Can Random Friends Seed More Buzz and Adoption? Leveraging the Friendship Paradox

Vineet Kumar
Yale School of Management, 165 Whitney Avenue, New Haven, CT 06511, vineet.kumar@yale.edu

K. Sudhir
Yale School of Management, 165 Whitney Avenue, New Haven, CT 06511, k.sudhir@yale.edu

A critical element of word of mouth (WOM) or buzz marketing is to identify seeds, often central actors with high degree in the social network. Seed identification typically requires data on the relevant network structure, which is often unavailable. We examine the impact of WOM seeding strategies motivated by the friendship paradox, which can obtain more central nodes *without knowing network structure*. Higher-degree nodes may be less effective as seeds if these nodes communicate less with neighbors or are less persuasive when they communicate; therefore whether friendship paradox motivated seeding strategies increase or reduce WOM and adoption remains an empirical question. We develop and estimate a model of WOM and adoption using data on microfinance adoption across village social networks in India. Counterfactuals show that the proposed strategies with limited seeds are about 14-30% more effective in increasing adoption relative to random seeding. These strategies are also on average 5-11% more effective than the firm's leader seeding strategy. We also find these strategies are relatively more effective when we have fewer seeds.

Key words: word of mouth, networks, seeding, friendship paradox, product adoption, diffusion*

1. Introduction

Firm-initiated and consumer-driven word of mouth (WOM) marketing (often referred to as buzz marketing), has received a lot of attention, and has proven effective in increasing adoption across a wide range of products and services. WOM has been examined both theoretically and empirically using a wide range of modeling approaches, to understand both the motivations to engage in it and its various impacts (Godes and Mayzlin 2009, Iyengar et al. 2011, Campbell et al. 2017, Berger and Iyengar 2013, Cai et al. 2015).

An important question in WOM marketing is how to choose appropriate seeds. There are a few broad approaches considered in the literature. The first approach uses network data on connections to identify central individuals (e.g. degree or eigenvector centrality) to provide the most WOM (Tucker 2008, Goldenberg et al. 2009, Libai et al. 2013). Recently, researchers have tried to combine

* We acknowledge helpful feedback from Eric Bradlow, Tat Chan, Khai Chiong, and participants in seminars at the Four-school Marketing Conference, Marketing Dynamics Conference, Massachusetts Institute of Technology, University of California at San Diego, University of Toronto, Yale Computational Social Science Seminar.

multiple networks among the same individuals to identify seeds with specific relationship types that can lead to higher adoption (Chen et al. 2017). The second approach uses individual characteristics to identify how opinion leaders can be used to seed networks (Iyengar et al. 2011). But opinion leaders are often highly context-specific and may not span multiple categories, e.g. an opinion leader in fashion might not be an opinion leader for consumer electronics or healthcare (King and Summers 1970). Another approach is to identify seeds based on local network properties and community characteristics to achieve higher diffusion (Yoganarasimhan 2012). But there might be tradeoffs in that network structures that enable high diversity of content might not be efficient at accelerating the flow of information (Aral and Van Alstyne 2011).[1]

Broadly, the emphasis in the recent literature has been to improve seed identification using more comprehensive network data. However, even with easier access to online social networks, data on the *relevant network for a particular purpose* is often unavailable. For example, even if one had access to the Facebook (or similar) social networks of everyone including physicians, the *relevant* physician-to-physician network data for seeding a new drug may be unavailable. Even within a specific context, there are many challenges in gathering accurate network data, including the time and effort required to obtain this data (Stark 2018). Moreover, the dynamically evolving nature of connections and relationships requires frequent updating of such data. Social media data, which are relatively easier to access also face the challenge that activity there maybe more of a substitute than a complement to offline or other social interactions and may not be effective in high-involvement applications (Borgatti et al. 2009). Given all of the above, a theoretically grounded seeding approach that can be used in the absence of complete data on network structure can be valuable.

The friendship paradox suggests such just such an approach to obtain higher degree seeds. The paradox is often stated as: "On average, your friends have more friends that you do," and proves that random friends are more highly connected (have higher degree) than random nodes (Feld 1991, **?**). We empirically investigate whether the friendship paradox can indeed be leveraged to choose seeds and generate higher WOM and adoption *when the relevant network structure information is unavailable*. Specifically, we address the following research questions:

1. Can friendship paradox based seeding strategies improve WOM and adoption relative to random seeding? Can it improve upon an opinion leader based strategy chosen by the firm?

2. Can hybrid approaches leveraging the friendship paradox along with leadership characteristics lead to higher adoption?

[1] There is a complementary literature in computer science inspired by Domingos and Richardson (2001) on approximate seeding algorithms for influence maximization with performance guarantees. These algorithms differ in the level of network information used. For example, Kempe et al. (2003) use full network information, Eckles et al. (2019) use partial network information and Wilder et al. (2018) consider algorithms when network information is unknown.

3. How does the extent of initial seeding (proportion of the network seeded) impact *absolute and relative performance* of the strategies?

Friendship Paradox Motivated Network Seeding Strategies. The friendship paradox statement "On average, your friends have more friends than you do" is based on a mathematical result that holds independent of network structure, because popular people are always over-represented in the set of friends (Feld 1991, Kumar et al. 2024). This strategy has been suggested for immunization of networks and sensors on networks (Cohen et al. 2003, Christakis and Fowler 2010). The basic intuition is simple: suppose we choose an initial node at random (so each node has an equal probability of being selected) and then choose one friend of that node at random. The chosen friend of the initial mode is likely to be more highly connected than average, since, by construction, a highly connected node will be in the friend set of more people, and therefore more likely to be nominated as a friend. For intuition, consider two extreme examples: (i) *a simple hub-spoke network* with a central node and several peripheral nodes, all connected *only* to the central node. Each node has equal probability of being initially selected, so we are very likely to get a peripheral node. When asked to nominate a friend, each of the peripheral nodes can only suggest the central node, who is their only friend; (ii) *an isolated node* without any connections. That node would never be chosen by anyone on the network as a friend.

The friendship paradox thus suggests potential strategies for sampling higher degree individuals (those with more friends) in any network, without knowing network structure. For example, one could select a random friend each for a set of randomly chosen individuals. This strategy only requires access to a set of randomly sampled individuals, and the ability to obtain a random friend from them. Further, one can easily obtain the relevant network, by choosing the list of relevant friends from which to sample for the particular seeding problem at hand, e.g.for physician influence networks we might ask a doctor to suggest the contact of a random physician friend with whom they discuss professional matters.

The theoretical results on the friendship paradox guarantee that individuals with higher than average degree are obtained *in expectation* no matter what the underlying network, allowing for potentially better seeds. However, even though the sampled individuals have higher expected degree, their use as seeds cannot guarantee greater WOM or product adoption, because the extent to which higher degree individuals communicate with friends in their network about the product is an empirical question. For instance, Kim et al. (2015) found that selecting the highest degree nodes did not yield greater adoption than with random seeding.

Challenges in evaluating network seeding strategies. A seemingly straightforward approach to empirically evaluate the effectiveness of alternative seeding challenges is to conduct a field experiment where different seeding strategies are assigned at random to different networks. However, it is a challenge to obtain credible, robust answers on the effectiveness of various seeding strategies using standard experimental approaches because effective matching of treatment and control groups at the level of network structure is not typically feasible through randomization.[2] Network structure plays a crucial role in diffusion processes, and even small changes in structure can make contagion cascades possible (Centola 2010, Katona et al. 2011). Therefore, merely comparing adoption differences across seeding strategies without appropriate controls for how network structure impacts communication and adoption within each network would not be credible or robust.

Kim et al. (2015) conducted a field experiment by randomizing seeding treatments for two health-related interventions across 32 Honduran villages and compare the average adoption performance of random seeding and friend of random individual seeding. They found mixed results about the effectiveness of friendship paradox based seeding in two health-based interventions—adoption of multivitamins and chlorine-based water purification. Beyond the challenges in assessing differences in treatment and control groups (discussed earlier), there are important differences in empirical contexts. First, target seeds in Kim et al. (2015) received an intervention (a product along with education), whereas in our case the seeds only decide whether to adopt the product. Second, seeds received tickets to distribute to their friends and the outcome measure was ticket redemption. Due to the focus on ticket redemption, word of mouth communication beyond neighbors of non-ticketed households and the role of non-adopters in information diffusion is ignored. But these aspects of communication that we model and account for are typically important in new product adoption.

