by first identifying the gossips and seeding them rather than choosing seeds randomly. However the seeding in the random and elder treatment villages does not exclude gossips. In fact in some random and elder treatment villages, gossip nominees were included in our seeding set by chance. On an average, 0.59 seeds were gossips in random villages. Another relevant question is to what extent information seeded to a gossip circulates more widely than information seeded to someone who is not a gossip.

Our next specification is thus to compare villages where "at least 1 gossip was hit," or "at least 1 elder was hit" (both could be true simultaneously) to those where no elder or no gossip was hit. Although the selection of households under treatments is random, the event that at least one gossip (elder) being hit is random only conditional on the number of potential gossip (elder) seeds present in the village. We thus include as controls in the OLS regression of number of calls on "at least 1 gossip (elder) seed hit". This specification should give us the causal effect of gossip (elder) seeding, but to assess its robustness, we also make directly use of the variation induced by the village level experiment, and instrument "at least 1 gossip (elder) seed hit" by the gossip (elder) treatment status of the village.

Therefore, we are interested in

$$y_j = \beta_0 + \beta_1 GossipReached_j + \beta_2 ElderReached_j + \beta_3 NumberSeeds_j + \beta_4 NumberGossip_j + \beta_5 NumberElder_j + \epsilon_j. \tag{2.2}$$

This equation is estimated both by OLS, and by instrumental variables, instrumenting $GossipReached_j$ with $GossipTreatment_j$ and $ElderReached_j$ with $ElderTreatment_j$. There the first stage equations are

$$GossipReached_j = \pi_0 + \pi_1 GossipTreatment_j + \pi_2 ElderTreatment_j + \pi_3 NumberSeeds_j + \pi_4 NumberGossip_j + \pi_5 NumberElder_j + \upsilon_j, \tag{2.3}$$

and

$$ElderReached_j = \rho_0 + \rho_1 GossipTreatment_j + \rho_2 ElderTreatment_j + \rho_3 NumberSeeds_j + \rho_4 NumberGossip_j + \rho_5 NumberElder_j + \nu_j. \tag{2.4}$$

Column 2 of Table 1 shows the OLS. The effect of hitting at least one gossip seed is 3.79 for the total number of calls ($p = 0.04$) ,which represents a 65% increase, relative to villages where no gossip seed was hit, or 0.95 ($p = 0.06$) calls per seed. Column 5 presents the IV estimates (Columns 3 and 4 present the first stage results for the IV). They are larger than the OLS estimates, and statistically indistinguishable, albeit less precise.

Given the distribution of calls, the results are potentially sensitive to outliers. We therefore present quantile regressions of the comparison between gossip/no gossip and Gossip treatment/Random villages in Figure 2. The specification that compares villages where gossips were either hit or not hit (Panel B) is more precise. The treatment effects are significantly greater than zero starting at the 35th percentile. Specifically, hitting a gossip significantly increases the median number of calls by 122% and calls at the 80th percentile by 71.27%.

This is our key experimental result: gossip nominees are much better than random seeds for diffusing a piece of information. Gossip seeds also lead to much more diffusion than elder seeds. In fact, the reduced form effect of seeding with an elder is negative, although it is not significant. This could BE specific to this application. Elders, like everybody else, are familiar with cell phones. Nonetheless they may have thought that this raffle was a frivolous undertaking, and did not feel they should circulate the information, whereas they might have circulated a more important piece of news. This is in fact a broader concern with the experimental setting. Since the information that was circulated was relatively anodyne, perhaps only people who really like to talk would take the trouble to talk about it. Recall that the nominations were elicited by asking for people who would be good at spreading news about, in part, an "event" or a fair, something social and relatively unimportant, similar to the piece of information that was actually diffused. We might have just selected the right people for something like that. The next policy question is thus whether gossip nominees are also good at circulating information on something more vital.

To find this out, we designed a second RCT on a subject that is both meaningful and potentially sensitive: immunization.

2.2. **Study 2: The Haryana Immunization RCT.** We conducted our second RCT in 2017 to apply the same idea to a setting of immediate policy interest: immunization.[12]

This RCT took place in Haryana, a state bordering New Delhi, in Northern India. J-PAL is collaborating with the government of Haryana on a series of initiatives designed to improve immunization rates in seven low immunization districts. 3116 villages, served by 140 primary Health centers and around 755 subcenters, are involved in the project. The project includes several components. In all villages, monthly immunization camps are held, and the government gave nurses tablets with a simple e-health application that the project team developed to keep track of all immunizations. The data thus generated is our main outcome.[13] In addition, J-PAL carried out several cross-randomized interventions in some or all of the villages: different types of small

[12] Prior research has shown that parent's vaccination choices are influenced by the perceptions and decisions of their neighbors (see, e.g., Brunson (2013)).

[13] We have completed over 5,000 cross-validation surveys, by visiting children at random and collecting information on their immunization status to cross-check with the data base. The administrative data is of excellent quality.

incentives for immunization,[14] a targeted SMS reminder campaign, and finally, the network seeding experiment.

2.2.1. *Experimental Design.* The "seeding" experiment took place in 517 villages. In all of those villages, six individuals (selected according to the protocol described below) were contacted in person a few weeks prior the launch of the tablet application and the incentives intervention (the seeds were contacted between June and August 2016, and the tablet application was launched in December 2016). They answered a short demographic survey and were asked to become ambassador for the program. If they agreed,[15] they gave us their phone number, and they agreed to receive regular reminders about upcoming immunization camps and to remind anyone they knew.

Specifically, the script to recruit them was as follows:

> *"Hello! My name is .... and I am from IFMR, a research institute in Chennai. We are conducting a research activity to disseminate information about immunization for children. We are conducting this study in several villages like yours to gather information, to help with this research activity. You are one of the people selected from your village to be a part of this experiment. Should you choose to participate, you will receive an SMS with information about immunization for children in the near future. The experiment will not cost you anything. We assure you that your phone number will only be used to send information about immunization and for no other purpose. Do you agree to participate?"*

And if they agreed, we used the following script at the end:

> *"You will receive an SMS on your phone containing information about immunization camps in the near future. When you receive the SMS, you can spread the information to your family, friends, relatives, neighbors, co-workers and any other person you feel should know about immunization. This will make them aware about immunization camps in*

[14] After each visit to an immunization camp, the caregiver receives a mobile credit on the phone. The incentives were randomized as (1) high incentives with flat payment (Rs. 90 per immunization); (2) high incentives with increasing pament (Rs. 50 for the first three, Rs. 100 for the fourth vaccination, and Rs. 200 for the fifth); (3) low incentives with flat payment (Rs. 50 per immunization); (4) low incentives with increasing payment (Rs. 10 for the first three, Rs. 60 for the fourth, and Rs. 160 for the fifth). There was a fifth control cell where no incentive was provided whatsoever.

[15] The refusal rate will be discussed in more detail below but it was around 18%. If a seed refused to participate they were not replaced, so there is some variation in the number of actual seed in each village, but all villages got some seeds

> *their village and will push them to get their children immunized. It is your choice to spread the information with whomsoever you want."*

The program launched in December 2016 and has been going on since then. The seeds have receive two monthly reminders, once by text message and once by voice message to encourage other to attend the immunization camp (they also received reminders about the incentive in incentives villages). The program has been going on for a year, and we have regular data since the beginning.

The seed villages were randomly assigned to 4 groups.

T1. Random seeds. In the random seeding group, we randomly selected six households from our census, and the seed was to be the head of the selected household.

In the three remaining groups, we first visited the village, and visited 17 randomly selected households. This was done in January and February 2016. We interviewed a respondent in the household asking them either of the gossip question. Note that in each village, we only asked one type of question, in order to keep the procedure simple to administer for the interviewer and to simulate real policy.

T2. Gossip seeds.

> *"Who are the people in this village, who when they share information, many people in the village get to know about it. For example, if they share information about a music festival, street play, fair in this village, or movie shooting many people would learn about it. This is because they have a wide network of friends/contacts in the village and they can use that to actively spread information to many villagers. Could you name four such individuals, male or female, that live in the village (within OR outside your neighborhood in the village) who when they say something many people get to know?"*

T3. Trusted seeds.

> *"Who are the people in this village that you and many villagers trust, both within and outside this neighborhood, trust? When I say trust I mean that when they give advice on something, many people believe that it is correct and tend to follow it. This could be advice on anything like choosing the right fertilizer for your crops, or keeping your child healthy. Could you name four such individuals, male or*

female, who live in the village (within OR outside your neighborhood in the village) and are trusted?"

T4. Trusted gossip seeds.

> *"Who are the people in this village, both within and outside this neighborhood, who when they share information, many people in the village get to know about it. For example, if they share information about a music festival, street play, fair in this village, or movie shooting many people would learn about it. This is because they have a wide network of friends/contacts in the village and they can use that to actively spread information to many villagers. Among these people, who are the people that you and many villagers trust? When I say trust I mean that when they give advice on something, many people believe that it is correct and tend to follow it. This could be advice on anything like choosing the right fertilizer for your crops, or keeping your child healthy. Could you name four such individuals, male or female, that live in the village (within OR outside your neighborhood in the village) who when they say something many people get to know and are trusted by you and other villagers?"*

Note that we specifically we asked about two things: fertilizers for crops and kids' health. This is in order to make sure that the trust question does not emphasize immunization. As in our previous experiment, the gossip question is centered purely on information transmission, and is phrased in a way to not flag any concerns about trust, while the trust questions explicitly asks about trust.

2.2.2. *Summary statistics.* Table 2 presents summary statistics about the number of seeds nominated in each (nomination) groups, the number of nominations received by the top 6 finalists (chosen as seed), the refusal rates, and the characteristics of the chosen seed in each of the group.

We received 19.9 gossip nominations per village (20.3 and 20.0 for trusted and trusted gossip nominations respectively). The top six nominees were selected per village, and the average number of nominations received per household was 11.2 for gossip seeding, 10.56 for trusted seeding, and 10.77 for trusted gossip seeding. Note that there is more consensus on the pure gossip than on the trusted seed: it is perhaps easier to know who is good at transmitting information than whom other people trust.

Most seeds agreed to be part of the experiment. The lowest refusal rate was among the gossip seeds (16.5%), followed by the trusted gossip (17.5%). The trusted and the random seeds were less likely to agree (22% and 19% refusals, respectively). This implies that we have slightly more active seeds in the gossip treatment, and this could account for part of the effect (but the difference is very small, not statistically significant, and every village had several active seeds).

Gossip seeds and trusted gossip seeds are very similar in terms of observable characteristics. They are slightly more likely to be female than random seeds (who are heads of households, and hence often male), although the vast majority are still male (12-13% females in gossip and trusted gossip groups). They are wealthier and more educated than the random seeds. They are much more likely to have some official responsibility in the village (*numberdhar* or *chaukihar*). Most notably, they are more likely to describe themselves as interactive. 46% of the gossips say that the interact very often with others, and that they participate frequently in community activities (the numbers are almost the same for trusted gossip), as against 26% for the random seeds and 37% for the trusted seeds. They are also more informed in the sense that they are more likely to know who the nurse in the local health subcenter is and that there is an immunization camp.

The trusted seeds are older, least likely to be female and Scheduled Castes, and tend to be wealthier than both gossip and random seeds. In terms of probability to hold an elected position, and of their level of interaction with the village, they are about halfway between the random seeds and the gossip or trusted gossip seeds.

2.2.3. *Impact on Immunization.* Our sample for analysis is restricted to the 517 villages for the seeding experiments, and the data is aggregated at the village×month level, which corresponds to the number of children who attended a monthly camp.[16] The dependent variable is the number of children in a village-month who got immunized against any particular disease, or for anything.

The empirical specification is as follows.

$$\begin{aligned}(2.5)\quad & y_{jt} = \theta_0 + \theta_1 GossipTreatment_j + \theta_2 TrustedTreatment_j + \\ & \qquad \theta_3 TrustedGossip_j + \theta_4 SlopeIncentive_j + \theta_5 FlatIncentive_i + D_k + M_t + \epsilon_{ji},\end{aligned}$$

where y_{jt} is the number of immunizations of each type received in the village, D_k is a set of seven district fixed effects and M_t is a set of month fixed effects. The standard

[16] Village month observations with zero child level observations are eliminated, since they were times with no camp.

errors are clustered at the sub-center level. For brevity, we do not report the incentive coefficients in the table.

The results are presented in Table 3. In a typical month, in the random seeding group, 17.07 children received at least one shot (column 6). In the gossip villages, four more children came every month for any immunization ($p = 0.09$). The results are not driven by any particular vaccine. There is a 25% increase in the number of children receiving each of the first three vaccines (BCG, penta 1 and penta 2) and a 28% increase for the two shots where the baseline levels tend to be lower (penta3 and measles). The increase of 0.96 children per village per month for measles is particularly important, as getting good coverage for measles immunization has proven very challenging in India.

These effects are somewhat smaller, in terms of proportions, than the results of the cell phone RCT (where we had an increase of 40%), but while that experiment was one-shot, this one continued for a year. Figure 3 shows a remarkable stability of the coefficient over time for the number of children receiving at least one shot in each month.

In term of point estimates, the impact of the trusted seed and the trusted gossip seed is about half that of the gossip, although given the standard errors we cannot reject either that for these two treatments there is no effect (compared to random seedings) or that the effect is as large as for the gossip seedings. At best, this suggests that there was no gain from explicitly trying to identify trustworthy people, even for a decision that probably requires some trust.

