Review of Economic Studies (2019) **86**, 2453–2490 doi:10.1093/restud/rdz008

Advance access publication 20 February 2019

Using Gossips to Spread Information: Theory and Evidence from Two Randomized Controlled Trials

ABHIJIT BANERJEE
MIT; NBER; J-PAL
ARUN G. CHANDRASEKHAR
Stanford University; NBER; J-PAL
ESTHER DUFLO
MIT; NBER; J-PAL
and
MATTHEW O. JACKSON
Stanford University; Santa Fe Institute

First version received June 2017*; Editorial decision October* 2018*; Accepted February* 2019(*Eds.*)

Can we identify highly central individuals in a network without collecting network data, simply by asking community members? Can seeding information via such nominated individuals lead to significantly wider diffusion than via randomly chosen people, or even respected ones? In two separate large field experiments in India, we answer both questions in the affirmative. In particular, in 521 villages in Haryana, we provided information on monthly immunization camps to either randomly selected individuals (in some villages) or to individuals nominated by villagers as people who would be good at transmitting information (in other villages). We find that the number of children vaccinated every month is 22% higher in villages in which nominees received the information. We show that people's knowledge of who are highly central individuals and good seeds can be explained by a model in which community members simply track how often they hear gossip about others. Indeed, we find in a third data set that nominated seeds are central in a network sense, and are not just those with many friends or in powerful positions.

Key words: Centrality, Gossip, Networks, Diffusion, Influence, Social learning.

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

1. INTRODUCTION

"The secret of my influence has always been that it remained secret."
— Salvador Dalí

The editor in charge of this paper was Adam Szeidl.

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, both in developing and developed economies. For example, Gmail was first diffused by invitations to leading bloggers and then via sequences of invitations that people could pass to their friends. The seeding of apps and other goods to central individuals is a common strategy of viral marketing campaigns (Hinz *et al.*, 2011; Aral *et al.*, 2013; Leduc *et al.*, 2017). 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 *et al.*, 2013; Beaman *et al.*, 2018).[1]

The central question of this article is how to find the most effective people to seed for a diffusion process. Starting with the classical studies of Simmel (1908), Katz and Lazarsfeld (1955), Coleman *et al.* (1966), a large body of work in sociology, marketing, economic theory and computer science studies the problem of identifying opinion leaders and key individuals in diffusing products and information. These studies include both theoretical and empirical research.[2] A key take away is that if the goal is to diffuse information by word of mouth, then the optimal seeds are those who are central, according to appropriate definitions.[3]

Moreover, as shown in Banerjee *et al.* (2013) and Beaman *et al.* (2018), even though many measures of centrality are correlated, successful diffusion requires seeding information via people who are central according to specific measures.[4] 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 article, 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. We explore a direct technique that turns out to be remarkably effective: simply asking a few individuals in the community which other members in the community would be best for spreading information.

Surprisingly, this is not a technique that had been examined in theory or tried in practice by many organizations in the field. In a review of 190 empirical studies of different seeding strategies, Valente and Pumpuang (2007) find that only four even considered asking network members. This is perhaps because there is reason to doubt if such a technique would work. There is a literature in sociology (Friedkin, 1983; Krackhardt, 1987; Casciaro, 1998; Krackhardt, 2014, among others)

1. 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 behaviours with peer effects (Ballester *et al.*, 2006). These examples include peer effects in schooling, networks of crime and delinquent behaviour, among other things. Similarly, the work of Paluck *et al.* (2016) shows that exposing "social referents" to an intervention that encouraged taking public stances against conflict at school was particularly effective.

2. See Rogers (1995), Krackhardt (1996), Kempe *et al.* (2003), Jackson (2008), Iyengar *et al.* (2010), Hinz *et al.* (2011), Katona *et al.* (2011), Jackson and Yariv (2011), Banerjee *et al.* (2013), Bloch *et al.* (2016), Jackson (2017) for some background.

3. An exception noted by Akbarpour *et al.* (2017) is that if a diffusion process is long-lasting (without a time limit) and individuals only need to be informed or reached by one contact in order to adopt or become infected (*i.e.* what is known as "simple" contagion), and the number of seeds tends to infinity, then random seeding can work as well as choosing seeds carefully. We will argue later in the article that none of these conditions are plausible in our context.

4. Bonacich (1987), Borgatti (2005, 2006), Ballester *et al.* (2006), Valente *et al.* (2008); Valente (2010), Valente (2012), Lim *et al.* (2015), Bloch *et al.* (2016), Jackson (2017) all explore what are appropriate measures of network centrality for various applications. Seeding several seeds optimally is a more complicated problem studied by Kempe *et al.* (2003, 2005).

that shows that people tend to have poor knowledge of their networks, and specific biases in their perceptions of the network. For example, people in a research centre or department in a university do not know what many of their colleagues work on, and so they tend to overestimate how many friends they have and underestimate the number of collaborators they have (Casciaro, 1998). In the rural Karnataka villages where part of this study was conducted, Breza *et al*. (2017) show that knowledge of the existence of a link between two other villagers is limited, that people think the network is more dense than it actually is, that knowledge of links is better for pairs of individuals that are "closer" to the respondent, that knowledge of links is better for central respondents, and finally that there are high error rates across the board. 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.

Our first contribution is empirical. In two different randomized controlled trials (RCTs) we show that, in practice, it is possible to cheaply identify influential seeds by asking community members.

The first RCT was run in 213 villages in Karnataka. We first asked a few villagers who would be a good diffuser of information (we refer to their nominees below as the "gossip nominees" or "gossips"). Then, in each of these villages, we seeded a simple piece of information to a few households: anyone who gave a free call to a particular phone number would be entered in a non-rival lottery in which they were guaranteed to win either a cash prize or a cell phone. In seventy-one "random seeding" villages, the seed households were randomly selected. In seventy-one "social status seeding" villages, they had status as "elders" in the village—leaders with a degree of authority in the community, who command respect. In the remaining seventy-one "gossip seeding" villages, the seeds were those nominated by others as being well-suited to spread information ("gossip nominees").

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. Using the initial random assignment as an instrument for actually hitting at least one gossip, we find that seeding at least one gossip yields an extra 7.4 calls, or nearly doubles 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. Secondly, targeting gossips might have been successful because the information was anodyne. Perhaps the elders or the randomly selected seeds would have more aggressively circulated a more relevant piece of information.

To investigate whether the success of the technique carries over to a policy-relevant setting, we conducted a second large-scale, policy-relevant RCT in the context of a collaboration with the Government of Haryana (India) on their immunization programme. Immunization is an important policy priority in Haryana, because it is remarkably low. This projects takes place in seven low-performing districts where full immunization rates were around 40% or less at baseline. We identified 521 villages for a "seed" intervention. Each of those villages were randomly assigned to one of four groups. In the first group ("gossip"), seventeen randomly selected households who is were surveyed and asked to identify who would be good diffusers of information; in the second group ("trust"), we asked seventeen randomly selected households who people in the village tend to trust; in the third one ("trusted gossip"), we asked the seventeen randomly selected households who is both good at diffusing information *and* trusted. In the fourth group, no nominations were elicited. We then visited the six individuals in each village with the most nominations in the first three groups, and the head of six randomly selected households in the fourth group, and asked them to become the programme's ambassadors. Throughout the year, they received regular SMSs and calls reminding them to spread information about immunization, and we tracked immunization

with administrative data over one year. The results of this RCT are consistent with those of the first. In the average monthly camp with random seeds, 18.11 children attended and received at least one shot. In villages with gossip seeds, the number was 23, or 27% higher. The other seeding strategies are in between: neither statistically different from random seeding (for most vaccines), nor statistically different from gossip seeding.

The two RCTs carried out in very different contexts illustrate that villagers can identify people who effectively spread information. This raises the question: How can individuals identify these effective spreaders, given their lack of knowledge of the network?

In our second contribution, we therefore go on to provide a theoretical 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. In particular, we model a process that we call "gossip" in which nodes generate pieces of information that are stochastically passed from neighbour to neighbour, 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.[5] We prove 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.[6] Thus, just by a simple process of counting, network members can have a good sense of others' diffusion centrality (DC), which may explain why they are able to identify those who would be good at transmitting information even with very little knowledge of the network.