Relatedly, the issue of better control and precision to detect differences in effectiveness of stochastic seeding strategies using field experimental data treated at the network level has been considered using Rubin (2005)'s potential outcomes framework (Chin et al. 2022). This is a useful approach, but there remain important practical challenges. First, the approach only works for stochastic seeding, while our approach works for both deterministic and stochastic seeding. In fact, for the data we use from Banerjee et al. (2013), the approach would not be applicable, because only leaders are chosen as seeds in all villages. Further, it is unclear if the technique can be adapted to answer the richer set of questions around leader and hybrid seeding strategies. Second, due to the nature of the algorithm, the improvements in precision occur only with relatively small seed sets (<5), which is unrealistic for many marketing settings except with very small network sizes. Substantively, unlike our results, they detect no difference between random friend and random strategies. This difference could be because the networks they consider have relatively small degree range—a factor that is known to make random friend strategies less effective (Kumar et al. 2024).[3]

[2] The number of possible networks structures grows exponentially in the number of nodes; for example if $N = 100$ nodes, there are $2^{\frac{N(N-1)}{2}} \approx 10^{1490}$ possible undirected network structures, and even more for directed networks.

[3] The data used in **?** is from the studies by Cai et al. (2015) and Paluck et al. (2016). The data in Cai et al. (2015) includes 185 village networks and states the following (with the exception of 2 villages): "The social network survey

Our approach. Our approach involves estimating a micro-founded structural model of diffusion of WOM and product adoption over networks and exploiting any experimental or quasi-experimental variation arising from seeding strategies across networks to estimate the model. Though the modeling approach is parametric, it allows for testing various assumptions about the diffusion process, and also allows for various practical challenges in randomization when conducting network based field experiments involving seeding to be accounted for in the model (rather than be assumed away). This aspect allows us to assess robustness of the results not only to differences in network structure, but also account for specific features of the data generation and WOM communication process. In terms of key modeling and estimation features, the model allows for a flexible relationship between degree and WOM — a critical ingredient to evaluating the benefits of increasing degree through the friendship paradox. Further, unlike typical diffusion models, which *assume* that all WOM arises from adopters, our model allows WOM from both adopters and non-adopters. We also provide a novel non-parametric identification argument that leverages the feature that leaders were chosen as seeds, along with the shape of the adoption trajectory to identify differential effects for leaders versus non-leaders.

Estimating such a WOM diffusion model is challenging because the necessary multi-network data is typically unavailable. Most diffusion models are estimated based on one product's time series of adoption through one market (or social network). Further, the original seeding is typically unobserved, and even if observed it is often not possible to identify the effect of different seeding without multiple diffusion paths across similar networks. Finally, the impact of WOM might be misidentified in the presence of advertising (Van den Bulte and Lilien 2001), attributing to WOM what was actually achieved by advertising.

We address these challenges using data on one product (microfinance) adoption across 43 independent and relatively isolated village social networks in India. The firm's seeding across the different villages leads to exogenous variation in network position and characteristics of seeds, which aids in identifying the impact of seeding. Also, there was no advertising or promotion activity by the firm that would confound WOM effects, which is known to bias estimates of impact (Van den Bulte and Lilien 2001). Based on the estimates, we simulate counterfactuals on WOM and product adoption across these villages as a function of alternative seeding strategies. Finally, we compare the effectiveness of the friendship paradox based Local Friend and hybrid seeding strategies relative to Random and opinion leader based strategies.

asked household heads to list five close friends, either within or outside the village, with whom they most frequently discuss rice production or financial issues." Similarly, Paluck et al. (2016) asks participants to list a maximum of 10 friends.

Findings. We find that higher degree nodes are less likely to communicate WOM among adopters, but there is no such difference for non-adopters. However, despite this negative correlation between degree and WOM among adopters, we estimate that the friendship paradox based Local Friend strategy provides a significant (15-24%) improvement over the Random strategy. It also improves effectiveness over a leader seeding strategy used by the firm by 5-13%. When the Local Friend strategy is used in conjunction with the Leader strategy, the hybrid provides a further marginal improvement. Finally, we find that informationally more demanding strategies like Top Degree and Top Diffusion perform better than other strategies, obtaining close to 44% improvement over the random baseline.

Contributions. Our paper makes a number of contributions to the literature on seeding strategies and diffusion. The present paper is among the first to empirically demonstrate the potential value of seeding a random friend over random nodes for word of mouth communication in a robust manner. We do so with a micro-model of household decision to adopt and communicate, that is flexible enough to permit higher degree nodes to communicate less per-capita with friends and distinguish WOM from leaders and non-leaders. Second, we compare leader seeding chosen by the firm to random friend seeding. To estimate differential effects of leaders from non-leaders, we develop a novel non-parametric identification strategy that uses the temporal trajectory of adoption diffusion. Finally, we evaluate counterfactual hybrid strategies using both friendship paradox ideas and leadership characteristics and detail which types of hybrid strategies obtain greater adoption. Overall, we demonstrate that our results around the superiority of informationally light random friend seeding is robust to various specification checks. We also find that when the proportion of seeds is lower, the relative effectiveness of the Local Friend strategy is greater.

2. Data

We use panel data collected by Banerjee et al. (2013) on the diffusion and adoption of microfinance across households belonging to 43 rural villages in southern India in combination with rich network data on the social connections among the households within each village.

The microfinance firm identified opinion leaders based on leader and social criteria in each village prior to entry and seeded information about the microfinance product among these individuals first. Table 1 provides the summary statistics of the village household networks. Households have an average of more than 4 individuals. Averaged across villages, 61% of households have private electricity, but only 28% of households have private latrines. There is relatively lower variation in the number of people relative to rooms or beds across the households.

Table 1 Village Network Statistics

Statistic	Mean	SD	Min	Max
Household Characteristics:				
Number of Households in Village	212.23	53.54	107.00	341.00
People in Household	4.77	0.37	4.20	5.69
Rooms in Household	2.31	0.41	0.75	2.94
Beds in Household	0.88	0.45	0.29	2.27
Proportion of Households with Electricity	0.61	0.16	0.11	0.89
Proportion of Households with Latrines	0.28	0.15	0.02	0.90
Proportion of Households with Leaders	0.13	0.03	0.07	0.21
Degree:				
Mean	9.66	1.64	6.82	13.59
Standard Deviation	7.09	1.32	5.18	11.02
Minimum	1.00	0.00	1.00	1.00
Maximum	39.72	13.01	23.00	90.00
Mean of Leaders	12.93	2.59	8.88	18.82

Note: Unit of analysis is a village network (N=43)

Table 2 Adoption across Household Types (%)

Statistic	Mean	SD	Min	Max
All Households	19.38	8.16	7.66	45.08
Leader Households	24.71	12.64	3.57	53.85
Follower Households	18.68	8.19	7.30	43.71
Non-electrified Households	23.68	10.48	6.94	55.46
Electrified Households	15.87	7.52	4.76	34.43
Non-latrine Households	21.74	9.87	7.03	51.25
Latrine Households	14.68	9.30	0.00	36.36

Note: Unit of analysis is a village network (N=43)

We use the union of all undirected network relationships detailed in the data.[4] There is considerable variation in the extent of relationships among households. Each village contains on average 212 households. Across villages, the mean degree (connections) of households is around 9, the mean of the standard deviation of degree for households at the village level is large at around 7.1, with the minimum and maximum reflecting wide variation. The mean degree of leaders is higher than the average and close to the maximum of average degree across villages. Opinion leaders have a much higher degree than average (34% more). We illustrate the network for Village 1 in Figure 1 as an example.

The primary performance comparison in our study is the adoption of microfinance by households across the villages. Table 2 summarizes adoption across household types. We find that 19.4% of

[4] We use the union since communication can happen during any type of interaction. Similarly, we convert directed ties (e.g., survey questions about about borrowing items like rice or fuel oil) to undirected ties since communication about the product can be bidirectional during any such asymmetric interaction.

