Submitted to *Information Systems Research*
manuscript

Long-range social influence in phone communication networks on offline adoption decisions

Yan Leng
McCombs School of Business, The University of Texas at Austin (yan.leng@mccombs.utexas.edu)

Xiaowen Dong
Department of Engineering Science, University of Oxford (xdong@robots.ox.ac.uk)

Esteban Moro
Department of Mathematics at Universidad Carlos III de Madrid and Media Lab at Massachusetts Institute of Technology (emoro@mit.edu)

Alex Pentland
Media Lab, Massachusetts Institute of Technology (pentland@mit.edu)

We utilize high-resolution mobile phone data with geolocation information and propose a novel technical framework to study how social influence propagates within a phone communication network and affects the offline decision to attend a performance event. Our fine-grained data is based on the universe of phone calls made in a European country between January and July 2016. We isolate social influence from observed and latent homophily by taking advantage of the rich spatial-temporal information and the social interactions available from the longitudinal behavioral data. We find that influence stemming from phone communication is significant and persists up to four degrees of separation in the communication network. Building on this finding, we introduce a new "influence" centrality measure that captures the empirical pattern of influence decay over successive connections. A validation test shows that the average influence centrality of the adopters at the beginning of each observational period can strongly predict the number of eventual adopters and has a stronger predictive power than other prevailing centrality measures such as the eigenvector centrality and state-of-the-art measures such as diffusion centrality. Our centrality measure can be used to improve optimal seeding strategies in contexts with influence over phone calls, such as targeted or viral marketing campaigns. Finally, we quantitatively demonstrate how raising the communication probability over each connection, as well as the number of initial seeds, can significantly amplify the expected adoption in the network and raise net revenue after taking into account the cost of these interventions.

Key words: Phone communication; Social influence; Behavioral matching; Long-range effect; Influence centrality

History:

1. Introduction

Social influence, mediated through various communication channels, plays an important role in influencing consumer behavior (Sundararajan et al. 2013, Mobius and Rosenblat 2014, Banerjee

et al. 2013). According to the media richness theory, different communication media vary in their ability to enable communication and information exchange and also in their ease of use (Dennis and Kinney 1998). Phone calls are an especially important communication channel through which social influence takes place; it offers a comparably high level of media richness as offline channels—thereby facilitating information flows and social influence—while also maintaining high ease of use (i.e., low cost) as online channels (see Table 1).

Table 1 Comparison of three communication channels in terms of media richness and ease of use.

		Phone	**Offline**	**Online**
Media richness (Dennis and Kinney 1998)	**Intimacy of relationship**	High	High	Low
	Non-verbal cues (e.g., emotion)	Yes	Yes	No
	Synchronicity	Yes	Yes	No
Ease of use	**Time cost**	Low	High	Low
	Location constraint	No	Yes	No
	Penetration rate	High	N/A	Medium

Despite phone communication's prevalence and importance, there exists limited studies in the information systems (IS) and the social influence literature that studies explicitly how phone communications mediate social influence[1]. A quantitative framework that credibly obtains estimates on social influence from phone communications is important in business and management settings because it can inform personalized mobile targeting (Ghose et al. 2019, Zhang et al. 2019) and viral marketing (Aral and Walker 2011) applications, which are increasingly commonly used in practice.

We fill this gap in the literature by developing a novel framework to quantitatively estimate the pattern of social influence via phone communications. First, we utilize high-resolution mobile phone data with geolocation information (call detail records, or CDRs) and propose a technical framework to study how social influence propagates within a phone communication network and affects the offline decision to attend a performance event. Our fine-grained data is based on the universe of phone calls made in a European country between January and July 2016. Our data contain the entire history of each mobile phone user's phone calls and geolocations (registered by the nearest cell towers). We measure network connections based on phone calls: two individuals are connected if there have been calls between them; additionally, we exploit the temporal dimension of phone calls to construct the sequence of dynamic communications over time—what we refer

[1] While studies have investigated the role of mobile phones in information exchange and the emergence of multiplex communication networks (Matous et al. 2014), they have not specifically examined the role of phone communication in social influence. In contrast, several studies investigate social influence through phone call data as a proxy for social connections (e.g., de Matos et al. (2014), Hu et al. (2019), Belo and Ferreira (2022)) or exposure to ring-back tones they hear (e.g., Ma et al. (2015), Zhang et al. (2018)), but these studies do not directly analyze social influence that is mediated through phone calls.

to as *communication cascades*. We measure adoption behavior using geolocation information—specifically, attending an offline performance event that occurred recurrently during July 2016. Social influence in our context is defined as the process by which a user, who has attended the event, influences another user to subsequently attend the same event via direct phone calls or indirectly through more than one degree of phone call separation in the communication cascades. We exploit the twenty-two occurrences of the performance: after each performance, we measure the impact of past attendees' phone calls on subsequent attendance by individuals receiving the call. By exploiting the temporal variation in the phone communication network and the repeated event occurrence, we construct a rolling window of "treated" individuals—those who have received calls from past attendees—and we estimate social influence based on comparing the behavior of these treated individuals and other non-treated individuals.

A key difficulty for credibly estimating social influence based on behavioral and network data is to control for homophily: two connected individuals may have correlated behaviors either because they have correlated preferences (homophily) or because one's behavior affects the other's (influence). The presence of homophily implies that the assignment of treatment is non-random. In this work, we introduce a novel technical framework to address this key challenge, by utilizing the rich mobility and network information in phone communication data. For observed homophily, we follow Eckles and Bakshy (2020) to adjust for behaviors highly relevant to the decision of interest using individuals' revealed preferences (i.e., mobility history). While observed homophily can be controlled for, latent homophily is driven by unobserved factors and is generally difficult to purge. We address this in two ways. First, we follow McFowland III and Shalizi (2021) and exploit the information contained in a *historical social network* (different from communication cascades) that captures the user's past network connection history, i.e., two individuals are connected in this network if reciprocal calls[2] exist between them in the month prior to the event. To the extent that any user-level characteristics simultaneously affect behavior and predict network connections—even if these characteristics are not observed—we can utilize the historical social network to control for such characteristics and thereby control for latent homophily and isolate social influence. Because network data is high-dimensional, to operationalize this strategy, we extract from the network data a low-dimensional, latent-feature representation of each individual using an efficient network representation learning approach, *node2vec* (Grover and Leskovec 2016), based on the user's historical social network. We then use the latent positions of each user as covariates to control for latent homophily. Second, we follow Belo and Ferreira (2022) to use the eventual adoption decisions of one's connections as a proxy for the focal individual's unobserved preferences toward the adoption decision. Controlling for such information, therefore, also helps control latent homophily.

[2] Reciprocity helps to reduce the possibility of including spam calls.

We use observed and latent homophily to create a matched control unit for each treated user, and we implement a matching-based difference-in-differences strategy to estimate social influence. We find that the influence stemming from phone communication is significant: a direct phone call with a past attendee raises the likelihood of future performance attendance by 87.61%, relative to the base adoption likelihood of 0.0098. The effect transmits over the network to second-degree neighbors of the past attendees and increases their likelihood of future attendance by 68.65%. Overall, we find that the effect persists up to four degrees of separation in the communication network: even being indirectly connected with a past attendee via a network path of length four significantly raises one's likelihood of future attendance.

Building on our empirical finding, we develop a new *influence centrality* measure that captures the empirical pattern of influence decay over successive connections. A node's influence centrality captures the expected increase in adoption in the network. Different from the standard Katz centrality, where the indirect influence decays exponentially at a common rate across successive connections, our influence centrality takes into account the empirically estimated separation-specific rates of decay and thus could be more relevant in empirical settings for increasing expected adoptions in the network. Our notion of influence centrality is useful for applications involving optimal seeding strategies in network contexts where social influence is present. We conduct two exercises to demonstrate this point. First, we conduct an in-sample test and show that the average influence centrality of those who have previously attended the event can significantly predict the number of eventual adopters. It has stronger predictive power than analogous measures constructed based on other prevailing centralities, such as the diffusion and Katz centralities. Second, we quantitatively demonstrate, in a simulated environment where high-centrality nodes are targeted to be the initial adopters. Raising the communication probability over each connection and the number of initial seeds can significantly amplify the overall expected adoption and may be desirable despite the cost of these interventions. This exercise can inform optimal seeding in viral and targeted marketing campaigns.

We summarize our contributions as follows. First, we develop a novel framework to estimate social influence, where we exploit the spatial-temporal information to control for observed and latent homophily using a matching-based difference-in-differences strategy. Second, we apply this framework using high-resolution call detail records (CDRs) and provide credible estimates of direct and long-range social influence over phone calls on offline behavior. We find social influence stemming from phone communications to be significant and persist up to four degrees of separation. Finally, we propose influence centrality, which is designed to capture the empirical pattern of influence decay over successive connections. The measure can be used to improve optimal seeding strategies in network contexts with social influence, such as targeted or viral marketing campaigns.

We quantitatively demonstrate how raising the communication probability over each connection and the number of initial seeds can significantly amplify the expected adoption in the network and may be desirable despite the cost of these interventions.

2. Literature review

2.1. Social influence identification in networks

Identifying social influence effect in observational studies can be challenging from a methodological standpoint (Shalizi and Thomas 2011). The reason is that individual decision-making in a social network can be affected by several factors, including homophily, exogenous factors, and social influence (Manski 1993). Various empirical approaches for studying social influence based on observational data have been adopted in the IS and social influence literature. First, in an instrumental variable approach, a standard instrumental variable might be the behavior of two-degree neighbors who are not neighbors of the focus user (de Matos et al. 2014). Next, propensity score matching has been applied in many empirical settings, including studies of the effects of instant messaging on the adoption of mobile applications (Aral et al. 2009), favoriting behavior on the songs individuals listen to (Dewan et al. 2017), and online content contributions (Rishika and Ramaprasad 2019). Finally, structural modeling, such as hierarchical Bayesian modeling, has been used to study the effects of social influence and latent homophily on dynamic and repeated consumer purchases (Ma et al. 2015).

The key to separating social influence from homophily and other exogenous variables lies in the use of effective control variables. Recent studies on social influence in statistics and IS provide promising solutions to partially address this issue using rich behavioral and network data that have become increasingly available on digital platforms. First, it has been shown that adjusting for high-dimensional behavioral data relevant to adoption behavior can remove the majority of the estimation (selection) bias, leading to statistically indistinguishable results from those obtained via a randomized experiment (Eckles and Bakshy 2020). Second, sufficient conditions for unbiased and consistent estimates of social influence have been established theoretically when controlling for estimated locations in a latent space, based on certain network generation processes (McFowland III and Shalizi 2021). Third, it has also been shown that eventual adoption decisions of neighbors may serve as a proxy for latent user preferences (Belo and Ferreira 2022). Inspired by these studies and exploiting the rich spatial-temporal information in CDRs, we propose innovative ways to operationalize and account for observed homophily with mobility data, as well as latent homophily with latent positions learned from the social network and neighbors' eventual adoption decisions.

2.2. Social influence using mobile phone data

An abundance of literature studies how online behaviors diffuse through IT-enabled social networks, as reviewed in Sundararajan et al. (2013). This literature has important managerial and strategic implications for online marketing and platform designs. However, because of the differences in the nature of the communication studied (summarized in Table 1), findings on online word of mouth (WOM) may not provide direct guidance on the situation of phone communication and the diffusion of influence through this different medium. Despite its importance as a medium of information exchange, how phone communications mediate social influence has scarcely been studied in the IS and social influence literature.