This is of course just a possibility result. There are other ways in which people can learn who is effective at diffusing information. The theory suggests one possible explanation for our empirical results but the empirical results are not a test of the theory to the exclusion of other explanations.

To provide more support for our interpretation, we test some more specific predictions of the theory. First, in thirty-three villages in which we had previously collected detailed network data, we collect gossip nomination data and find that individuals nominate highly diffusion central people. We also show that the nominations are not simply based on the nominee's leadership status, degree, or geographic position in the village, but are significantly correlated with diffusion centrality even after controlling for these characteristics.

Even this does not prove that the specific mechanism we highlight is at play. For instance, individuals may believe that more gregarious individuals are better disseminators of information. By their very nature, gregarious individuals may be more central, but the mechanism would then not be about the network structure per se, but about gregariousness (which is unobserved by us). In the next step, we use these data to examine a much more specific implication of our model. If the listening and learning about others' centralities process only runs for a short time, agents can have different rankings of others' centralities. We show that, conditional on individual fixed effects and even controlling for the diffusion centrality of the agent being assessed, a respondent is more likely to nominate an agent who is of higher rank in the respondent's *personal* finite-time ranking calculated using our model. This result is much more difficult to explain with alternative theories than the basic reduced form results, and therefore provides a somewhat stronger test of the theory.

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

5. 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" (Bloch *et al.*, 2014).

6. 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 Online Appendix B.

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. Additionally, information diffuses more extensively when we hit at least one gossip. However, when we include both an indicator for hitting at least one high diffusion centrality seed *and* an indicator for hitting at least one gossip in the regression, the coefficient of hitting at least one gossip 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. Of course it remains possible 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 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. The previous literature had highlighted the ignorance and biases that people have about their own network. We show that people are able to name central people in their network even if they do not know the specific links between people. This theory provides some guidance as to when asking people to name seeds may or may not work. For instance, in settings where topics tend to be discussed over longer periods of time or where there is very little social fractionalization and short average distances, gossip nominations will be more correlated with overall network centrality than in contexts where topics tend to be short-lived or there are large social cleavages.[7]

The remainder of the article 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 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 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 programme.

2.1. *Study 1: The cell phone and cash raffle RCT*

We conducted an RCT in 213 villages in Karnataka (India) with 196.5 households on average 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,

7. In this second setting, people can still learn about who is central in their own community within the network.

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 roughly its cash value (Rs. 3,000) to villagers. 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). 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 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 three and five seeds on our outcome variable (the number of calls received).

We randomly divided the sample of 213 villages into three sets of seventy-one, where the k seeds were selected as follows. A few days before the experiment, we interviewed up to fifteen households in every village (selected uniformly at random via circular random sampling using the right-hand rule method) to identify "elders" and "gossips".[8] 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 to the fifteen households to identify the gossip nominees was:

> 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?

We randomly, and evenly, divided the sample of 213 villages into three treatment arms:

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

8. 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.

T2. Gossip: k households were chosen from the list of gossip nominees obtained one week prior.
T3. Elder: k households were chosen from the list of village elders obtained one week prior.

In each T2 and T3 village, we randomly selected gossip and elder seeds from the set of nominees in that village.

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 non-deterministic polynomial-time (NP)-complete one, see Kempe *et al.* (2003, 2005)). Here, we randomly selected seeds from the set of nominees, which if anything biases the test against the gossip treatment. Instead, 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 were well-spaced in the network.

The main outcome variable that we are interested in is the number of unique households that called. This includes any call from the seed themselves, and so in the main specification we control for fixed effects for the number of seeds. This represents the number of people who heard about the promotion and wanted to participate. The mean number of calls per village is 9.35 (with standard deviation 15.64). The median number of villagers who participated is three across all villages. In 80.28% of villages, we received at least one call, and the 95th percentile is thirty-nine. It is debatable whether these are large or small numbers for a marketing campaign, but for our purpose what is important is that there is enough 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 Online Appendix E. They are qualitatively similar: the Ordinary Least Squares (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 those in the elder and random villages. 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.

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? We estimate

$$y_j = \theta_0 + \theta_1 GossipTreatment_j + \theta_2 ElderTreatment_j + \theta_3 NumberSeeds_j + \theta_4 NumberGossip_j + \theta_5 NumberElder_j + u_j, \tag{2.1}$$

where y_j is the number of calls received in village j or the number of calls per seed in village j, $GossipTreatment_j$ 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

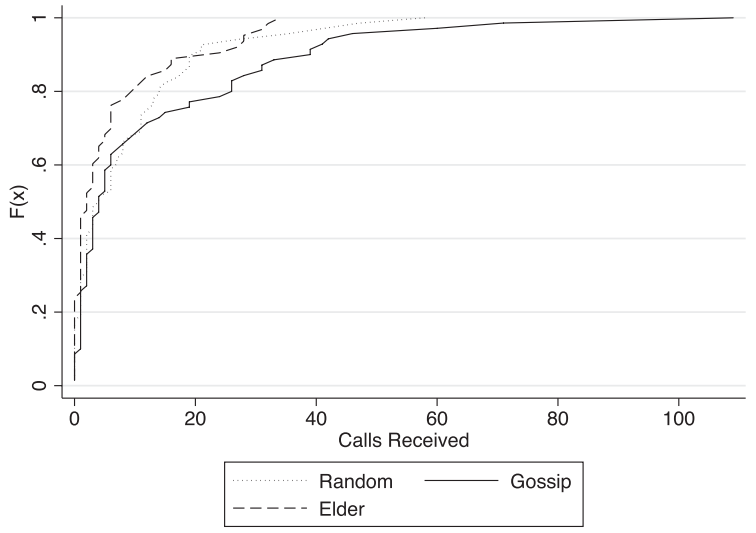


FIGURE 1
Distribution of calls received by treatment in the Karnataka cell phone RCT.

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 Online Appendix E. Column 1 shows the reduced form of equation (2.1). In control (random) villages, we received 8.08 calls, or an average of 1.97 per seed. In gossip treatment villages, we received 3.65 more calls ($p=0.19$) in total or 1.05 additional calls per seed ($p=0.13$).

This exercise is of independent interest since it is the answer to the policy question of how much a policymaker 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 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 thus compares 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 use of the variation induced by the village level experiment directly, and instrument "at least 1 gossip (elder) seed hit" by the gossip (elder) treatment status of the village.

Therefore, we are interested in

$$y_j = \beta_0 + \beta_1 GossipReached_j + \beta_2 ElderReached_j + \beta_3 NumberSeeds_j + \beta_4 NumberGossip_j + \beta_5 NumberElder_j + \epsilon_j. \quad (2.2)$$

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

TABLE 1
Calls received by treatment

	(1) RF Calls received	(2) OLS Calls received	(3) IV 1: First Stage At least 1 gossip	(4) IV 2: First stage At least 1 elder	(5) IV: Second stage 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) RF Calls received / Seeds	(2) OLS Calls received / Seeds	(3) IV 1: First stage At least 1 gossip	(4) IV 2: First stage At least 1 elder	(5) IV: Second stage Calls received
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 data set. 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.

equations are

$$GossipReached_j = \pi_0 + \pi_1 GossipTreatment_j + \pi_2 ElderTreatment_j + \pi_3 NumberSeeds_j + \pi_4 NumberGossip_j + \pi_5 NumberElder_j + v_j, \quad (2.3)$$

and

$$ElderReached_j = \rho_0 + \rho_1 GossipTreatment_j + \rho_2 ElderTreatment_j + \rho_3 NumberSeeds_j + \rho_4 NumberGossip_j + \rho_5 NumberElder_j + \nu_j. \quad (2.4)$$

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 from them, 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 at the 35th percentile by 122% and calls at the 80th percentile by 71.27%.

This is our key experimental result: gossip nominees are significantly 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 spreading this type of information. 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.

This RCT took place in Haryana, a state bordering New Delhi, in Northern India. J-PAL was collaborating with the government of Haryana on a series of initiatives designed to improve immunization rates in seven low-immunization districts. Overall, 3,116 villages were involved in the project. The project included several components. In all villages, monthly immunization camps were held, and the government gave nurses tablets with a simple e-health application that the project team developed to track immunizations. The data thus generated is our main outcome.[9]

9. We have completed over 5,000 cross-validation surveys, visiting children at random and collecting information on their immunization status, to cross-check this administrative data. The administrative data is of excellent quality.

