Appendix for Long-Range Social Influence in Phone Communication Networks on Offline Adoption Decisions

A. A brief overview of *node2vec*

Node2vec is a machine learning algorithm designed for node embedding, developed by Grover and Leskovec (2016). The key objective of this algorithm is to map nodes in a graph into vectors within a lower-dimensional space, ensuring that similar nodes end up being represented closely in this new space. The inner workings of *node2vec* are based on a random walk procedure, which generates a series of node sequences. This model treats the node sequences as sentences and aim to predict the context (neighboring nodes) given a target node. By training on these node sequences, *node2vec* learns representations (embeddings) for each node that encode the structural properties of the graph.

The optimization technique employed by *node2vec* is designed to maximize the log-probability of observing a specific network neighborhood, denoted as $\mathcal{N}^b(i)$, for a given node i, in light of its corresponding mapping function $f(i)$. The objective can be formally expressed as:

$$\max_f \sum_{i \in \mathcal{V}} \log Pr(\mathcal{N}^b(i)|f(i))$$

where $f(i)$ signifies the function that projects node i into a lower-dimensional vector, represented as $\mathbf{c}_i$; $Pr(\mathcal{N}^b(i)|f(i))$ is the likelihood of observing a neighborhood for a node i conditioned on its features $f(i)$; $\mathcal{V}$ is the node set.

Additionally, *node2vec* includes a random walk mechanism that is characterized by several parameters: a return parameter q^r, which influences the probability of an immediate revisit to a node during the walk; an in-out parameter q^{io}, which distinguishes between "inward" and "outward" nodes; the length of the walk l; and the dimensionality of the latent representations d_c. The tuning of these parameters aims to maximize their predictive power on adoption decisions, with the optimal parameters found to be $q^r = 0.25, q^{io} = 2, l = 27$, and $d_c = 16$.

B. Measurement errors in using phone data to estimate social influence

In this discussion, we primarily address two main sources of measurement errors associated with the use of phone data to estimate social influence. The first substantial source of measurement error stems from the challenging task of accurately identifying adoption decisions. There are several scenarios that could complicate this process. For instance, adopters may not use their phones when attending the event. Alternatively, some individuals might be near the event venue but not participate in the event. Both scenarios could result in inaccuracies when identifying initial adopters and their adoption decisions within both the treatment and control groups. These measurement errors consequently impact two significant variables: 1) the adoption decisions of individuals in the treatment and control groups z_{sjt}; and 2) the identification of the initial adopters, which leads to errors in the treatment D_{sj}. The second primary source of measurement error is attributed to the constraints of our 24-hour observation window to construct communication cascades. Any communication involving initial adopters occurring directly or indirectly beyond this window will introduce additional measurement errors to the treatment variable, D_{sj}.

In summary, these two sources of measurement errors can be considered as inaccuracies in the outcome (left-hand side, attributed to the first source) and predictor (right-hand side, attributed to both sources) variables within the regression model, as outlined in Equation (1). In econometrics, it is well-documented that the OLS estimate is downward biased in the case of a mismeasured predictor, while a mismeasured outcome does not lead to a bias. This result is based on classical mismeasurement assumptions, which assumes that the error of measurement is not correlated with the true variables and is not correlated with stochastic disturbance in the regression specification (Hausman 2001, Lewbel 2007). We discuss both factors in detail below.

Mismeasurement in the adoption outcomes. It is reasonable to assume that the patterns of individuals' phone usage, particularly the probability of them not using their phones while attending the event, are equivalent for both the treatment and control groups, after accounting for latent positions derived from the historical social network and the eventual adoption decisions of neighbors. Therefore, any mismeasurement in the outcome variable is subsumed into the error term, ϵ_{it}, of Equation (1). This mismeasurement does not influence the estimation of the treatment effect.[17] In other words, random errors in the dependent variable do not compromise the consistency of the treatment effect estimate, γ_h. We formally demonstrate this with the derivation of an OLS

[17] For clarity, we adopt i instead of sj as the subscript in this section.

estimator.[18] Consider a mismeasured adoption decision z_{it}, in which u_i^z is the measurement error. We then have the following equation:

$$z_{it} = z_{it}^* + u_i^z = \mathbf{x}_i'\boldsymbol{\alpha}_x + \mathbf{c}_i'\boldsymbol{\alpha}_c + \mathbf{f}_i'\boldsymbol{\alpha}_f + \gamma_h D_i \text{ after}_{it} + \pi_{D_i=1} + \pi_{\text{after}_{it}=1} + \eta_s + \nu_t + \underbrace{(\epsilon_{it} + u_i^z)}_{\epsilon_{it}^*}, \quad (9)$$

where z_{it}^* is the ground-truth outcome. The new error term, $\epsilon_{it}^* = \epsilon_{it} + u_i^z$, is a zero-mean random variable because both ϵ_{it} and u_i^z are random variables having a zero mean. Therefore, a random measurement error in the adoption outcome does not introduce bias in the treatment effect estimation γ_h.