The results thus confirm that a simple procedure to identify key actors, namely interviewing a random set of households about who are good people to convey general information, leads to more diffusion of information over a long period of time, in a policy relevant context involving a serious and important decision with real consequences, relative to the seeding of a random person.

3. Network Communication and Knowledge of Centrality

On the one hand, these results may appear to be common sense. In order to find something out about a community, for example who is influential, why not just ask the community members? While this may seem obvious, this is not a strategy that is commonly employed by organizations in the field: they tend to rely on demographic or occupation characteristics, or on the judgement of a single extension officer (usually not from the village), rather than interview a few people and ask. One possible reason is that it is not in fact so obvious that they would know. Even in small communities, like the Karnataka villages where we conducted the cell phone RCT, people have a very dim idea of the network. Breza et al. (2017) show that 47% of randomly selected individuals are unable to offer a guess about whether two others in their village share a link and being one step further from the pair corresponds to a 10pp increase in the probability of mis-assessing link status. There is clearly considerable uncertainty over network structure among the villagers, but then how is it possible that they are able to nominate the right people in the network from the point of view of diffusing information? The goal of our theoretical section is to provide an answer to this question: we show that it is in fact entirely plausible that even a boundedly rational agent knows who is influential, even if they know almost nothing about the network.

3.1. A Model of Network Communication.

We consider the following model.

3.1.1. *A Network of Individuals.* A society of n individuals are connected via a directed and weighted network, which has an adjacency matrix $\mathbf{w} \in [0,1]^{n\times n}$. The ij-th entry is the relative probability with which i tells something to j. This relation does not have to be reciprocal.

Unless otherwise stated, we take the network $\mathbf{w}$ to be fixed and let $v^{(R,1)}$ be its first (right-hand) eigenvector, corresponding to the largest eigenvalue λ_1.The first eigenvector is nonnegative and real-valued by the Perron–Frobenius Theorem. Throughout what follows, 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.[17]

Two concepts, *diffusion centrality and network gossip* are central to the theory.

3.1.2. *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, based on a process that underlies many models of contagion.[18]

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

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

[18] See Jackson and Yariv (2011) for background and references on models of diffusion and contagion, and Bloch and Tebaldi (2016); Jackson (2017) for how diffusion centrality compares with some other centrality measures. A continuous time version of diffusion centrality was subsequently defined in Lawyer (2014).

of neighbors and history, each informed node i informs each of its neighbors j 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 good reasons to allow T to be finite. For instance, a new piece of information may only be relevant for a limited time. Also, after some time, boredom may set in or some other news may arrive and the topic of conversation may change.

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

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

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

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

So, $DC(\mathbf{w}, T)_i$ 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.

By allowing the network to be weighted and directed, this generalizes the notion of Diffusion Centrality that we defined in Banerjee, Chandrasekhar, Duflo, and Jackson (2013). That definition applied to the spacial case in which $\mathbf{g}$ is a (possibly directed) adjacency matrix taking on values in $\{0, 1\}$, and $\mathbf{w} = q\mathbf{g}$ for some communication probability $q \in (0, 1]$. That case, with corresponding hearing and diffusion centrality measures are $\mathbf{H}(q\mathbf{g}, T)$ and $DC(q\mathbf{g}, T)$, is useful in our empirical work. For the theory, we impose no requirement that the probabilities be similar across pairs of nodes, or even that two nodes reciprocate.

In Banerjee et al. (2013), we showed that diffusion centrality of the initially informed members of a community was a statistically significant predictor of the spread of information – in that case, about a microfinance program.

As we claimed in Banerjee et al. (2013), diffusion centrality nests three of the most prominent and widely used centrality measures: degree centrality, eigenvector centrality, and Katz–Bonacich centrality.[19] Diffusion centrality thus provides a foundation

[19] Let $d(\mathbf{w})$ denote (out) degree centrality: $d_i(\mathbf{g}) = \sum_j w_{ij}$. Eigenvector centrality corresponds to $v^{(R,1)}(\mathbf{w})$: the first eigenvector of $\mathbf{w}$. Also, let $GKB(\mathbf{w})$ denote a "generalized" version of

for these measures, but importantly it can behave very differently in the gap between them.

In Appendix B we prove that for the general class of weighted and directed networks:

(i) if $T = 1$, then diffusion centrality is proportional to (out) degree centrality, while

(ii) if T tends to ∞ then

   (a) if $\lambda_1(\mathbf{w}) < 1$, diffusion centrality coincides with Generalized Katz–Bonacich centrality, and and

   (b) if $\lambda_1(\mathbf{w}) > 1$, diffusion centrality approaches eigenvector centrality.

Part iib is the part that requires nontrivial proof, whereas the other parts are direct.

In view of the above results, the choice of parameters q, T make a difference when we operationalize $DC(q\mathbf{g}, T)$ for our empirical investigations.

The threshold of $q = 1/\lambda_1(\mathbf{g})$ (i.e., then $\lambda_1(\mathbf{w}) = 1$) is key, even when T is finite. In the Appendix B, we prove that diffusion centrality behaves fundamentally differently depending on whether $\lambda_1(\mathbf{w})$ is above or below 1. Intuitively, if the communication probabilities in $\mathbf{w}$ are small (when $\lambda_1(\mathbf{w}) < 1$), then very limited diffusion takes place even for large T; and if those probability are large (when $\lambda_1(\mathbf{w}) > 1$), then knowledge saturates the network. The threshold of $q = 1/\lambda_1(\mathbf{g})$ is thus the point at which information has a chance of reaching all nodes, but does not overly saturate.

We also show that diffusion centrality behaves quite differently depending on whether T is smaller or bigger than the diameter of the graph. If T is below the diameter, news from some nodes does not have a long enough time to other nodes. In contrast, once T exceeds the diameter of the graph, then many of the weighted walks counted by $\mathbf{w}^T$ have "echoes" in them: they visit some nodes multiple times. For instance, news passing from node 1 to node 2 to node 3 then back to node 2 and then to node 4, etc.

Thus, from the point of view of the empirical exercises that are at the heart of this paper, these results suggest that the threshold case of $q = 1/\mathrm{E}[\lambda_1(\mathbf{g})]$ and $T = \mathrm{E}[Diam(\mathbf{g})]$ provides a natural benchmark value for q and T, at which information can diffuse, but does not over-saturate a network. This allows us to assign numerical values to $DC(q\mathbf{g}, T)_i$. We use this throughout in our empirical analysis.

Katz–Bonacich centrality to account for possibly weighted and directed networks – defined when $\lambda_1(\mathbf{w}) < 1$ by $GKB(\mathbf{w}) := \left(\sum_{t=1}^{\infty} (\mathbf{w})^t\right) \cdot \mathbf{1}$.

3.1.3. *Network Gossip.* Diffusion centrality considers diffusion from the *sender's* perspective. Next, consider the same 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, i may tell j that he has a new car. Then j may tell k that "i has a new car," and then k may tell ℓ that "i has a new car." i may also have told u that he thinks house prices will go up, and u could have told ℓ that "i thinks that house prices will go up." In this model, ℓ keeps track of the cumulative number of times bits of information that originated from i reach her and compares it with the number of times she hears bits of 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 "i" – and not what the information is about. The first piece of news originating from i could be about something he has done ("bought a car"), but the second could just be an opinion ("i thinks house prices will go up"). ℓ keeps track of how often she hears of things originating from i. Then ℓ ranks people based on how often she hears about them. ℓ ranks i, j, k, and so on, just based on the frequency that she hears things that originated at each one of them.[20]

Recall that

$$\mathbf{H}(\mathbf{w}, T) = \sum_{t=1}^{T} (\mathbf{w})^t,$$

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

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

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 $NG(\mathbf{w}, T)_{ij}$ is twice as high as $NG(\mathbf{w}, T)_{kj}$ 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.

[20] Of course, one can imagine other gossip processes and could enrich the model along many dimensions. The point here is simply to provide a "possibility" result - to understand how it could be that people can easily learn information about the centrality of others. Noising up the model could noise up people's knowledge of others' centralities, but this benchmark gives us a starting point.

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.

3.2. **Relating Diffusion Centrality to Network Gossip.**

We now turn to the first of our main theoretical results. The main point we make here is that individuals in a society can easily estimate who is diffusion central simply by counting how often they hear gossip that originated with others.

3.2.1. *Identifying Central Individuals.*

We first show that, on average, individuals' rankings of others based on NG_j, the amount of gossip that j has heard about others, is positively correlated with diffusion centrality for any $\mathbf{w}, T$.

THEOREM **1.** *For any matrix of passing probabilities* $\mathbf{w}$ *and finite time* T*,*

$$\sum_j \operatorname{cov}(DC(\mathbf{w}, T), NG(\mathbf{w}, T)_j) = \operatorname{var}(DC(\mathbf{w}, T)).$$

Thus, in any network with differences in diffusion centrality among individuals, the average covariance between diffusion centrality and network gossip is positive.

We emphasize that although network gossip and diffusion centrality are both based on the same sort of information process, they are 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. Indeed, the reason that Theorem 1 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 relatively distant nodes, and moreover, they might be biased towards their local neighborhoods.[21]

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 – not just ordinally, but *cardinally.*

[21] 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. 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 $\sum$ in Theorem 1 is important

THEOREM **2.** *If $\lambda_1(\mathbf{w}) > 1$ and $\mathbf{w}$ is aperiodic, then as $T \to \infty$ every individual j's ranking of others under $NG(\mathbf{w}, T)_j$ converges to be proportional to diffusion centrality, $DC(\mathbf{w}, T)$, and hence according to eigenvector centrality, $v^{(R,1)}$.*

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

Note that in particular when we are interested in eliciting information as to which members of a network would be the most central, all this requires is that respondents track which individuals tend to be mentioned very often. They need not even track the counts or rankings of those who tend not to be mentioned frequently. Thus the computational burden is quite minimal.

More sophisticated strategies in which individuals try to infer network topology could accelerate learning. Nonetheless, what our result underscores is that 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 $\lambda_1(\mathbf{w}) > 1$ is important. When $\lambda_1(\mathbf{w})$ falls below 1, some people can hear about some others with vanishing frequency, and network distance between people influences whom they think is the most important.

Also, in our definition of network gossip, NG, nodes are similar in how frequently they generate new information or gossip - we weight the information passing but not its initial production. However, provided the generation rate of new information is positively related to nodes' centralities, the results still hold, though of course if the rate of generation of information about nodes were negatively correlated with their position, then our results would be attenuated. Regardless, the result is still of interest.

We have not discussed the possibility of hearing about people in other ways than through communication with friends: information only travels through edges in the network. This is not really an issue for two reasons. First, this is realistic in the contexts we study. Second, things like media outlets are easily treated as nodes in the network that receive and broadcast information, especially given that our analysis allows for arbitrarily weighted and directed networks.

Also, here we do not model any quality of information: there is no notion of trust nor endorsement. It could be, for example, that gossips are people who love to talk but are not necessarily reliable. In that case, their friends may resist passing on information originating from them. This is of interest in some settings and is an important issue for further research.

4. Additional Evidence: Who are the Gossips?

Although the model provides an explanation for why people are able to name good diffusers, even though they may have little network knowledge, there are alternative explanations. For example, people might simply be identifying individuals who talk a lot, or know many people, instead of highly central people in a diffusion centrality sense.

We return to data from Karnataka to present more evidence consistent with the more specific channel proposed in the model. We show that individuals nominate people who are significantly more central than the average, and especially in terms of diffusion centrality.

4.1. **Data Collection.** As an empirical study of people's ability to nominate central individuals, we use a rich network data set that we gathered from villages in rural Karnataka (India). We collected detailed network data in 2006 and again in 2012 in order to study the diffusion of microfinance as well as how networks changed in response to microfinance (Banerjee, Chandrasekhar, Duflo, and Jackson, 2013, 2018). We use that more recent and complete data here. To collect the network data (as described in detail in Banerjee et al. (2013, 2018)), we asked adults to name those with whom they interact in the course of daily activities.[22] 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, and (12) with whom he or she goes to pray (e.g., at a temple, church or mosque).

[22] 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 the second wave of data.

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 any selection bias associated with data-mining to find the most predictive subnetworks. The resulting objects are undirected, unweighted networks at the household level.

Table 4 provides summary statistics. The networks are typically sparse: the average number of households in a village is 196 with a standard deviation of 61.7, while the average degree per household 17.7 with a standard deviation of 9.8.

We combine that network data with "gossip" information from a subset of 33 villages. After the network data were collected, to collect the 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?*

We asked two questions to check whether there was any difference depending on what people thought was to be diffused. It made no difference, as is clear from the results below.[23]

Half of the households responded to our "gossip" questions. This is in itself intriguing. Some people may have been reluctant to offer an opinion if they are unsure of the answer.[24] In Appendix F we show that the patterns of who is more likely to offer a guess is consistent with our model above. In particular, in that appendix we show that people whose network position provides them with more accurate information about other people's diffusion centrality are more likely to offer an opinion.

Conditional on naming someone there is substantial concordance of opinion - in that people's nominations tend to coincide. 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

[23] The correlation between being nominated for a loan and an event is substantial: 0.76 (and 0.877 for the correlation between the number of times nominated under each category).