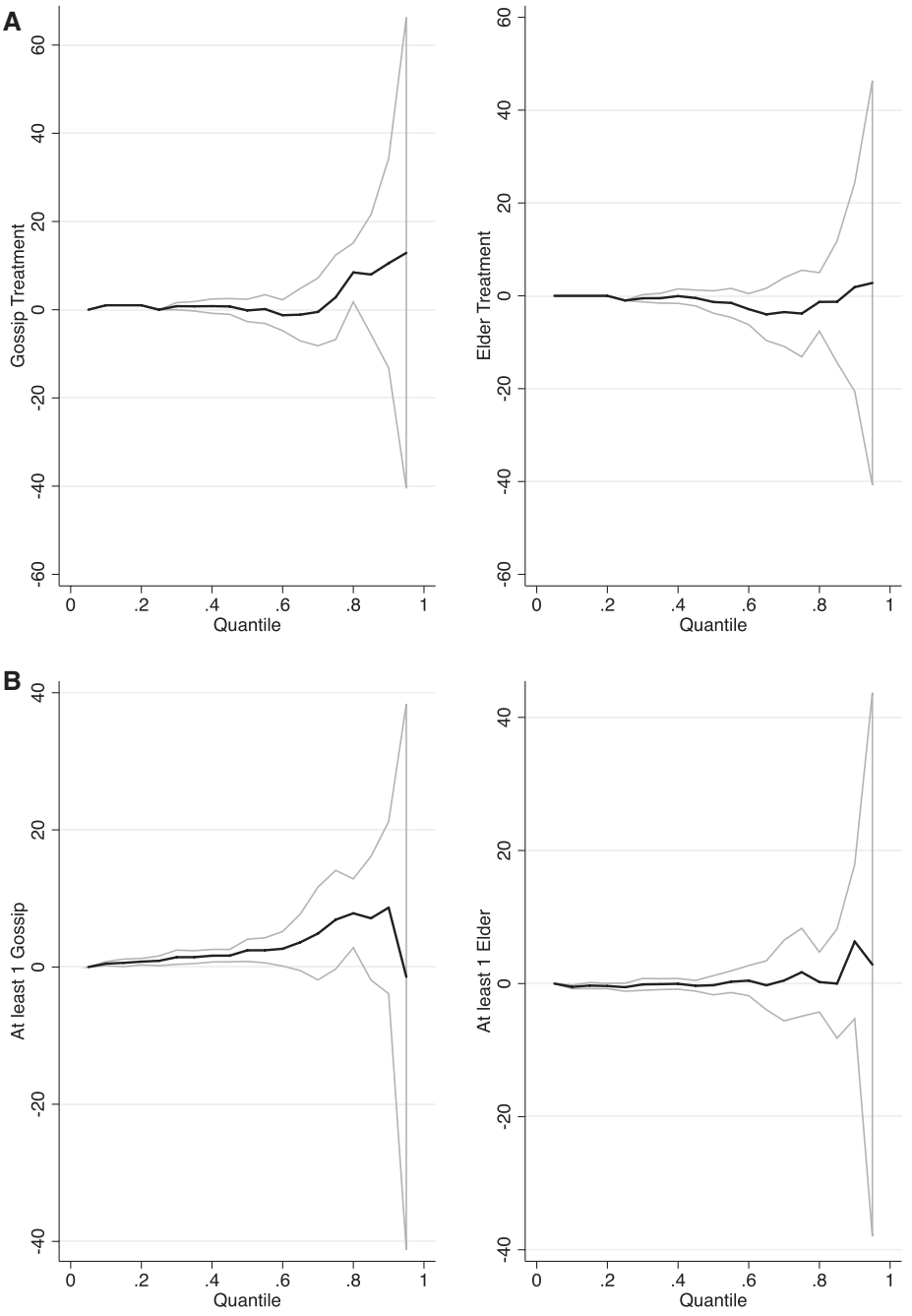


FIGURE 2

Quantile treatment effects where for $j \in \{Gossip, Elder\}$, $\widehat{\beta}_j(u)$ is computed for $u = \{0.05, \ldots, 0.95\}$. (A) Quantile treatment effect by treatment—Reduced Form. (B) Quantile treatment effect by hitting at least one gossip or elder.

Note: The intercept $\alpha(u)$ (not pictured) in each case is the omitted category corresponding to the random treatment.

In addition, J-PAL carried out several cross-randomized interventions designed to increase the demand for immunization. In some randomly selected villages, small incentives for immunization were offered. We cross-randomized whether households received personalized SMS reminding them of the next vaccine that their child was due for. Depending on the village, the reminders were sent to either no households, or to 33% or 66% of all households that had previously attended an immunization camp. In the SMS "blast" villages that received reminders, recipients were randomly selected from the database of attendees.

These experiments are analysed in Banerjee *et al.* (2018a). Finally, in 521 of these villages (drawn equally from all of the other treatments strata), we conducted the seeding experiment that is the focus of this article.

2.2.1. Experimental Design. In all of the villages that were part of the seeding experiment, six individuals (selected according to the protocol described below) were contacted in person a few weeks prior to the launch of the tablet application and the incentives intervention (the seeds were initially approached and asked if they wished to participate between June and August 2016, the tablet application was launched in December 2016, and the seeds were re-contacted and reminded to spread information starting in February 2017). When first contacted, potential seeds answered a short demographic survey and were asked to become ambassadors for the programme. If they agreed,[10] they gave us their phone number and they received two monthly reminders (one text and one phone call). These notifications reminded them about upcoming immunization camps, what shots their child was due for, and any incentive if one existed, and asked them to spread that information.

Specifically, the script used to recruit them was the Hindi translation of what 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 in order to help us 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 of recruitment:

> 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 of immunization camps in their village and will push them to get their children immunized. It is your choice to spread the information with whomsoever you want.

10. The refusal rate was around 18%. If a potential seed refused to participate they were not replaced, so there is some variation in the number of actual seeds in each village, but all villages have at least some seeds.

A village-level treatment assignment determined the type of seeds chosen in each village. In particular, the seed villages were randomly assigned to one of four groups:

T1. Random seeds. In the random seeding group, we randomly selected six households from our census, and the seed was the head of the selected household.
In the three remaining groups, we first visited the village, and visited seventeen randomly selected households. This was done in January and February 2016. We interviewed a respondent in the household, asking them one of the following three questions to elicit nominations for seeds. The question asked, and therefore the type of people nominated, was determined by the village-level treatment assignment. Note that in each village, we only asked one of the questions, 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 such that 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 these individuals 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 of 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, both within and outside of this neighborhood, 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, who live in the village (within or outside of your neighborhood in the village) and are trusted?

T4. Trusted gossip seeds.

> Who are the people in this village, both within and outside of this neighborhood, such that 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 these individuals 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 (either within or outside of your neighborhood) such that when they say something, many people get hear about it, and are trusted by you and other villagers?

TABLE 2
Summary Statistics of Haryana immunization RCT

	(1) Random seed	(2) Gossip seed	(3) Trusted seed	(4) 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 six 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 Auxiliary Nurse Midwife (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

Note that we specifically asked about two things when eliciting nominations for trusted seeds: fertilizers for crops and children's health. That was done to ensure that the trust question did not emphasize immunization. As in our previous experiment, the gossip question is centred purely on transmission of relatively unimportant information, and, in contrast to the trust questions, does not flag any concerns about trust.

In the T2, T3, and T4 villages, we selected the six individuals with the highest number of nominations as seeds.

2.2.2. Summary statistics. Table 2 presents summary statistics about the number of seeds nominated in each (nomination) group, the number of nominations received by the top six

finalists (chosen as seeds), the refusal rates, and the characteristics of the chosen seeds in each of the groups.

On average, we received 19.9 nominations per gossip village, 20.3 per trusted village, and 20.0 nominations per trusted gossip village. The top six nominees were selected in each village, and the average number of nominations received per household was 11.2 for gossip seeds, 10.56 for trusted seeds, and 10.77 for trusted gossip seeds.