Phone call data have been used in studying the purchase of caller ring-back tones (Ma et al. 2015, Zhang et al. 2018), switching of mobile carriers (Hu et al. 2019) and the use of phone plans for unlimited calls (Belo and Ferreira 2022), and adoption of new mobile phone models (de Matos et al. 2014). Our study differs from these papers in several aspects. First, Hu et al. (2019) and de Matos et al. (2014) only use phone call relationships to construct a proxy of social networks. Our work relates to theirs, but we focus on the social influence that travels through phone communications. Specifically, our study utilizes significantly richer panel data—for each phone call, we observe the time stamp and location of both the call initiator and receiver, and our analysis is designed to fully utilize the richness of the data, both spatial and temporal. By contrast, only static networks are constructed based on collapsed, cross-sectional data using phone calls aggregated over a period of time (9 months in Hu et al. (2019) and 11 months in de Matos et al. (2014)). It is precisely because of these differences and the additional details we can observe that we can estimate social influence mediated via phone calls. Second, while Ma et al. (2015) and Zhang et al. (2018) study the influence of exposure to caller ring-back tones resulting from phone calls, they do not explicitly examine how phone conversations mediate social influence. In contrast, our study focuses specifically on how social influence spreads through phone conversations, leveraging their unique media richness compared to other communication channels as discussed in Table 1. By examining this specific channel, our study provides new insights into the ways in which social influence operates and the behaviors it influences, beyond the scope of previous studies that focused solely on caller ring-back tones. Third, all these studies focus on adoption decisions directly related to phone use, while our focus is on offline adoption behavior, which is arguably a more general type of behavior that may be influenced through phone communication. Offline decisions are common and of obvious interest in marketing applications; indeed, many important behaviors pertain to offline settings and entail a certain degree of effort, such as voter turnout (Bond et al. 2012), receiving immunizations (Banerjee et al. 2019), and healthy habits (Christakis and Fowler 2013). Our paper extends this IS and social influence literature and constitutes one of the first studies investigating the effect of social influence through phone calls on an offline adoption decision using the large-scale CDR data.

2.3. Indirect social influence in network environment

In the study of social influence in a network environment, one may consider both the direct influence (i.e., influence on one's immediate neighbors in the network) and the indirect influence (i.e., influence beyond one's immediate neighbors) on adoption decisions. The IS literature predominantly examines the direct influence on different types of technology adoption decisions (Aral et al. 2009, Katona et al. 2011, Rishika and Ramaprasad 2019, Dewan et al. 2017, de Matos et al. 2014). However, studies have found only limited and inconsistent evidence that positive influence may extend beyond direct neighbors in the social network. On the one hand, indirect influence was initially found to be more effective than direct influence in medical innovation (Burt 1987, den Bulte and Lilien 2001). More recently, it was shown that online messages play a role in political mobilization and have an effect on two-degree neighbors in the Facebook friendship network (Bond et al. 2012). Similarly, it was shown that indirect (i.e., two-hop) neighbors, like direct neighbors, exert influence in the context of caller ring-back tone adoption decisions (Zhang et al. 2018). On the other hand, this effect can be negative beyond immediate neighbors, e.g., it was shown that the likelihood of individuals taking deworming was reduced if their direct first-order or indirect second-order social contacts were exposed to it (Kremer and Miguel 2007). Contrasting with these views, several studies show that influence was restricted to immediate neighbors in the social network, such as the case of cooperative behavior in local public goods games (S. Suri 2011) or decisions to get vaccination against influenza (Mobius and Rosenblat 2014, Rao et al. 2017).

Our paper aims to extend and enrich this literature on indirect social influence, and examine for the first time the potential cascading effect of influence through the medium of phone communication. This research question is interesting to study in the phone communication medium for two reasons. On the one hand, because of the personal and persuasive nature of phone communication and the ease with which it is established, social influence via phone calls may extend beyond immediate neighbors in the communication network. For example, an individual who has heard a colleague's positive impression of an event that she attended may be eager further to pass on that impression to her own neighbors. On the other hand, phone calls are a form of synchronous oral communication that allows for less time for contemplation and fewer opportunity for selective self-presentation, which may reduce the impact of the communication on behavior (Berger and Iyengar 2013) and hence its propagation in the network. Addressing this research question has direct implications for IS and marketing research. Indeed, if the effect of phone communication is restricted to immediate neighbors, then businesses should mainly target individuals who have many direct connections in their communication network. Otherwise, businesses should instead consider targeting individuals who have many indirect neighbors to capitalize on the cascading effect of influence.

2.4. Network centrality and application in seeding in social networks

There is a rich body of literature on centrality measures of nodes in a network (Bloch et al. 2019, Leng et al. 2020). Intuitively, centrality quantifies how "central" a node is, according to different criteria, and therefore captures how important the node is in the network. This structural importance of nodes is a crucial concept in network science with many applications; in particular, it has been widely applied to identify key individuals in social networks, bottleneck locations in infrastructure networks, or super-spreaders of epidemics (Newman 2018).

Defining node centrality (i.e., the importance of nodes) for a certain application has generated much theoretical interest in IS and network science research (Sundararajan et al. 2013). In business research, one notable application of network centrality is seeding. The key idea is to target a small subset of individuals in the network for intervention, leading to a maximal spread of information or adoption decisions. In contrast to approaches based on influence or utility maximization (Kempe et al. 2003, Li et al. 2018, Dou et al. 2013, Mallipeddi et al. 2022), which often involve computationally intensive procedures, especially in large networks, centrality-based seeding is computationally efficient and easy to interpret. Different centralities have been proven effective in a variety of contexts, for example, betweenness centrality (Jackson 2008), Katz-Bonacich centrality (Ballester et al. 2006), diffusion centrality (Banerjee et al. 2013), eigenvector centrality (Golub and Jackson 2010), and degree centrality (Jackson 2019). A commonality of these centrality measures is the focus on the context of information spreading or diffusion in the network, where the implication for adoption remains implicit. In contrast, in this work, we propose influence centrality that is designed to capture the empirical pattern of influence decay over successive connections. Thus, influence centrality is directly related to social influence and thereby contributes more explicitly to increasing overall expected adoption. In addition, different from the standard Katz/eigenvector centrality or the state-of-the-art diffusion centrality, where the indirect influence decays exponentially at a common rate across successive connections, our influence centrality establishes different weights for neighbors at different geodesic distances from the focal user using empirically-estimated separation-specific rates of decay. This is another notable difference from existing centrality measures which we will discuss from a technical perspective in Section 5.

3. Technical framework

We develop a technical framework for studying the social influence that happens through phone communications and its effect on offline decisions using CDRs. The proposed technical framework consists of three steps: (1) identify the adoption decision based on the visitation or attendance inferred from the phone user's mobility; (2) use phone data to construct communication cascades, identifying individuals who have direct phone calls or are indirectly connected with adopters;

and (3) isolate social influence via phone calls from homophily, the measurement of the latter is operationalized using the mobility and phone call data.

3.1. Setting

We consider a large-scale mobile phone data set, call detail records (CDRs), collected in a small European country. The data set includes individual phone usage records (i.e., phone calls, text messages, and internet activities using the data service), as well as the location of the cell tower with which each record was associated. The mobile carrier we collaborate with is the only network provider in the country, meaning that the activity of all individuals who have been connected to any cell tower in the country has been recorded. The data set covers seven months, from January 2016 to July 2016. The cultural event under consideration took place 22 times in July 2016 on most weekdays (plus a few weekend days). The historical data from January to June 2016 are used to collect user behavior indicators as appropriate controls, which we discuss in more detail in Section 3.2.1. Table 2 shows the statistics indicating daily average phone use for every individual in the mobile phone data set used in this study.

We consider the offline adoption behavior of attending an international cultural performance in the country. Although the performance venue was located in a city park, the event took place late in the evening, which reduced the chance of passers-by being mistakenly identified as adopters. Three cell towers are located within a 500-meter distance and cover an area of radius of about 0.25 kilometers to 1.5 kilometers. We assume that the individuals who were connected to any of these three cell towers during the event period (with a buffer time of 30 minutes before and after) are the ones who attended the event (i.e., made the adoption decision). We discuss the statistical implication of violating this assumption at the end of this section. For notational convenience, we call the cultural performance the "product" and the attendees "adopters".

Table 2 Basic statistics about the mobile phone data set (daily average per person).

	Mean	Std. Dev.	Min	25th PCTL	50th PCTL	75th PCTL	Max
Number of calls	2.84	3.64	0.00	0.92	2.00	3.65	68.90
Number of texts	3.04	5.73	0.00	0.67	1.81	4.00	45.65
Number of activities using data	29.07	65.00	0.00	0.00	3.20	36.90	3355.10
Number of total activities	34.95	66.40	0.00	4.00	11.00	43.08	3359.58

We divide the overall data set into non-overlapping observation periods; each observation period is defined as $T = [s, s + l]$, where $T \in \Psi$, $s \in \mathcal{S}$, and l is the length of each period. Here, Ψ is the set of all observation periods, and $\mathcal{S}$ is the set of the starting time instances of each period. For each performance day, we choose the observation period T in Figure 1a to be a period of $l = 24$ hours, starting with the beginning of the performance each day. The motivation for choosing this

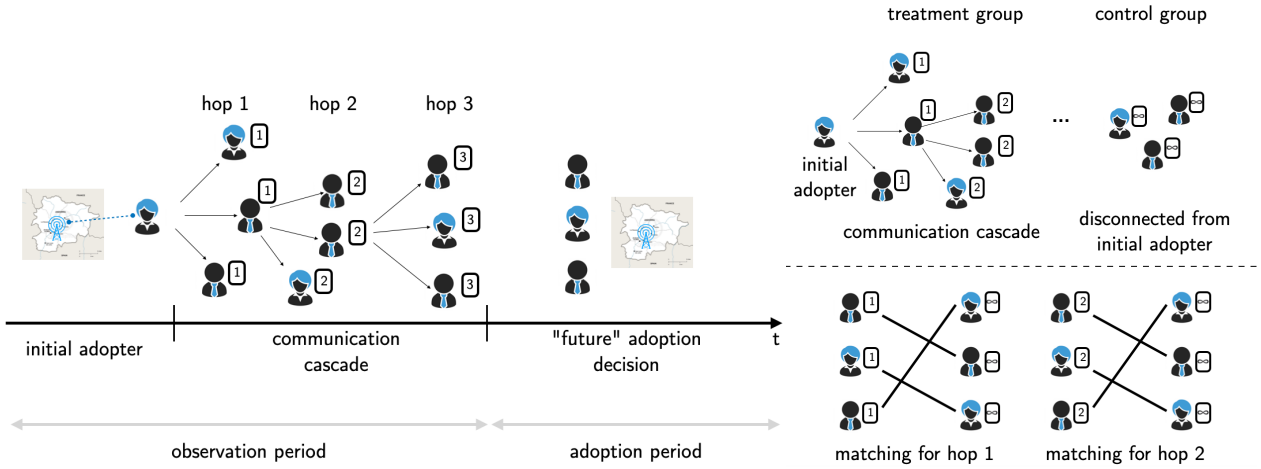


Figure 1 **Illustration of the framework. (a) Each observation period is separated into two parts: 1) identifying the initial adopter and 2) constructing the phone communication cascade. After this observation period, we evaluate the eventual adoption decisions in the adoption period. (b) Identifying treatment and control groups. In the upper panel, we show how to construct the treatment group and the control group. Individuals connected to the initial adopter directly (labeled as hop 1) and indirectly (labeled as hop 2 and higher) in the communication cascade $\mathcal{C}_T$ are categorized in the treatment group during the observation period T. Individuals who are disconnected from any initial adopters through the information cascade $\mathcal{C}_T$ across all observation periods are labeled as the control group. As demonstrated in the lower panel of (b), we aggregate all observation periods and perform empirical analysis separately for each hop index group.**

threshold is that the cultural performance took place each evening, and we would like to keep the observation periods non-overlapping (so that the communication cascades defined below will not interfere with each other).

We now introduce several key concepts in this paper in Figure 1a. First, we define $\mathcal{D}_T$ as the set of *initial adopters* for the observation period T. We identify individuals as initial adopters if they were connected to one of the three cell towers nearest the performance venue during a time interval at the beginning of the period T in Figure 1a, where the time interval is defined as the time window of the performance (with a buffer time of $\pm$ 30 minutes). This strategy is similar to the one in Toole et al. (2015), which uses connections made to three cell towers near an auto-parts manufacturing plant to label whether individuals worked at the plant.

Second, we construct a *communication cascade* as a directed graph $\mathcal{C}_T = (\mathcal{I}_T, \mathcal{E}_T)$, where the node set $\mathcal{I}_T = \{1, 2, 3, ..., n\}$ is a set of n individuals who have at least one mobile phone activity in T; meanwhile, the edge set $\mathcal{E}_T = \{(i, j)\}$ is a collection of ordered node pairs (i, j), conditioned such that $i \in \mathcal{I}_T$ has information about the product when the communication with $j \in \mathcal{I}_T$ takes place and that i will spread the information to j. We cannot obtain the actual content of the

communication because of privacy considerations. The assumption that information of interest has been transmitted through the observed communication channel has been adopted in prior studies (Aral et al. 2009).