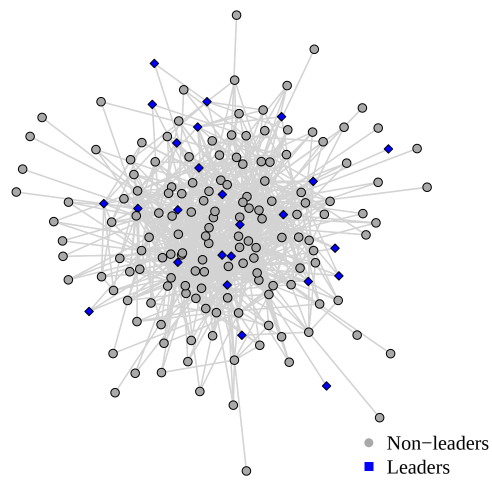

Figure 1 Social Network of Households in Village 1

households adopt microfinance, with significant variation across the villages. Opinion leaders are more likely to adopt than followers, perhaps a feature of the information propagation chosen by the firm, which targeted these leaders in each village. Adoption is correlated with household characteristics; electrified households are less likely to adopt compared to non-electrified, and households with a latrine are less likely to adopt than those without. Broadly, these statistics reflect the reality that microfinance is used by poorer households in emerging markets.

3. Model and Estimation

We use a model of WOM and product adoption across a social network. Using network terminology, households are *nodes* and connections between them are *edges*. Households need to be informed about the product in order to adopt. Households who are informed communicate with their neighbors probabilistically, even if they have not adopted. We build upon the model of Banerjee et al. (2013), with key adaptations required to study our research question related to the friendship paradox. First, we allow the WOM probability from a node to differ by degree, reflecting the idea that WOM effort may depend on this factor. Second, we allow the WOM probability from a node to differ for those identified as leaders by the firm. Banerjee et al. (2013) allow the probability of WOM to depend on adoption status but not on the number of connections (degree) or leader status. Our extensions are specifically motivated by the strategies examined here. Recall that our approach based on Friendship Paradox obtains higher degree nodes than average. Thus, we chose a conservative approach, allowing for the idea that whereas high degree nodes may be better due to their degree, they *might* also be less likely to communicate with their friends or network neighbors. If we did not account for that, then we could be biasing the results *in favor of the friendship paradox strategy.* Similarly, accounting for differences in WOM among firm-designated "leaders" is critical to assess the effectiveness of leader strategies.

Baseline Model

Word of Mouth Communication: WOM occurs in the network when a household receives information (only) from its *informed* neighbors. We allow WOM probability $p^s(D)$ to depend on adoption status s and degree D.

$$p^s(D) = q^s_{min} + (q^s_{max} - q^s_{min}) \left[\frac{D - D_{min}}{D_{max} - D_{min}} \right] \tag{1}$$

The WOM probability $p^s(D)$ refers to the probability that a node with degree D and adoptions status s (either adopter or non-adopter) communicates with each of its network neighbors. Thus, q^s_{min} represents the WOM probability for a node with minimum degree ($D = D_{min}$), whereas q^s_{max} represents the WOM probability for the highest degree node ($D = D_{max}$) and adoption status is denoted s. These quantities are based on the minimum and maximum degrees across all networks. Both parameters depend on the adoption status $s \in \{NA, A\}$ of the node, with NA indicating "Not Adopted" and A indicating "Adopted." The specification in Banerjee et al. (2013) is a special case of this model when $q_{min} = q_{max} = q$, such that WOM is independent of degree. Nodes continue communicating with neighbors in periods after they become informed.

The WOM process details are provided in §EC.3 in the supplement. To summarize, we model households as belonging to Uninformed (U), Informed (I), Adopter (A) and Non-adopter (NA). Initially, all households are uninformed. The initial seeds become informed (I) due to the firm's communication. In each period, all informed nodes communicate probabilistically with each of their neighbors, and the probability of such communication is $p^s(D)$. Such communication occurs at all time periods, motivated by in-person interactions that households have with their friends (network connections). We model these events as being independent draws across the set of friends of a household. Households become fully informed following a WOM communication received from any friend. Once a household becomes newly informed (transition from U to I state), they make a decision on adoption, and they do not revisit the adoption choice in subsequent periods.

Adoption: When a household becomes aware of the product at time t, the household's decision of whether to adopt, $y \in \{0, 1\}$, is modeled as a standard logit choice with observed heterogeneity. The utility of household i from adoption and non-adoption is:

$$\begin{aligned} u_i(y = 1) &= \beta_0 + \beta X_i + \epsilon_{i,1} \\ u_i(y = 0) &= \epsilon_{i,0} \end{aligned} \tag{2}$$

X_i represents the vector of leader characteristics of household i, β the vector of coefficients, and $\epsilon_{i,s}$ are distributed as Type I EV random variables.

After a node becomes *informed* either as an initial seed or through a neighbor, further WOM from others does not impact the likelihood of adoption. Thus, WOM is purely informational rather than persuasive in this baseline. While the baseline model provides a useful benchmark, it leads to the question of whether there are more complex or sophisticated decision processes for communication and adoption, which we examine and model below.

Endorsement or Persuasion

In the endorsement or persuasion model, (termed "complex contagion" by Centola and Macy (2007)), likelihood of adoption varies based on whether WOM is received from more friends. Following Banerjee et al. (2013), the utility of adoption is:

$$u_i(y=1) = \beta_0 + \beta X_i + \lambda F_{it} + \epsilon_{i,1} \tag{3}$$

where F_{it} is the fraction of neighbors who have informed i about microfinance and λ is the endorsement parameter. The utility of non-adoption remains unchanged.

Leader Effects

Leaders selected as seeds by the firm may have unobserved individual characteristics (leadership) that lead to higher probability of WOM relative to non-leaders, over and above their higher degree. Further, firms may have provided specific information to their selected leader seeds, which may make their WOM more effective.[5] To capture such differences, we extend the baseline model to allow for differential probability of WOM for leaders:

$$p_i^s(D) = q_{min}^s + (q_{max}^s - q_{min}^s)\left[\frac{D - D_{min}}{D_{max} - D_{min}}\right] + q_\ell \mathbf{1}[i \in Leaders] \tag{4}$$

Thus, if leaders are especially effective in spreading WOM, we would find the parameter q_ℓ to be positive, whereas a negative value would indicate leaders are less effective than non-leaders.[6]

Nonlinear Effect of Degree

Finally, we allow WOM likelihood to be nonlinear in degree by allowing a quadratic effect, which can also capture potential non-monotonicity with respect to degree.

$$p^s(D) = q_{min}^s + (q_{max}^s - q_{min}^s)\left[\frac{D - D_{min}}{D_{max} - D_{min}}\right] + q_Q\left[\frac{D - D_{min}}{D_{max} - D_{min}}\right]^2 \tag{5}$$

[5] Our model does not distinguish between incidence of WOM and its effectiveness, but so long as both those effects do not change in the counterfactual, the strategy comparisons remain valid.

[6] We note that since all initial seeds are "leaders," it aids the leader fixed effect identification as any impact of leader fixed effect will be stronger in the initial periods and can be therefore identified off the adoption trajectory. More details about the identification of the Leader Fixed Effect is detailed in §EC.2.

where q_Q represents the parameter corresponding to the quadratic term.

We examine a number of models, combining these modeling elements, as summarized in Table 3. Overall, we have 8 specifications. The first 4 models have no endorsement or persuasion effect (denoted by superscript $\mathbf{E=0}$). In $\mathbf{M_1^{E=0}}$, the WOM probability does not depend on degree. This model is identical to the model in Banerjee et al. (2013). In $\mathbf{M_2^{E=0}}$, the WOM probability depends on degree. $\mathbf{M_3^{E=0}}$ incorporates a differential effect for leaders to the prior model specification. $\mathbf{M_4^{E=0}}$ allows for a nonlinear relationship between WOM probability and degree with a quadratic function. The next four models are identical to the first four, but with an endorsement effect (denoted by superscript $\mathbf{E=1}$).

Table 3 Summary of WOM Model Components

	$\mathbf{M_1^{E=0}}$	$\mathbf{M_2^{E=0}}$	$\mathbf{M_3^{E=0}}$	$\mathbf{M_4^{E=0}}$	$\mathbf{M_1^{E=1}}$	$\mathbf{M_2^{E=1}}$	$\mathbf{M_3^{E=1}}$	$\mathbf{M_4^{E=1}}$
Endorsement	×	×	×	×	✓	✓	✓	✓
Degree-dependent WOM	×	✓	✓	✓	×	✓	✓	✓
Leader Differential WOM	×	×	✓	✓	×	×	✓	✓
Nonlinear Effect: WOM and Degree	×	×	×	✓	×	×	×	✓

Estimation

The model estimation proceeds in three steps similar to Banerjee et al. (2013), with specific differences. The estimation procedure is detailed in Supplement §EC.3. Note that we use optimization algorithms for estimation rather than grid search. Here we provide a high level description of the three steps.