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

nominated *nine* times.[25] This is a first indication that the answers are meaningful, since if people are good at identifying central individuals, we would expect their nominations to coincide.

We label as "leaders" households that contain shopkeepers, teachers, and leaders of self-help groups – almost 12 percent of households fall into this category. This was how the Microfinance Institution (MFI) in our microfinance study defined leaders, who were identified as people to be seeded with information about their product (because it was believed they would be good at transmitting the information). The MFI's theory was that such "leaders" were likely to be well-connected in the villages and thereby would be good seeds for the diffusion of microfinance.[26]

Table 5 shows that there is some overlap between leaders and gossip nominees. We refer to the nominees as "gossips." Overall, 86% of the population were neither gossips nor leaders, just 1% were both, 3% were nominated but not leaders, and 11% were leaders but not nominated. This means that 9% of leaders were nominated as a gossip under the event question whereas 91% where not nominated. Similarly, 27% of nominated gossips under the event question were leaders, whereas 73% were not. The loan question demonstrates very similar results, and Figure 4 presents this information.

4.2. **Do individuals nominate central nodes?**

Our theoretical results suggest that people can learn others' diffusion centralities simply by tracking news that they hear through the network, and therefore should be able to 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.*

As motivating evidence, Figure 4 shows the distribution of diffusion centrality (normalized by its standard deviation across the sample for interpretability) across households that were nominated for the event question, those who were nominated as leaders, and those who were named for both or neither.[27] Clearly, the distribution of

[25] 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.

[26] 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.

[27] Recall from our discussion in Section 3.1.2, we set $q = 1/\mathrm{E}[\lambda_1]$ and $T = \mathrm{E}[Diam(\mathbf{g})]$ throughout our empirical analysis, based on our theoretical results.

centrality of those who are both nominated and are also leaders first order stochastically dominates the other distributions. Moreover, the distribution of centralities of those who are nominated but not leaders dominates the distribution of those who are leaders but were not nominated. Finally, those who are neither nominated nor a leader exhibit a distribution that is dominated by the rest. Taken together, this shows that individuals who are both nominated and leaders tend to be more central than those who are nominated but not leaders, who are in turn more central than those who are not nominated but are leaders.

Figure 5 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, then individuals might tend to nominate only households with whom they are directly connected. Panel A of Figure 5 shows that fewer than 13% of individuals nominate someone with whom they are linked in the network, compared to there being about 9% of households with whom a typical household is linked. At the same time, over 28% of nominations come from a network distance of at least three or more (41% of nodes are in this category). Therefore, although respondents do tilt nominations towards 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 the most central people tend to have more 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.

Panel B of Figure 5 shows that the average diffusion centrality in percentile terms of those named at distance 1 is higher than of those named at distance 2, which is higher than of those named at distance 3 or more. This suggests that individuals have more accurate information about central individuals who are closer to them, and when they don't, they are more reluctant to nominate (recall that fewer than half of the households nominate anyone, Appendix F).

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. For example, individuals may nominate the person who has the most friends, and people with many friends tend to be more diffusion central than those with fewer friends (i.e., diffusion centrality and degree centrality tend to be positively correlated). Alternatively, it may be that people simply nominate the "leaders" within their village, or people who are central geographically, and these could also correlate with diffusion centrality. There are 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 expected these people to be 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. For that reason, since 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.[28]

Although the correlations below do not constitute proof that the causal mechanism is indeed gossip, they do help to rule out these confounding factors.

Again, recall that to operationalize our analysis we use $DC\ (1/\mathrm{E}[\lambda_1]\mathbf{g}, \mathrm{E}[Diam(\mathbf{g}(n,p))])$ as our measure of diffusion centrality, as discussed in Section 3.1.2.

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.

[28] 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.

Given the large choice sets, it is convenient to 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 Table 6. A parallel OLS specification leads to the same conclusion, and is presented in Appendix C.

We begin with a number of bivariate regressions in Table 7. 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 repeat the exercise with two other network statistics (degree and eigenvector centrality), the the "leader" dummy, and an indicator for geographic centrality. All of these variables, except for geographic centrality, significantly 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 7. Degree is insignificant, and does not affect the coefficient of diffusion centrality. Eigenvector centrality is quite correlated with diffusion centrality (as it should be, since they converge to each other with enough time periods), and hard to distinguish from it. Introducing it cuts the effect of diffusion centrality by about 50%, though it remains significant. The leader dummy is close to being 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.

These results provide suggestive evidence that a key driver of the nomination decision involves diffusion centrality with $T > 1$, although it may be more difficult to separate eigenvector centrality and diffusion centrality from each other, which is not surprising since they are closely related concepts.

To confirm this pattern, in the last column, we introduce all the variables together and perform a LASSO analysis, which "picks" out the variable that is most strongly associated with the outcome variable, the number of nominations. 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 variable (the number of nominations). In the second step, a standard Poisson regression is run on the support selected in the first stage.[29,30]

As we did before, we consider the variables diffusion centrality, degree centrality, eigenvector centrality, leadership status, and geographic centrality in the standard LASSO to select the support. For the event nomination, LASSO picks out only one predictor: diffusion centrality. The post-LASSO coefficient and standard error thus exactly replicate the Poisson regression of using just diffusion centrality. This confirms that diffusion centrality is the key predictor of gossip nomination at least within the set of alternatives we have considered. For the loan nomination, the LASSO picks out both degree and diffusion centrality as relevant, though degree is insignificant. We repeat the analysis with OLS instead of Poisson regression in Appendix C, with identical qualitative results.

Thus, it appears that villagers to nominate people who tend to be diffusion central (and not some other obvious network characteristics). Of course, it does not provide a proof that they in fact track all the gossip they hear. It could be that they pick people with some characteristics (e.g., someone who is very talkative, which is something we do not observe) that is correlated with centrality, and they can easily pick up on.

Finally, we test a much more specific implication of the model, that relies not on j's characteristics but on j's relationship with i in the network. Observe that the theory suggests that a given individual i should be relatively more likely to nominate j as a gossip, conditional on the diffusion centrality of j, if NG_{ji} is higher. As discussed above, this captures the expected number of times i hears about news originating from j. In Table 8, we regress whether j was nominated by i on the (percentile) of j in i's network gossip assessment.[31] We include in specifications both individual i fixed effects and flexibly control for j's diffusion centrality or include j fixed effects. We find that the network gossip of j as evaluated by i is positively associated with i nominating j, conditional on individual level fixed effects and the diffusion centrality of j or j fixed effects. Specifically, being at the 99th percentile as compared to the

[29] 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 the usual OLS (in our case, Poisson) on the variables selected in the first stage in a second step will recover consistent estimates for the parameters of interest.

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

[31] Recall that network gossip for node i, NG_i is the i-th column vector of the hearing matrix $\mathbf{H}$ described in Section 3.1.3. As before, to compute this we set $q = 1/\mathrm{E}[\lambda_1]$ and $T = \mathrm{E}[Diam(\mathbf{g})]$.

1st percentile of $NG_{\cdot,i}$ corresponds to about an 84% increase in the probability of j being nominated by i ($p = 0.00$).

4.3. **Reintepreting the cell phone RCT results: does diffusion centrality capture gossip seed diffusion?**

To what extent is the greater diffusion of information in the Karnataka cell phone RCT 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 network data in 69 villages in which seeds were randomly selected (2 of the 71 villages were not accessible at the time). In these villages, by chance, some seeds happened to be gossips and/or elders. We create a measure of centrality that parallels the gossip dummy and elder dummy by forming a dummy for "high diffusion centrality." We defined a household has "high diffusion centrality" if its diffusion centrality is at least one standard deviation above the mean. With these measures, in our 69 villages, 13% of households are defined to have "high diffusion centrality", while 1.7% were nominated as seeds, and 9.6% are "leaders." Twenty-four villages have exactly one high diffusion centrality seed and 14 have more than one. Twenty-three villages have exactly one gossip seed, and 8 have more than one.[32]

Column 1 of Table 9 runs the same specification as in Table 1 but in the 68 random villages. In these villages, hitting a gossip by chance increases the number of calls by 6.65 (compared to 3.78 in the whole sample). In column 3, we regress the number of calls on a dummy for hitting = a high DC seed: high DC seeds do increase the number of calls (by 5.18 calls). Finally, we regress number of calls received only on dummy of hitting a high DC seed, and we see that the number of calls increase by 5.18. In column 2, we augment the specification in column 1 to add the dummy for "at least one DC central seed". Since DC and Gossip are correlated, the regression is not particularly precise. The point estimate of gossip, however, only declines slightly.

[32] We continue to exclude the one village in which a gossip seed broadcasted information. The results including that village are in Appendix E.2. They reinforce the conclusion that diffusion centrality does not capture everything about why gossips are good seeds, since this particular gossip seed had low diffusion centrality. With this village in, the coefficient of hitting at least one gossip does not decline when we control for diffusion centrality, and in fact diffusion centrality, even on its own, is not significantly associated with more diffusion.

The point estimates suggest that diffusion centrality captures part of the impact of a gossip nomination, but likely not all of it. Gossip seeds tend to be highly central, and information does spread more from highly central seeds. This accounts for some part of the reason why information diffuses more extensively from gossip nominated seeds. At the same time, it is also appears that the model does not capture the entire reason as to why gossip seeds are best for diffusing information: even controlling for their diffusion centrality, gossip seeds still lead to greater diffusion.

There are at least two reasons why we might expect this. First, it is likely that our measures of the network are imperfect, and so part of the extra diffusion from the gossip nominations could reflect that villagers have better estimates of diffusion centrality from their network gossip than we do from our surveys. Second, it also could be that the gossip nomination is a richer proxy for information diffusion than model-based centrality measure. For instance, 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 may be good at capturing those factors. However, the standard errors do not allow us to pinpoint how much of the extra diffusion coming from being nominated as a gossip is explained by network centrality.

5. Conclusion

In a specially designed RCT, we find that nominated individuals are indeed much more effective at diffusing a simple piece of information than other individuals, even village elders. Motivated by this evidence, we designed and implemented a large scale policy RCT to encourage the take up of immunization. We find very consistent results: there is an increase of over 20% in immunization visits when the seeds are a gossip nominees.

A simple network information model rationalizes these results, and illustrates that it should be easy for even very myopic and non-Bayesian (as well as fully rational) agents, simply by counting, to have an idea of who is central in their community – according to fairly complex measures of centrality. Motivated by this, we asked villagers to identify good diffusers in their village. They do not simply name locally central individuals (the most central among those they know), but actually name people who are *globally* central within the village. This suggests that people can use simple observations to learn valuable things about the complex social systems within which they are embedded, and that researchers and others who are interested in diffusing information have an easy and direct method of identifying highly central seeds.

Although 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 not only highly central but who are even more successful at diffusing information than the average highly central individual.

Our findings have important policy implications, since such nominations are easy to collect and therefore can be used in a variety of contexts, either on their own or combined with other easily collected data, to identify effective seeds for information diffusion. Thus, using this sort of protocol may be a cost-effective way to improve diffusion and outreach, as demonstrated in the Haryana immunization RCT.

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.

There are two limitations that are worth highlighting and discussing. First, this paper focuses on the pure transmission of information – simple knowledge that is either known or not. In some applications, people may not only need to know of an opportunity but may also be unsure of whether they wish to take advantage of that opportunity, and thus may also rely on endorsements of others. In those cases, trust in the sender will also matter in the diffusion process. We focus, for most of the paper, on the spread of simple sorts of information, and in the cell phone RCT, the piece of information we seeded was designed not to require trust in order to participate. Although issues of trust are certainly relevant in some applications, pure lack of information is often a binding and important constraint, and is therefore worthy of study. In addition, in our work on microfinance (Banerjee, Chandrasekhar, Duflo, and Jackson, 2013), for example, we could not reject the hypothesis that the role of the social network in the take up of microfinance was entirely mediated by information transmission, and that endorsement played no role. The example of immunization in this paper shows that even when the final outcome involved an important and personal decision, pure information "gossips" are also effective at expanding diffusion and while trust-based seeding strategies deliver very noisy, not statistically significant from zero, results.

Second, our experiments here are limited to communities on the order of a thousand people. It is clear that peoples' abilities to name highly central individuals may not scale fully to networks that involve hundreds of thousands or millions of people. Nonetheless, our work still demonstrates that people are effective at naming central people within reasonably sized communities. There are many settings, in both the developing and developed world, in which person-to-person communication within a community, company, department, or organization of limited scale is important. Our model and empirical findings are therefore a useful first step in a broader research agenda.[33]

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*.

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

Aral, S., L. Muchnik, and A. Sundararajan (2013): "Engineering social contagions: Optimal network seeding in the presence of homophily," *Network Science*, 1, 125–153.

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

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

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

——— (2018): "Changes in social network structure in response to exposure to formal credit," .

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

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

[33] More generally, one may want to choose many seeds in a large society, some within in each of various sub-communities, in which case the techniques developed here would still be useful.

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.

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

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

Bindlish, V. and R. E. Evenson (1997): "The impact of T&V extension in Africa: The experience of Kenya and Burkina Faso," *The World Bank Research Observer*, 183–201.