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 seeds (17.5%). The trusted and the random seeds were less likely to agree (22% and 19% refusal rates, respectively). This implies that we have slightly more active seeds in the gossip treatment, but the difference is 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 more likely to have some official responsibility in the village (*e.g.* panchayat member, *numberdaar*, or *chaukidaar*). Most notably, they are more likely to describe themselves as interactive. A 46% of the gossips and a similar number of trusted gossips say that they interact very often with others, and that they participate frequently in community activities, in contrast to 26% of the random seeds and 37% of 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.

Relative to both gossip and random seeds, the trusted seeds are older and less likely to be female or a member of a Scheduled Caste. In terms of their probability of holding 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 521 villages from the Haryana project in which the seeding experiment was conducted. The data are aggregated at the village × month level, reflecting the number of children in a village who attended a monthly camp. The intervention lasted for one year (February 2017 to March 2018). The dependent variable is the number of children in a village-month who got immunized against each of a set of particular diseases or for any disease whatsoever.[11]

The empirical specification is as follows:[12]

$$y_{jt} = \theta_0 + \theta_1 GossipTreatment_j + \theta_2 TrustedTreatment_j + \theta_3 TrustedGossip_j + \theta_4 SlopeIncentive_j + \theta_5 FlatIncentive_j + D_{kt} + \epsilon_{jt}, \tag{2.5}$$

where y_{jt} is the number of immunizations of each type received in village j in month t, D_{kt} is a set of district-by-month fixed effects. The standard errors are clustered at the village level. For brevity, we do not report the incentive coefficients in the table.

The results are presented in (the upper portion of) Table 3. In a typical month, in a random seeds village, 18.11 children received at least one shot (column 5). In the gossip villages, 4.9

11. Village × month observations with zero child-level observations are eliminated, since those were months where no camp was held in the village.

12. The variables SlopeIncentive and FlatIncentive control for the additional randomization into two different treatments in terms of incentive payments that people received for multiple vaccinations: flat or varying with the marginal immunization.

TABLE 3
Haryana immunization programme, communication treatment effect

	(1) Children received Penta1	(2) Children received Penta2	(3) Children received Penta3	(4) Children received measles	(5) Children attended session
Gossip	1.017	1.022	1.030	1.078	4.903
	(0.603)	(0.561)	(0.523)	(0.500)	(2.503)
Trusted	0.261	0.302	0.490	0.439	1.849
	(0.486)	(0.448)	(0.418)	(0.408)	(2.047)
Trusted gossip	0.479	0.526	0.514	0.444	2.376
	(0.470)	(0.429)	(0.396)	(0.376)	(1.917)
Observations	6697	6697	6697	6697	6712
Villages	521	521	521	521	521
Mean (Random seeds)	4.31	4.06	3.71	3.53	18.11
Gossip = Random (pval.)	0.092	0.069	0.049	0.032	0.051
Gossip = Trusted (pval.)	0.176	0.168	0.268	0.182	0.192
Gossip = Trusted Gossip (pval.)	0.343	0.338	0.281	0.166	0.271

	(1) Children received Penta1	(2) Children received Penta2	(3) Children received Penta3	(4) Children received Measles	(5) Children attended session
Gossip	1.056	1.056	1.060	1.099	5.052
	(0.604)	(0.563)	(0.525)	(0.501)	(2.509)
Trusted	0.250	0.295	0.486	0.436	1.821
	(0.486)	(0.449)	(0.419)	(0.409)	(2.052)
Trusted gossip	0.474	0.535	0.528	0.452	2.423
	(0.471)	(0.432)	(0.400)	(0.378)	(1.934)
SMS blast 33%	0.799	0.801	0.750	0.516	3.507
	(0.547)	(0.515)	(0.484)	(0.456)	(2.293)
SMS blast 66%	0.024	0.144	0.184	0.111	0.723
	(0.535)	(0.504)	(0.478)	(0.466)	(2.338)
Observations	6697	6697	6697	6697	6712
Villages	521	521	521	521	521
Mean (Random seeds)	4.31	4.06	3.71	3.53	18.11
Gossip = SMS Blast 33% (pval.)	0.746	0.725	0.656	0.382	0.643
Gossip = SMS Blast 66% (pval.)	0.214	0.236	0.226	0.153	0.211
Gossip = Random (pval.)	0.081	0.061	0.044	0.029	0.045
Gossip = Trusted (pval.)	0.153	0.148	0.241	0.168	0.169
Gossip = Trusted Gossip (pval.)	0.309	0.319	0.272	0.16	0.256

Notes: This table uses data from the Haryana immunization programme. It reports estimates of the communication treatment effects. The outcomes are the number of children who 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 village level) are reported in parentheses.

additional children came every month for any immunization ($p = 0.05$). The results are not driven by any particular vaccine. There is a 24–25% increase in the number of children receiving each of the first two vaccines to be given at the camp (penta 1 and penta 2) and a 28–31% increase for the two shots whose baseline take-up levels tend to be lower (penta 3 and measles). The increase of 1.1 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 the gossip treatment effect on the number of children receiving at least one shot per month is remarkably stable over time.

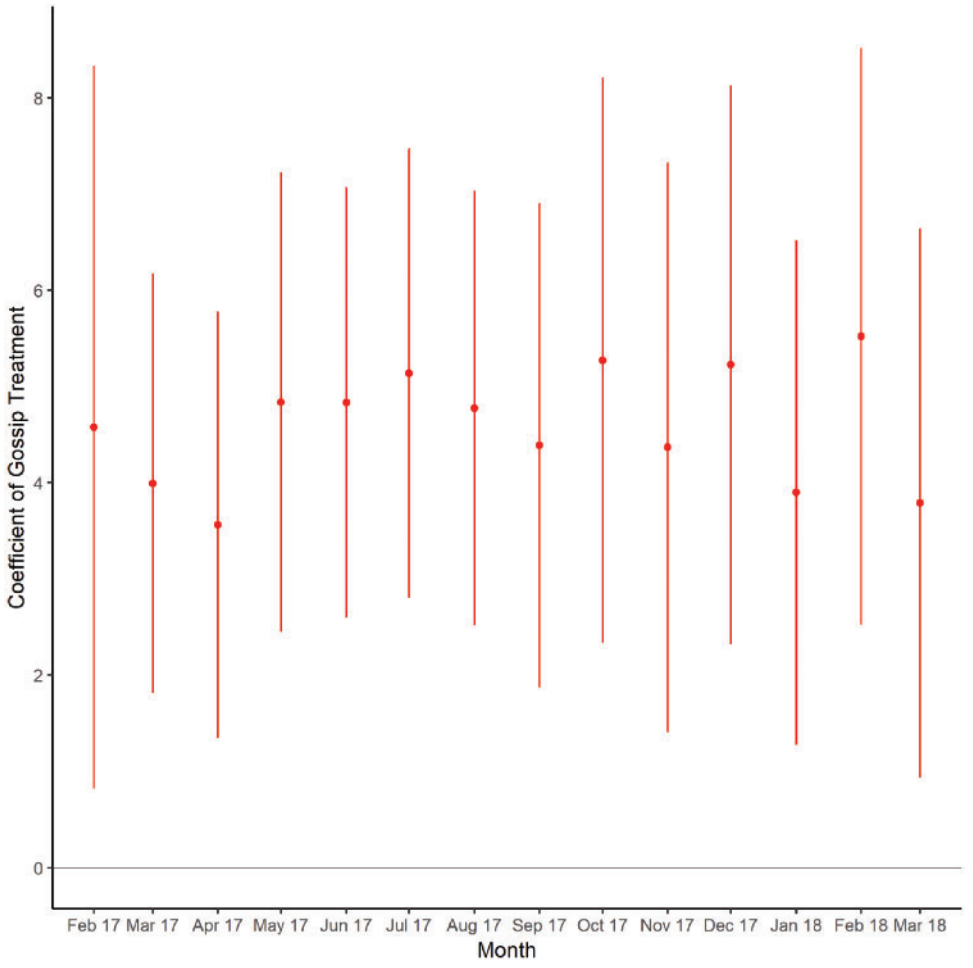


FIGURE 3

Effect of the "gossip" treatment on the number of children who attended an immunization session by month in the Haryana immunization RCT.

Trusted and trusted gossip seeds lead to 1.8 and 2.4, respectively, more children receiving at least one vaccine than random seeds. This is about half the effect of gossip seeds (4.9 more children receiving at least one vaccine than with random seeds). Given the standard errors, we cannot reject that there is no effect of either of these two treatments (compared to random seeds) or that the effect is as large as for the gossip seeds. 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.