Step 1: Adoption Process. We estimate the adoption process parameters β with a logistic regression using the adoption decisions of only the initially seeded individuals based on equation (2).

Step 2: WOM Process. We estimate the WOM process parameters $(q_{min}^{NA}, q_{max}^{NA}, q_{min}^{A}, q_{max}^{A})$ as well as endorsement (λ), leader effect (q_ℓ) and quadratic effect (q_Q) using the method of simulated moments (MSM). We use the same set of cross sectional moments used in Banerjee et al. (2013), supplemented by time series moments, all listed in Table 4. Overall, the moments capture key aspects of diffusion within a network, both globally over the entire network and locally across connections. The first moment is global, matching overall adoption levels in the network. Moments 2-4 are local moments that fit household level adoption as a function of adoption characteristics of their neighbors, and help identify communication probabilities for non-adopters and adopters respectively. Moments 5 and 6 are also local moments in that they capture covariance in adoption between a household and its first and second degree neighbors respectively.[7] Next, we include time

[7] We provide precise specification of the moments and the rationale for using them in §EC.3.

Table 4 List of Moments

#	Description
1.	Proportion of seeds adopting
2.	Proportion of households with no adopting neighbors who have adopted
3.	Proportion of neighbors of adopting seeds who have adopted
4.	Proportion of neighbors of non-adopting seeds who have adopted
5.	Covariance between a household's adoption and average adoption of its first degree neighbors
6.	Covariance between a household's adoption and average adoption of its second degree neighbors
7,8,9	Cumulative adoption upto time $t = 1, 2, 3$ (Time series moments)

series moments that have not been used in Banerjee et al. (2013). Moments 7-9 characterize the temporal trajectory of adoption within villages, which helps us in the identification of the leader fixed effect, as detailed in §EC.2. We detail how each moment informs the estimation of each parameter, i.e. the sensitivity of parameter estimates to each of the cross sectional and time series moments based on the approach of Andrews et al. (2017) in §EC.5.6.

The objective function for the parameter vector θ is defined as in Banerjee et al. (2013):

$$S(\theta) = \left(\frac{1}{S} \sum_{s=1}^{S} \left[m^S(\theta) - m^D \right]' \right) \mathbf{W} \left(\frac{1}{S} \sum_{s=1}^{S} \left[m^S(\theta) - m^D \right] \right) \tag{6}$$

where $m^S(\theta)$ represents the vector of model simulated moments, m^D denotes the vector of data moments. W is the weighing matrix, which can either be estimated with a two–stage approach or be set to be the identity matrix to obtain consistent estimates. The estimator is then defined as:

$$\hat{\theta} = \arg\min_{\theta} S(\theta) \tag{7}$$

Step 3: Standard Errors. We estimate the standard errors using a block-bootstrap resampling procedure of sampling with replacement, treating each network as a block.

4. Results

We first detail the results from the adoption model in Table 5. The number of beds in the household and the rooms per person are negatively associated with adoption probability, and access to a private latrine in the home and beds per person has a negative impact. The estimates are consistent with the idea that microfinance is used by relatively poor households without access to traditional banking services.[8]

For estimation of WOM parameters, we use data from the villages where microfinance was introduced. Table 6 reports the estimates for the 8 WOM models.

[8] We report a variety of adoption models in Section EC.5.5; the results presented here is for the best fitting (lowest AIC) model.

Table 5 Adoption: DV: Microfinance Adoption (1=yes, 0=no).

Variable	Estimate	SE
Constant	-1.210^{***}	(0.322)
Rooms	0.007	(0.085)
Beds	-0.283^{**}	(0.143)
(No) Electricity	0.156	(0.123)
(No) Latrine	0.179^{**}	(0.080)
Rooms per person	-1.023^{***}	(0.392)
Beds per person	1.147^{*}	(0.656)
Log Likelihood	-603.093	
Note:	*p<0.1; **p<0.05; ***p<0.01	

We use the SMM (or Simulated Method of Moments) for model estimation. We use the model specifications $\mathbf{M_2^{E=0}}$ (without endorsement) and $\mathbf{M_2^{E=1}}$ (with endorsement) as our primary specifications for discussion, although the results are provided for all models. We note that our primary results of interest continue to hold qualitatively across all the model specifications. For some models, it might be more relevant to use the temporal variation.[9]

Table 6 Model Estimates

		Model Specification: Estimates (Standard Errors)							
		No Endorsement				*With Endorsement*			
Parameter	Symbol	$\mathbf{M_1^{E=0}}$	$\mathbf{M_2^{E=0}}$	$\mathbf{M_3^{E=0}}$	$\mathbf{M_4^{E=0}}$	$\mathbf{M_1^{E=1}}$	$\mathbf{M_2^{E=1}}$	$\mathbf{M_3^{E=1}}$	$\mathbf{M_4^{E=1}}$
Non-adopter lowest degree	q_{min}^{NA}	0.186	0.177	0.123	0.157	0.016	0.211	0.137	0.136
		(0.075)	(0.0845)	(0.0866)	(0.0865)	(0.111)	(0.124)	(0.0592)	(0.078)
Non-adopter highest degree	q_{max}^{NA}	0.186	0.030	0.116	0.078	0.016	0.054	0.078	0.084
		(0.075)	(0.113)	(0.122)	(0.0958)	(0.111)	(0.109)	(0.0423)	(0.0639)
Adopter lowest degree	q_{min}^{A}	0.297	0.382	0.345	0.389	0.351	0.387	0.375	0.352
		(0.0921)	(0.113)	(0.0876)	(0.0992)	(0.0703)	(0.0846)	(0.0728)	(0.0693)
Adopter highest degree	q_{max}^{A}	0.297	0.305	0.383	0.271	0.351	0.275	0.315	0.249
		(0.0921)	(0.104)	(0.0787)	(0.0859)	(0.0703)	(0.0944)	(0.0831)	(0.0844)
Leader Effect	q_ℓ	–	–	-0.091	0.046	–	–	0.054	0.067
				(0.0907)	(0.0814)			(0.0554)	(0.0612)
Quadratic Effect	q_Q	–	–	–	-0.010	–	–	–	-0.108
					(0.0407)				(0.0396)
Endorsement	λ	–	–	–	–	0.417	0.312	0.134	0.157
						(0.0281)	(0.0419)	(0.0206)	(0.048)

In Models $\mathbf{M_1^{E=0}}$ and $\mathbf{M_1^{E=1}}$, grayed out parameters are not estimated since $q_{min}^s = q_{max}^s$.

We first interpret the parameter estimates of the preferred model specifications $\mathbf{M_2^{E=0}}$ and $\mathbf{M_2^{E=1}}$. We begin with the case of no endorsement. First, the WOM probability for adopters is much higher

[9] We estimate various other models to test different specifications of the data generating process. We mention a few here. First, we allowed a "broadcast process" to consider the effect of an initial village meeting by seeds to communicate to all households. Second, we allowed for a "leader certification" effect to test for an incremental effect of leader endorsement on adoption. Third, we allowed opinion leader seeds to be chosen by occupation, where leader effects are conferred on all members with that occupation. These specifications do not fit the data better than $\mathbf{M_2^{E=0}}$. The counterfactual performance under *all* of the models are provided in §EC.4.3). We also discuss additional model fit metrics in §EC.3.1, evaluating both in-sample and out-of-sample fit.

than that of non-adopters, by an order of magnitude ($q^{A}_{min} \gg q^{NA}_{min}$). Next, we examine degree dependence.

For non-adopters, the WOM probability depends on household degree. For low-degree households, the WOM probability is significant, whereas for high-degree households, it lacks significance ($q^{NA}_{min} > q^{NA}_{max}$). Thus, while low-degree households are open to communicating with their neighbors without adopting the product, their more connected neighbors are not. We note that this communication is a one-on-one process.

For adopters too, high-degree households are less likely to communicate with each of their peers relative to low-degree households ($q^{A}_{max} < q^{A}_{min}$), but the difference is not as great as non-adopters. We note that high-degree households communicate more overall since they have more connections.

From $\mathbf{M_3^{E=0}}$ and $\mathbf{M_4^{E=0}}$, we find no differential effect of leaders; the parameter q_ℓ is small in magnitude and not statistically significant, implying that leaders do not communicate more than others. Similarly, for the quadratic effect, we do not find q_Q to be statistically significant.

Next is the set of models ($E = 1$) with an endorsement effect. We find the effect to be positive and significant across all models. The other parameters are qualitatively the same as for the models without endorsement. Specifically, in all cases, for both non-adopters and adopters, lower-degree households communicate more than higher-degree households. Again, we find the leader effect to be not significant across all of these models. The quadratic effect is marginally significant in this case.