We define the third concept, *hop index*, for an individual i in $\mathcal{I}_T$, as the length of the path from individual i to an individual $j \in \mathcal{D}_T$. Therefore, an individual i of hop index h is h-degrees of separation from an initial adopter in $\mathcal{D}_T$. In our analysis, we define treatment groups as individuals who have not yet made an adoption decision and have been connected via a single path to an initial adopter[3] in only one observation period. In addition, we define one treatment group for each hop h as the group of individuals (from any observation period) with a finite hop index h. Note that if i is an isolated node in $\mathcal{C}_T$ for all observation periods (i.e., if i is not connected directly or indirectly to any adopter), then the hop index would be infinity; hence, we use these nodes as the control group, as in Figure 1b.

Finally, we define the *adoption period* to be the period that starts immediately after the observation period in which one received a treatment[4], till the last day of the performance. Figure 1a illustrates the adoption period in connection with the observation period.

Our data set includes 19 observation periods. We do not construct the observation periods for the last three days of the performance because individuals who received information through phone communications on these days did not have enough time to attend the event. For each observation period T, we construct a communication cascade and compute the hop index for each individual appearing in $\mathcal{C}_T$. We remove individuals who had less than five observations in the past six months to ensure that we have sufficient information to control for. To reduce the chance that individuals are communicating through other information channels, we exclude the following data: 1) Phone calls between parties (i.e., the caller and the receiver) that were served by the same cell tower, thereby reducing the chance of including face-to-face communications. 2) Individuals whose network geodesic distance from the initial adopter in the communication network (with reciprocal phone communication) from the two months prior to the event was shorter than their hop indices; thus, we avoid inflation on hop indices computation. 3) Individuals who were disconnected from any individuals in the historical social network. After removing individuals using these criteria, our data set include 23,581 individuals across four treatment groups. Another 21,652 individuals who were disconnected from all communication cascades were in a single control group.

[3] We do so to avoid the difficulty in disentangling the multiplication effect of social influence.

[4] For a member of the control group, her adoption period is the same as that of the matched individual in a treatment group (more details are provided in Section 3.2 when we construct the panel data).

Table 3 Research design of the empirical strategy

	Identification strategy	Sensitivity analysis and robustness checks
1. Observed homophily ($\mathbf{x}$) 2. Latent homophily ($\mathbf{c}, \mathbf{f}$)	PSM	1. Rosenbaum sensitivity test 2. Balance in propensity scores and covariates 3. Other matching strategies and Post-Lasso estimation
1. Observed homophily ($\mathbf{x}$) 2. Latent homophily ($\mathbf{c}, \mathbf{f}$) 3. Trend before and after treatment ($\pi_{\text{after}_{sjt}=1}$) 4. Pre-treatment difference ($\pi_{D_{sj}=1}$) 5. Matched pair fixed effect (η_s) 6. Time-varying common shocks (ν_t)	DID + PSM	Shuffle test

3.2. Difference-in-differences in combination with behavioral matching

Identifying the social influence effect is challenging, especially when using observational data. That individual decision-making (e.g., adoption behavior) in a social network can be affected by a number of factors is widely recognized. The first factor is the correlation or homophily effect (McPherson et al. 2001), which suggests that individuals tend to become neighbors (connected in the network) because of a shared background or interest, which in turn leads to the adoption by both individuals. The second set of factors is exogenous factors (i.e., external causes common to network neighbors (Manski 1993), such as marketing campaigns). The third set of factors is peer effects (i.e., social influence), which states that one's adoption is either directly or indirectly affected by communication with one's neighbors who have adopted the behavior.

We use a difference-in-differences (DID) model in combination with propensity score matching (PSM) (Rishika et al. 2013, Li 2016, Dewan et al. 2017, Jung et al. 2019). We control for homophily using behavioral variables: For observed homophily, we use visited locations in mobility history; for latent homophily, we use latent positions inferred from a historical social network and neighbors' eventual adoption behaviors. We also perform sensitivity analysis and robustness checks on the results. Table 3 summarizes the empirical strategy of this paper.

3.2.1. Behavioral matching based on observed and latent homophily To ensure similarities between the treated group and the control group in the DID analysis, we first adopt a matching-based estimation framework to assemble a matched sample of the treated and control units. Deciding on which variables to use to match individuals is a critical question. Existing studies rely primarily on socio-demographic information (de Matos et al. 2014, Jung et al. 2019), but this approach has three shortcomings: 1) Such information is not always available; 2) it does not capture the latent preferences (e.g., latent homophily (Ma et al. 2015)); and 3) it cannot adapt to changes in individual tastes and preferences. To address these issues, we design a behavioral matching framework based on observed homophily (visited locations in mobility history) and latent homophily (latent preferences inferred from the historical social network, and neighbors' eventual adoption decisions). We demonstrate these three types of behavioral covariates, computed using the mobile phone data, in Figure 2.

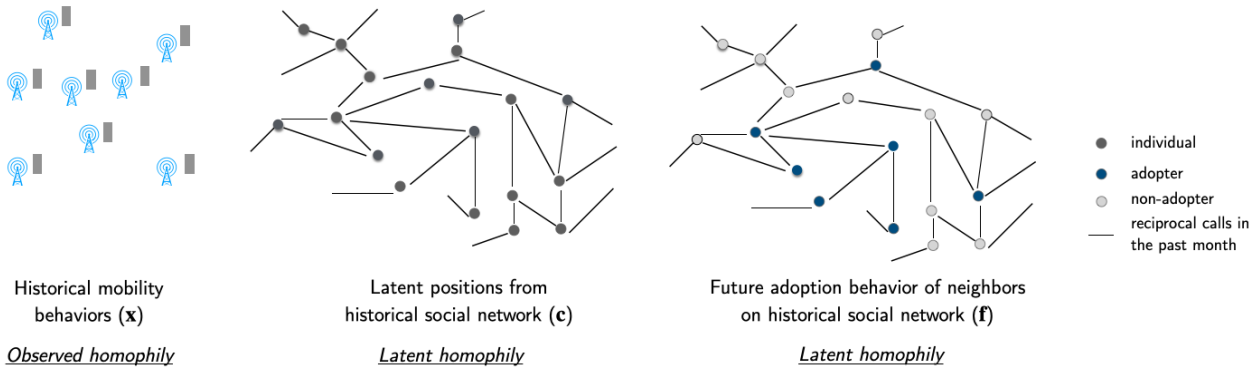


Figure 2 **Illustration of the three types of behavioral covariates extracted from the mobile phone data to approximate high-dimensional observed and latent homophilous covariates.**

Observed homophily: revealed preferences from mobility history. We first use individuals' history of visited locations to control for observed homophily. The theoretical foundations for using spatial locations are revealed preference theory and consumer behavioral theory, which together suggest that consumer choices, serving as revealed preferences, are indicative of consumer preferences (Samuelson 1938, McFadden 2001). Furthermore, co-occurrence of locations and mobility trajectory similarities between individuals have been demonstrated to reveal similarities in preferences (Ghose et al. 2019). Thus, we use individual mobility histories on weekends (i.e., the frequency with which individuals visit different places) as data for the revealed preference. We specifically use mobility behaviors on weekends because behavior in one's spare time offers a better proxy for individuals' preferences. In addition, it has been shown that adjusting for behavioral covariates relevant to the adoption decision of interest reduces the estimation bias substantively and yields an estimate that is statistically indistinguishable from what is obtained through Random Controlled Trials (RCTs) in a Facebook context (Eckles and Bakshy 2020). In our case, mobility behaviors are highly relevant to the adoption decision of interest, which also is measured using location visits.

We consider an individual-location matrix $\mathbf{M}$, where the i-th row and k-th column correspond to the i-th individual and k-th location (i.e., of the k-th cell tower), respectively, and where m_{ik} represents the number of times that individual i has visited location k during a six-month period prior to the performance month. We then apply principal component analysis (PCA) and project $\mathbf{M}$ onto a subspace established by the top eigenvectors of its covariance matrix to obtain an eigen-preference matrix in which the i-th column, $\mathbf{x}_i$, represents the latent preferences of individual i. We choose 19 principal components (PCs) ($\mathbf{x}_i \in \mathbb{R}^{d_x}$ where $d_x = 19$) in the adoption behavior of attending the cultural performance, such that they explain more than 90% of the variance in $\mathbf{M}$.

Latent homophily: neighbors' eventual adoption decisions and latent preferences learned from their historical social network. In this section, we explain how we control for latent homophily (Ma et al. 2015). We control for two sources of latent homophily: (1) using neighbors' eventual adoption decisions as a proxy for user fixed effects, following (Belo and Ferreira 2022); and (2) latent positions learned from the historical social network. We discuss how we control for these two sources of latent homophily in sequence.

First, we follow Belo and Ferreira (2022) to control for the adoption behaviors of neighbors (in the historical social network) as a proxy for individuals' interest in and attitude toward the adoption decision. The rationale behind this proxy is that, as a result of homophily, adopters are more likely to be connected to adopters and non-adopters to non-adopters. These connections lead to a positive correlation between being an adopter and having neighbors who also are adopters. Consequently, neighbors' eventual adoption decisions (observed by the end of their respective adoption periods) are a direct reflection of the focal individuals' interests and preferences.[5] Therefore, adding these variables helps partially control for latent homophily. We specifically use two measures, the number and percentage of neighbors who ended up adopting the behavior, as the control variables (denoted as $\mathbf{f}_i \in \mathbb{R}^2$ for individual i).

Second, we use latent positions learned from the historical social network to further control for latent homophily. Social networks can be informative about latent characteristics of individuals resulting from homophily (McPherson et al. 2001). McFowland III and Shalizi (2021) establish sufficient conditions under which controlling for estimated locations in a latent space leads to asymptotically unbiased and consistent social influence estimates, assuming a certain network formation process (e.g., either a stochastic block model or a continuous latent space model)[6]. We adapt their approach to controlling for latent covariates encoded in the historical social networks in order to reduce bias due to latent homophily; specifically, we propose using an efficient network representation learning approach, *node2vec* (Grover and Leskovec 2016), to learn feature representations for the individuals using the historical social network. Although this approach corresponds to a relaxation of the specific assumptions in McFowland III and Shalizi (2021) in terms of the network formation process, the principle behind *node2vec* remains that the network is homophilous, i.e., nodes with similar characteristics and preferences will be more likely to form a link. In other

[5] Simulation studies have demonstrated the effectiveness of this proxy over a wide range of parameters, independent of the network's structure, and with varying levels of homophily and the product's baseline level of adoption (Belo and Ferreira 2022).

[6] More formally, the assumptions made in McFowland III and Shalizi (2021) are: (1) for the underlying network models, all links in the historical social network are conditionally independent of each other, given the latent positions for each individual; and (2) in observations of the whole network, adoption provides no additional information about an individual's latent positions.

words, individuals that have similar network positions (e.g., they connect to one another or to the same others, or they lie in the same social community) remain close in a low-dimensional latent space. Therefore, latent positions computed using *node2vec* ($\mathbf{c}_i \in \mathbb{R}^{d_c}$ for individual i where $d_c = 16$ in our case) can be used as covariates in a regression to control for latent homophily. We include more details on *node2vec* and how the parameter d_c is determined in Appendix A.