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

Bloch, Francis and, M. O. and P. Tebaldi (2016): "Centrality Measures in Networks," *http://ssrn.com/abstract=2749124.*

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

Bonacich, P. (1987): "Power and centrality : a family of measures," *American Journal of Sociology*, 92, 1170–1182.

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

——— (2006): "Identifying sets of key players in a social network," *Computational & Mathematical Organization Theory*, 12, 21–34.

Breza, E., A. Chandrasekhar, and A. Tahbaz-Salehi (2017): "Seeing the forest for the trees? An investigation of network knowledge," .

Brunson, E. K. (2013): "The impact of social networks on parents? vaccination decisions," *Pediatrics*, peds–2012.

Centola, D. (2010): "The Spread of Behavior in an Online Social Network Experiment," *Science*, 329: 5996, 1194–1197, DOI: 10.1126/science.1185231.

——— (2011): "An experimental study of homophily in the adoption of health behavior," *Science*, 334(6060), 1269–1272.

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

Coleman, J., E. Katz, and H. Menzel (1966): *Medical Innovation: A Diffusion Study*, Indianapolis, Ind.: Bobbs-Merrill.

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

FELD, S. L. (1991): "Why Your Friends Have More Friends Than You Do," *American Journal of Sociology*, 96:6, 1464–1477.

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

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

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.

HINZ, O., B. SKIERA, C. BARROT, AND J. U. BECKER (2011a): "Seeding strategies for viral marketing: An empirical comparison," *Journal of Marketing*, 75, 55–71.

——— (2011b): "Seeding Strategies for Viral Marketing: An Empirical Comparison," *Journal of Marketing*, 75:6, 55–71.

IYENGAR, R., C. V. DEN BULTE, AND T. W. VALENTE (2010): "Opinion Leadership and Social Contagion in New Product Diffusion," *Marketing Science*, 30:2, 195–212.

JACKSON, M. O. (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.

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

——— (2016): "The Friendship Paradox and Systematic Biases in Perceptions and Social Norms," *Journal of Political Economy, forthcoming.*

——— (2017): "A Typology of Social Capital and Associated Network Measures," *SSRN http://ssrn.com/abstract=3073496.*

JACKSON, M. O. 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.*

KATONA, Z., P. P. ZUBCSEK, AND M. SARVARY (2011): "Network Effects and Personal Influences: The Diffusion of an Online Social Network," *Journal of Marketing Research*, 48:3, 425–443.

KATZ, E. AND P. LAZARSFELD (1955): *Personal influence: The part played by people in the flow of mass communication*, Free Press, Glencoe, IL.

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.

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

KIM, D. A., A. R. HWONG, D. STAFF, D. A. HUGHES, A. J. OÂĂŹMALLEY, J. H. FOWLER, AND N. A. CHRISTAKIS (2015): "Social network targeting to maximise population behaviour change: a cluster randomised controlled trial," *Lancet*, 386, 145–153.

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

——— (1996): "Strucutral Leverage in Marketing," in *Networks in Marketing*, ed. by D. Iacobucci, Sage, Thousand Oaks, 50–59.

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

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.

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

LIM, Y., A. OZDAGLAR, AND A. TEYTELBOYM (2015): "A Simple Model of Cascades in Networks," *mimeo.*

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

PALUCK, E. L., H. SHEPHERD, AND P. M. ARONOW (2016): "Changing climates of conflict: A social network experiment in 56 schools," *Proceedings of the National Academy of Sciences*, 113, 566–571.

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

SIMMEL, G. (1908): *Sociology: Investigations on the Forms of Sociation*, Leipzig: Duncker and Humblot.

VALENTE, T. W. (2010): *Social networks and health: Models, methods, and applications*, vol. 1, Oxford University Press New York.

——— (2012): "Network interventions," *Science*, 337, 49–53.

Valente, T. W., K. Coronges, C. Lakon, and E. Costenbader (2008): "How Correlated Are Network Centrality Measures?" *Connect (Tor).*, 28, 16–26.

Figures

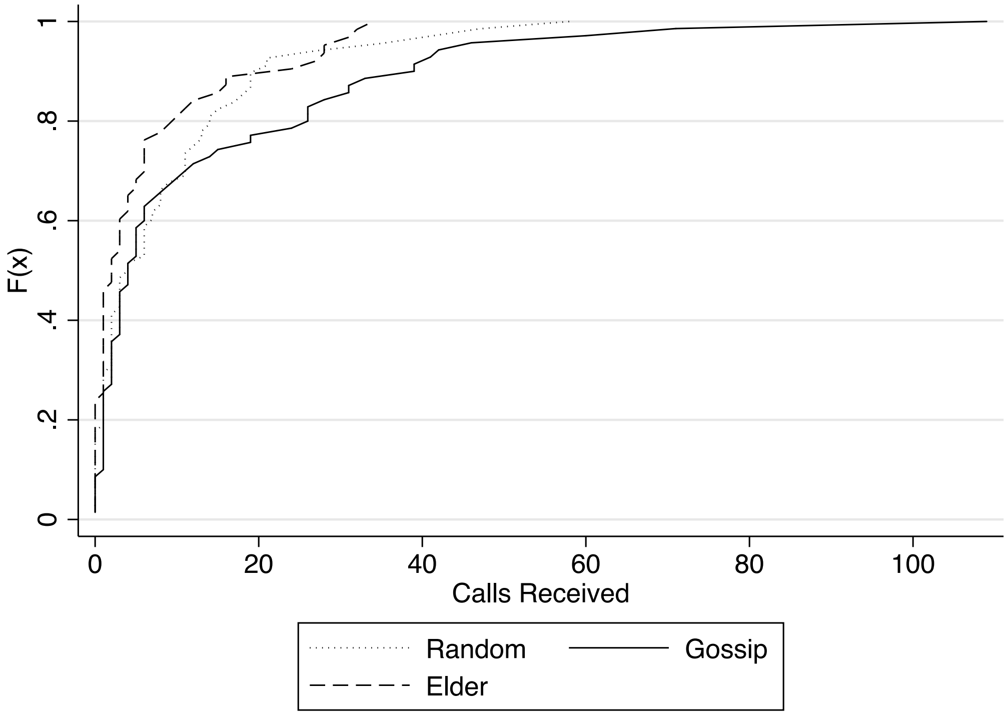
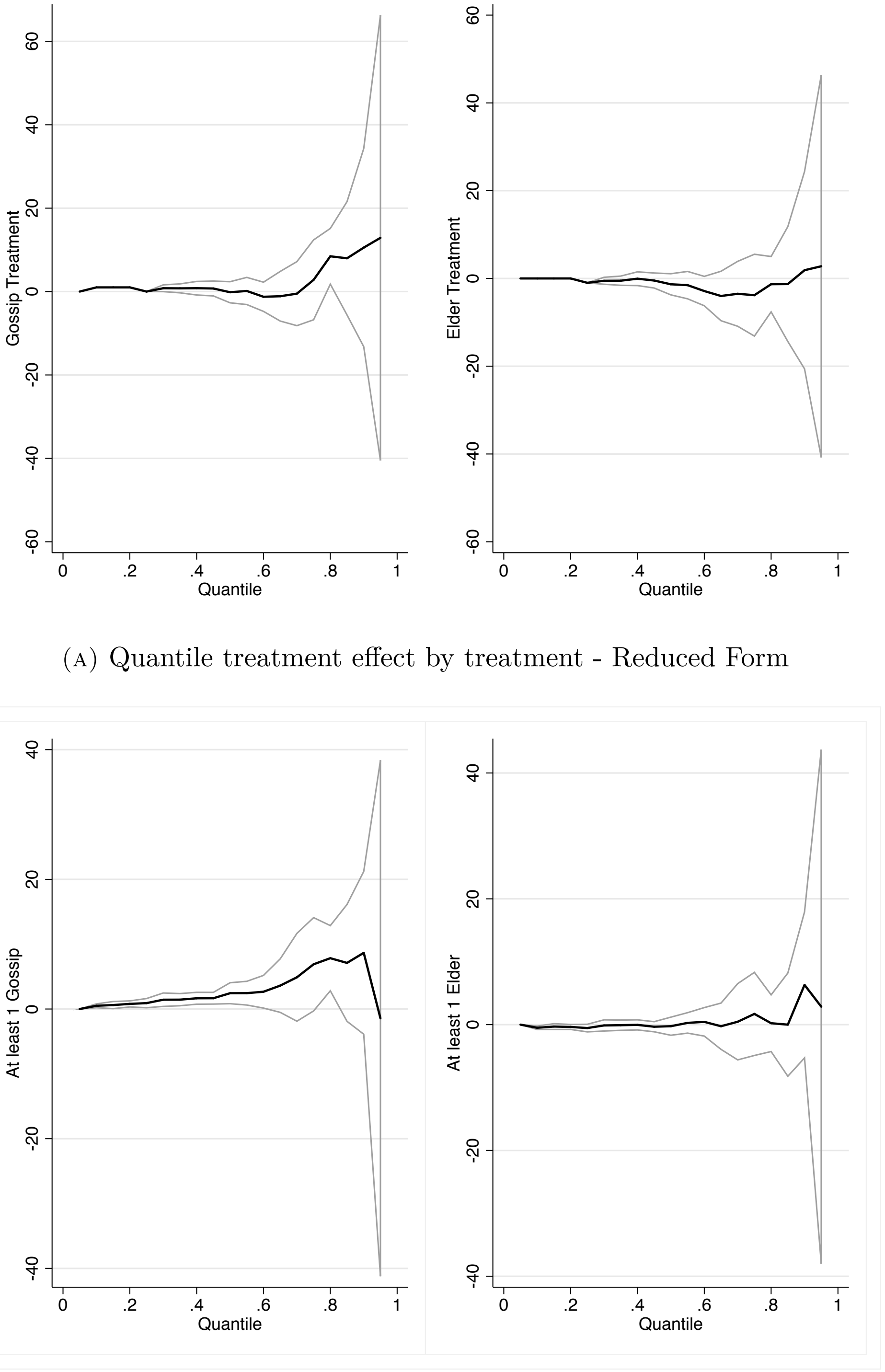


Figure 1. Distribution of calls received by treatment in the Karnataka cell phone RCT.



(A) Quantile treatment effect by treatment - Reduced Form

(B) Quantile treatment effect by hitting at least one gossip or elder

FIGURE 2. Quantile treatment effects where for $j \in \{Gossip,\ Elder\}$, $\widehat{\beta}_j(u)$ is computed for $u = \{0.05, ..., 0.95\}$. The intercept $\alpha(u)$ (not pictured) in each case is the omitted category corresponding to the random treatment.

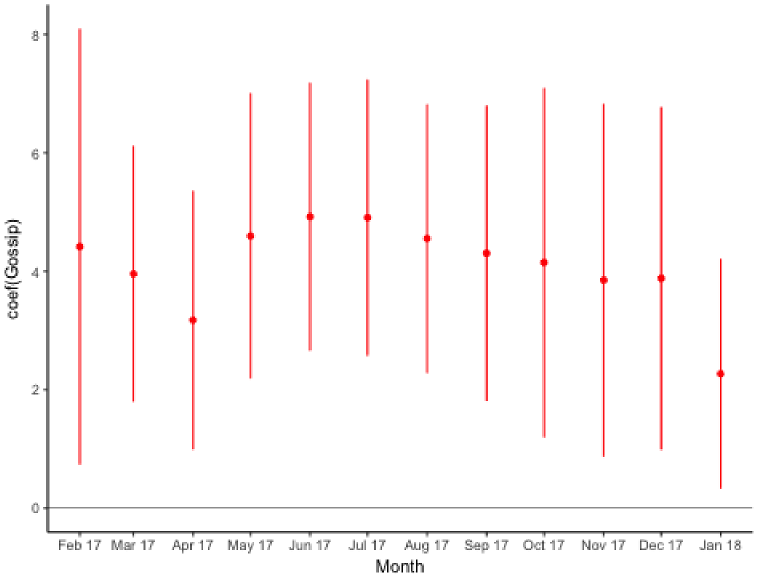
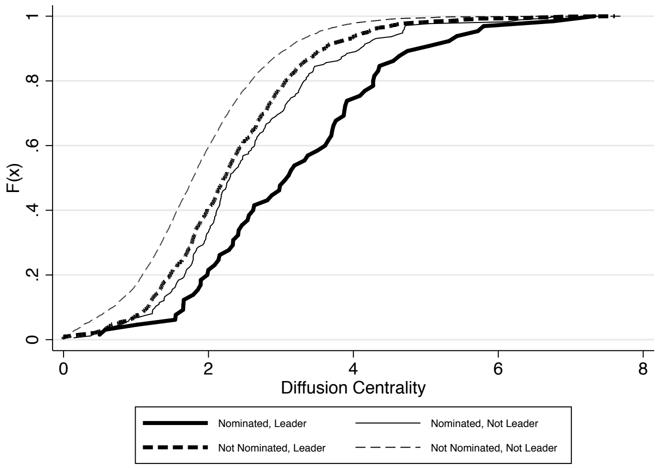
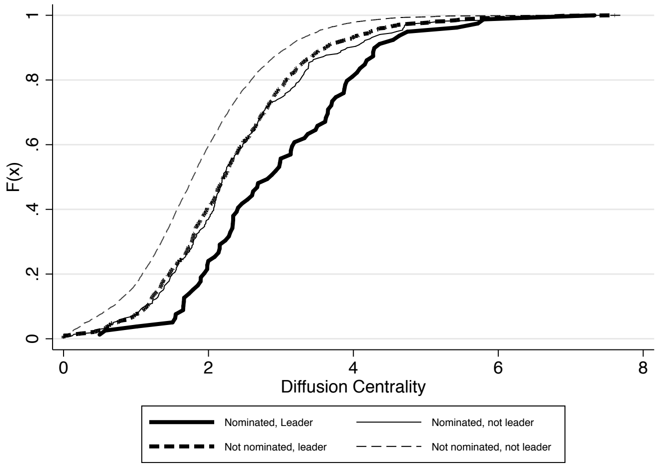


FIGURE 3. Number of kids receiving at least 1 vaccination per month in the Haryana Immunization RCT.