In the supplement §EC.5, we consider a number of other additional models of the data generating process.[10] In all cases, our baseline model indeed does fit best relative to other models, so we use that as the primary specification for the counterfactuals.

5. Counterfactuals

We use counterfactuals to evaluate various seeding strategies based on Friend, Leader, Hybrid and Network Information categories described in Table 7. Within the Friend category, we examine the impact of the Local Friend strategy, which samples on friends of randomly chosen network nodes (households) to obtain seeds. In Leader, we examine both the "(Firm's) Leader" strategy, using the original leaders that were designated for seeding by the microfinance firm, and "Like Leader," which chooses as seeds leader-like nodes who have similar network positions as leaders. We use three dimensions to measure network position: degree, eigenvector and power centrality (Bonacich 1987).

[10] We consider a benchmark where there are only broadcasts (§EC.5.1), but no communication through networks. In §EC.5.2, we consider a model where there is an initial broadcast by seeds, to model an initial village meeting described in Banerjee et al. (2013). We also consider the case where leaders may be present outside the initial seed set (§EC.5.3), and where leaders may have specific certification ability (§EC.5.4). Specifically, we show in Figure EC.3 that there is reasonable probability that seeds chosen by any of our counterfactual strategies overlap with the seeds chosen by the firm.

This helps evaluate whether the impact of seeding is due to the network position or due to the differential impact by individual characteristics of leaders. Hybrid strategies combine the features of sampling on friends along with information on opinion leaders. We examine two different hybrid strategies: choosing a random *Friend of Leader* household (weak hybrid) or choosing a random *Leader Friend of Leader* household (strong hybrid).

We also evaluate two network information benchmarks, Top Degree and Top Diffusion. Unlike the above strategies, these network strategies require complete knowledge on who is connected to whom (network structure) or the degree distribution, i.e. the number of connections of each node. If highly connected nodes are likely to be better in accelerating adoption, the idea is to choose from the set of highly connected nodes in the Top Degree strategy. Top Diffusion is an approach proposed by Banerjee et al. (2013) to identify nodes with high centrality for the purpose of information diffusion. It requires the social network structure (adjacency matrix), but does not require knowledge of the parameters of the diffusion process. Seeds are randomly chosen from the set of top 15% of nodes for both top degree and top diffusion strategies. Further details about the strategies, including informational requirements, are provided in Section EC.4.2.

Table 7 Seeding Strategies and Implementation

Category	**Strategy**	**Implementation Procedure (for each of m seeds)**
Friendship	Local Friend	Select node at random from list. Obtain one randomly chosen friend of node as a seed.
Leader	(Firm's) Leader	Select node from list of leaders
	Like Leader	Select leader node ℓ at random. Select the non-leader node most similar to ℓ in terms of network properties.
Hybrid	Friend of Leader (Weak Hybrid)	Select a random **leader** from list of leaders. Obtain one randomly chosen friend of this leader as a seed.
	Leader Friend of Leader (Strong Hybrid)	Select a random **leader** from list of leaders. Obtain one randomly chosen friend who is also a **leader** to be seed.
Network Information	Top Degree	Select a seed node at random from the list of top (Top 15%) degree (most connected) nodes .
	Top Diffusion	Select a seed node at random from the list of top (Top 15%) diffusion nodes (proposed by Banerjee et al. (2013), and defined in Table EC.4).

We use the estimated parameters from $\mathbf{M_2^{E=1}}$ for the counterfactual simulations below. In the Supplement, we provide a comparison of the counterfactual results of all the different model specifications summarized in Table 3. We set seeding level as a percentage of households in the village, so the number of households seeded varies across villages as a function of village populations. We examine the sensitivity of the results to different seeding levels (0.5%, 1%, 5%) in Section 5.1 below.

Table 8: Comparison of Strategies (5% seeding)

Strategy	**Informed (%)** Mean	SD	**Adopted (%)** Mean	SD	**ΔInformed(%)** over Random	**ΔAdopted(%)** over Random
Random	38.77	42.19	7.56	25.58	—	—
Local	42.98	43.28	8.54	27.12	10.86	13.08
(Firm's) Leader	41.70	43.03	8.25	26.70	7.54	9.13
Like Leader	41.57	42.96	8.20	26.60	7.21	8.48
Hybrid Strategies:						
Friend of Leader	43.75	43.52	8.74	27.42	12.85	15.67
Leader Friend of Leader	40.85	42.73	8.08	26.39	5.37	6.91
Network Information Strategies:						
Top Degree	47.37	44.65	9.62	28.78	22.18	27.28
Top Diffusion	46.78	44.35	9.51	28.56	20.65	25.91

Note: Parameter Estimates from model $\mathbf{M_2^{E=1}}$ used for counterfactuals.

We evaluate seeding effectiveness in terms of proportion of informed households and adoption generated by the seeding strategies as the performance measure. We start with a higher level of seeding (5%) to be conservative, since the advantage of the proposed strategies is greater with fewer seeds. Table 8 reports the aggregate statistics on the proportion of households informed about the microfinance service and the proportion adopting microfinance. The improvement for Local Friend over Random is about 13.1%, while the improvement over Random for Leader is about 9.1%. We also find that the Hybrid strategy Friend of Leader performs the best with a 15.7% improvement over Random, suggesting that the two broad approaches of leveraging network structure (using friendship paradox) and leadership or other demographic characteristics (using Leader indicator) can be combined to achieve higher performance. However, we note that using the Local Friend strategy alone without any information about the network structure or leader information can generate much of this performance benefit. However, there is a risk in applying a strict criterion requiring the friend to also be a leader. Overall, the Local Friend and Hybrid strategies do better than the Leader strategy without data on the full network structure, suggesting that they are viable approaches to seeding WOM with unknown networks.

As we might expect, the network information strategies, which require global knowledge of the network structure lead to much greater adoption than the above strategies. Interestingly, the top degree approach performs better than the diffusion centrality based approach, although the difference is minimal. It's likely that for these relatively small village networks, the overlap in seed sets among the top degree and top diffusion strategies is quite high, hence the similarity. In larger networks, e.g. Twitter we might see larger differences between them. Overall, we observe that the Local Friend strategy is able to obtain about 88.7% (and the weak hybrid achieves about 91%) of the performance of the best informationally demanding network information strategy. This finding characterizes the tradeoff between the amount of information required and the effectiveness of the strategy in driving adoption. In cases where it is impractical to obtain the *relevant* network information, the Local Friend or hybrid strategies could be profitably used as an alternative.

We report the pairwise comparison between strategies Table 9. The Local Friend strategy is better than Random and leads to improved adoption in most of the villages. The Local Friend strategy also outperforms the Leader strategy across a majority of villages. The (Firm's) Leader strategy does worse than Random in about 14% of of the villages. The weak hybrid Friend of Leader strategy is also better than random in about 93% of villages, but the strong hybrid Leader Friend of Leader actually performs worse than random in about 35% of the villages. This implies that *it matters how the hybrid strategy is implemented*, and whether the condition of leadership is required for not just the initial node but also for the nominated friend. The results suggest reduced effectiveness of seeding when we require that the nominated friend also be a leader. Finally, and as expected, the Like Leader strategy is the most similar in performance to the Leader strategy.

Table 9 Pairwise Comparison of Strategies (5% seeding)

	Local	Leader	Like Leader	Friend of Leader	Leader Friend of Leader	Top Degree	Top Diffusion
Random	97.67	86.05	81.40	93.02	65.12	100.00	100.00
Local		32.56	30.23	55.81	27.91	90.70	90.70
(Firm's) Leader			53.49	67.44	41.86	100.00	100.00
Like Leader				74.42	39.53	95.35	97.67
Friend of Leader					20.93	90.70	88.37
Leader Friend of Leader						97.67	97.67
Top Degree							37.21

Note: Number in cell indicates % of villages where **column** strategy achieves higher adoption than **row** strategy.

5.1. How does Extent of Seeding Impact Performance of Strategies?

The idea of word-of-mouth marketing is to choose a small number of seeds to help spread information about a product or service. We summarize in Table 10 how the performance of the seeding strategies varies with the proportion of nodes seeded, at 0.5%, 1%, and 5% of nodes seeded. For full results across all model specifications, see Supplement Section EC.4.3.