Behavioral matching. As mentioned previously, there are multiple treatment groups for each period T, and one for each finite hop index (see Figure 1b). For each treatment group, every individual is matched to one individual in the control group. Thus, we use PSM to control for observed homophily, drawn from mobility histories, and for latent homophily, drawn from latent positions from the historical social network and neighbors' eventual adoption decisions. The propensity score for being treated in hop h is defined as the conditional probability of being connected to the initial adopter via h hops, which we estimate based on individuals' latent preferences using the logistic regression. We estimate the propensity score model for each treatment group and for the control group. Specifically, for each hop index, we compute,

$$\log\Big(\frac{\mathbb{P}(D_i=1)}{\mathbb{P}(D_i=0)}\Big) = \alpha_0^{ps} + \mathbf{x}_i'\boldsymbol{\alpha}_x^{ps} + \mathbf{c}_i'\boldsymbol{\alpha}_c^{ps} + \mathbf{f}_i'\boldsymbol{\alpha}_f^{ps} + \xi_i,$$

where $(\,)'$ is the transpose operation; $\boldsymbol{\alpha}_x^{ps} \in \mathbb{R}^{d_x}$ is the coefficient vector for observed homophily; $\boldsymbol{\alpha}_c^{ps} \in \mathbb{R}^{d_c}$ is the coefficient vector for latent positions; $\boldsymbol{\alpha}_f^{ps} \in \mathbb{R}^2$ is the coefficient vector for neighbors' eventual adoption decisions at the end of their respective adoption periods; $\alpha_0^{ps} \in \mathbb{R}$ is the intercept; and ξ_i is the error term. We use the estimated coefficients to predict the time-invariant propensity scores of each user and match individuals using the predicted propensity scores.

3.2.2. Difference-in-differences on matched samples. The DID approach compares the changes in the adoption decisions of the treated units before and after the treatment (i.e., the communication) to the adoption decisions in the control units over the same period of time. The behavioral matching framework in the previous section helps substantially improve the similarity between the treatment and the control group and to account for (both observed and latent) homophily, thereby enhancing the inference related to the DID analysis and improving the consistency of the estimates (Stewart and Swaffield 2008). We conduct the analysis on the matched samples separately for each of the treatment groups. In other words, we apply the DID model to the treatment group associated with each hop index and to the corresponding matched units in the control group.

The DID model takes a panel data set as its input; we illustrate this structure in Figure 3. Following the standard in constructing panels to measure diffusion processes, individuals leave the panel after they adopt the behavior. Consider a matched pair s, as shown in Figure 3; the treated

Bob

day	z_{s1t}	received call	D_{s1}	after_{s1t}	$D_{s1}\text{after}_{s1t}$
1	0	0	1	0	0
2	0	0	1	0	0
3	0	1	1	0	0
4	0	0	1	1	1
5	1	0	1	1	1

Anne

day	z_{s0t}	received call	D_{s0}	after_{s0t}	$D_{s0}\text{after}_{s0t}$
1	0	0	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	1	0
5	0	0	0	1	0

Figure 3 Illustration of the panel structure for the DID model, showing the data structure for a treated individual (Bob) and a control individual (Anne).

individual, Bob (for hop index 1, without loss of generality), is on the left panel, and the matched control individual, Anne, is on the right panel. Assume that Bob was treated (i.e., he received a phone call) on day 3 and attended the event on day 5. We add a series of 1s after the treatment day for after_{sjt}. Because Bob adopted on day 5, we remove the dates after day 5. Anne, the matched control individual, neither received a call nor adopted; hence, the columns of "adoption" and "received call" are filled with 0s. Because Anne is matched with Bob, we let $\text{after}_{s0t} = \text{after}_{s1t}$ for Anne.

We use a linear probability model with a binary outcome variable as follows:

$$z_{sjt} = \underbrace{\overbrace{\mathbf{x}'_{sj}\boldsymbol{\alpha}_x}^{\text{observed homophily}} + \overbrace{\mathbf{c}'_{sj}\boldsymbol{\alpha}_c + \mathbf{f}'_{sj}\boldsymbol{\alpha}_f}^{\text{latent homophily}}}_{\text{homophily}} + \underbrace{\gamma_h D_{sj}\ \text{after}_{sjt}}_{\text{phone communications}} + \pi_{D_{sj}=1} + \pi_{\text{after}_{sjt}=1} + \eta_s + \nu_t + \epsilon_{sjt}. \quad (1)$$

In Equation (1), t is the index for a day in the time period during which the event took place; $(\,)'$ is the transpose operation; s indexes a matched pair of treated and control units; j denotes a treated ($j = 1$) or a control ($j = 0$) unit; and ϵ_{sjt} is the error term. The dependent variable z_{sjt} is the adoption behavior of the (treated or control) unit j in the matched pair s at time t, where $z_{sjt} = 1$ indicates adoption and $z_{sjt} = 0$ indicates non-adoption. D_{sj} is a treatment dummy variable that equals 1 if the unit is in the treatment group and 0 if it is in the control group; after_{sjt} is a dummy variable that equals 1 for the time period after the treatment (e.g., direct or indirect communication) and 0 for the time period before the treatment.

The main parameter of interest is γ_h, which measures the change in the likelihood of adoption if the individual had been included in the treatment group corresponding to hop h (i.e., if he or she had received the phone call during the observation period with h hop distances from the initial adopter). We use $\pi_{D_{sj}=1} \in \mathbb{R}$ to denote the pre-treatment difference in the two groups, which turns on for the treatment unit in the matched pair s. We use $\pi_{\text{after}_{sjt}=1} \in \mathbb{R}$ to denote the time trend in the control group before and after the treatment is received. This variable turns on after the treatment

is received and varies across the matched pairs s. We further use the fixed effect (η_s) at the level of matched pairs s to capture the potential, unobserved, time-invariant heterogeneity.[7] Finally, to control for common shocks over time that affect the adoption behavior (e.g., a discount for an event occurring at a certain time period t), we include the time fixed effect ν_t which is specific for each time period t, thus addressing the possible time-varying common shocks. Note that the time fixed effect ν_t differs from $\pi_{\text{after}_{sjt}=1}$, because the former is fixed effect at t and is the same across different units, while the latter differs across matched pairs. We use $\boldsymbol{\alpha}_x \in \mathbb{R}^{d_x}$, $\boldsymbol{\alpha}_c \in \mathbb{R}^{d_c}$, $\boldsymbol{\alpha}_f \in \mathbb{R}^2$ to represent the coefficients for the observed mobility covariates ($\mathbf{x}_{sj}$), latent positions ($\mathbf{c}_{sj}$), and the neighbors' eventual adoption behaviors by the end of their respective adoption periods ($\mathbf{f}_{sj}$).

We provide a final remark on potential measurement errors. The first source of measurement errors in using phone data to estimate social influence is identifying adoption decisions. Adopters may not actually use their phones when attending the event, or individuals may pass by the performance venue without attending it. These measurement errors can affect two variables: 1) the adoption decisions of individuals in the treatment group and the control group z_{sjt}, and 2) the identification of initial adopters, leading to errors in the treatment D_{sj}. Second, each observation period in our setting is limited to 24 hours, and any phone calls made with initial adopters directly of indirectly beyond this period can generate a measurement error in the treatment variable D_{sj}. The classic result in the econometrics literature shows that: 1) a mismeasured outcome z_{sjt} does not lead to a bias, and 2) a mismeasured predictor (e.g., the treatment variable D_{sj}) will bias the effect toward zero (Lewbel 2007). In other words, a mismeasured adoption outcome does not bias our estimate of the social influence effect. However, if we have measurement errors in identifying the initial adopters or if some treatments are missing after the observation period, this will lead to an underestimation of the social influence effect. Nevertheless, our results remain valid even with these types of measurement errors. We discuss the impact of measurement errors on our social influence estimates in more detail in Appendix B, closely following Theorem 1 of Lewbel (2007). Overall, despite potential measurement errors, our results remain valid.

4. Empirical results

4.1. Long-range effect of social influence via phone communication

We use the technical framework developed in Section 3 to quantify the long-range effect of social influence via phone calls based on CDRs. The summary statistics of all control variables we use are included in Appendix C. As described in the previous section, our identification strategy consists of matching followed by DID analysis. To visually demonstrate how the DID estimator works, in

[7] Note that, in understanding diffusion of adoption decisions, not adding the individual-level fixed effects is customary, because such fixed effects would capture the adoption perfectly and thereby absorb the effect of interest (Belo and Ferreira 2022).

Figure D1 of Appendix D, we plot the over-time survival rate separately for the treatment and the control groups. As we can see, starting from the date of the treatment—having received the phone call—the two groups exhibit a widening gap in survival rates (one minus the probability of adoption), as the treated group becomes differentially more likely to attend the event and therefore are subsequently dropped out of the sample. The differential surviving rate quantitatively reflects the cumulative effects of social influence over time.

We are now ready to present our main empirical results. We present the main estimates on the change in the adoption likelihood (i.e., attending the event) due to social influence through phone communication in Table 4. The detailed estimation results are presented in Tables E1-E2 in Appendix E. In Figure 4, we present the estimates, with respect to different hop indices, relative to the adoption likelihood of the control group. Our analysis reveals that being a direct contact of an initial adopter increases the likelihood of attending the event by 87.61%[8]. For individuals who are two degrees of separation away from the initial adopter, the increase in adoption likelihood is 68.65%. The effect of social influence on adoption likelihood weakens as the degree of separation increases. Individuals who are three degrees of separation away from the initial adopter have a 53.10% increase in adoption likelihood, while those who are four degrees of separation away have a 46.71% increase. Interestingly, we find that the increase in adoption likelihood from direct neighbors to two-degree indirect neighbors decreases by 21.65% ($-\frac{\gamma_2-\gamma_1}{\gamma_1}$). The increase in adoption likelihood for three-degree neighbors is further reduced by 22.64% ($-\frac{\gamma_3-\gamma_2}{\gamma_2}$) compared to the increase observed for two-degree neighbors. Overall, we observe a significant positive effect of influence through phone communication from hop one to hop four, demonstrating the long-range impact of social influence via phone communications. This finding suggests the potential of viral and seeded marketing designs using phone communications. Although the treatment effect for hop five is also significant, the estimate is not robust to unobserved confounders, as confirmed by the Rosenbaum sensitivity analysis (see Section 4.2 for details). Therefore, we limit our analysis to hops one to four, representing four degrees of separation.

Our empirical results on the long-range and decaying social influence motivate us to better understand this observation in two aspects: 1) whether the results are reliable according to different robustness checks; 2) what might be the mechanism behind the long-range effect. We address the first point in Section 4.2 and the second in Section 4.3.

[8] For the ease of readability in the paper's description, we round the numbers to four digits, but in the calculation, we use the eight-digit decimals provided in Table E1 of Appendix E.

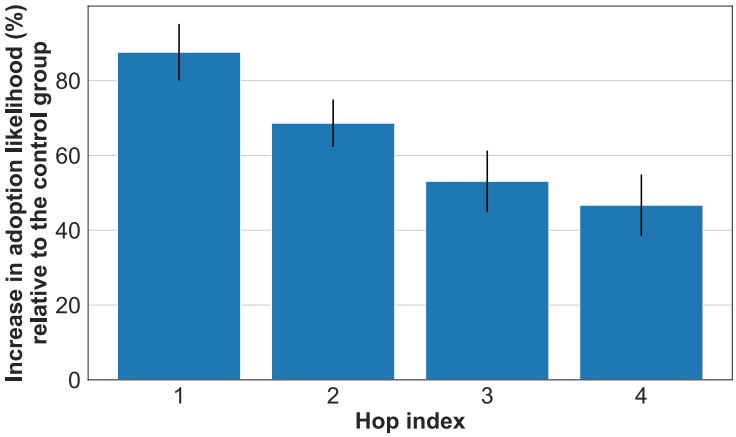


Figure 4 **Change in adoption likelihood from social influence through phone communications (relative to the matched control group). The vertical line corresponds to the 95% confidence interval.**

Table 4 **Social influence estimate (γ_h) from Equation** (1)

	Dependent variable: Adoption			
	hop 1	hop 2	hop 3	hop 4
D_{sj} after$_{sjt}$	0.0086***	0.0067***	0.0052***	0.0046***
	(0.0004)	(0.0003)	(0.0006)	(0.0006)
Time fixed effect (ν_t)	✓	✓	✓	✓
Pair fixed effect (η_s)	✓	✓	✓	✓
Time-trend ($\pi_{\text{after}_{sjt}=1}$)	✓	✓	✓	✓
Pre-treatment difference ($\pi_{D_{sj}=1}$)	✓	✓	✓	✓
Observations	360,226	368,000	60,398	49,680
Residual Std. Error	0.0386 (df = 348,243)	0.0341 (df = 355,980)	0.0227 (df = 58,007)	0.0206 (df = 47,680)

Note: *p<0.1; **p<0.05; ***p<0.01
Robust standard errors in parentheses.