Mismeasurement of the treatment. Measurement errors among the initial adopters or direct/indirect communication with them after the observational period can lead to misclassification of the treatment variable. Formally, let us consider a scenario where a measurement error occurs in the treatment variable. Instead of observing the the ground-truth variable $D_i \text{ after}_{it}^*$, we observe $D_i \text{ after}_{it}$. Recalling that $\mathbf{z}$ represents the adoption behavior, and $\mathbf{x}$, $\mathbf{c}$, and $\mathbf{f}$ are the control covariates, we define $\mathbf{B}$ as the concatenation of all these control variables. We outline two necessary assumptions, following assumptions A1 and A2 in Lewbel (2007),

ASSUMPTION 1. *There exists* $\mathbb{E}(\mathbf{z}|\mathbf{B}, D \textit{ after}^*, D \textit{ after}) = \mathbb{E}(\mathbf{z}|\mathbf{B}, D \textit{ after}^*)$.

To enhance clarity, we omit the $\mathbf{B}$ and b argument in the following discussion. Assumption 1 implies that $\mathbf{z}$ is mean independent of $D \text{ after} - D \text{ after}^*$, conditional on $\mathbf{B}$ and $D \text{ after}^*$. This means that the misclassification in the treatment does not impact the true expected adoption behavior. However, this assumption would be violated if misclassification occurs due to misperception or deceit by individuals in either the treatment or control group. Fortunately, the mechanical misclassification of initial adopters, such as individuals who did not use phones during the performance, does not influence the true adoption behaviors of other individuals in both the treatment and control groups. In our scenario, the misclassification of initial adopters is unobservable to individuals in both groups, eliminating the possibility of misperception or deceit. Consequently, this assumption, akin to the classical assumption of independent measurement errors, is reasonable in our setting.

Before we proceed to the next assumption, we define some useful functions. We first define function $r^*(b)$ as the ground-truth conditional probability (on $\mathbf{B} = b$) of receiving the treatment:

$$r^*(b) = \mathbb{E}(D \text{ after}^*|\mathbf{B} = b) = \mathbb{P}(D \text{ after}^* = 1|\mathbf{B} = b).$$

We similarly define $r(b)$ by replacing $D \text{ after}^*$ with $D \text{ after}$:

$$r(b) = \mathbb{E}(D \text{ after}|\mathbf{B} = b) = \mathbb{P}(D \text{ after} = 1|\mathbf{B} = b).$$

[18] The proof in this section is conducted using the OLS estimator because the fixed-effect models can be estimated using the so-called entity-demeaned OLS (i.e., by removing the time-specific and matched-pair-specific means).

We define $\beth_1(b)$ and $\beth_0(b)$ as the conditional probabilities of misclassifying the treated individuals and the control individuals, respectively. We have:

$$\beth_d(b) = \mathbb{P}(D \text{ after} = 1 - d | \mathbf{B} = b, D \text{ after}^* = d).$$

We define the ground-truth conditional mean adoption decision, given $\mathbf{B}$ and D after* as:

$$h^*(\mathbf{B}, D \text{ after}^*) = \mathbb{E}(\mathbf{z}|\mathbf{B}, D \text{ after}^*) = h_0^*(\mathbf{B}) + \gamma^*(b) D \text{ after}^*, \tag{10}$$

where $h_0^*(\mathbf{B}) = h^*(\mathbf{B}, 0)$. We define $\gamma^*(b)$ as the ground-truth conditional average treatment effect,

$$\gamma^*(b) = h^*(b, 1) - h^*(b, 0).$$

And $\gamma(b)$ is the estimated (and mismeasured) conditional average treatment effect:

$$\gamma(b) = h(b, 1) - h(b, 0),$$

where $h(\mathbf{B}, D \text{ after}) = \mathbb{E}(\mathbf{z}|\mathbf{B}, D \text{ after})$.

ASSUMPTION 2. *There exists* $\beth_0(b) + \beth_1(b) < 1$ *and* $0 < r^*(b) < 1$ *for all* $b \in$ *support* $(\mathbf{B})$.

The first inequality indicates that the sum of the misclassification probabilities is less than 1, suggesting that, on average, observations of D after are more accurate than random guesses. Therefore, this assumption can be easily satisfied in practice. The second part of this assumption requires the presence of at least one individual in both the treatment and control groups. Under these two assumptions, Proposition 1 follows exactly from Theorem 1 of Lewbel (2007). It illustrates that the mismeasurement error in the treatment variable biases the estimate towards zero.

PROPOSITION 1. *(Lewbel 2007). If Assumption 1 is satisfied, then there exists a function* $\mu(b)$ *with* $|\mu(b)| \leq 1$*, such that* $\gamma(b) = \gamma^*(b)\mu(b)$*. If, in addition, Assumption 2 is satisfied, then* $\mu(b) > 0$*.*

Proposition 1 suggests that measurement errors in our binary treatment variable introduce an attenuation bias in the estimated treatment effect. The magnitude of the mismeasured treatment effect estimate $\gamma(b)$ provides a lower bound on the true treatment effect $\gamma^*(b)$, and when Assumption 2 is also satisfied, the sign of the mismeasured effect $\gamma(b)$ matches the sign of the true effect $\gamma^*(b)$.