We define the performance metric as leverage, in terms of how well a proposed seeding strategy s performs relative to the Random strategy (whose leverage is 1 by definition):

$$Leverage(s) = \frac{\text{\# Households Adopting under Strategy } s}{\text{\# Households Adopting under Random Strategy}}$$

Table 10 Leverage for Counterfactual Strategies

		No Endorsement			With Endorsement		
Strategy	*Seeding at:*	0.500%	1.000%	5.000%	0.500%	1.000%	5.000%
Local		1.140	1.064	1.008	1.315	1.307	1.131
(Firm's) Leader		1.094	1.045	1.005	1.186	1.178	1.091
Like Leader		1.080	1.038	1.002	1.188	1.126	1.085
Hybrid Strategies:							
Friend of Leader		1.145	1.071	1.012	1.355	1.303	1.157
Leader Friend of Leader		1.089	1.049	1.006	1.208	1.216	1.069
Network based Strategies:							
Top Degree		1.241	1.108	1.012	1.713	1.621	1.273
Top Diffusion		1.242	1.100	1.010	1.675	1.618	1.259

The following observations are noteworthy. First, the (Firm's) Leader strategy always outperforms the Random strategy and the Local Friend strategy always outperforms the Leader. Thus, our main results hold across the range of seeding proportions examined for models with and without the endorsement effect. Second, the weak hybrid strategy dominates across most model specifications, whereas the strong hybrid consistently underperforms the Local Friend strategy. Third, Like Leader performs very similar to leader, indicating that performance of the leader strategy is not driven by the differential leader effects, but rather the network position of leaders. Fourth, while the full information network-based strategies perform the best, we also find that the friendship paradox based strategies get upto half of the differential benefit without requiring network structure. Finally, leverage for all strategies decreases as the number of seeds increases, implying that their performance benefit is greater under more constrained circumstances, e.g., when product samples are limited or expensive, or the seeding process requires intensive education or interaction.

6. Conclusion

We estimate a model of network-mediated WOM and product adoption and evaluated the effectiveness of alternative seeding strategies that leverage the friendship paradox. The proposed friendship-paradox based strategies, which are *informationally light* and require little knowledge of network structure significantly improve WOM seeding and product adoption relative to not just random seeding, but also relative to the firm's opinion leader seeding. Specifically, we find about a 15% average improvement with Local Friend seeding in both information spread and adoption compared with Random, and about 5-10% improvement over the Firm's Leader seeding, which is based on pre-selected occupations. Further, Local Friend seeding is typically better than Random across multiple villages with varying network structures, whereas we find that the Leader strategy can be worse than Random in a significant number of village networks. We note that this result is based on expected performance and can vary based on network structure. Further, it should be clear that seeding strategies that use detailed network information can improve adoption relative to the Local Friend strategy, with the tradeoff that much more information is required.

We find that the relative advantage of both Local Friend and hybrid strategies relative to the Random strategy is inversely related to the proportion of nodes seeded. Thus, when we have fewer seeds, these strategies become even more advantageous in expectation. This result is practically useful in cases where the target population is large, and seeding is either financially costly or practically challenging due to time constraints or other operational limitations. The combination of the battery of robustness checks and the monotonic improvement in performance with greater network information suggests that the inability to detect robust gains from using friend seeding strategies in past research using field experiments (e.g., Kim et al. 2015) may be either due to lack of adequate control for network characteristics or inadequate variation in network degree.

As discussed, the data across multiple village networks and explicit knowledge of who were chosen as initial seeds has many advantages for studying the current seeding problem. Yet, common with much research on diffusion, a limitation is that communication is unobserved. While we model communication as a latent process, and show that our key counterfactual claims about friend based seeding are robust to many alternative models of communication, it would be useful to study this problem in setting where communication is directly observable. Further, even though the seeding process is observed, the firm always seeded on "leaders" based on a certain set of occupations. We also provide a novel non-parametric identification argument that does not depend on specific functional forms to estimate the leader fixed effect with our current data where only leaders are used as seeds. Future work may evaluate the effect of leaders on adoption by randomizing seeds with leaders in some villages and non-leaders in others.

It would also be useful to explore whether the benefits of seeding using the friendship paradox generalize to other contexts and different network structures. Microfinance has certain adoption and WOM communication features that may differ from other products. For example, the poor need microfinance and may be more persuaded about its benefits than the rich; the poor may also have systematically different social network structures. Hence it would be valuable to assess whether the benefits of friend-based seeding remain robust for other traditional products. Future research may also evaluate how network structural characteristics may systematically impact friend based seeding effectiveness.

We also suggest some broader issues to explore in future research. First, it would be useful to consider the potential tradeoffs in cost and time in using our informationally light seeding strategies relative to investing in identifying (even limited) network information prior to seeding. Second, rather than use opinion leaders, it may be useful to seed individuals nominated by others as "gossipers" to assess their impact on diffusion and higher overall adoption (Stephen and Lehmann 2016, Banerjee et al. 2014). Finally, it would be useful to consider whether seeding approaches proposed here need to be adapted for highly asymmetric networks, where directional ties are significant (Ben Sliman and Kohli 2018).

References

Andrews I, Gentzkow M, Shapiro JM (2017) Measuring the sensitivity of parameter estimates to estimation moments. *The Quarterly Journal of Economics* 132(4):1553–1592.

Aral S, Van Alstyne M (2011) The diversity-bandwidth trade-off. *American Journal of Sociology* 117(1):90–171.

Banerjee A, Chandrasekhar AG, Duflo E, Jackson MO (2013) The diffusion of microfinance. *Science* 341(6144):1236498.

Banerjee A, Chandrasekhar AG, Duflo E, Jackson MO (2014) Gossip: Identifying central individuals in a social network. Technical report, National Bureau of Economic Research.

Bass FM (1969) A new product growth for model consumer durables. *Management science* 15(5):215–227.

Ben Sliman M, Kohli R (2018) Asymmetric relations and the friendship paradox. *Columbia Business School Research Paper* (18-73).

Berger J, Iyengar R (2013) Communication channels and word of mouth: How the medium shapes the message. *Journal of Consumer Research* 40(3):567–579.

Bonacich P (1987) Power and centrality: A family of measures. *American Journal of Sociology* 92(5):1170–1182, URL `http://dx.doi.org/10.1086/228631`.

Borgatti SP, Mehra A, Brass DJ, Labianca G (2009) Network analysis in the social sciences. *Science* 323(5916):892–895.

Cai J, De Janvry A, Sadoulet E (2015) Social networks and the decision to insure. *American Economic Journal: Applied Economics* 7(2):81–108.

Campbell A, Mayzlin D, Shin J (2017) Managing buzz. *The RAND Journal of Economics* 48(1):203–229.

Centola D (2010) The spread of behavior in an online social network experiment. *science* 329(5996):1194–1197.

Centola D, Macy M (2007) Complex contagions and the weakness of long ties. *American journal of Sociology* 113(3):702–734.

Chen X, van der Lans R, Phan TQ (2017) Uncovering the importance of relationship characteristics in social networks: Implications for seeding strategies. *Journal of Marketing Research* 54(2):187–201.

Chin A, Eckles D, Ugander J (2022) Evaluating stochastic seeding strategies in networks. *Management Science* 68(3):1714–1736.

Christakis NA, Fowler JH (2010) Social network sensors for early detection of contagious outbreaks. *PloS one* 5(9):e12948.

Cohen R, Havlin S, Ben-Avraham D (2003) Efficient immunization strategies for computer networks and populations. *Physical review letters* 91(24):247901.

Domingos P, Richardson M (2001) Mining the network value of customers. *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, 57–66.

Eckles D, Esfandiari H, Mossel E, Rahimian MA (2019) Seeding with costly network information. *Proceedings of the 2019 ACM Conference on Economics and Computation*, 421–422.

Feld SL (1991) Why your friends have more friends than you do. *American Journal of Sociology* 96(6):1464–1477.

Godes D, Mayzlin D (2009) Firm-created word-of-mouth communication: Evidence from a field test. *Marketing Science* 28(4):721–739.

Goldenberg J, Han S, Lehmann DR, Hong JW (2009) The role of hubs in the adoption process. *Journal of Marketing* 73(2):1–13.

Guadagni PM, Little JD (1983) A logit model of brand choice calibrated on scanner data. *Marketing science* 2(3):203–238.

Iyengar R, Van den Bulte C, Valente TW (2011) Opinion leadership and social contagion in new product diffusion. *Marketing Science* 30(2):195–212.