(A) Event question

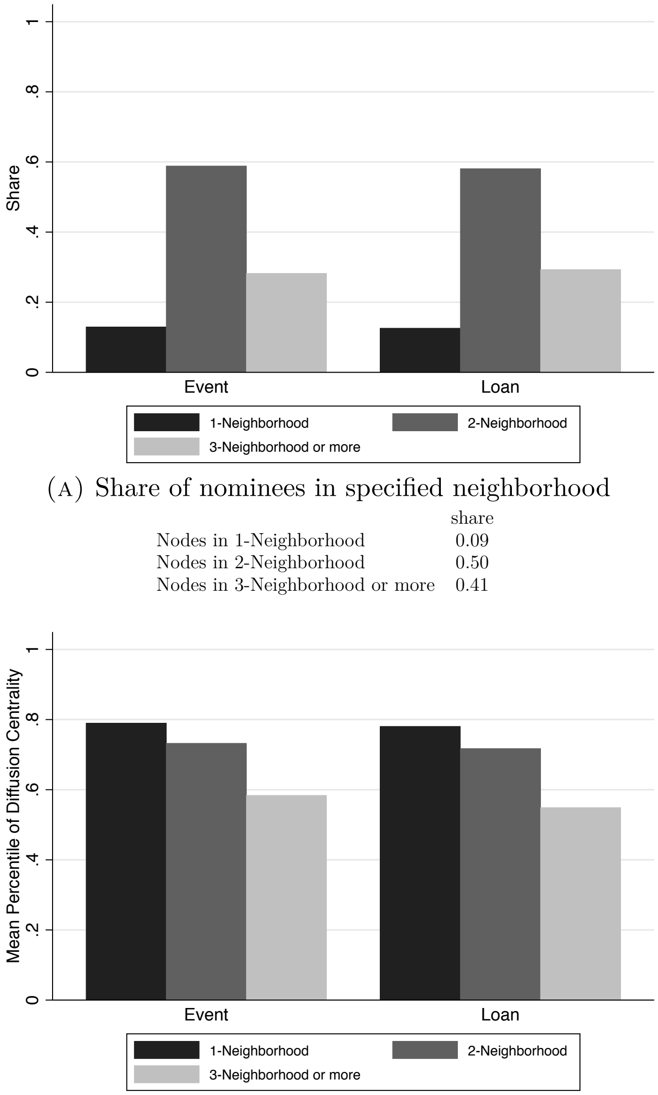
	population share
nominated, leader (event)	0.01
not nominated, leader (event)	0.11
nominated, not leader (event)	0.03
not nominated, not leader (event)	0.86



(B) Loan question

	population share
nominated, leader (loan)	0.01
not nominated, leader (loan)	0.10
nominated, not leader (loan)	0.03
not nominated, not leader (loan)	0.85

Figure 4. This figure uses the Karnataka microfinance village (wave 2) dataset. It presents CDFs of the (normalized) diffusion centrality, diffusion centrality divided by the standard deviation, conditional on classification (whether or not it is nominated under the event question in Panel A and the loan question in Panel B and whether or not it has a village leader).



(A) Share of nominees in specified neighborhood

	share
Nodes in 1-Neighborhood	0.09
Nodes in 2-Neighborhood	0.50
Nodes in 3-Neighborhood or more	0.41

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

FIGURE 5. Distribution of nominees and their diffusion centrality by network distance in the Karnataka microfinance village (wave 2) dataset.

Tables

Table 1. Calls received by treatment

	(1)	(2)	(3)	(4)	(5)
	RF	OLS	IV 1: First Stage	IV 2: First Stage	IV: Second Stage
VARIABLES	Calls Received	Calls Received	At least 1 Gossip	At least 1 Elder	Calls Received
Gossip Treatment	3.651		0.644	0.328	
	(2.786)		(0.0660)	(0.0824)	
Elder Treatment	-1.219		0.230	0.842	
	(2.053)		(0.0807)	(0.0509)	
At least 1 Gossip		3.786			7.436
		(1.858)			(4.266)
At least 1 Elder		0.792			-3.475
		(2.056)			(2.259)
Observations	212	212	212	212	212
Control Group Mean	8.077	5.846	0.391	0.184	5.805
Gossip Treatment=Elder Treatment (pval.)	0.0300		0	0	
At least 1 Gossip=At least 1 Elder (pval.)		0.330			0.0300

	(1)	(2)	(3)	(4)	(5)
	RF	OLS	IV 1: First Stage	IV 2: First Stage	IV: Second Stage
VARIABLES	$\frac{\text{Calls Received}}{\text{Seeds}}$	$\frac{\text{Calls Received}}{\text{Seeds}}$	At least 1 Gossip	At least 1 Elder	$\frac{\text{Calls Received}}{\text{Seeds}}$
Gossip Treatment	1.053		0.644	0.328	
	(0.698)		(0.0660)	(0.0824)	
Elder Treatment	-0.116		0.230	0.842	
	(0.518)		(0.0807)	(0.0509)	
At least 1 Gossip		0.952			1.979
		(0.501)			(1.071)
At least 1 Elder		0.309			-0.677
		(0.511)			(0.588)
Observations	212	212	212	212	212
Control Group Mean	1.967	1.451	0.391	0.184	1.317
Gossip Treatment=Elder Treatment (pval.)	0.0400		0	0	
At least 1 Gossip=At least 1 Elder (pval.)		0.410			0.0400

Notes: This table uses data from the Karnataka cell phone RCT dataset. Panel A uses the number of calls received as the outcome variable. Panel B normalizes the number of calls received by the number of seeds, 3 or 5, which is randomly assigned. For both panels, Column (1) shows the reduced form results of regressing number of calls received on dummies for gossip treatment and elder treatment. Column (2) regresses number of calls received on the dummies for if at least 1 gossip was hit and for if at least 1 elder was hit in the village. Columns (3) and (4) show the first stages of the instrumental variable regressions, where the dummies for “at least 1 gossip” and “at least 1 elder” are regressed on the exogenous variables: gossip treatment dummy and elder treatment dummy. Column (5) shows the second stage of the IV; it regresses the number of calls received on the dummies for if at least 1 gossip was hit and if at least 1 elder was hit, both instrumented by treatment status of the village (gossip treatment or not, elder treatment or not). All columns control for number of gossips, number of elders, and number of seeds. For columns (1), (3), and (4) the control group mean is calculated as the mean expectation of the outcome variable when the treatment is “random”. For columns (2) and (5) the control group mean is calculated as the mean expectation of the outcome variable when no gossips or elders are reached. The control group mean for the second stage IV is calculated using IV estimates. Robust standard errors are reported in parentheses.

Table 2. Summary Statistics of Haryana Immunization RCT

	(1)	(2)	(3)	(4)
	Random Seed	Gossip Seed	Trusted Seed	Trusted Gossip Seed
Nominations Statistics (per village)				
Number of Nominations	.	19.915	20.313	19.993
	.	(8.585)	(8.670)	(11.351)
Nominations for top 6 individuals	.	11.217	10.560	10.769
	.	(4.576)	(4.265)	(5.575)
Seed Characteristics				
Refused to Participate	0.186	0.165	0.219	0.175
	(0.389)	(0.372)	(0.414)	(0.380)
Age	49.233	48.569	52.040	48.890
	(14.617)	(14.347)	(14.130)	(14.082)
Female	0.067	0.129	0.070	0.119
	(0.250)	(0.336)	(0.256)	(0.324)
Education (years)	6.980	8.499	8.116	8.753
	(4.280)	(3.966)	(4.073)	(3.930)
Owns Land	0.586	0.675	0.680	0.687
	(0.493)	(0.469)	(0.467)	(0.464)
Wealth index from assets	0.183	0.218	0.217	0.226
	(0.098)	(0.121)	(0.114)	(0.120)
Hindu	0.866	0.876	0.876	0.892
	(0.341)	(0.330)	(0.330)	(0.311)
Muslim	0.103	0.107	0.103	0.086
	(0.305)	(0.310)	(0.304)	(0.281)
Scheduled Caste/ Tribe	0.231	0.200	0.173	0.200
	(0.422)	(0.400)	(0.378)	(0.400)
Other Backwards Caste	0.237	0.253	0.246	0.209
	(0.426)	(0.435)	(0.431)	(0.407)
Panchayat Member	0.106	0.320	0.259	0.300
	(0.308)	(0.467)	(0.438)	(0.459)
Numberdaar or Chaukidaar	0.112	0.353	0.261	0.326
	(0.316)	(0.478)	(0.439)	(0.469)
Interacts with Others: Very Often	0.263	0.455	0.371	0.444
	(0.441)	(0.498)	(0.483)	(0.497)
Paricipates in Community Activities: Very Often	0.264	0.457	0.371	0.445
	(0.441)	(0.499)	(0.483)	(0.497)
Aware of Immunization Camps	0.687	0.758	0.689	0.762
	(0.464)	(0.428)	(0.463)	(0.426)
Aware of ANMs	0.432	0.646	0.574	0.622
	(0.496)	(0.479)	(0.495)	(0.485)
Aware of Ashas	0.605	0.794	0.706	0.780
	(0.489)	(0.404)	(0.456)	(0.415)
Observations	570	648	712	674

TABLE 3. Haryana immunization program, communication treatment effect

	(1)	(2)	(3)	(4)	(5)	(6)
	Children	Children	Children	Children	Children	Children
VARIABLES	received BCG	received Penta1	received Penta2	received Penta3	received Measles	received at least one vaccine
Gossip	2.091	1.969	1.657	1.522	1.026	4.054
	(1.092)	(1.035)	(0.841)	(0.71)	(0.503)	(2.273)
Trusted	1.371	1.311	1.204	1.117	0.656	2.476
	(0.982)	(0.922)	(0.746)	(0.628)	(0.44)	(2.007)
Trusted Gossip	1.144	1.078	0.944	0.912	0.625	2.631
	(0.937)	(0.874)	(0.703)	(0.584)	(0.413)	(1.894)
Observations	5558	5558	5558	5558	5558	5558
Villages	516	516	516	516	516	516
Mean (Random seeds)	8.225	7.827	6.359	5.231	3.616	17.034
Gossip=Random (pval.)	0.055	0.057	0.049	0.032	0.041	0.075
Gossip=Trusted (pval.)	0.531	0.542	0.605	0.589	0.483	0.509
Gossip=Trusted Gossip (pval.)	0.39	0.385	0.389	0.383	0.418	0.535

Notes: This table uses data from the Haryana Immunization RCT. It reports estimates of the communication treatment effect. The outcomes are the number of children that received a vaccine by month in a village. Regressions include incentive treatment and the interaction between month and district fixed effects. Standard errors (clustered at the subcenter level) are reported in parentheses.

Table 4. 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 distnace between nodes in a village	2.37	0.33
fraction in the giant component	0.98	0.01
is a leader	0.12	0.32
nominated someone for event	0.38	0.16
nominated someone for loan	0.48	0.16
was nominated for event	0.04	0.2
was nominated for loan	0.05	0.3
number of nominations received for event	0.34	3.28
number of nominations received for loan	0.45	3.91

Notes: This table presents summary statistics from the Karnataka microfinance village (wave 2) dataset: 33 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 5. Leader Gossip Overlap

	share
leaders who are nominated (loan)	0.11
nominated who are leaders (loan)	0.27
leaders who are not nominated (loan)	0.89
nominated who are not leaders (loan)	0.73
leaders who are nominated (event)	0.09
nominated who are leaders (event)	0.27
leaders who are not nominated (event)	0.91
nominated who are not leaders (event)	0.73

Notes: This table presents the overlap between "leaders" in the sample and those nominated as gossips (for loan and event).

Table 6. Factors predicting nominations

	(1) Event	(2) Event	(3) Event	(4) Event	(5) Event
Diffusion Centrality	0.607				
	(0.085)				
Degree Centrality		0.460			
		(0.078)			
Eigenvector Centrality			0.605		
			(0.094)		
Leader				0.915	
				(0.279)	
Geographic Centrality					-0.082
					(0.136)
Observations	6,466	6,466	6,466	6,466	6,466
	(1) Loan	(2) Loan	(3) Loan	(4) Loan	(5) Loan
Diffusion Centrality	0.625				
	(0.075)				
Degree Centrality		0.490			
		(0.067)			
Eigenvector Centrality			0.614		
			(0.084)		
Leader				1.013	
				(0.263)	
Geographic Centrality					-0.113
					(0.082)
Observations	6,466	6,466	6,466	6,466	6,466

Notes: This table uses data from the Karnataka microfinance village (wave 2) dataset. It reports estimates of Poisson regressions where the outcome variable is the expected number of nominations. Panel A presents results for the event question, and Panel B presents results for the loan question. Degree centrality, eigenvector centrality, and diffusion centrality, $DC\,(1/\mathrm{E}[\lambda_1]\mathbf{g}, \mathrm{E}[Diam(\mathbf{g}(n,p))])$, are normalized by their standard deviations. Standard errors (clustered at the village level) are reported in parentheses.