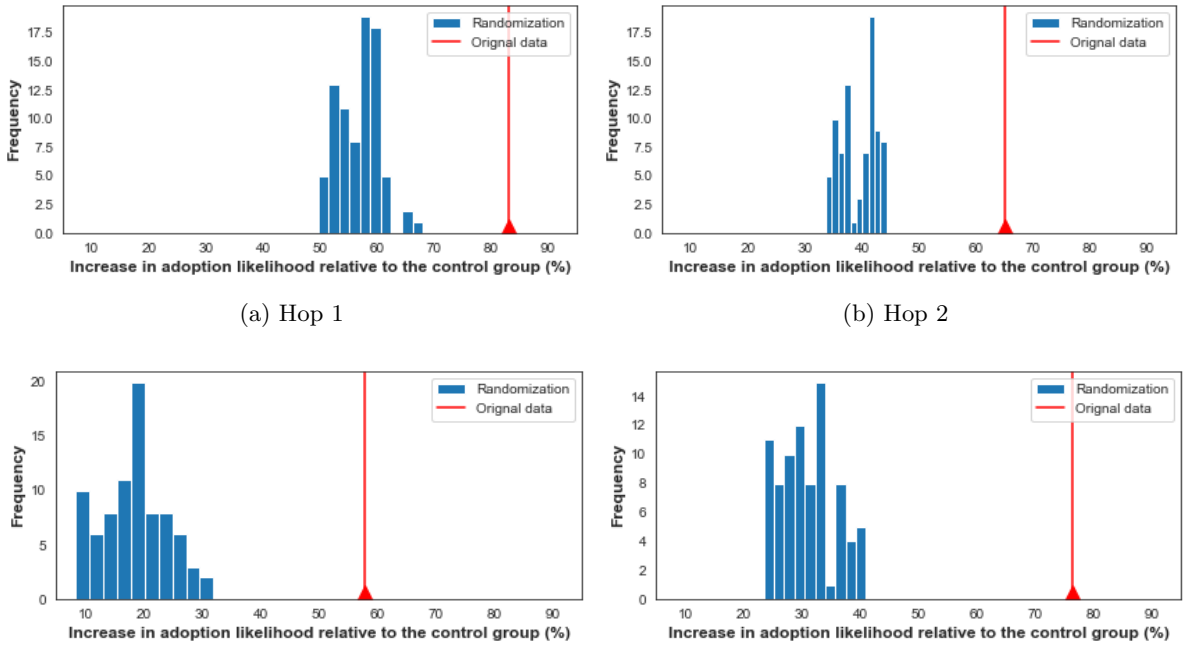
4.2. Robustness check

Balance between the treatment group and the control group. Checking covariate and propensity score imbalance post-matching is important to assess the quality of the matching technique. In our study, the standardized differences in the covariates of the treatment group and the control group after matching are far below the rule-of-thumb value (Figure F1 of Appendix F1). After matching, we achieved substantial reductions in the differences between treatment and control groups for all latent homophily-related covariates and most observed homophily-related covariates, as indicated by significant coefficients in Table F1 of Appendix F1. We observe that the distributions of the propensity scores for the control and treated groups are similar and have a significant post-matching overlap (using Figure F2 and Table F2 in Appendix F1). Both robustness checks in the covariates and the propensity score demonstrate that the matched pairs in the treatment and control groups are well-balanced.

Sensitivity analysis toward unobserved confounders. As the treatment assignments in our study (i.e., the phone calls) are not randomized, there may still be some level of bias in our analysis, despite our efforts to control for observed and latent homophily. We analyze the sensitivity with respect to the selection on unobservables using the Rosenbaum bounds approach (Rosenbaum 2005). It evaluates the extent to which unobserved variables might affect an individual's assignment into the treatment or control group and, therefore, the inference. We use the odds ratio of treatment assignment (Γ) to quantify the amount of bias from unobserved variables required to change the results qualitatively. Our results, as shown in Figure F3 in Appendix F2, indicate that the critical level of Γ at which we would question the validity of the PSM is 8.5 (hop 1), 7.4 (hop 2), 2.0 (hop 3), and 2.0 (hop 4). Specifically, for hop one, when $\Gamma = 8.5$, the upper bound p-value is larger than 0.05, indicating that the confidence interval for the social influence effect would include zero if an unobserved confounder caused the odds ratio of the treatment assignment to differ between the treatment and control groups by 8.5. This interpretation applies to other hops as well. While there is no clear consensus on a rule-of-thumb value for Γ, some studies have suggested that anything above $\Gamma = 1.5$ indicates substantial insensitivity to unobserved confounders (Sen 2014, Ransbotham et al. 2019). Our Γ values are sufficiently larger than this value across four hops, indicating that our results demonstrate substantial insensitivity to hidden bias and strong support for the existence of social influence through phone communications up to four degrees of separation. However, beyond the fourth hop, the results are no longer robust to unobserved confounders, and therefore, we exclude them from our analysis. In summary, our findings, as shown in Figure 4, are robust to a plausible range of unobserved selection bias, up to the fourth hop.

Shuffle test. To further validate our findings on the impact of social influence through phone communications, we perform the "shuffle test" introduced by Anagnostopoulos et al. (2008). This shuffle test aims to exclude the effect of social influence while retaining other factors, such as homophily and other unobserved confounders. This method, adapted by Belo and Ferreira (2022), provides a lower bound in absolute terms for the effect of social influence (see Appendix E of Belo and Ferreira (2022)).

To conduct this test, we shuffle the dates of the phone calls (hence, the treatment) within each treatment group (for each hop index) so that the overall adoption rate and the adoption curve (by time) remain the same. We further constrain the shuffling to include only the individuals that were treated in the same week, similar to the approach in Belo and Ferreira (2022). This restriction addresses the concern that the adoption dates may conceal unobserved effects leading to adoption. Specifically, unrestricted shuffling may not be desirable in the presence of temporal clustering in the adoption pattern. For instance, it could lead to the assignment of late adoption dates to early adopters. We then use the same DID strategy on the matched pairs (according to observed and



(a) Hop 1 (b) Hop 2

(c) Hop 3 (d) Hop 4

Figure 5 **Distribution of estimates over 100 shuffles of adoption dates. The x-axis is the change in adoption likelihood resulting from phone communication. The y-axis is the frequency of the estimates over the 100 shuffles. The red vertical line represents the coefficient obtained using the original data.**

latent homophily) to compute the change in adoption likelihood on the shuffled data. Afterward, we compute the empirical distribution of the effect of social influence using the shuffled data, and we compare this distribution with the effect of social influence from the original data. We can reject the null hypothesis of no social influence if the estimates from the original data fall outside the 95% confidence interval of the parameter obtained from the randomized data. Figure 5 shows that the estimates from the original data are outside the 95% confidence interval of the estimates obtained from the shuffling procedure. Additionally, the estimates obtained from the shuffled data are significantly lower than those obtained from the original data for all hop indices. Hence, we reject the null hypothesis that $\gamma_h = 0$ for $h \in \{1, 2, 3, 4\}$ and conclude that social influence increases the likelihood of adoption up to four degrees of separation. Given that randomization provides a lower bound for the effect of social influence, this test indicates that the observed patterns of social influence up to four degrees of separation are not likely to be driven entirely by the effects of homophily or other unobserved confounders.

Other observational analysis methods. We test a series of methods (including coarsened exact matching, subclassification, Mahalanobis distance matching, and Post-Lasso estimation), with results shown in Figure G1 in Appendix G. All methods present the long-range social influence

effect with a decay pattern as the degree of separation increases, suggesting the robustness of our findings with respect to the observational methods.

4.3. Mechanism: information loss along phone communication cascade

The empirical findings motivated us to investigate the potential mechanism that leads to the decay of social influence along the hop indices in the communication cascades. To this end, we adopt a simple structural Bayesian approach that models a sequential update process through information sharing, following Zhang (2010). In this process, information about the quality of the event (i.e., the subject of adoption) is shared through WOM communication via phone calls.

In the following paragraphs, we discuss the utility function and the Bayesian learning process. Let $u_i(\mathcal{S}_{it})$ denote the utility of user i to adopt the decision at time t, based on state variables contained in $\mathcal{S}_{it} = \{I_{it}, \zeta_{it}\}$, where I_{it} is a set of signals i received up to t and ζ_{it} is the idiosyncratic utility shock to individual i. Following Zhang (2010), we have[9]:

$$u_i(\mathcal{S}_{it}) = \alpha\theta_t - \alpha\rho\theta_t^2 + \zeta_{it}, \tag{2}$$

where θ_t characterizes any unobservable quality component of the product at time t; α is the associated utility weight; ρ captures i's risk-averse tendency. Because of the time and monetary costs of attending the event (i.e., relative to resharing content on social media or downloading an app), we assume that individuals are risk-averse. We follow Zhang (2010) and introduce the quadratic term $\alpha\rho\theta_t^2$ to capture this tendency, allowing for a positive risk-averse tendency ρ. Based on this utility, individuals then make an adoption decision using a sigmoid function:

$$\mathbb{P}\big(u_i(\mathcal{S}_{it})\big) = \frac{1}{1 + e^{-u_i(\mathcal{S}_{it})}}. \tag{3}$$

We assume that individuals have prior knowledge about the distribution of θ_t, which is assumed to be *i.i.d.* normal with fixed mean μ and variance σ_θ^2: $\theta \sim \mathcal{N}(\mu, \sigma_\theta^2)$. In our context, such prior knowledge might be obtained from television or offline advertisements of the event. In addition, user i might receive a private signal s_{it} of the unobserved quality θ_t.

We next describe two types of information update processes. The signal $\mathcal{S}_{it}$ might be derived from the experience of attending the event (for initial adopters) or from communicating with their neighbors (for non-initial adopters). In addition to their prior knowledge and their own private signals (available if they have attended the event), individuals can gather private signals from individuals with whom they communicate via phone calls. That is, compared to individuals in the

[9] In this utility function, without loss of generality and following the setting in the literature of Bayesian learning (Acemoglu et al. 2011), we consider individuals to be homogeneous and therefore do not include user covariates.

control group, those in the treatment groups can fine-tune their quality signals if they also receive private signals from phone communications.

According to Bayes's rule, the expectation of the posterior distribution of θ_t is a weighted average of the posterior mean μ and the private signal, which follows a normal distribution with mean s_{it} and standard deviation σ_s. If one's private signal is the only information available (e.g., in the case of initial adopters, after they attended the event), then the rule for updating the expectation of θ_t is (following Equation (8) of Zhang (2010)):

$$\mathbb{E}(\theta_t|I_{it}) = \frac{\sigma_\theta^2 s_{it} + \sigma_s^2 \mu}{\sigma_\theta^2 + \sigma_s^2},\ I_{it} = \{s_{it}\}. \tag{4}$$

On the other hand, if an individual i receives r private signals by communicating with others (e.g., in the case of any non-initial adopters from hop 1 onward in the cascade), the expectation of the posterior distribution of θ_t is a weighted average of the prior mean μ and the sample average of these signals (following Equation (9) of Zhang (2010)):

$$\mathbb{E}(\theta_t|I_{it}) = \frac{\sigma_\theta^2 \sum_{j=1}^{r} s_{jt} + \sigma_s^2 \mu}{\sigma_\theta^2 + \sigma_s^2},\ I_{it} = \{s_{1t}, ..., s_{rt}\}. \tag{5}$$

We then use simulation to understand two elements. The first element is information loss along the communication chain from hop 1 to subsequent hops, which is represented by the difference in the expectation of the posterior probability on θ_t (when simulation stops) between the initial adopter and individuals in later hops. We let

$$\text{Information loss} = \Big|\frac{1}{|\mathcal{I}_0|}\sum_{i\in\mathcal{I}_0}\mathbb{E}(\theta_t|I_{it}) - \frac{1}{|\mathcal{I}_h|}\sum_{j\in\mathcal{I}_h}\mathbb{E}(\theta_t|I_{jt})\Big|, \tag{6}$$

where $\mathcal{I}_0$ is the set of initial adopters and $\mathcal{I}_h$ is the set of individuals in hop h.