Recall that some villages were randomized to be in one of two reminder groups. In those villages, either 33% or 66% of all households whose children had previously attended at least one camp were randomly selected to receive targeted reminders of the next vaccine their children were due for. This allows us to study whether essentially adding many more seeds (here considering those reminded individuals as seeds) generates much more diffusion than, say, gossip-based seeding.

The selection of those who received reminders, however, was different from our targeted seeding treatments in three important respects. First, many more people got information about the camp (given the average attendance in the camp, on average, 34 (68) families per village received the reminders in the 33% (66%) blast villages). This means that there were effectively many more individuals who can be thought of as de facto seeds, but these individuals are not necessarily targeted gossips. Secondly, recipients were parents of children who were likely to need a shot, so we might expect a direct effect, which was not typically the case for the seeds. Thirdly, they were randomly selected among parents of young children who had attended at least one session. It is thus likely that they were less central than the gossip seeds, but on the other hand, they may have had more friends who were responsive to information on child immunization.

In the lower portion of Table 3, we add dummy variables for SMS 33% and 66% blast villages to the regression. There are several interesting results. First, the coefficient of Gossip, Trusted, and Trusted Gossip do not change relative to Panel A. Secondly, the two SMS blast dummies are insignificant, suggesting that the SMS blasts did not lead to greater adoption as compared to seeding information with just a few seeds (*e.g.* six gossips). This is despite the fact that, on average, at least 34 (68) people for whom the information was potentially relevant were directly informed; even in the absence of any information diffusion, these individuals should be affected. Even if power is too low in this sample to distinguish the effects, it is worth noting that the point estimates of the effect of the SMS blasts are much smaller than that of the effect of gossip.

These results run counter to the suggestion in Akbarpour *et al.* (2017) that provided enough seeds are reached, their position in the network should not matter. The SMS blasts reached a large fraction of the population for whom the information was directly relevant but did not lead to any significantly increased diffusion or change in behaviour; in contrast, reaching only a few seeds for whom the information may or may not have been directly relevant, but who were nominated as good diffusers of information, significantly increased immunization.[13] This is not particularly surprising. The results in Akbarpour *et al.* (2017) require that four conditions are met: (1) there is no complex contagion—a single contact suffices to prompt adoption and to diffuse to others, (2) the density of the network is sufficient so that it has a giant component relative to the contagion so that if any node in that component is infected then it infects all the others, (3) information circulates sufficiently long, and (4) the number of seeds grows without bound. Under these conditions, with enough randomly "infected" nodes, the information will hit the giant component, which being dense enough ensures that the information spread fully regardless of where it started. Our results suggest that not all of these conditions are met in our context.

Indeed, we have some direct evidence that condition (1) is violated. The SMS blast was sent only to people who needed to act upon it, but we cannot even reject the hypothesis that none of those who got the SMS responded to it, let alone drawing other people in with them. Along the same lines, going from a blast that reached 33% to one that reached 66% also generated no extra take up. This evidence suggests that message is not enough to affect behaviour, suggesting that we need to think of the diffusion as a complex contagion.[14] This is consistent with what is seen in other development contexts. For instance, Beaman *et al.* (2018) argue that in an experiment with agricultural technology adoption, behaviour is consistent with complex contagion, but not with simple contagion. Our measure of diffusion centrality, discussed in detail below, embodies the total expected number of times that others are reached starting from a seed rather than just how

13. Banerjee *et al.* (2018c) also find, in the context of the Indian demonetization, that considerably greater social learning occurred when the same information was seeded with several gossips rather than broadcast to everyone in the village.

14. When contagion is complex, random seeding can perform arbitrarily worse than targeted seeding, as shown in Jackson and Storms (2018).

many people are reached, and as we shall see, the gossip nominated seeds have high centrality in this sense. Thus, the seeding by nominations may be good at finding seeds that are not only central for simple diffusion, but also central for more complex diffusions. We have no direct evidence on condition (3), which is about how long people continue to talk about issues, in our context. But it is plausible that people do not talk about immunization for very long. Indeed, the folk wisdom on this is that the public seems to lose interest in most topics that are of interest to policymakers rather too quickly.[15] Even in a setting as prominent as the Indian demonetization, conversations among villagers about the confusing policies about what they can do with their cash declined drastically in a matter of weeks (Banerjee *et al*., 2018c).

These results thus confirm that a simple procedure to identify key actors, namely interviewing a random set of households about who are good diffusers of information, can be used to select seeds that generate more wide-spread take-up. Providing information to gossip nominees rather than to a set of random seeds produces greater diffusion. In contrast, providing information to many more households, even to households for whom the information is directly relevant, does not significantly lead to more diffusion. The effect of seeding gossip nominees is large in magnitude, and affects an important and policy-relevant decision with real consequences.

3. NETWORK COMMUNICATION AND KNOWLEDGE OF CENTRALITY

On the one hand, these results may appear to be common sense. In order to find out something about a community, who is influential for example, why not just ask the community members? While this may seem obvious, this is not, as we noted, 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 on information provided by community members. One possible reason is that it is not in fact so obvious that community members 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 10 percentage point 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.

15. We show below that in fact diffusion centrality with finite T (set equal to the expected diameter, which is typically 3) out-performs the $T = \infty$ case at predicting nominations, which then predicts diffusion. Further, in our earlier work on the diffusion of microfinance (Banerjee *et al*., 2013), we showed that diffusion centrality with finite T predicts the extent of diffusion when controlling for all other centralities, including the $T = \infty$ case, as well as all available demographic controls. So T is best thought of as finite and not very large.

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 non-negative 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. Note that everything that we say applies to components of the network.

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

3.1.2. Diffusion Centrality. In Banerjee *et al.* (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.[16]

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 of neighbours and history, each informed node i informs each of its neighbours 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 $\mathbf{H}$, $H(\mathbf{w},T)_{ij}$, 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, we generalize the notion of Diffusion Centrality that we defined in Banerjee *et al.* (2013). That definition is applied to the special case in which $\mathbf{g}$ is a (possibly directed) adjacency matrix taking on values in $\{0,1\}$, where $\mathbf{w} = q\mathbf{g}$ for some communication probability $q \in (0,1]$. That case, with corresponding hearing and diffusion centrality measures, $\mathbf{H}(q\mathbf{g},T)$ 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 about a microfinance programme.

16. See Jackson and Yariv (2011) for background and references on models of diffusion and contagion, and Bloch *et al.* (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).

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.[17] Diffusion centrality thus provides a foundation for these measures, but, importantly, it can behave very differently in the gap between these extreme parameter values of diffusion centrality that these prominent measures take.

In Online 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, (at $T=\infty$) and
(b) if $\lambda_1(\mathbf{w})>1$, diffusion centrality approaches eigenvector centrality.

Part ii(b) 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 Appendix A, 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 limited diffusion takes place even for large T; if those probabilities 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 reach 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 article, these results suggest that the threshold case of $q=1/\mathrm{E}[\lambda_1(\mathbf{g})]$ and $T=\mathrm{E}[Diam(\mathbf{g})]$ provides natural benchmark values for q and T; at these values, information can diffuse, but does not oversaturate a network. This allows us to assign numerical values to $DC(q\mathbf{g},T)_i$. We use this throughout our empirical analysis.

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 housing prices will

17. Let $d(\mathbf{w})$ denote (out) degree centrality: $d_i(\mathbf{w})=\sum_j w_{ij}$. Eigenvector centrality corresponds to $v^{(R,1)}(\mathbf{w})$: the first eigenvector of $\mathbf{w}$. Also, let $GKB(\mathbf{w})$ denote a "generalized" version of Katz–Bonacich centrality to account for possibly weighted and directed networks—defined when $\lambda_1(\mathbf{w})<1$ by $GKB(\mathbf{w}):=\left(\sum_{t=1}^{\infty}(\mathbf{w})^t\right)\cdot\mathbf{1}$.

go up, and u could have told ℓ that "i thinks housing 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 ("i has a new car"), but the second could just be an opinion ("i thinks housing prices will go up"). ℓ keeps track of how often she hears of things originating from i. ℓ ranks i, j, k, and so on, just based on the frequency that she hears things that originated at each one of them.[18]

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.