Table 7. Factors predicting nominations

	(1) Event	(2) Event	(3) Event	(4) Event	(5) Event	(6) Event
Diffusion Centrality	0.642	0.354	0.567	0.606	0.374	0.607
	(0.127)	(0.176)	(0.091)	(0.085)	(0.206)	(0.085)
Degree Centrality	-0.039				-0.020	
	(0.101)				(0.101)	
Eigenvector Centrality		0.283			0.281	
		(0.186)			(0.186)	
Leader			0.535			
			(0.301)			
Geographic Centrality				-0.082		
				(0.142)		
Observations	6,466	6,466	6,466	6,466	6,466	6,466
Post-LASSO						✓

	(1) Loan	(2) Loan	(3) Loan	(4) Loan	(5) Loan	(6) Loan
Diffusion Centrality	0.560	0.431	0.578	0.624	0.339	0.560
	(0.122)	(0.130)	(0.081)	(0.075)	(0.170)	(0.122)
Degree Centrality	0.070				0.088	0.070
	(0.086)				(0.084)	(0.086)
Eigenvector Centrality		0.219			0.231	
		(0.138)			(0.138)	
Leader			0.623			
			(0.288)			
Geographic Centrality				-0.115		
				(0.089)		
Observations	6,466	6,466	6,466	6,466	6,466	6,466
Post-LASSO						✓

Notes: This table uses data from the Karnataka microfinance village (wave 2) dataset. It reports estimates of Poisson regressions where the outcome variable is the expected number of nominations under the event question. Panel A presents results for the event question, and Panel B presents results for the loan question. Degree centrality, eigenvector centrality, and diffusion centrality, $DC\,(1/\mathrm{E}[\lambda_1]\mathbf{g}, \mathrm{E}[\mathit{Diam}(\mathbf{g}(n,p))])$, are normalized by their standard deviations. Column (6) 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.

Table 8. Does network gossip differentially predict nominations?

VARIABLES	(1) Nominated	(2) Nominated	(3) Nominated	(4) Nominated	(5) Nominated	(6) Nominated
Percentile of Network Gossip j, i	0.256	0.245	0.348	0.356	0.068	0.080
	(0.090)	(0.105)	(0.049)	(0.057)	(0.030)	(0.032)
Observations	665,301	665,301	665,301	665,301	665,301	665,301
Dep. var mean	0.382	0.382	0.382	0.382	0.382	0.382
Respondent FE		✓		✓		✓
Rankee FE					✓	✓
Flexible Controls for DC			✓	✓		

Notes: This table uses data from the Karnataka microfinance village (wave 2) dataset. The data consists of an individual level panel and the outcome variable is whether a given respondent i nominated j or not under the lottery gossip question.The key regressor is the percentile of j's network gossip for i. We present OLS regressions where columns 2 and 4 include individual fixed effects, columns 3 and 4 control flexibly for a third-degree polynomial of diffusion centrality of j, column 5 includes rankee (j level) fixed effects, and column 6 has both i and j level fixed effects. Standard errors (clustered at the village level) are reported in parentheses.

Table 9. Calls received by seed type

VARIABLES	(1) Calls Received	(2) Calls Received	(3) Calls Received	(4) $\frac{\text{Calls Received}}{\text{Seeds}}$	(5) $\frac{\text{Calls Received}}{\text{Seeds}}$	(6) $\frac{\text{Calls Received}}{\text{Seeds}}$
At least 1 Gossip	6.645	5.574		1.637	1.370	
	(3.867)	(4.119)		(0.949)	(0.992)	
At least 1 Elder	0.346	0.0566		0.245	0.173	
	(3.602)	(3.576)		(0.926)	(0.912)	
At least 1 High *DC* Seed		3.663	5.183		0.916	1.312
		(2.494)	(2.383)		(0.623)	(0.649)
Observations	68	68	68	68	68	68
Control Group Mean	5.586	5.586	5.719	1.353	1.353	1.402
At least 1 Gossip=At least 1 Elder (pval.)	0.260	0.340		0.310	0.400	
At least 1 Gossip=At least 1 High *DC* Seed (pval.)		0.730			0.720	
At least 1 Elder=At least 1 High *DC* Seed (pval.)		0.420			0.480	

Notes: This table uses data from the Karnataka cell phone RCT and follow-up network dataset. It presents OLS regressions of number of calls received (and number of calls received normalized by the number of seeds, 3 or 5, which is randomly assigned) on characteristics of the set of seeds. High *DC* refers to a seed being above the mean by one standard deviation of the centrality distribution. All columns control for total number of gossips, number of elders, and number of seeds. For columns (1), (2), (4), and (5), the control group mean is calculated as the mean expectation of the outcome variable when no gossips or elders are reached. For columns (3) and (6), the control group mean is calculated as the mean expectation of the outcome variable when no high *DC* seeds are reached. Robust standard errors are reported in parentheses.

Appendix A. Threshold Parameters (q, T) for Diffusion Centrality

We present two new theoretical results about diffusion centrality: Theorem A.1 and Corollary A.1. We explicitly demonstrate that there are natural intermediate parameters associated with diffusion centrality at which it is distinct from the two boundary cases in which it simplifies to other well-known centrality measures.

We can think of overall the number of times everyone is informed about information coming from a seed as being composed of direct paths (the seed, i, tells j), indirect natural paths (i tells j who tells k who tells l and each are distinct), and echoes or other cycles (i tells j who tells k who tells j who tells k). If there is only one round of communication, then information never travels beyond the seed's neighborhood. In that case diffusion centrality just counts direct paths and is coincides with degree centrality. On the other hand, if there are infinite rounds of communication (and the probability of communicating across a link is high enough), diffusion centrality converges to eigenvector centrality, and by capturing arbitrary walks is partly driven by echoes and cycles as well as potentially long indirect paths. Our proposed intermediate benchmark captures direct paths and indirect natural paths and involves fewer cycles (which then become endemic as T goes to infinity). Theorem A.1 and Corollary A.1 below are theoretical results that provide network-based guidance on what intermediate parameters achieve this goal of mostly stripping out echoes.

For the formal analysis we limit ourselves to a sequence of Erdos–Renyi networks, as those provide for clear limiting properties. These properties extend to more general classes of random graph models by standard arguments (e.g., see Jackson (2008a)), but an exploration of such models takes us beyond our scope here.

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 A.1. *If T is not too large ($T = o(pn)$),*[34] *then the expected diffusion centrality of any node converges to $npq\frac{1-(npq)^T}{1-npq}$. That is, for any i,*

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

[34] 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.

Theorem A.1 provides a precise expression for how diffusion centrality behaves in large graphs. Provided that T grows at a rate that is not overly fast[35], then we *expect* the diffusion centrality of a typical node to converge to $npq\frac{1-(npq)^T}{1-npq}$. Of course, individual nodes vary in the centralities based on the realized network, but this result provides us with the extent of diffusion that is expected from nodes, on average.

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

Corollary **A.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* $\varepsilon > 0$[36] *and any corresponding* $T = o(pn)$*. Then for any node* i*:*

(1) $1/\mathrm{E}[\lambda_1]$ *is a threshold for* q *as to whether diffusion reaches a vanishing or expanding number of nodes :*
  (a) *If* $q = o(1/\mathrm{E}[\lambda_1])$*, then* $\mathrm{E}\left[DC\left(q\mathbf{g}(n,p),T\right)_i\right] \to 0$.
  (b) *If* $1/\mathrm{E}[\lambda_1] = o(q)$*, then* $\mathrm{E}\left[DC\left(q\mathbf{g}(n,p),T\right)_i\right] \to \infty$.[37]

(2) $\mathrm{E}[Diam(\mathbf{g}(n,p))]$ *is a threshold relative for* T *as to whether diffusion reaches a vanishing or full fraction of nodes:*[38]
  (a) *If* $T < (1-\varepsilon)\mathrm{E}[Diam(\mathbf{g}(n,p))]$ *for some* $\varepsilon > 0$*, then* $\frac{\mathrm{E}\left[DC(q\mathbf{g}(n,p),T)_i\right]}{n} \to 0$.
  (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* $\frac{\mathrm{E}\left[DC(q\mathbf{g}(n,p),T)_i\right]}{n} = \Omega(1)$.

Putting these results together, we know that $q = 1/\mathrm{E}[\lambda_1]$ and $T = \mathrm{E}[Diam(\mathbf{g})]$ are the critical values where the process transitions from a regime where diffusion is expected (in a large network) to reach almost nobody to one where it will saturate the network. At the critical value itself, diffusion reaches a non-trivial fraction of the network but not everybody in it.

This makes $DC\left(1/\mathrm{E}[\lambda_1]\mathbf{g}, \mathrm{E}[Diam(\mathbf{g}(n,p))]\right)$ an interesting measure of centrality, distinct from other standard measures of centrality at these values of the parameters. This fixes q and T as a function of the graph so that the centrality measure no longer

[35] Note that T can still grow at a rate that can tend to infinity and in particular can grow faster than the growth rate of the diameter of the network – T can grow up to pn, which will generally be larger than $\log(n)$, while diameter is proportional to $\log(n)/\log(pn)$.

[36] 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).

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

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

has any free parameters – enabling one to compare it to other centrality measures without worrying that it performs better simply because it has parameters that can be adjusted by the researcher. As per our discussion in subsection 3.1.2, we set $q = 1/\mathrm{E}[\lambda_1]$ and $T = \mathrm{E}[Diam(\mathbf{g})]$ throughout our empirical analysis.

For Online Publication

Appendix B. Proofs

B.1. Relation of Diffusion Centrality to Other Measures.

We prove all of the statements for the case of weighted ($w_{ij} \in [0,1]$) and directed networks. Thus, $\mathbf{w} \in [0,1]^{n\times n}$ to allow for full heterogeneity in communication. For instance, w_{ij} and w_{ik} could both be positive, and yet differ from each other, or one could be positive and the other zero, or both zero, etc.

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, $v^{(L,k)} = v^{(R,k)}$.

Let $d(\mathbf{w})$ denote (out) degree centrality (so $d_i(\mathbf{w}) = \sum_j w_{ij}$). Eigenvector centrality corresponds to $v^{(R,1)}(\mathbf{w})$: the first eigenvector of $\mathbf{w}$. Also, let $GKB(\mathbf{w})$ denote generalized Katz–Bonacich centrality – defined for $\lambda_1(\mathbf{w}) < 1$ by:[39]

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

It is direct to see that (i) diffusion centrality is proportional to out degree centrality at the extreme at which $T = 1$, and (ii) if $\lambda_1 < 1$, then diffusion centrality coincides with generalized Katz–Bonacich centrality if we set $T = \infty$. We now show that when $\lambda_1 > 1$ 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 λ_1 is sufficiently small so that limited diffusion takes place even in the limit of large T, or whether λ_1 is sufficiently large so that the knowledge saturates the network and then it is only relative amounts of saturation that are being measured.[40]

Theorem **B.1.**

[39] See (2.7) in Jackson (2008b) for additional discussion and background. This is a generalization of a measure first discussed by Katz, and corresponds to Bonacich's definition if the network is unweighted and all passing probabilities are the same, and then both of Bonacich's parameters are set to q.

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

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

$$DC(\mathbf{w}, 1) = d(\mathbf{w}).$$

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

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

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

$$DC(\mathbf{w}, \infty) = KB(\mathbf{w}).$$

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 B.1. We show the second statement as the others follow directly.

First, consider any irreducible and aperiodic nonnegative (and hence primitive) $\mathbf{w}$. If the statement holds for any arbitrarily close positive and diagonalizable $\mathbf{w}'$ (which are dense in a nonnegative neighborhood of $\mathbf{w}$), then since $\frac{DC(\mathbf{w},T)}{\sum_{t=1}^{T}(\lambda_1)^t}$ is a continuous function (in a neighborhood of a primitive $\mathbf{w}$, which has a simple first eigenvalue), as is the first eigenvector, then the statement also holds at $\mathbf{w}$.[41] Thus, it is enough to prove the result for a positive and diagonalizable $\mathbf{w}$.