The second element is how information loss affects adoption decisions and, in turn, social influence. The strength of social influence is computed as:

$$\text{Strength of social influence} = \mathbb{P}\big(u_i(\mathcal{S}_{i,t=1})\big) - \mathbb{P}\big(u_i(\mathcal{S}_{i,t=0})\big). \tag{7}$$

In this setup, the effect of social influence is quantified by the difference in the probability of adoption before and after receiving the information via phone calls. In the simulation, we let $\sigma_\theta = 0.1$, $\sigma_s = 0.1$, $\alpha = 1$, $\rho = 0.1$, and $\zeta_{it} \sim \mathcal{N}(0, 0.1)$. The prior mean for all individuals is set at $\mu = 0.5$. Given the reputation of the event, the initial adopter is likely to receive a private signal reflecting the high quality of the event, in which case we set $s_{it} = 0.9$ if $i \in \mathcal{I}_0$. In our empirical setting, we consider only a single-path communication cascade; therefore, the private signal would come only from the individual who communicated with i.

The simulation process aims to mimic social learning in phone communication using the following steps in sequence:

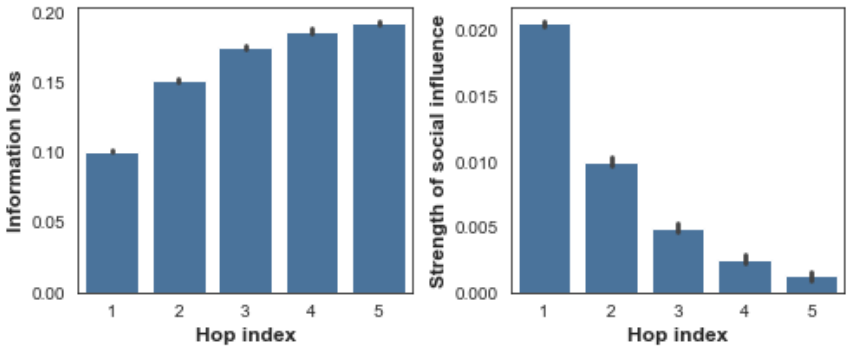


Figure 6 **Information loss in a Bayesian learning process (according to Equation (6)) can lead to social influence decay (according to Equation (7)).**

1. The initial adopter computes the posterior using Equation (4), after receiving their private signals through attending the event.

2. The initial adopter communicates with and sends a private signal (drawn from the posterior from Equation (4)) to the individuals in hop 1 (which is set as the current hop index).

3. Individuals in the current hop update their posterior probability, according to Equation (5).

4. Compute the adoption probability of the individual (who received the information signal) using Equation (3).

5. Repeat steps 3 and 4 to start the next hop and all subsequent iterations up to the fifth iteration[10].

We simulate the process 1,000 times over a single branch of a hypothetical communication cascade. The results are presented in Figure 6. As shown in the left panel, we see that information loss increases as the hop index increases, which provides evidence that information is lost along the communication cascade. The right panel shows social influence based on the difference in adoption probability, computed using the prior signal and the posterior signal. That is, for each user in the communication cascade, we compute the difference in the adoption probability before and after the phone communication. We see that social influence decreases along the hop indices because of the information loss. This simple Bayesian model provides a mechanism that may explain the empirically observed decay of social influence in Section 4.1.

5. Influence centrality

Centrality is an important characteristic for nodes in a network. It has been widely used in network-based systems—for example, in seeding for marketing purposes. Defining node centrality for a given application has generated substantial theoretical interest in IS, network science, and economics research (Sundararajan et al. 2013, Liu 2019). Because centrality measures can be used to

[10] We do not present the result after the fifth hop because the value of interest converges.

understand how diffusion processes on digital and information networks may alter a wide variety of economic outcomes, centrality's definition and quantification may vary, depending on the substantive settings. One immediate implication of the long-range effect of social influence in Section 4.1 is the development of a new context-dependent centrality measure, which we call *influence centrality*. *Influence centrality* is designed to quantify the structural importance of nodes, relating the overall increase in expected adoption to a node that serves as the "injection" node (i.e., the individual who is seeded to diffuse a certain product or behavior).

5.1. Influence centrality: A new centrality based on social influence effect

Consider a marketing firm or a public health agency that aims to use WOM through mobile phone communications to spread information about a product or health-related behavior—for example, an offline event in the former case or the benefits of immunization in the latter. To what degree does the overall increase in expected adoption rely on who the firm or agency approaches first (e.g., to offer free tickets or free trials)? Given that we are interested in social influence, this question is different from questions about increasing the spread of information in the network, which is the motivation behind many widely applied centrality measures. Our centrality measure answers this question by quantifying the importance of a user in the network with regard to the *increase in expected adoption* if this user is the only one initially informed.

The measure is defined in a random walk fashion using an independent cascade model (Easley and Kleinberg 2010). Assume that each informed individual calls a neighbor in the social network with probability p , and that call from the seeded individual increases the adoption likelihood of immediate neighbors by γ_1. As social influence propagates, it reaches longer distances over the social network; we assume that this initial seed increases the adoption likelihood of each of the second-degree neighbors (who has been called with probability p^2) by γ_2, of each third-degree neighbor (who has been called with probability p^3) by γ_3, and so on. Recall that the adjacency matrix of the historical social network is $\mathbf{A}$. The two-hop adjacency matrix, which captures individuals who can reach one another by two hops, is $\mathbf{A}^2$. Similarly, the h-hop adjacency matrix $\mathbf{A}^h$ measures the expected number of walks of length h between each pair of individuals. The diffusion process continues until a fixed degree of separation is reached. We can therefore define *influence centrality*, abbreviated as IC, as follows:

$$IC(\mathbf{A}; p, \boldsymbol{\gamma}, H) = \sum_{h=1}^{H} \gamma_h (p\mathbf{A})^h \cdot \mathbf{1}, \tag{8}$$

where H is the maximal reach of the social influence (we choose $H = 4$ as informed by our empirical results in Section 4.1), and $\mathbf{1}$ is an all-one vector. The values of $\boldsymbol{\gamma} = \{\gamma_1, ..., \gamma_H\}$ are estimated empirically using our framework and account for the decaying strength of social influence.

IC bears similarities to and generalizes prevailing centrality measures. The main difference between IC and other centrality measures is its focus on increasing social influence and expected adoption, as well as its capability of accounting for heterogeneity between neighbors at different hop indices (via $\boldsymbol{\gamma}$). From this perspective, IC can be perceived as assigning a weight γ_h to edges in a random walk matrix $(p\mathbf{A})^h$, where the edge weights can be estimated empirically through the technical framework in Section 3. If $\boldsymbol{\gamma} = \{1\}_{h=1}^{H}$, $p = 1$, and $H = 1$, IC is proportional to the degree centrality. If $\boldsymbol{\gamma} = \{1\}_{h=1}^{H}$ and $H = \infty$, IC becomes proportional to either the Katz centrality or the eigenvector centrality, depending on whether p is smaller than the inverse of the largest eigenvalue of $\mathbf{A}$ or not.[11] Finally, IC is most similar to the diffusion centrality proposed in Banerjee et al. (2013) among all centrality measures. However, the focus of the former is to amplify the influence on adoption (hence different γ_h for different h) while the latter (and indeed most centralities in the literature) is on the spread or diffusion of information (hence γ_h is homogeneous across different h). In all these approaches, the communication (diffusion) probability p can be estimated either empirically or using simulations to test for robustness. For example, we estimate p from the historical data: On each day of the month prior to the event, each individual communicated, on average, with 7% of the individuals they communicated with during the whole month. Hence p is set to be 0.07 in our analysis.

Following the evaluation procedure described in Banerjee et al. (2013), we conduct regression analyses to evaluate the predictive power of mean and median centrality of those who have information about the product initially (e.g., those who attended the event) on the number of eventual adopters for each observation period. To perform the analyses, we first compute degree centrality, eigenvector centrality, Katz centrality, diffusion centrality and *influence centrality* using the historical social network[12]. Then, for each observation period, we compute the mean and median centrality of initial adopters using any centrality measure above (as the independent variables), and record the total number of individuals that eventually adopted (as the dependent variable) after excluding initial adopters. This processing step gives us a pair of data points for each of the 19 observation periods. To test the predictive power of different centralities toward the number of adoptions, we then regress the dependent variable (total number of adopters) on the independent variables (mean and median of a certain centrality of initial adopters) and show the coefficient

[11] For both the eigenvector centrality and the Katz centrality, p needs to be smaller than the inverse of the largest eigenvalue of the adjacency matrix $\mathbf{A}$.

[12] We compute degree centrality by summing the number of contacts of each initial adopter, normalized by $N-1$, where N is the number of individuals who appear in the historical social network $\mathbf{A}$. The eigenvector centrality is based on the leading eigenvector of $\mathbf{A}$. We set the diffusion probability p in diffusion centrality, Katz centrality, and influence centrality to be 0.07. We compute the diffusion centrality using H as the diameter of the largest connected component of the historical social network.

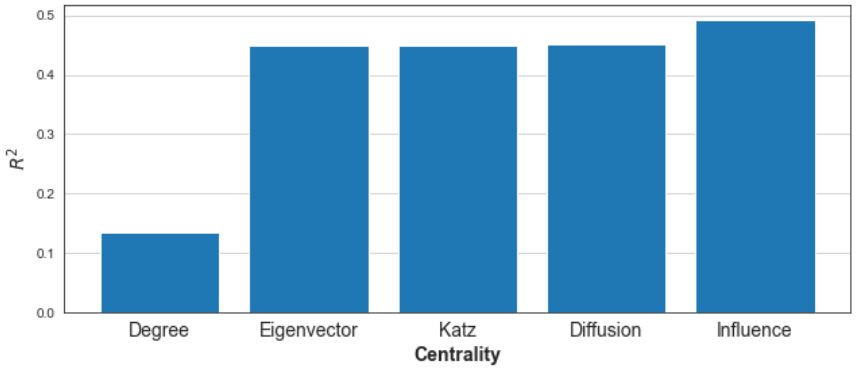


Figure 7 **The proportion of variance in the total number of adopters explained by the centrality of the initial adopters.**

of determination (R^2) of the regression model in Figure 7. We can see that IC outperforms the other centrality measures in predicting the number of adopters; hence, it is a stronger predictor of adoption behavior, thanks to accounting for quantitative estimates of social influence over successive connections. This experiment suggests the possibility of IC being used to inform strategies for commercial firms in promoting product adoption (Aral et al. 2009), for the government in encouraging voter turnout (Bond et al. 2012), and for public health agencies in promoting immunization programs (Banerjee et al. 2019).

5.2. Policy implication in seeding and viral marketing

The proposed influence centrality, combined with the empirical results in Section 4.1, has a number of policy implications because managers and campaigners can use such knowledge to improve their decision-making as they promote new products or behaviors. In this section, we quantitatively demonstrate how raising the communication probability over each connection (thus promoting phone communications) and the number of initial seeds can significantly amplify the expected adoption. First, IC can be applied in seeded marketing to identify a set of individuals for targeted interventions that can maximize overall likelihood of adoption.[13]. Second, we quantitatively demonstrate in a simulated environment where high-centrality nodes are targeted to be the initial adopters, raising the communication probability over each connection and the number of initial seeds can significantly amplify the overall expected adoption and may be desirable despite the cost of these interventions. This exercise can inform optimal seeding in viral and targeted marketing campaigns.

[13] Although marketing agencies can solve an optimization problem to determine the optimal set of seeds, this calculation is seldom done in practice. The reason is that the influence maximization problem, using an independent cascade model, is a non-deterministic polynomial-time hardness (NP-hard) problem for which approximations are needed but computationally expensive given the large size of social networks.

To further investigate these policy implications, we conduct two analyses. The simulation procedure used in these analyses is summarized in the following steps:

1. Compute IC on the historical social network $\mathbf{A}$.

2. Rank the centrality measures and select the top m individuals as seeds (initial adopters), where m is a pre-determined number.