Katona Z, Zubcsek PP, Sarvary M (2011) Network effects and personal influences: The diffusion of an online social network. *Journal of marketing research* 48(3):425–443.

Kempe D, Kleinberg J, Tardos É (2003) Maximizing the spread of influence through a social network. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 137–146.

Kim DA, Hwong AR, Stafford D, Hughes DA, O'Malley AJ, Fowler JH, Christakis NA (2015) Social network targeting to maximise population behaviour change: a cluster randomised controlled trial. *The Lancet* 386(9989):145–153.

King CW, Summers JO (1970) Overlap of opinion leadership across consumer product categories. *Journal of Marketing Research* 43–50.

Kumar V, Krackhardt D, Feld S (2024) On the friendship aradox and inversity: A network property with applications to privacy-sensitive network interventions. *Proceedings of the National Academy of Sciences* 121(30).

Libai B, Muller E, Peres R (2013) Decomposing the value of word-of-mouth seeding programs: Acceleration versus expansion. *Journal of Marketing Research* 50(2):161–176.

Paluck EL, Shepherd H, Aronow PM (2016) Changing climates of conflict: A social network experiment in 56 schools. *Proceedings of the National Academy of Sciences* 113(3):566–571.

Rubin DB (2005) Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association* 100(469):322–331.

Stark TH (2018) Collecting social network data. *The Palgrave Handbook of Survey Research*, 241–254 (Springer).

Stephen AT, Lehmann DR (2016) How word-of-mouth transmission encouragement affects consumers' transmission decisions, receiver selection, and diffusion speed. *International Journal of Research in Marketing* 33(4):755–766.

Tucker C (2008) Identifying formal and informal influence in technology adoption with network externalities. *Management Science* 54(12):2024–2038.

Van den Bulte C, Lilien GL (2001) Medical innovation revisited: Social contagion versus marketing effort. *American journal of sociology* 106(5):1409–1435.

Wilder B, Immorlica N, Rice E, Tambe M (2018) Influence maximization with unknown network structure. *Artificial Intelligence and Social Work* 77.

Yoganarasimhan H (2012) Impact of social network structure on content propagation: A study using youtube data. *Quantitative Marketing and Economics* 10(1):111–150.

Electronic Companion Supplement

EC.1. Mathematical Notation

In Table EC.1 below, we define the terms used in networks. These terms are helpful when we define network properties and in the moment conditions. We illustrate a village network in **??**, with the black nodes indicating leaders.

Table EC.1 Table of Notation

Characteristic	Description	Definition
Nodes Degree	Number of connections (edges) of i	D_i
Edge	Connection between nodes i and j	$e_{ij} \in \{0,1\}$
Adjacency (Edge) Matrix	Connection between nodes i and j	$\mathbf{E}, E_{i,j} \in \{0,1\}$
Node Set	Set of all N nodes in Network	$\mathcal{V} = \{1,2,\ldots,N\}$
Edge Set	Set of all edges in Network	$E = \{(i,j) : e_{ij} = 1\}$
Network Edge Count	Number of undirected connections	$e = \sum_{i\in\mathcal{V},j>i} e_{ij}$
Seeds	Set of all nodes chosen as seeds	$\mathcal{S}$
Adopters	Set of all nodes which have adopted	$\mathcal{A}$
Reachable Set	Nodes with adoption status $s \in \{A, NA\}$ reachable from i in k steps	$E_i^s(k)$
Proportion of adopting neighbors	Fraction of adopting nodes among those reachable from node i in k steps	$z_i(k) = \dfrac{\lvert E_i^A(k)\rvert}{\lvert E_i^A(k)\rvert + \lvert E_i^{NA}(k)\rvert}$
Vector of above	Vector of adopting proportion of neighbors for each node	$z(k) = [z_1(k), \ldots, z_N(k)]$
Minimum Distance	Distance of Shortest Path between i and j	$\delta_{ij} = \min_k s.t. E^k_{(i,j)} > 0$

EC.2. Identification of Leader Fixed Effect

We demonstrate below that the WOM communication probability for leaders q_L is separately identified from the word of mouth communication probability q for non-leaders. While the argument itself is non-parametric and does not rely on a specific functional form, our demonstration model uses a simple parametric representation consistent with the paper. For this argument, we choose to add a leader fixed effect to the simplest model (Model 1) from the paper.

Suppose we had only static adoption data, we would not be able to identify the fixed effect. However, (i) the availability of time series aggregate adoption data and (ii) the presence of multiple networks allows us to identify the fixed "leader" effect.

First, we note that using only the final adoption levels *will not allow* leader fixed effect q_L to be identified separately from just overall propensity to communicate q. Increases in each of these parameters will result in higher final adoption levels in a network. It is straightforward to see that a relatively low level of q in conjunction with a high level of q_L might result in the same adoption level as a high level of q and a low level of q_L.

However, the curvature of the adoption trajectory over time provides variation that permits identification of the leader effect q_L separately from q. Intuitively, if q_L is higher, the adoption trajectory shows a steeper increase in the earlier periods, since only leaders are communicating initially, and only in subsequent periods do non-leaders communicate. Thus, the proportion of communication attributable to leaders is highest at the beginning and decreasing over time. Thus, the impact of a higher q_L will be greatest in earlier periods as opposed to later periods. In contrast, the impact of a higher q will be lower in the initial periods, since few non-leaders are informed, and it has proportionally greater impact on adoption in later periods.

While the above argument is non-parametric and does not rely on specific functional forms for identification, for the purpose of illustration, we use a parametric model below.

Simplified Model

We provide a highly simplified version of the model similar to Model 1 in the paper, for the specific purpose of examining identification and making the required variation transparent. The main features of this model are:

1. A few leader nodes are informed initially (similar to the main model).
2. In each period, each informed node communicates with probability (that depends on the node's leadership status). Thus, non-leaders communicate with probability q and leaders communicate with probability q_L with each of its neighbors. Note that in this simplified model, adoption status *does not* impact communication probability.[11]
3. When nodes are newly informed, they have the ability to adopt a product with probability $\gamma = 0.2$. (We don't have any covariates impacting adoption here, unlike in the main model, and do not require the variation obtainable from these covariates).

The WOM communication probability for node i is specified as:

$$p_i = \begin{cases} q, & \text{if } i \text{ is not a leader} \\ q_L = q + q_\ell, & \text{if } i \text{ is a leader} \end{cases}$$

where q_ℓ is the leader fixed effect. Recall that the leader fixed effect is the difference between the WOM communication probabilities of leaders and non-leaders.

We demonstrate in Figure EC.1 precisely the variation that is required for this identification. There are several sources of possible variation in the network data. First, we observe that both adoption trajectories for (a) $q = 0.01, q_\ell = 0.08, q_L = q + q_\ell = 0.09$ (red curve) and (b) $q = 0.13, q_\ell = -0.12, q_L = q + q_\ell = 0.01$ (green curve) end up after $T = 5$ periods at the same overall adoption

[11] Even though this additional variation based on adoption status might prove useful as a separate source of identification, our identification argument does not require it.

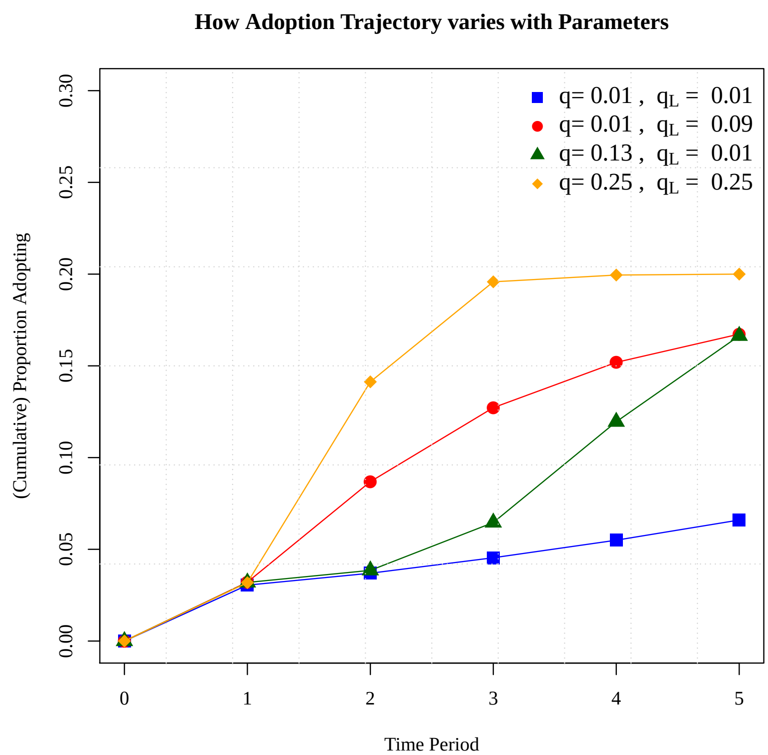

Figure EC.1 Identification and Adoption Time Trajectory

level, i.e. 0.165 or 16.5%. Thus, just having the final adoption levels, it would not be possible to separately identify q and q_L.