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 at each other node.

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, are positively correlated with others' diffusion centralities for any $\mathbf{w}, T$.

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

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

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

18. 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.

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.[19]

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*.

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.

Also, in our definition of network gossip, NG, nodes are similar in how frequently they generate new information or gossip; we weighted 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. 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. Secondly, 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.

The theory provides some guidance regarding when the strategy of asking members of the network who are the right seeds will work. First, the restriction to $\lambda_1(\mathbf{w}) > 1$ is important. When

19. 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 further 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.

$\lambda_1(\mathbf{w})$ falls too far 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. Thus, if news does not spread at a high enough rate, people will not have a good idea of who the central people in the overall network are. In a network where people do not exchange information with each other, it will be doubly difficult to use word of mouth to transmit new ideas. Information will in general not circulate very widely and it will be difficult to identify the right seeds.

Secondly, network gossip converges to diffusion centrality as the number of periods of communication increases. Thus, if there are some pieces of news that people talk about for a long time, people will have a better idea of who the central people may be.

Thirdly, for moderate values of q and T, fractionalization in the network—where there are many connections within groups and few across groups (*e.g.* caste)—implies that people will have mostly heard about people from their part of their subnetwork.[20] For instance, people will name central people within their caste accurately but be less likely to nominate people from other castes even if they are in fact central. It will also mean, however, that any seeded information will circulate much more in the subnetwork of each seed. Thus, to apply our gossip nomination strategy in fractionalized societies, one needs to pay attention to ask for nominations within each of the different subnetworks (*e.g.* to reach people in different castes, or hamlets) and be mindful to seed in each community.[21]

Fourthly, we do not model the 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 particular case, their friends may resist passing on information originating from them. The theory suggests that the strategy is more likely to work when the objective is to diffuse information, rather than to model a behaviour. Of course in any given application, it may not be obvious whether the key barrier is information or values. It is interesting that in the case of demand for immunization, it seems the barrier was information; seeding gossips, who are likely to spread information, increases immunization, while seeding trusted individuals or trusted gossips, who may be perceived as more reliable, does not.

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 additional 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. Furthermore, we show that the specific pairwise ranking of centrality (the ranking of j by i) also determines nomination, even after accounting for all observed and unobserved characteristics of i and j with individual fixed effects.

4.1. *Data collection*

We use a rich network dataset 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

20. As a corollary to our main result, if q and T are large, fractionalization does not matter.
21. See Jackson and Storms (2018) for an algorithm for effective seeding with multiple communities.

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) data set: thirty-three villages of the Banerjee *et al*. (2013) networks data set 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.

microfinance as well as how networks changed in response to microfinance (Banerjee *et al*., 2013, 2018b). We use the 2012 data here. To collect the network data (as described in detail in Banerjee *et al*. (2013, 2018b)), we asked adults to name those with whom they interact in the course of daily activities.[22] We have data concerning twelve 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).

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 member of the other household in at least one of the twelve ways. Individuals can communicate if they interact in any of the twelve 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, weighted 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 is 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?*

22. In our Karnataka microfinance village (wave 2) data set, we have network data from 89.14% of the 16,476 households based on interviews with 65% of all adult individuals aged 18–55 years.

We asked two questions to check whether there was any difference depending on what people thought was to be diffused. In practice, the correlation between being nominated for a loan and an event is substantial (0.76) and it made no difference (which is why we collapsed the two questions in the subsequent RCT).

Only 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.[23] In Online 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 nominated *nine* times.[24] This is a first indication that the answers are meaningful; 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% 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.[25]

We refer to the nominees as "gossips". Panel A of Figure 4 shows that under the event question, 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. Accordingly, Table 5 shows that 9% of leaders were nominated as a gossip under the event question whereas 91% were not nominated. Similarly, Panels B of Figure 4 and Table 5 present similar results for the loan question. For instance, 27% of nominated gossips under the event question were leaders, whereas 73% were not.

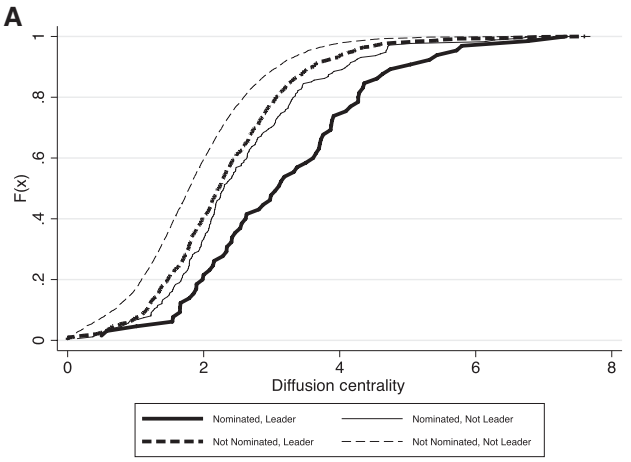
4.2. *Do individuals nominate central nodes?*

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. The first key finding is that villagers are much more likely to nominate people who are central, and they do not just nominate their friends or people who have many friends or positions of influence. A second finding is consistent with a more specific prediction of the model when T is finite. The model gives us specific predictions as to how relatively likely each i would be to nominate each j, when processes run for a short time. We find this to hold in the data. Even controlling for all the characteristics of both i and j (with individual fixed effects), j is more likely to nominate i than i' when i has a higher network gossip measure than i' from the perspective of j.

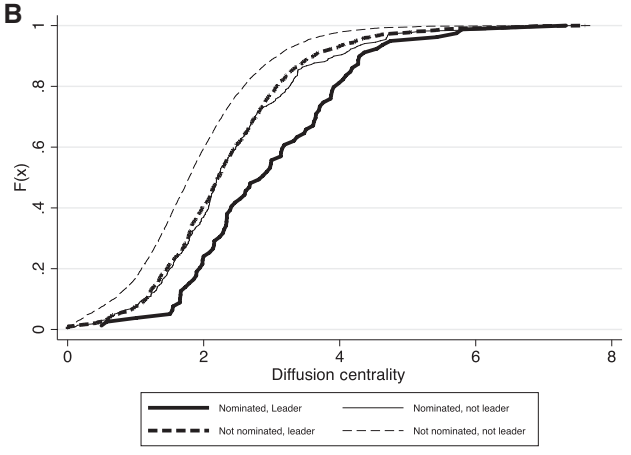
23. See Alatas *et al*. (2014) for a model that builds on this idea.

24. 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.

25. 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.



	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



	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) data set.

Notes: It presents Cumulative Distribution Function (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 (A) and the loan question in (B) and whether or not it has a village leader).

4.2.1. Do households nominate only their friends? 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. 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

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 (under loan and event questions, separately).

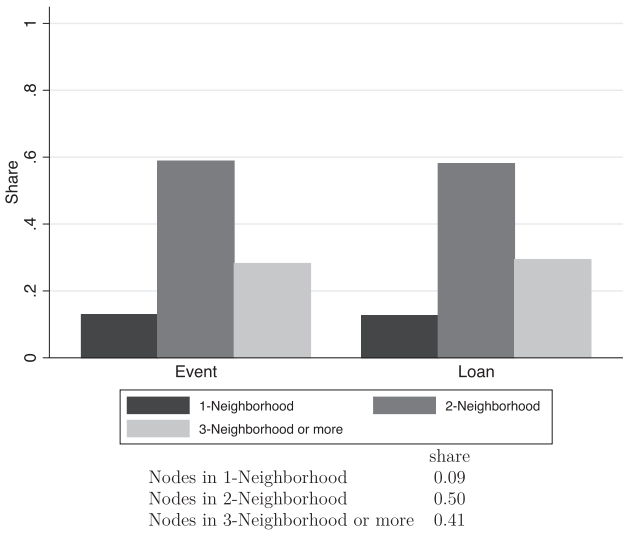


FIGURE 5
Distribution of nominees by network distance in the Karnataka microfinance village (Wave 2) data set.

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, 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" (Feld, 1991; Jackson, 2016)), 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.