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

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

- If $\lambda_1 > 1$, then

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

- If $\lambda_1 = 1$, then

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

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 first show the statement for the case where $\lambda_1 = 1$. It is sufficient to show

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

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

$$DC_i\left(\mathbf{w},T\right) = \sum_j \sum_{t=1}^{T} \sum_k v_i^{(R,k)} v_j^{(L,k)} \lambda_k^t.$$

Thus,

$$\begin{aligned}
\left| \frac{DC_i\left(\mathbf{w},T\right)}{T} - v_i^{(R,1)} \right| &= \left| \frac{\sum_j \sum_{t=1}^{T} \sum_{k=1}^{n} v_i^{(R,k)} v_j^{(L,k)} \lambda_k^t}{T} - v_i^{(R,1)} \right| = \\
&= \left| \frac{1}{T} \sum_j \sum_{t=1}^{T} \sum_{k=2}^{n} v_i^{(R,k)} v_j^{(L,k)} \lambda_k^t \right| \leq \frac{1}{T} \sum_{t=1}^{T} \sum_{k=2}^{n} 1 \cdot \underbrace{\left| \sum_{j=1}^{n} v_j^{(L,k)} \right|}_{\leq 1} \cdot \left| \lambda_k^t \right| \\
&\leq \frac{n}{T} \sum_{t=1}^{T} \left| \lambda_2^t \right| = \frac{n}{T} \frac{\left|\lambda_2\right|}{1-\left|\lambda_2\right|} \left(1 - \left|\lambda_2\right|^T\right) \to 0.
\end{aligned}$$

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 $\lambda_1 > 1$. We show

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

By similar derivations as above,

$$\begin{aligned}
\left|\frac{DC_i\left(\mathbf{w},T\right)}{\sum_{t=1}^T \lambda_1^t} - v_i^{(R,1)}\right| &= \left|\frac{\sum_j \sum_{t=1}^T \sum_{k=1}^n v_i^{(R,k)} v_j^{(L,k)} \lambda_k^t}{\sum_{t=1}^T \lambda_1^t} - v_i^{(R,1)}\right| \\
&= \left|\frac{\sum_j \sum_{t=1}^T \sum_{k=2}^n v_i^{(R,k)} v_j^{(L,k)} \lambda_k^t}{\sum_{t=1}^T \lambda_1^t} + \frac{\sum_j \sum_{t=1}^T v_i^{(R,1)} v_j^{(L,1)} \lambda_1^t}{\sum_{t=1}^T \lambda_1^t} - v_i^{(R,1)}\right| \\
&= \left|\frac{\sum_j \sum_{t=1}^T \sum_{k=2}^n v_i^{(R,k)} v_j^{(L,k)} \lambda_k^t}{\sum_{t=1}^T \lambda_1^t} + \frac{\sum_{t=1}^T v_i^{(R,1)} \lambda_1^t}{\sum_{t=1}^T \lambda_1^t} - v_i^{(R,1)}\right| \\
&= \left|\frac{1}{\sum_{t=1}^T \lambda_1^t} \sum_j \sum_{t=1}^T \sum_{k=2}^n v_i^{(R,k)} v_j^{(L,k)} \lambda_k^t\right| \\
&\leq \frac{1}{\sum_{t=1}^T \lambda_1^t} \sum_{t=1}^T \sum_{k=2}^n 1 \cdot \left|\sum_{j=1}^n v_j^{(L,k)}\right| \cdot \left|\lambda_k^t\right| \\
&\leq \frac{n}{\sum_{t=1}^T \lambda_1^t} \sum_{t=1}^T \left|\lambda_2^t\right|.
\end{aligned}$$

Note that this last expression converges to 0 since $\lambda_1 > 1$, and $\lambda_1 > \lambda_2$.[42] which completes the argument. ■

B.2. **Other Proofs.**

Proof of Theorem A.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

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

[42] Note that it is important that $\lambda_1 \geq 1$ for this claim. In that case, observe that

$$\frac{\sum_{t=1}^T |\lambda_2|^t}{\sum_{t=1}^T \lambda_1^t} = \frac{\lambda_2}{\lambda_1} \cdot \frac{1-\lambda_1}{1-\lambda_2}$$

by the properties of a geometric sum, which is of constant order. Thus, higher order terms (λ_2, etc.) persistently matter and are not dominated relative to $\sum_t^T \lambda_1^t$.

If all the indexed $g_{..}$'s were distinct, the right hand side of this equation would simply be $n^{t-1}p^t$. However, in the summand sometimes terms repeat. For example, if there were exactly x repetitions, the probability of getting the walk would be p^{t-x} instead of p^t. 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 generated by setting a loose upper bound on 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 1 . Recall that $\mathbf{H} = \sum_{t=1}^{T} (\mathbf{w})^t$ and $DC = \left(\sum_{t=1}^{T} (\mathbf{w})^t\right) \cdot \mathbf{1}$ and so

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

Additionally,

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

Thus

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

implying

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

which completes the proof. ■

Proof of Corollary A.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 A.1 yields the result under (a) and (b), respectively.

To see (2), we consider the case in which $q > 1/(\mathrm{E}[\lambda_1])^{1-\varepsilon}$, which of course is equivalent to $npq > (np)^{\varepsilon}$. 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

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

So, to have

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

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 $\mathrm{E}\left[DC\left(q\mathbf{g}(n,p),T\right)\right]_i \leq kn$ for any k, and so (a) follows. ■

Proof of Theorem 2. Again, we prove the result for a positive diagonalizable $\mathbf{w}$, noting that it then holds for any (nonnegative) $\mathbf{w}$.

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

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

It then follows that we can write

$$\mathbf{H} = \sum_{t=1}^{T} (\mathbf{w})^t = \sum_{t=1}^{T} \left(\sum_{k=1}^{n} v_i^{(R,k)} v_j^{(L,k)} \lambda_k^t \right).$$

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

$$\begin{aligned} \mathbf{H}_{\cdot,j} &= \sum_{t=1}^{T} \left[v^{(R,1)} v_j^{(L,1)} \lambda_1^t + v^{(R,2)} v_j^{(L,2)} \lambda_2^t + O\left(|\lambda_2|^t \right) \right] \\ &= \sum_{t=1}^{T} \left[v^{(R,1)} v_j^{(L,1)} \lambda_1^t + O\left(|\lambda_2|^t \right) \right] \\ &= v^{(R,1)} v_j^{(L,1)} \sum_{t=1}^{T} \lambda_1^t + O\left(\sum_{t=1}^{T} |\lambda_2|^t \right). \end{aligned}$$

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

$$\frac{\mathbf{H}_{\cdot,j}}{\sum_{t=1}^{T} \lambda_1^t} = v^{(R,1)} v_j^{(L,1)} + O\left(\frac{\sum_{t=1}^{T} |\lambda_2|^t}{\sum_{t=1}^{T} \lambda_1^t} \right).$$

Note that the last expression converges to 0 since $\lambda_1 > 1$, and $\lambda_1 > \lambda_2$. Thus,

$$\frac{\mathbf{H}_{\cdot,j}}{\sum_{t=1}^{T} \lambda_1^t} \to v^{(R,1)} v_j^{(L,1)}$$

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.[43] 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 j's ranking 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. ■

[43] The discussion in Footnote 42 clarifies why $\lambda_1 > 1$ is required for the argument.

Appendix C. Extension of Microfinance Village (wave 2) Network results

This section extends the descriptive analysis from the Microfinance Village (wave 2) network data on 33 villages. We repeat all of our analyses with OLS specifications instead of Poisson specifications. Additionally, we include a Post-LASSO estimation which conducts a LASSO to select which variables best explain our outcome of interest (number of nominations) and then does a post-estimation to recover consistent parameter estimates.

Table C.1. Factors predicting nominations

	(1) Event	(2) Event	(3) Event	(4) Event	(5) Event
Diffusion Centrality	0.285				
	(0.060)				
Degree Centrality		0.250			
		(0.061)			
Eigenvector Centrality			0.283		
			(0.064)		
Leader				0.436	
				(0.168)	
Geographic Centrality					-0.025
					(0.038)
Observations	6,466	6,466	6,466	6,466	6,466

	(1) Loan	(2) Loan	(3) Loan	(4) Loan	(5) Loan
Diffusion Centrality	0.391				
	(0.071)				
Degree Centrality		0.367			
		(0.065)			
Eigenvector Centrality			0.378		
			(0.074)		
Leader				0.653	
				(0.224)	
Geographic Centrality					-0.045
					(0.029)
Observations	6,466	6,466	6,466	6,466	6,466

Notes: This table uses data from the microfinance village (wave 2) dataset. It reports estimates of OLS regressions where the outcome variable is the expected number of nominations under the event question. Panel A presents results for the event question, and Panel B presents results for the loan question. 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 by their standard deviations. Standard errors (clustered at the village level) are reported in parentheses.

Table C.2. Factors predicting nominations

	(1) Event	(2) Event	(3) Event	(4) Event	(5) Event	(6) Event
Diffusion Centrality	0.303	0.161	0.269	0.285	0.173	0.285
	(0.091)	(0.087)	(0.061)	(0.060)	(0.107)	(0.060)
Degree Centrality	-0.020				-0.013	
	(0.066)				(0.068)	
Eigenvector Centrality		0.138			0.137	
		(0.095)			(0.095)	
Leader			0.294			
			(0.174)			
Geographic Centrality				-0.026		
				(0.039)		
Observations	6,466	6,466	6,466	6,466	6,466	6,466
Post-LASSO						✓

	(1) Loan	(2) Loan	(3) Loan	(4) Loan	(5) Loan	(6) Loan
Diffusion Centrality	0.310	0.266	0.366	0.391	0.175	0.310
	(0.112)	(0.089)	(0.071)	(0.071)	(0.124)	(0.112)
Degree Centrality	0.091				0.098	0.091
	(0.079)				(0.079)	(0.079)
Eigenvector Centrality		0.138			0.144	
		(0.089)			(0.087)	
Leader			0.461			
			(0.229)			
Geographic Centrality				-0.045		
				(0.030)		
Observations	6,466	6,466	6,466	6,466	6,466	6,466
Post-LASSO						✓

Notes: This table uses data from the microfinance village (wave 2) dataset. It reports estimates of OLS regressions where the outcome variable is the expected number of nominations. Panel A presents results for the event question, and Panel B presents results for the loan question. 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 by their standard deviations. Column (6) 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.

Appendix D. Extension of experiment analysis

This section extends the analysis of the experiment results to using four instruments.

Table D.1. Calls received by treatment

	(1)	(2)	(3)	(4)	(5)
	RF	OLS	IV 1: First Stage	IV 2: First Stage	IV: Second Stage
	Calls Received	Calls Received	At least 1 Gossip	At least 1 Elder	Calls Received
Gossip Treatment	4.559		0.795	0.430	
	(3.121)		(0.0753)	(0.108)	
5 Gossip Seeds	-1.785		-0.303	-0.206	
	(5.290)		(0.110)	(0.153)	
Elder Treatment	2.279		0.370	0.872	
	(2.424)		(0.106)	(0.0685)	
5 Elder Seeds	-6.798		-0.272	-0.0578	
	(3.487)		(0.149)	(0.100)	
At least 1 Gossip		3.786			8.063
		(1.858)			(3.845)
At least 1 Elder		0.792			-3.684
		(2.056)			(2.266)
Observations	212	212	212	212	212
Control Group Mean	8.019	5.846	0.389	0.183	5.496

	(1)	(2)	(3)	(4)	(5)
	RF	OLS	IV 1: First Stage	IV 2: First Stage	IV: Second Stage
	$\frac{\text{Calls Received}}{\text{Seeds}}$	$\frac{\text{Calls Received}}{\text{Seeds}}$	At least 1 Gossip	At least 1 Elder	$\frac{\text{Calls Received}}{\text{Seeds}}$
Gossip Treatment	1.593		0.795	0.430	
	(1.030)		(0.0753)	(0.108)	
5 Gossip Seeds	-1.083		-0.303	-0.206	
	(1.348)		(0.110)	(0.153)	
Elder Treatment	0.622		0.370	0.872	
	(0.770)		(0.106)	(0.0685)	
5 Elder Seeds	-1.430		-0.272	-0.0578	
	(0.912)		(0.149)	(0.100)	
At least 1 Gossip		0.952			2.169
		(0.501)			(1.043)
At least 1 Elder		0.309			-0.676
		(0.511)			(0.578)
Observations	212	212	212	212	212
Control Group Mean	1.953	1.451	0.389	0.183	1.186

Notes: This table uses data from the cell phone RCT dataset. Panel A uses the number of calls received as the outcome variable. Panel B normalizes the number of calls received by the number of seeds, 3 or 5, which is randomly assigned. For both panels, Column (1) shows the reduced form results of regressing number of calls received on dummies for gossip treatment and elder treatment. Column (2) regresses number of calls received on the dummies for if at least 1 gossip was hit and for if at least 1 elder was hit in the village. Columns (3) and (4) show the first stages of the instrumental variable regressions, where the dummies for "at least 1 gossip" and "at least 1 elder" are regressed on the exogenous variables: gossip treatment dummy, 5 gossip seeds dummy, elder treatment dummy, 5 elder seeds dummy. Column (5) shows the second stage of the IV; it regresses the number of calls received on the dummies for if at least 1 gossip was hit and if at least 1 elder was hit, both instrumented by treatment status of the village (gossip treatment or not, elder treatment or not) and seed number dummies for the village (5 gossip seeds or not, 5 elder seeds or not). All columns control for number of gossips, number of elders and number of seeds. For columns (1), (3), and (4) the control group mean is calculated as the mean expectation of the outcome variable when the treatment is "random". For columns (2) and (5), the control group mean is calculated as the mean expectation of the outcome variable when no gossips or elders are reached. The control group mean for the second stage IV is calculated using IV estimates. Robust standard errors are reported in parentheses.

Appendix E. Experiment Analysis with Broadcast Village

This section repeats our main experimental analyses but includes the broadcast village where the poster was made by one of the seeds.