However, their adoptions differ in their time trajectories. For (a) (red curve), with a higher leader fixed effect $q_L = 0.09$, we see the **early period trajectory is steeper** than the case (b) (green curve). On the other hand, with (b), the later period trajectory is steeper than in (a).

In general, for different combinations of (q, q_L) that *obtain the same level of final overall adoption*, the area under the adoption trajectory curve will be greater for combinations of (q, q_L) with higher levels of q_L and lower levels of q.

Does exclusive seeding by leaders help or hinder identification of leader fixed effect? There are two reasons why leader seeding (in contrast to random seeding) is helpful to answering our research question.

First, it may appear that our context in which the firm exclusively used leaders by the for initial seeding makes it more challenging to separately identify the leader fixed effect. But in fact, our explanation above should clarify that this exclusive use of leaders for initial seeding aids identification of the leader fixed effect and allows us to disentangle q_L and q. This is because the exclusive use of leaders for initial seeding guarantees that a higher leader fixed effect will increase the earlier adoption trajectory relative to later. Therefore if the seeding had been random, it would not be feasible to separately identify the effects as one cannot use this identification argument.

Second, leader seeding avoids a specific kind of bias in leader effects. Suppose we only have random seeding, but there are leaders present in the data. If leaders have different (higher or lower) degree on average than others, and if they have differential communication, it would not be possible to identify any leader specific communication effect. For instance, if Leaders have higher degree, the Local friend strategy could result in more leaders on average. The counterfactual results would then be biased to find lower effects for the Local strategy than would be obtained in reality. Due to leader seeding in our data, we can identify and characterize the leader fixed effect (separately from non-leaders), and thus avoid this potential bias.

EC.3. Model Details and Estimation

First, we detail the estimation of the adoption process, followed by the WOM communication process, and finally detail the block bootstrap to obtain standard errors. We simulated $N_{sim} = 150$ diffusion paths with seeds chosen stochastically corresponding to each seeding level and using each of the seeding strategies. The reported WOM communication parameters are based on the average of the simulated diffusion paths.

Adoption Process

The adoption parameter vector is $\beta = (\beta_0, \ldots, \beta_6)$. The logistic regression specification for the adoption decision follows from the utility specification. The log likelihood for household i is $l_i(\beta|X_i)$ and for all households in the network is $l(\beta|X)$

$$l(\beta|X) = \sum_{i=1}^{N} l_i(\beta|X_i) = \sum_{i=1}^{N} \log P(y_i = 1|X_i) = \sum_{i=1}^{N} \log \left[\frac{\exp(\beta X_i)}{1 + \exp(\beta X_i)} \right] \tag{EC.1}$$

(EC.2)

The adoption process is estimated by maximum likelihood estimation.

WOM Process

Given adoption parameters β, the WOM process is simulated separately for each village network. We track two states for each household: its information state and its adoption state. The information states are *uninformed* (U) and *informed* (I), whereas the adoption states are *Not-adopted* (NA) and *Adopted (A)*. Both the Informed and Adopted states are absorbing states, during which nodes can communicate with their neighbors.

An *informed household* with adoption status $s \in \{NA, A\}$ (i.e. non-adopting or adopting) will communicate with any of its neighbors in a single time period with probability $p^s(D)$. This is a dynamically evolving process over time, and depends on the informed status of all households in the network. We have formalized these details further below using additional notation. Let $p^s(D)$ be the probability that an *informed household* with adoption status at the beginning of time t $s^j(t) \in \{U, NA, A\}$ (i.e. uninformed, non-adopting or adopting) of degree D will communicate with any of its neighbors *in a single time period.* Uninformed households do not communicate. During time period t, an uninformed household i becomes informed if it receives a communication from any of its network neighbors $\mathcal{N}_i$. This event happens with probability $p_{it} = 1 - \Pi_{j\in\mathcal{N}_i}(1 - p^{s_j(t)}(D_j))$.

The WOM process for each of the N_{sim} simulations begins with Step (0) and then proceeds through Steps (1)-(3) for each time period.

(0) Each household (node) in the network is initially in an uninformed (U) information state. In initial period $t = 0$, the seed nodes are chosen in each network based on the seeding strategy. In the actual data, the seed nodes in each village were chosen based on the opinion leadership criterion. In the counterfactual scenarios, seed nodes are chosen based on an alternative strategy (Random, Local Friend etc.). In all cases, the information state of the seed nodes changes from Uninformed (U) $\longrightarrow$ Informed (I).

The following process (1) – (3) process then takes place in each period $t \in \{1, 2, \ldots, T_v\}$ for village v.[12]

(1) Each household that has become informed decides whether to adopt.

(2) Then, an informed household can probabilistically communicate about the microfinance product with each of its network neighbors. The probability of such communication $p^s(D)$ may depend on both its degree D, i.e. the number of neighbors the informed household has, as well as the adoption status $s \in \{A, NA\}$ of the informed household. We separate out the probabilities $p^{NA}(D)$ and $p^A(D)$ as detailed in §3 of the paper.

(3) When this communication takes place, each neighbor receiving information changes its information state from Uninformed (U) $\longrightarrow$ Informed (I). If the neighbor node has already been informed earlier, there is no change in its state.

For each simulation and for each village v, we compute 6 cross-sectional moments according to Table EC.2 at the end of T_v periods of simulation, and 3 time series moments. Thus, for the 43 villages with microfinance adoption, we have $N_{moments} = 9 \times 43 = 301$ moments across the villages. We then minimize the MSM objective function $S(\theta)$ detailed in equation (7) from §3 in the $[0,1]^K$ region to obtain the probability parameter estimates presented in Table 6 in §4 of the paper. For the MSM objective, we start with the initial weight matrix set to the identity matrix to obtain consistent estimates. Since we obtain standard errors through bootstrap, a consistent estimator is all that is needed.

[12] The number of time periods varies across villages in the data, with a mean of 6.5 and SD of 1.83.

Standard Errors with Bootstrap Estimation

We obtain standard errors for the communication probability parameters using a bootstrap procedure detailed below. First, we obtain $N_R = 2,000$ draws using a random grid for the communication probability vector $\theta = (q_0^{NA}, q_1^{NA}, q_0^A, q_1^A) \in [0,1]^4$. The parameter is characterized appropriately based on the model specification.

We proceed through Steps (a) – (c) below for each of the N_{sim} draws to obtain moments for each village v.

(a): We choose seeds corresponding to the Leader strategy used in the data.

(b): We compute the simulated WOM Process detailed above for T_v periods for each draw of the parameter vector θ.

(c): We use the cross-section and time series adoption status data to compute the moments detailed in Table EC.2 separately for each village.

Compute $B = 10,000$ bootstrap estimates using the moments obtained from the samples above. For $b = 1, 2, \ldots, B$ do Steps (d) – (f) below.

(d): Resample with replacement from moments from the set of villages showing microfinance activity.

(e): Compute the objective function with the resampled moments at each of the N_R points evaluated above.

(f): Choose the parameter vector with the minimum objective as the estimate $\beta^{(b)}$ to be used in the bootstrap.

The distribution of $\beta^{(b)}$, with $b = 1, 2, \ldots, B$ provides the bootstrap estimate distribution for computing standard errors.

Moment Conditions for Estimation

In this section, we describe the rationales for the moments listed in Table EC.2 that we use in our estimation. The required mathematical notation is defined in §EC.1.

In general, all moments are informative in the estimation of all parameters. However, the connections between some moments and parameters are more intuitive. The time series moments, and more generally the temporal trajectory are especially important for identification when there are differential effects for leaders. We describe the moments and the obvious associated links with parameters below.

First, we detail the cross-sectional moments MC1 to MC6. (MC1) is the proportion of seeds that have adopted. Since the seeds are guaranteed to be informed outside the WOM process, this allows us to estimate the parameters impacting adoption probability without relying on the communication process. In contrast, (MC2) is the proportion of households with no adopting