3. For neighbors of each seeded individual (for the first iteration) or neighbors of individuals who have received influence in the previous iteration but has not yet diffused influence (for subsequent iterations), decide whether each specific neighbor receives the phone call (hence influence), according to a pre-determined communication probability p. If the individual receives the phone call, then the adoption likelihood for that individual will increase by γ_h, where h is the individual's distance to the initial adopter in the network. Keep track of the total increase (from multiple iterations) in adoption likelihood for each individual who has received influence. The upper bound of the total increase in the adoption likelihood is 1.

4. Start subsequent iterations by repeating step 3 up to the fourth iteration (i.e., up to hop-4 neighbors of initial adopters, inspired by our empirical results).

5. Compute the increase in *expected adoption* by summing up the effect of social influence ($\boldsymbol{\gamma}$) on all individuals who receive the treatment (based on their distance to the initial adopter). The hop index of a specific individual who has received the information is based on the degree of separation from an initial adopter.

Two factors may play a role in the total increase in expected adoption: the number of seeds m and the communication probability p. In the following sections, we first examine how these two factors are related to the increase in expected adoption (Section 5.2.1); we then examine two marketing strategies based on these two factors, analyzing the cost and benefit of these strategies (Section 5.2.2).

5.2.1. Expected adoption with respect to seeding and communication probability. Figure 8 shows the increase in expected adoption as a function of the two factors we consider: the number of seeds and the phone communication probability. In Figure 8(a), we see that the increase in communication probability leads to an increase in the expected adoption, as expected. However, this effect is not linear; instead, two phases of transition indicate where the increase is more significant: The increase (on a logarithmic scale) is the most significant before reaching the probability of $p = 0.2$, and it slows down up to $p = 0.4$. After this point, the effect of seeding becomes saturated. A similar pattern can be observed in Figure 8(b), which shows the effect of the number of seeds. Indeed, the initial 20 to 30 seeds lead to a significant increase in expected adoption (again on a logarithmic scale) before the effect saturates. This saturation is understandable because the

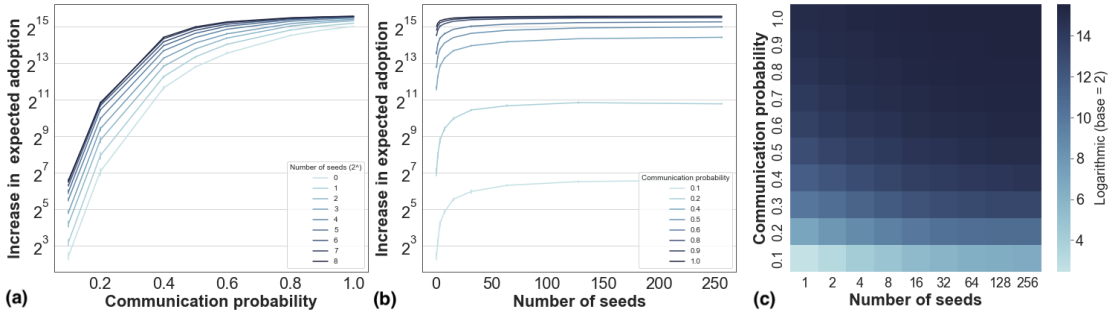


Figure 8 **Overall increase in expected adoption based on the different (a) communication probability, and (b) number of initial seeds. Panel (c) shows the joint effect of these two factors, with color in the heatmap indicating the increase in expected adoption.**

seeds were chosen according to decreasing influence centrality. We could also investigate the joint effect of the two strategies, i.e., increasing the number of seeds (m) and promoting the probability of communication (p), which is illustrated in Figure 8(c). These results confirm that both the number of seeds and the communication probability play a role in the increase in expected adoptions. The former is related to seeded marketing, and the latter to viral marketing. This analysis motivates us to examine two marketing strategies based on these two factors in the following section.

5.2.2. Cost-benefit analysis in seeding and viral marketing. To analyze the cost-effectiveness of the two factors discussed in the previous section, we examine two marketing strategies in seeded WOM and viral marketing. We first describe the strategies as follows.

1. [**Strategy 1**] Seed m individuals, through promotional offers and/or free tickets. This strategy has a direct effect on increasing expected adoption.

2. [**Strategy 2**] Seed m individuals; in addition, promote phone communication across the network by designing viral features into the advertising content (Aral and Walker 2011). Either the marketing team of an advertising firm or a third-party content marketing firm can design such features. As a result of this action, the communication probability p increases in all iterations of step 3 of the simulation, resulting in an increase in expected adoption.

For strategy 1, we consider a base communication probability of $p = 0.07$, which is similarly estimated from historical data as in Section 5.1. For strategy 2, we consider three levels of marketing services that involve engineering viral content. These levels lead to adoption probabilities of $p = 0.12$ (Basic), $p = 0.15$ (Pro), and $p = 0.17$ (Diamond), respectively.[14], respectively.

We set the cost of adding a seed to be c_s and the cost of designing viral features to increase p to be c_v (this cost increases as we go from the Basic level to the Diamond level). Thus, the cost of the two strategies is as follows:

[14] Note that we use these specific probabilities as illustrative examples. Marketing firms can estimate these costs according to their context.

1. **[Strategy 1]**: $c_s \times m$.
2. **[Strategy 2]**: $c_s \times m + c_v$.

We collect statistics from real-world data to make the simulation more realistic. We set the benefit of every 100% increase[15] in expected adoption at \$88 (the average ticket price for the offline event in this study) and compute the total benefit as this amount times the increase in units of expected adoption. We set $c_s = \$22$ ($\frac{1}{4}$ of the ticket price) and $c_v \in \{\$202, \$337, \$560\}$ for the three levels of services for viral content design[16]. By subtracting the cost from the benefit, we can analyze the cost-effectiveness of the strategies under different numbers of seeds m.

Figure 9 shows the net benefit (benefit minus cost) of the strategies as a function of the number of seeds. Note that the "no viral marketing" scenario corresponds to Strategy 1, while the other three scenarios correspond to Strategy 2. As expected, the net benefit increases as the number of seeds increases for all levels of viral marketing efforts. The addition of viral features leads to an increased net benefit in all scenarios; the increase is most pronounced when the number of seeds is small, although this effect becomes saturated when more seeds are added. In addition, for the same level of net benefit, one can either improve the viral content or increase the number of seeds. For example, using the "Pro" service and seeding eight individuals generates roughly the same net benefit, compared with using the "Basic" service and seeding thirty-two individuals. These findings demonstrate that both a seeding strategy based on IC and promoting communications in the social network using viral features are effective in promoting adoption. Engineering viral content to increase communication probability is particularly effective when the number of seeds is small (toward the left end of Figure 9), e.g., practical constraints limit the ability to expand seeding. The marketing strategies may both be profitable, despite their costs, and a cost-benefit analysis can be used to estimate their net benefit.

6. Discussion

Phone communications play a crucial role in facilitating information exchange due to their unique characteristics, as shown in Table 1. The availability of large-scale and longitudinal mobile phone communication data and the associated mobility information from CDRs has allowed us to identify social influence on one's immediate and distant neighbors in the phone communication network. In this study, we propose a new technical framework to investigate how social influence spreads

[15] This 100% increase in expected adoption could result from, for example, 30% increase in expected adoption for an individual A and 70% for another B.

[16] We obtained these reference prices from the following content marketing platform: `https://z3i.zerys.com/#/pricingcalculator`. Figure H1 of Appendix H shows the prices for the three tiers of content marketing services to create viral content. These prices are adopted for illustrative purposes and marketing firms can estimate them based on their own context.

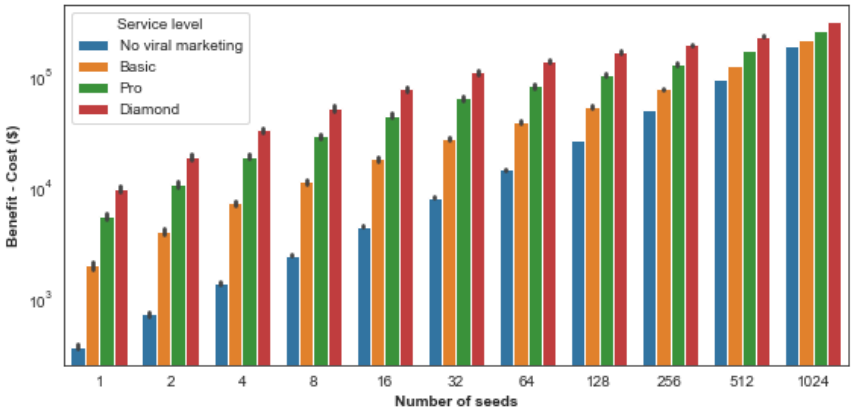


Figure 9 **Policy simulations using the two marketing strategies, with different service levels and number of seeds.**

through this communication channel. Our findings demonstrate that social influence through phone communication can impact an offline adoption decision up to four degrees of separation in the phone communication network. This finding improves our fundamental understanding of how social influence spreads through an under-explored phone communication channel. Moreover, our empirical results have inspired the development of a new centrality measure, *influence centrality*, which evaluates the structural importance of nodes in amplifying expected adoption. This centrality measure offers a new perspective on leveraging the complex structure of social networks for marketing purposes via mobile phone communications, thus expanding the existing literature in network science. This measure provides a fresh perspective on using the complex structure of social networks for marketing purposes through mobile phone communications, expanding the existing literature in network science.

6.1. Theoretical and managerial implications

A quantitative framework for studying social influence via phone communication. Despite the widespread use of mobile phones and their potential for mobile advertising campaigns, understanding social influence on adoption behaviors through phone communication has been hampered by the lack of practical tools for identifying influence in large-scale networks. Our study proposes a technical framework for studying the impact of social influence mediated through phone communications using CDRs, which have become increasingly accessible in recent years (see Appendix I). Our framework has several potential and practical implications: 1) Our methodology for isolating social influence from homophily (and in particular, both observed and latent homophily) using social interaction and behavioral data can be helpful in empirical IS research when socio-demographic information is not available; 2) Our framework can be applied to other adoption decisions and other types of social interaction data, such as Facebook, Twitter, and Yelp,

or communication media, such as video calls or text messages. Overall, our analysis demonstrate the potential of combining large-scale spatial-temporal data and network mining with econometric models to better understand and quantify social influence. As discussed in Section 5.2, this understanding and quantitative estimate can lead to more effective strategies in seeded and viral marketing.

Seeded WOM and viral marketing. Seeded WOM and viral marketing are popular techniques used in the advertising industry, as well as in public health campaigns and government initiatives. Despite their effectiveness, identifying the right individuals to seed remains a challenge. In network contexts, centrality measures are often used to select influential seeds. However, existing centrality measures focus on information spread and diffusion, while neglecting the importance of social influence. To address this gap, we propose a new centrality measure called Influence Centrality (IC). Unlike existing measures, IC focuses on amplifying social influence, leading to an increase in expected adoption. Our empirical findings on the decaying patterns of social influence reveal the heterogeneity of influence across different hop indices from a focal individual's perspective. Additionally, IC's context-dependent nature and ability to capture heterogeneity can lead to more effective marketing strategies for commercial firms, as well as for successful campaigns for humanitarian and public health goals. Our study offers new perspectives on developing targeted seeding strategies and identifying influential individuals in social networks. By incorporating IC into seeding and viral marketing strategies, organizations can more effectively harness the power of social influence to achieve their goals.

Extending hyper-contextual mobile targeting to phone communication networks. Mobile targeting enables personalized advertising based on hyper-contextual insights derived from mobile phone usage data, including location (where), time (when), search behavior (how and what), and co-presence with others (with whom). Our study builds on this theory by extending the concept of co-presence (with whom) to include phone communication networks. We demonstrate that social influence can spread through phone communication networks, allowing firms to target individuals who have interacted with recent product adopters, even indirectly. Our findings highlight the importance of considering phone communication networks when designing hyper-contextual mobile targeting strategies.