4.2.2. Are central individuals more likely to be nominated? Those who have higher diffusion centrality are indeed more likely to be nominated. This is evident in Panel A of Figure 4, which pictures the distributions of diffusion centrality (normalized by its standard deviation across the sample for interpretability) separately for households that were nominated for the event question, those who the organization considered to be leaders, and those who were named

for both or neither.[26] The two distributions of nominated individuals (leaders and not) first-order stochastically dominate the two distributions of those who are not nominated. Leaders are also more central than those who are neither leaders nor nominees. Panel B displays similar results for the loan question.

While these results are consistent with the prediction of the model, there are several plausible alternative interpretations. 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. 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. 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, 2016) and one might therefore expect geographic data to be a useful predictor of centrality. For that reason, since we have detailed GPS coordinates for every household in each village, we include these in our analysis below as controls.[27]

To help rule out these confounding factors, 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), μ_v is a village fixed effect, and ϵ_{ijv} is a Type-I extreme value distributed disturbance.

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 Online Appendix C.

We begin with a number of bivariate regressions in Table 6. 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 2–5 repeat the exercise with two other network statistics (degree and eigenvector centrality), the "leader" dummy, and 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 with one another. To investigate whether diffusion centrality remains a predictor of gossip nomination after controlling for the

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

27. To operationalize geographic centrality, we use two measures. The first uses the centre of mass. We compute the centre of mass and then compute the geographic distance for each agent i from the centre of mass. Geographic centrality is the inverse of this distance, which we normalize by the standard deviation of this measure village-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. We use this weighted graph to compute the eigenvector centrality measure associated with this network. Results are robust to either definition.

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) data set. 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.

other measures, we start by introducing them one by one as controls in columns 1–4 of Table 7 and the three network measures together in column 5. 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 eigenvector centrality cuts the effect of diffusion centrality by about 50%, though diffusion centrality remains significant. The leader dummy is close to being significant, but the coefficient of diffusion centrality remains strong and significant. The geographic centrality variable in column 4 now has a negative coefficient, and does not affect the 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 given they are closely related concepts).

To confirm this pattern, in the last column, we introduce all the variables together and perform a least absolute shrinkage and selection operator (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

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) data set. 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. 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.

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.[28]

We consider the variables diffusion centrality, degree centrality, eigenvector centrality, leadership status, and geographic centrality in the standard LASSO to select the support (*i.e.* the set of relevant variables). 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 that includes only 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 Online Appendix C and find results that are qualitatively similar.

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

TABLE 8
Does network gossip differentially predict nominations?

	(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) data set. 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 in i's network gossip assessment. 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.

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

Therefore, 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.[29] 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 respondent i fixed effects, diffusion centrality of j, or j fixed effects. Specifically, being at the 99th percentile as compared to the 50th percentile of $NG_{\cdot,i}$ corresponds to about an 46.6% increase in the probability of j being nominated by i ($p=0.057$, column 4) and an 10.5% increase conditional on both i and j fixed effects ($p=0.032$, column 6).

4.3. *Re-interpreting 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 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 sixty-nine villages in which seeds were randomly selected (two of the seventy-one 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 define a household as having "high diffusion centrality" if its diffusion centrality is at least one standard deviation above the mean. Under this definition, in our sixty-nine villages, 13% of households are considered to have "high diffusion

29. 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 $T=\mathrm{E}[Diam(\mathbf{g})]$.

TABLE 9
Calls received by seed type

	(1) Calls received	(2) Calls received	(3) Calls received	(4) Calls received / Seeds	(5) Calls received / Seeds	(6) Calls received / 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 data set. 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.

centrality", while 1.7% of households were nominated as seeds, and 9.6% were "elders". Twenty-four villages have exactly one randomly chosen seed that is of high diffusion centrality fourteen have more than one. Twenty-three villages have exactly one randomly chosen seed that is a gossip, and eight have more than one.[30]

Column 1 of Table 9 runs the same specification as in Table 1 but in the sixty-eight random seed villages. In these villages, hitting a gossip by chance increases the number of calls by 6.65 (compared to 3.79 in the whole sample). Having at least one gossip more than doubles the number of calls received than random (statistically significant), and having at least one elder only increases the number of calls by 0.35 relative to random. The results are thus very similar to the results in the full sample, although power is of course lower with only sixty-eight villages, and we cannot separate that gossip and elder is different in this restricted sample ($p=0.26$). In column 3, we regress the number of calls on a dummy for hitting at least one high *DC* seed. High *DC* seeds do increase the number of calls; the point estimate on at least one high *DC* seed is 5.18. In column 2, we augment the specification in column 1 to add the dummy for at least one high *DC* seed. Since *DC* and gossip nomination are correlated, the regression is not particularly precise. The point estimate of at least one gossip, however, only declines slightly.

The point estimates suggest that diffusion centrality captures part of the impact of a gossip nomination, but likely not all of it: even controlling for their diffusion centrality, gossip seeds still lead to greater diffusion.

30. We continue to exclude the one village in which a gossip seed broadcasted information. The results including that village are in Online 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. In fact diffusion centrality, even on its own, is not significantly associated with more 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 villagers having better estimates of diffusion centrality from their network gossip than we do from our surveys. Secondly, it also could be that gossip nomination is a richer proxy for information diffusion than a 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 (*e.g.*, altruism, interest in the information, etc.) and villagers may be good at capturing those factors as well. 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 many settings, identifying central members in a social network and seeding information to them can accelerate the diffusion of information. But collecting network data is expensive and not always practical, and it is therefore important to find cheaper ways to identify the central people. In this article, we ask whether villagers, though they have a poor image of the social network, are able to identify the most central individuals in the network.

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. Results of the latter RCT are consistent with those of the former: there is an increase of over 20% in immunization visits when the seeds are gossip nominees.

A simple network information model rationalizes these results. It 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. Indeed, when asked for nominations, villagers 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 information 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. This is because 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, they 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.

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. 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 our work on microfinance (Banerjee *et al.*, 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 article shows that even when the final outcome involved an important and personal decision, pure information "gossips" are effective at increasing diffusion and take-up, while trust-based seeding strategies deliver very noisy results that statistically do not differ from zero.

Secondly, our experiments here are limited to communities whose size is 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. There are many settings, in both the developing and developed worlds, 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 developing strategies to do this. Even in large societies, one may want to choose many seeds, some within in each of various sub-communities, in which case the techniques developed here would be useful.

Beyond these applications, the work presented here opens a rich agenda for further research: exploring which other aspects of agents' social environments can be learned in simple ways. Given the fact that people seem to know little about the network around them, how do they navigate it in their daily decisions? That is, how do people decide whom to ask to find important information, and whom to tell in order to help spread information to others? Social learning in general is far from a passive activity: people make choices in both how to acquire information and how to spread it. Our theory provides one part of a foundation for a more general theory and investigation into how such "active social learning" occurs.[31] This has important policy implications, as people's knowledge of their networks determines the efficiency of such active social learning, and any distortions that it may exhibit.

Acknowledgments. 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 the editor, four referees, Yann Bramoullé, 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.

Supplementary Data

Supplementary data are available at *Review of Economic Studies* online.

REFERENCES

AKBARPOUR, M., MALLADI, S. and SABERI, A. (2017), "Diffusion, Seeding, and the Value of Network Information" Available at SSRN: https://ssrn.com/abstract=3062830.

ALATAS, V., BANERJEE, A., Chandrasekhar, A. G., *ET AL.* (2014), "Network structure and the aggregation of information: Theory and evidence from Indonesia" (Working Paper NBER).

AMBRUS, A., MOBIUS M. and SZEIDL, A. (2014), "Consumption Risk-Sharing in Social Networks", *American Economic Review*, **104**, 149–182.

ARAL, S., MUCHNIK, L. and SUNDARARAJAN, A. (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., CALVÓ-ARMENGOL, A. and ZENOU, Y. (2006), "Who's Who in Networks, Wanted: The Key Player", *Econometrica*, **74**, 1403–1417.

31. Another thread that can be thought of as part of such an investigation is the seminal work of Katz and Lazarsfeld (1955) on opinion leaders, and the long literature that followed, including some in economics (Galeotti and Goyal, 2010).