Table E.1. Calls received by treatment

	(1)	(2)	(3)	(4)	(5)
	RF	OLS	IV 1: First Stage	IV 2: First Stage	IV: Second Stage
VARIABLES	Calls Received	Calls Received	At least 1 Gossip	At least 1 Elder	Calls Received
Gossip Treatment	2.266		0.636	0.331	
	(3.116)		(0.0660)	(0.0821)	
Elder Treatment	-2.809		0.220	0.846	
	(2.577)		(0.0807)	(0.0502)	
At least 1 Gossip		5.005			6.122
		(2.210)			(4.532)
At least 1 Elder		-0.619			-4.914
		(2.472)			(2.628)
Observations	213	213	213	213	213
Control Group Mean	9.534	6.277	0.400	0.180	7.971
Gossip Treatment=Elder Treatment (pval.)	0.0300		0	0	
At least 1 Gossip=At least 1 Elder (pval.)		0.160			0.0300

	(1)	(2)	(3)	(4)	(5)
	RF	OLS	IV 1: First Stage	IV 2: First Stage	IV: Second Stage
VARIABLES	$\frac{\text{Calls Received}}{\text{Seeds}}$	$\frac{\text{Calls Received}}{\text{Seeds}}$	At least 1 Gossip	At least 1 Elder	$\frac{\text{Calls Received}}{\text{Seeds}}$
Gossip Treatment	0.591		0.636	0.331	
	(0.841)		(0.0660)	(0.0821)	
Elder Treatment	-0.646		0.220	0.846	
	(0.738)		(0.0807)	(0.0502)	
At least 1 Gossip		1.359			1.535
		(0.644)			(1.179)
At least 1 Elder		-0.162			-1.164
		(0.691)			(0.748)
Constant				0.109	
				(0.160)	
Observations	213	213	213	213	213
Control Group Mean	2.452	1.595	0.400	0.180	2.048
Gossip Treatment=Elder Treatment (pval.)	0.0400		0	0	
At least 1 Gossip=At least 1 Elder (pval.)		0.190			0.0400

Notes: This table uses data from the cell phone RCT dataset. Panel A uses the number of calls received as the outcome variable. Panel B normalizes the number of calls received by the number of seeds, 3 or 5, which is randomly assigned. For both panels, Column (1) shows the reduced form results of regressing number of calls received on dummies for gossip treatment and elder treatment. Column (2) regresses number of calls received on the dummies for if at least 1 gossip was hit and for if at least 1 elder was hit in the village. Columns (3) and (4) show the first stages of the instrumental variable regressions, where the dummies for “at least 1 gossip” and “at least 1 elder” are regressed on the exogenous variables: gossip treatment dummy and elder treatment dummy. Column (5) shows the second stage of the IV; it regresses the number of calls received on the dummies for if at least 1 gossip was hit and if at least 1 elder was hit, both instrumented by treatment status of the village (gossip treatment or not, elder treatment or not). All columns control for number of gossips, number of elders, and number of seeds. For columns (1), (3), and (4) the control group mean is calculated as the mean expectation of the outcome variable when the treatment is “random”. For columns (2) and (5), the control group mean is calculated as the mean expectation of the outcome variable when no gossips or elders are reached. The control group mean for the second stage IV is calculated using IV estimates. Robust standard errors are reported in parentheses.

Table E.2. Calls received by seed type

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Calls Received	Calls Received	Calls Received	$\frac{\text{Calls Received}}{\text{Seeds}}$	$\frac{\text{Calls Received}}{\text{Seeds}}$	$\frac{\text{Calls Received}}{\text{Seeds}}$
At least 1 Gossip	12.89	13.02		3.751	3.871	
	(7.225)	(8.157)		(2.282)	(2.584)	
At least 1 Elder	-3.371	-3.321		-1.012	-0.962	
	(5.155)	(4.946)		(1.547)	(1.456)	
At least 1 High *DC* Seed		-0.485	2.262		-0.478	0.342
		(4.803)	(3.834)		(1.515)	(1.189)
Observations	69	69	69	69	69	69
Control Group Mean	4.840	4.840	8.828	1.101	1.101	2.433
At least 1 Gossip=At least 1 Elder (pval.)	0.150	0.170		0.190	0.200	
At least 1 Gossip=At least 1 High *DC* Seed (pval.)		0.270			0.270	
At least 1 Elder=At least 1 High *DC* Seed (pval.)		0.580			0.710	

Notes: This table uses data from the cell phone RCT and follow-up network dataset. The table presents OLS regressions of number of calls received (and number of calls received normalized by the number of seeds, 3 or 5, which is randomly assigned) on characteristics of the set of seeds. High *DC* refers to a seed being above the mean by one standard deviation of the centrality distribution. All columns control for total number of gossips, number of elders, and number of seeds. For columns (1), (2), (4), and (5), the control group mean is calculated as the mean expectation of the outcome variable when no gossips or elders are reached. For columns (3) and (6), the control group mean is calculated as the mean expectation of the outcome variable when no high *DC* seeds are reached. Robust standard errors are reported in parentheses.

Appendix F. When People Don't Nominate Anyone

Here we look at whether the odds that someone refuses to nominate anyone can be explained by our model.

Suppose that people get disutility from reporting incorrect guesses to a surveyor.[44]

Our theoretical results on network gossip converging to diffusion centrality are asymptotic results - with any finite number of iterations of gossip passing, an individual's ranking of others' centralities can be noisy.

In particular, after T periods, i's perceptions are not quite at their expectations:

$$\widehat{NG}_{ji} = NG_{ji} + \varepsilon_{ji}.$$

where ε_{ji} will depend on T, $\mathbf{w}$, and i and j's positions in the network - eventually vanishing (in proportion to NG_{ji}) when T becomes large.

One way that i can assess how accurate their ranking is, would be looking at how extreme $\widehat{NG}_{ji}$ is compared to the average number of times that i has heard about other people. If this is more extreme, then it is less likely to reverse. Then if i' 99th percentile $\widehat{NG}_{ji}$ is more extreme in terms of numbers of times heard compared to the average, then i can distinguish this tail better from the rest of the distribution, even if there is noise. That is, the ability to distinguish $NG_{ji}^{99\%}$ from $\bar{NG}_i$ should increase in $NG_{ji}^{99\%}$ holding ε fixed.

From this perspective, our model predicts that the larger the extreme quantiles of i's network gossip distribution are, controlling for the average (note that is just proportional to DC_i), then i should be more likely to nominate and less likely to answer that he does not know. This comes from the fact that he should be better able to distinguish between alternatives.

For this estimation, as in the rest of the paper we work with T equal to the diameter of the network and q equal to the inverse of the first eigenvalue.

The results in Table F.1 are consistent with this story. A one standard deviation increase in the 99th percentile of network gossip of j as perceived by i corresponds to, holding fixed the average network gossip of others as perceived by i, a 1.3 percent increase in the probability of i nominating anyone. Across specifications we have p values of $0.11, 0.08, 0.18, 0.6$, respectively, across columns.

[44] See Alatas et al. (2014) for such an example of where this happened in practice when individuals were more likely to report that they don't know rather than offer a guess when trying to rank other villagers in terms of wealth.

Table F.1. Does the tail of network gossip drive nominations?

	(1)	(2)	(3)	(4)
VARIABLES	Nominated Anyone	Nominated Anyone	Nominated Anyone	Nominated Anyone
99th percentile of NG_{ji}	3.739	8.134		
	(2.348)	(4.651)		
98th percentile of NG_{ji}			3.499	6.174
			(2.620)	(4.409)
99th Percentile	✓	✓		
Village FE		✓		✓
98th Percentile			✓	✓

Notes: This table uses data from the microfinance village (wave 2) dataset. The data consists of a individual level observations and the outcome is whether the individual nominated anyone in response to the lottery gossip question.The key regressor is the value of the person who is at the 99th (or 9th) percentile from the distribution network gossip for i. Columns 2 and 4 include village fixed effects, estimated by a conditional logit. Standard errors (clustered at the village level) are reported in parentheses.

Next we look at the demographics of those who choose to nominate versus those who choose not to nominate in Table F.2. We look at a number of demographics: caste, occupation, household amenities, respondent gender, and geography. We also include social status variables such as leadership (as defined from our previous work based on the lender's designations) and the number of nominations the individual's household received under loan and event questions. Finally we include the diffusion centrality of the respondent's household.

Table F.2 presents the result. Being more diffusion central is positively and significantly associated with the respondent being willing to nominate someone. This is consistent with our model. Meanwhile, almost none of the other demographics or social status variables matter. In fact, there is no other statistically significant variable for loan nominations. For event nominations, the number of times the respondent's household was nominated under the event question matters as does whether the respondent owns their own house. Beyond that, no other variable matters.

Table F.2. Demographics of those who choose to nominate

	(1)	(2)
VARIABLES	Nominates Someone (Loan)	Nominates Someone (Event)
Diffusion Centrality (Standardized)	0.024	0.015
	(0.008)	(0.008)
No. of Nominations (Loans)	-0.000	-0.001
	(0.003)	(0.003)
No. of Nominations (Events)	0.004	0.008
	(0.003)	(0.003)
Leader	0.004	-0.007
	(0.020)	(0.018)
SCST	-0.010	0.006
	(0.026)	(0.022)
Electrified	-0.031	-0.002
	(0.031)	(0.028)
Private Electrification	0.013	-0.003
	(0.017)	(0.017)
Own House	-0.036	-0.049
	(0.026)	(0.028)
No. of Rooms	-0.002	0.002
	(0.005)	(0.005)
Land Owner	-0.020	-0.013
	(0.028)	(0.026)
Farm Laborer	-0.032	-0.014
	(0.022)	(0.021)
Business Owner	-0.020	-0.014
	(0.027)	(0.025)
GPS Centrality	-0.007	-0.008
	(0.008)	(0.006)
Female Respondent	0.009	0.014
	(0.015)	(0.013)
Observations	5,707	5,707

Notes: This table uses data from the microfinance village (wave 2) dataset. The data consists of a individual level observations and the outcome is whether the individual nominated anyone in response to the lottery gossip question. Standard errors (clustered at the village level) are reported in parentheses and all specifications include village fixed effects.

Appendix G. Characteristics of Gossips, Elders, and Random

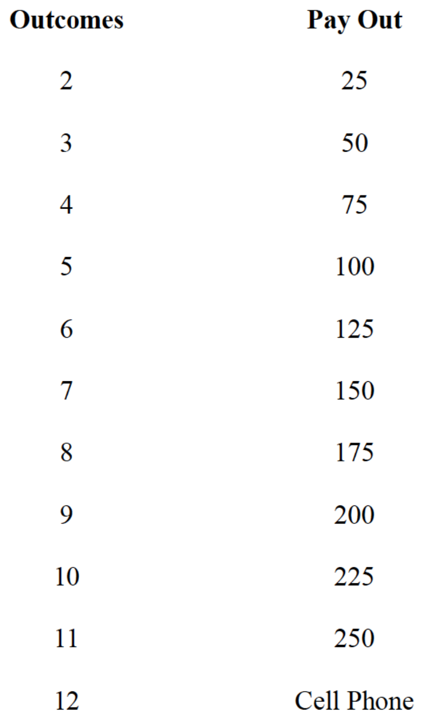
Table G.1. Characteristics of gossip, elder, and random households

VARIABLES	(1) SCST	(2) Laborer	(3) Land Owner	(4) Electrified	(5) Private Electricity	(6) Own House	(7) No. of Rooms
Gossip Nominee	-0.0278	-0.0729	0.0793	0.0173	0.0455	0.0197	0.229
	(0.0258)	(0.0189)	(0.0241)	(0.00637)	(0.0173)	(0.00810)	(0.0492)
Elder Nominee	-0.107	-0.217	0.291	0.0196	0.0903	0.0262	0.687
	(0.0250)	(0.0215)	(0.0279)	(0.00636)	(0.0227)	(0.00744)	(0.0849)
Observations	13,660	13,660	13,660	13,660	13,660	13,590	13,590
Random Household Mean	0.377	0.406	0.275	0.962	0.727	0.948	2.846
Gossip = Elder p-val	0.0382	3.32e-05	2.03e-06	0.820	0.174	0.577	6.02e-05

Notes: This table uses data from the Karnataka cell phone RCT dataset. The data consists of a individual level observations and the outcome is whether the individual nominated (as a gossip or elder, omitted is random) has the characteristic noted. Column 1 is whether the individual is SCST, column 2 is whether the primary occupation of the household is farm labor, column 3 is whether the primary income comes from land ownership, column 4 is whether the household is electrified, column 5 is whether electrification is from private purchase, column 6 is whether they own their house, column 7 is the number of rooms in the house. Standard errors (clustered at the village level) are reported in parentheses.

Appendix H. Karnataka Cellphone Experiment Payment Schedule

In the Karnataka cell phone RCT, every individual who called in and therefore was eligible for a prize was able to do the following. They simply rolled two dice. The outcome of the roll – some number between 2 and 12 – corresponded to a prize. This was independent across all participants. Every roll had some cash prize and the cell phone was worth approximately Rs. 3000. Note that if every participant rolled a 12, then every participant would win a cell phone.



Outcomes	Pay Out
2	25
3	50
4	75
5	100
6	125
7	150
8	175
9	200
10	225
11	250
12	Cell Phone

Figure 6. Prizes as a function of the roll of two dice in the Karnataka cell phone RCT.