6.2. Limitations and future work

Our study has several limitations that provide avenues for future research. First, although we measure observed and latent homophily by analyzing detailed behavioral information, the CDR data are not comprehensive and cannot capture social interactions that take place through other communication channels (for example, online or email interactions). This limitation in observability is

a general concern for most, if not all, social influence studies using data collected from one digital platform (such as online social media (Bond et al. 2012) or messaging apps (Aral and Walker 2014)): Due to ethical and privacy considerations, and the technical challenge in merging social interactions from multiple communication media, most studies only obtain social interactions from one medium. Consequently, our method, relying on the CDR data, establishes upper bounds on influence estimates when communications through other channels are unobserved. Second, we do not observe the content of phone communications due to data privacy and confidentiality reasons. As a result, the social influence effect on event attendance that we intend to measure may not have taken place through phone calls. Future studies might use surveys, similar to that in Lovett et al. (2013), to assess the probability of relevant information being spread through phone communications. Third, due to the lack of sufficient data, we investigate the treatment effects of a single communication path between the initial adopter and an individual a certain distance away in the communication work. Future studies might consider multiplicative effects of social influence with multiple communication paths. Fourth, our empirical context focuses on attending an offline performance event, so the generalizability of the findings is limited to similar offline behaviors. Future studies may apply the proposed framework to investigate which real-world offline behaviors are amenable to phone communication. Our framework might also be used to study heterogeneity in the effects of social influence (e.g., with respect to factors such as tie strength) or how social influence varies as time elapses.

Acknowledgments

The authors would like to thank the senior editor, associate editor, and the anonymous reviewers for their constructive comments and suggestions throughout the review process. The authors would also like to thank Ernest Liu and Francis DiTraglia for helpful discussion and suggestions. Additionally, Yan Leng acknowledges the support provided by the NSF grant IIS-2153468.

References

Acemoglu D, Dahleh MA, Lobel I, Ozdaglar A (2011) Bayesian learning in social networks. *The Review of Economic Studies* 78(4):1201–1236.

Anagnostopoulos A, Kumar R, Mahdian M (2008) Influence and correlation in social networks. *Proceedings of 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 7–15 (ACM).

Aral S, Muchnik L, Sundararajan A (2009) Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences* 106(51):21544–21549.

Aral S, Walker D (2011) Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Science* 57(9):1623–1639.

Aral S, Walker D (2014) Tie strength, embeddedness, and social influence: A large-scale networked experiment. *Management Science* 60(6):1352–1370.

Ballester C, Calvó-Armengol A, Zenou Y (2006) Who's who in networks. Wanted: The key player. *Econometrica* 74(5):1403–1417.

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 (2019) Using gossips to spread information: Theory and evidence from two randomized controlled trials. *The Review of Economic Studies* 86(6):2453–2490.

Barbour E, Davila CC, Gupta S, Reinhart C, Kaur J, González MC (2019) Planning for sustainable cities by estimating building occupancy with mobile phones. *Nature Communications* 10(1):1–10.

Belloni A, Chernozhukov V, Wei Y (2013) Honest confidence regions for a regression parameter in logistic regression with a large number of controls. Technical report, Cemmap working paper, Centre for Microdata Methods and Practice.

Belo R, Ferreira P (2022) Free-riding in products with positive network externalities: Empirical evidence from a large mobile network. *Management Information Systems Quarterly* 46(1):401–430.

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.

Bloch F, Bloch F, Jackson MO, Tebaldi P (2019) Centrality measures in networks. *Available at SSRN: https://ssrn.com/abstract=2749124* .

Bond RM, Fariss CJ, Jones JJ, Kramer AD, Marlow C, Settle JE, Fowler JH (2012) A 61-million-person experiment in social influence and political mobilization. *Nature* 489(7415):295–298.

Burt RS (1987) Social contagion and innovation: Cohesion versus structural equivalence. *American journal of Sociology* 92(6):1287–1335.

Christakis NA, Fowler JH (2013) Social contagion theory: examining dynamic social networks and human behavior. *Statistics in Medicine* 32(4):556–577.

Cohen J (1988) Statistical power analysis for the behavioral sciences. *NJ: Lawrence Earlbaum Associates* 2.

de Matos MG, Ferreira P, Krackhardt D (2014) Peer influence in the diffusion of iphone 3g over a large social network. *Management Information Systems Quarterly* 38(4):1103–1133.

den Bulte CV, Lilien GL (2001) Medical innovation revisited: Social contagion versus marketing effort. *American Journal of Sociology* 106(5):1409–1435.

Dennis AR, Kinney ST (1998) Testing media richness theory in the new media: The effects of cues, feedback, and task equivocality. *Information Systems Research* 9(3):256–274.

Dewan S, Ho YJ, Ramaprasad J (2017) Popularity or proximity: Characterizing the nature of social influence in an online music community. *Information Systems Research* 28(1):117–136.

Dou Y, Niculescu MF, Wu D (2013) Engineering optimal network effects via social media features and seeding in markets for digital goods and services. *Information Systems Research* 24(1):164–185.

Easley D, Kleinberg J (2010) *Networks, crowds, and markets: Reasoning about a highly connected world* (Cambridge University Press).

Eckles D, Bakshy E (2020) Bias and high-dimensional adjustment in observational studies of peer effects. *Journal of the American Statistical Association* 1–11.

Ghose A, Li B, Liu S (2019) Mobile targeting using customer trajectory patterns. *Management Science* 65(11):5027–5049.

Golub B, Jackson MO (2010) Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics* 2(1):112–149.

Grover A, Leskovec J (2016) node2vec: Scalable feature learning for networks. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 855–864.

Hanck C, Arnold M, Gerber A, Schmelzer M (2019) Introduction to econometrics with *r*. *University of Duisburg-Essen* 1–9.

Hausman J (2001) Mismeasured variables in econometric analysis: problems from the right and problems from the left. *Journal of Economic Perspectives* 15(4):57–67.

Hu M, Yang S, Xu Y (2019) Understanding the social learning effect in contagious switching behavior. *Management Science* 65(10):4771–4794.

Iacus SM, King G, Porro G (2012) Causal inference without balance checking: Coarsened exact matching. *Political Analysis* 1–24.

Imbens GW, Rubin DB (2015) *Causal inference in statistics, social, and biomedical sciences* (Cambridge University Press).

Jackson MO (2008) *Social and economic networks* (Princeton University Press).

Jackson MO (2019) The friendship paradox and systematic biases in perceptions and social norms. *Journal of Political Economy* 127(2):777–818.

Jones KH, Daniels H, Heys S, Ford DV (2018) Challenges and potential opportunities of mobile phone call detail records in health research. *JMIR mHealth and uHealth* 6(7):e9974.

Jung J, Bapna R, Ramaprasad J, Umyarov A (2019) Love unshackled: Identifying the effect of mobile app adoption in online dating. *Management Information Systems Quarterly* 43:47–72.

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 (ACM).

Kremer M, Miguel E (2007) The illusion of sustainability. *The Quarterly Journal of Economics* 122(3):1007–1065.

Leng Y, Babwany NA, Pentland A (2021a) Unraveling the association between socioeconomic diversity and consumer price index in a tourism country. *Humanities and Social Sciences Communications* 8(1):1–10.

Leng Y, Noriega A, Pentland A (2021b) Tourism event analytics with mobile phone data. *ACM/IMS Trans. Data Sci.* 2(3), ISSN 2691-1922, URL `http://dx.doi.org/10.1145/3479975`.

Leng Y, Sella Y, Ruiz R, Pentland A (2020) Contextual centrality: going beyond network structure. *Scientific Reports* 10(1):1–10.

Leng Y, Zhao J, Koutsopoulos H (2021c) Leveraging individual and collective regularity to profile and segment user locations from mobile phone data. *ACM Transactions on Management Information Systems (TMIS)* 12(3):1–22.

Lewbel A (2007) Estimation of average treatment effects with misclassification. *Econometrica* 75(2):537–551.

Li X (2016) Could deal promotion improve merchants' online reputations? the moderating role of prior reviews. *Journal of Management Information Systems* 33(1):171–201.

Li Y, Fan J, Wang Y, Tan KL (2018) Influence maximization on social graphs: A survey. *IEEE Transactions on Knowledge and Data Engineering* 30(10):1852–1872.

Liu E (2019) Industrial policies in production networks. *The Quarterly Journal of Economics* 134(4):1883–1948.

Lovett MJ, Peres R, Shachar R (2013) On brands and word of mouth. *Journal of Marketing Research* 50(4):427–444.

Ma L, Krishnan R, Montgomery AL (2015) Latent homophily or social influence? an empirical analysis of purchase within a social network. *Management Science* 61(2):454–473.

Mallipeddi RR, Kumar S, Sriskandarajah C, Zhu Y (2022) A framework for analyzing influencer marketing in social networks: selection and scheduling of influencers. *Management Science* 68(1):75–104.

Manski CF (1993) Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies* 60(3):531–542.

Matous P, Todo Y, Ishikawa T (2014) Emergence of multiplex mobile phone communication networks across rural areas: An ethiopian experiment. *Network Science* 2(2):162–188.

McFadden D (2001) Economic choices. *American economic review* 91(3):351–378.

McFowland III E, Shalizi CR (2021) Estimating causal peer influence in homophilous social networks by inferring latent locations. *Journal of the American Statistical Association* 1–12.

McPherson M, Smith-Lovin L, Cook J (2001) Birds of a feather: Homophily in social networks. *Annual Review of Sociology* 27:415–444.

Mobius M, Rosenblat T (2014) Social learning in economics. *Annual Review of Economics* 6(1):827–847.

Newman M (2018) *Networks* (Oxford University Press).

Oliver N, Lepri B, Sterly H, Lambiotte R, Deletaille S, De Nadai M, Letouzé E, Salah AA, Benjamins R, Cattuto C, et al. (2020) Mobile phone data for informing public health actions across the covid-19 pandemic life cycle. *Science Advances* 6(23):eabc0764.

Ransbotham S, Lurie NH, Liu H (2019) Creation and consumption of mobile word of mouth: how are mobile reviews different? *Marketing Science* 38(5):773–792.

Rao N, Mobius M, Rosenblat T (2017) Social networks and vaccination decisions. *FRB of Boston Working Paper No. 07-12* .

Rishika R, Kumar A, Janakiraman R, Bezawada R (2013) The effect of customers' social media participation on customer visit frequency and profitability: an empirical investigation. *Information Systems Research* 24(1):108–127.

Rishika R, Ramaprasad J (2019) The effects of asymmetric social ties, structural embeddedness, and tie strength on online content contribution behavior. *Management Science* 65(7):3398–3422.

Rosenbaum PR (2005) *Observational Study* (John Wiley & Sons, Ltd), ISBN 9780470013199, URL `http://dx.doi.org/https://doi.org/10.1002/0470013192.bsa454`.

S Suri DW (2011) Cooperation and contagion in web-based, networked public goods experiments. *PLoS ONE* 6(3):e16836.

Salah AA, Pentland A, Lepri B, Letouzé E (2019) *Guide to Mobile Data Analytics in Refugee Scenarios* (Springer).

Samuelson PA (1938) A note on the pure theory of consumer's behaviour. *Economica* 5(17):61–71.

Sen M (2014) How judicial qualification ratings may disadvantage minority and female candidates. *Journal of Law and Courts* 2(1):33–65.

Shalizi CR, Thomas AC (2011) Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research* 40(2):211–239.

Stewart MB, Swaffield JK (2008) The other margin: do minimum wages cause working hours adjustments for low-wage workers? *Economica* 75(297):148–167.

Stuart EA, Lee BK, Leacy FP (2013) Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *Journal of clinical epidemiology* 66(8):S84–S90.

Sundararajan A, Provost F, Oestreicher-Singer G, Aral S (2013) Research commentary – information in digital, economic, and social networks. *Information Systems Research* 24(4):883–905.

Toole JL, Lin YR, Muehlegger E, Shoag D, González MC, Lazer D (2015) Tracking employment shocks using mobile phone data. *Journal of The Royal Society Interface* 12(107):20150185.

Ucar I, Gramaglia M, Fiore M, Smoreda Z, Moro E (2021) News or social media? socio-economic divide of mobile service consumption. *Journal of The Royal Society Interface* 18(185):20210350.

Zhang B, Pavlou PA, Krishnan R (2018) On direct vs. indirect peer influence in large social networks. *Information Systems Research* 29(2):292–314.

Zhang J (2010) The sound of silence: Observational learning in the us kidney market. *Marketing Science* 29(2):315–335.

Zhang Y, Li B, Luo X, Wang X (2019) Personalized mobile targeting with user engagement stages: Combining a structural hidden markov model and field experiment. *Information Systems Research* 30(3):787–804.