BANERJEE, A., CHANDRASEKHAR, A., DUFLO, E., *ET AL.* (2018a), "*TW10.1119 Improving Full Immunization Rates in Haryana, India: Evaluating Incentives and Communication Methods*", Grantee Final Report, International Initiative For Impact Evaluation.
BANERJEE, A., CHANDRASEKHAR, A., DUFLO, E., *ET AL.* (2013), "Diffusion of Microfinance", *Science*, **341**, DOI: 10.1126/science.1236498.
BANERJEE, A., CHANDRASEKHAR, A., DUFLO, E., *ET AL.* (2018b), "Changes in Social Network Structure in Response to Exposure to Formal Credit", Available at SSRN: https://ssrn.com/abstract=3245656.
BANERJEE, A. V., BREZA, E., CHANDRASEKHAR, A. G., *ET AL.* (2018c), "When Less is More: Experimental Evidence on Information Delivery During India's Demonetization" (Mimeo: Stanford University).
BEAMAN, L., BENYISHAY, A., MAGRUDER, J., *ET AL.* (2018), "Can Network Theory based Targeting Increase Technology Adoption?" (No. w24912). National Bureau of Economic Research.
BELLONI, A. and CHERNOZHUKOV, V. (2009), "Least Squares after Model Selection in High-Dimensional Sparse Models" (Working Paper MIT Department of Economics).
BINDLISH, V. and EVENSON, R. E. (1997), "The Impact of T&V Extension in Africa: The Experience of Kenya and Burkina Faso (English)", *The World Bank Research Observer*, **12**, 183–201.
BLOCH, F., DEMANGE, G. and KRANTON, R. (2014), "Rumors and Social Networks" (Working paper 2014 - 15 Paris School of Economics).
BLOCH, F., JACKSON, M. O. and TEBALDI, P. (2016), "Centrality Measures in Networks", *http://ssrn.com/abstract=2749124*.
BOLLOBAS, B. (2001), *Random Graphs*, (Cambridge, UK: 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.
BORGATTI, S. P. (2006), "Identifying Sets of Key Players in a Social Network", *Computational & Mathematical Organization Theory*, **12**, 21–34.
BREZA, E., CHANDRASEKHAR, A. and TAHBAZ-SALEHI, A. (2017), "Seeing the Forest for the Trees? An Investigation of Network Knowledge" (Mimeo, Stanford).
CASCIARO, T. (1998), "Seeing Things Clearly: Social Structure, Personality, and Accuracy in Social Network Perception", *Social Networks*, **20**, 331–351.
CHANDRASEKHAR, A. and LEWIS, R. (2016), "Econometrics of Sampled Networks" (Mimeo, Working paper Stanford).
COLEMAN, J., KATZ, E. and MENZEL, H. (1966), *Medical Innovation: A Diffusion Study* (Indianapolis, IN: Bobbs-Merrill).
FAFCHAMPS, M. and GUBERT, F. (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**, 1464–1477.
FRIEDKIN, N. E. (1983), "Horizons of Observability and Limits of Informal Control in Organizations", *Social Forces*, **61**, 54–77.
GALEOTTI, A. and GOYAL, S. (2010), "The Law of the Few", *American Economic Review*, **100**, 1468–1492.
GUIMARAES, P., FIGUEIRDO, O. and WOODWARD, D. (2003), "A Tractable Approach to the Firm Location Decision Problem", *Review of Economics and Statistics*, **85**, 201–204.
HINZ, O., SKIERA, B., BARROT, C., *ET AL.* (2011), "Seeding Strategies for Viral Marketing: An Empirical Comparison", *Journal of Marketing*, **75**, 55–71.
IYENGAR, R., DEN BULTE, C. V. and VALENTE, T. W. (2010), "Opinion Leadership and Social Contagion in New Product Diffusion", *Marketing Science*, **30**, 195–212.
JACKSON, M. O. (2008), "Average Distance, Diameter, and Clustering in Social Networks with Homophily", in Papadimitriou, C. and Zhang S. (eds), *the Proceedings of the Workshop in Internet and Network Economics (WINE 2008), Lecture Notes in Computer Science, also: arXiv:0810.2603v1* (Berlin Heidelberg: Springer-Verlag).
JACKSON, M. O. (2016), "The Friendship Paradox and Systematic Biases in Perceptions and Social Norms", *Journal of Political Economy*. https://www.journals.uchicago.edu/doi/abs/10.1086/701031?journalCode=jpe.
JACKSON, M. O. (2017), "A Typology of Social Capital and Associated Network Measures", Available at SSRN http://ssrn.com/abstract=3073496.
JACKSON, M. O. and STORMS, E. C. (2018), "Behavioral Communities and the Atomic Structure of Networks", Available at SSRN: https://ssrn.com/abstract=3049748.
JACKSON, M. O. and YARIV, L. (2011), "Diffusion, Strategic Interaction, and Social Structure", *Handbook of Social Economics*, San Diego: North Holland, edited by Benhabib, J. Bisin, A. and Jackson, M.O. (eds).
KATONA, Z., ZUBCSEK, P. P. and SARVARY, M. (2011), "Network Effects and Personal Influences: The Diffusion of an Online Social Network", *Journal of Marketing Research*, **48**, 425–443.
KATZ, E. and LAZARSFELD, P. (1955), *Personal Influence: The Part Played by People in the Flow of Mass Communication* (Glencoe, IL: Free Press).
KEMPE, D., KLEINBERG, J. and TARDOS, E. (2003), "Maximizing the Spread of Influence through a Social Network", *Proc. 9th Intl. Conf. on Knowledge Discovery and Data Mining*, 137–146.
KEMPE, D., KLEINBERG, J. and TARDOS, E. (2005), "Influential Nodes in a Diffusion Model for Social Networks", *In Proc. 32nd Intl. Colloq. on Automata, Languages and Programming*, 1127–1138.
KRACKHARDT, D. (1987), "Cognitive Social Structures", *Social Networks*, **9**, 109–134.

KRACKHARDT, D. (1996), "Strucutral Leverage in Marketing", in *Networks in Marketing*, ed. by D. Iacobucci, Sage, Thousand Oaks, 50–59.
KRACKHARDT, D. (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)*, **58**, 253–266.
LAWYER, G. (2014), "Understanding the Spreading Power of All Nodes in a Network: A Continuous-Time Perspective", *arXiv:1405.6707v2*.
LEDUC, M. V., JACKSON, M. O. and JOHARI, R. (2017), "Pricing and Referrals in Diffusion on Networks", *Games and Economic Behavior*, **104**, 568–594.
LIM, Y., OZDAGLAR, A. and TEYTELBOYM, A. (2015), "A Simple Model of Cascades in Networks" (Mimeo, University of Oxford).
PALMGREN, J. (1981), "The Fisher Information Matrix for Log Linear Models Arguing Conditionally on Observed Explanatory Variables", *Biometrika*, 563–566.
PALUCK, E. L., SHEPHERD, H. and ARONOW, P. M. (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.
VALENTE, T. W. (2012), "Network Interventions", *Science*, **337**, 49–53.
VALENTE, T. W., CORONGES, K., LAKON, C., *ET AL.* (2008), "How Correlated Are Network Centrality Measures?" *Connect (Tor).*, **28**, 16–26.
VALENTE, T. W. and PUMPUANG, P. (2007), "Identifying Opinion Leaders to Promote Behavior Change", *Health Education & Behavior*, **34**, 881–896.

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 neighbourhood. In that case diffusion centrality just counts direct paths and it 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 see, e.g., standard arguments (Jackson, 2008), 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 ($T = o(pn)$),*[32] *then the expected diffusion centrality of any node converges to $npq\frac{1-(npq)^T}{1-npq}$. That is, for any i,*

32. 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.

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

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[33], 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* $\varepsilon>0$[34] *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* $\mathrm{E}\left[DC\left(q\mathbf{g}(n,p),T\right)_i\right]\to\infty$.[35]

(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:*[36]

(a) If $T<(1-\varepsilon)\mathrm{E}[Diam(\mathbf{g}(n,p))]$ *for some* $\varepsilon>0$, *then* $\frac{\mathrm{E}\left[DC\left(q\mathbf{g}(n,p),T\right)_i\right]}{n}\to 0$.
(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* $\frac{\mathrm{E}\left[DC\left(q\mathbf{g}(n,p),T\right)_i\right]}{n}=\Omega(1)$.

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 $DC(1/\mathrm{E}[\lambda_1]\mathbf{g},\mathrm{E}[Diam(\mathbf{g}(n,p))])$ 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 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.

33. 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 $\log(n)/\log(pn)$.

34. 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).

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

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