Proof for Proposition 1 (Theorem 1 in Lewbel (2007)). We refer readers to the proof of Theorem 1 in Lewbel (2007) for more details.

Define

$$p_d(\mathbf{B}) = \mathbb{E}(D \text{ after}^* | \mathbf{B}, D \text{ after} = d) = \mathbb{P}(D \text{ after}^* = 1 | \mathbf{B}, D \text{ after} = d).$$

We suppress the $\mathbf{B}$ and b argument for clarity. By Bayes' rule,

$$p_0 = \frac{\beth_1 r^*}{1-r} \text{ and } p_1 = \frac{(1-\beth_1)r^*}{r}. \tag{11}$$

Also,

$$r = \mathbb{E}(D \text{ after}) = \sum_{d\in\{0,1\}} \mathbb{E}(D \text{ after}|D \text{ after}^* = d)\mathbb{P}(D \text{ after}^* = d) = (1-\beth_1)r^* + \beth_0(1-r^*), \tag{12}$$

which gives $r = \beth_0$ when $\beth_0 + \beth_1 = 1$; otherwise

$$r^* = \frac{r-\beth_0}{1-\beth_0-\beth_1} \text{ and } 1-r^* = \frac{1-\beth_1-r}{1-\beth_0-\beth_1}. \tag{13}$$

Based on Assumption 1 and Equation (10),

$$\mathbb{E}(\mathbf{z}|D \text{ after}^*, D \text{ after}) = h_0^* + \gamma^* D \text{ after}^*.$$

By law of iterated expectations, this gives,

$$\mathbb{E}(\mathbf{z}|D \text{ after} = d) = h_0^* + \gamma^* p_d$$

Because $\gamma = \mathbb{E}(\mathbf{z}|D \text{ after} = 1) - \mathbb{E}(\mathbf{z}|D \text{ after} = 0)$, we obtain

$$\gamma = (p_1 - p_0)\gamma^* = \gamma^* \mu.$$

The μ in Proposition 1 equals $p_1 - p_0$. Because μ equals the difference between the two probabilities, then $-1 \le \mu \le 1$.

Based on Equation (11),

$$\mu = p_1 - p_0 = \frac{r^*}{(1-r)r}(1-\beth_1-r),$$

and using Equation (13) for $1-r^*$:

$$(1-r)r\mu = (1-r^*)r^*(1-\beth_0-\beth_1).$$

Because probabilities r and r^* lie between 0 and 1, then $\mu > 0$ when Assumption 2 holds. □

C. Data summary statistics

Table C1 presents the statistics of all control variables, including the mean, standard deviation, minimum value, 25th percentile, 50th percentile (median), 75th percentile, and maximum value.

Table C1 Data statistics for the control variables used in behavioral matching and DID.

	Mean	Standard deviation	Minimum	25th percentile	50th percentile	75th percentile	Maximum
$\mathbf{x}_1$	0.000	3315.870	-1045.555	-1016.075	-909.806	-105.592	177716.884
$\mathbf{x}_2$	0.000	2484.175	-104151.427	-69.455	127.305	139.723	135271.576
$\mathbf{x}_3$	0.000	2426.733	-33269.196	-301.625	-295.754	-241.110	163151.808
$\mathbf{x}_4$	-0.000	2118.413	-8512.189	-270.335	-267.812	-227.533	151609.627
$\mathbf{x}_5$	-0.000	2093.425	-5572.187	-207.957	-205.477	-153.883	150347.958
$\mathbf{x}_6$	0.000	1961.369	-55575.096	-283.942	-278.329	-165.060	153474.042
$\mathbf{x}_7$	-0.000	1892.127	-53795.043	-132.826	-100.791	-69.239	113301.895
$\mathbf{x}_8$	-0.000	1517.756	-12457.799	-173.908	-170.007	-155.343	70394.145
$\mathbf{x}_9$	0.000	1469.649	-28428.331	-56.958	-55.842	-46.624	92583.819
$\mathbf{x}_{10}$	0.000	1431.712	-88502.543	-138.733	-121.221	-90.011	88909.405
$\mathbf{x}_{11}$	-0.000	1335.346	-1862.969	-51.971	-51.660	-44.747	111304.677
$\mathbf{x}_{12}$	-0.000	1250.929	-47449.051	-99.935	-88.995	-53.808	55097.260
$\mathbf{x}_{13}$	-0.000	1228.536	-36863.835	-58.163	-43.748	-43.027	70186.154
$\mathbf{x}_{14}$	0.000	1007.812	-9384.834	-41.941	-32.855	-30.590	67042.285
$\mathbf{x}_{15}$	0.000	952.018	-11033.319	-58.717	-53.144	-40.690	86802.956
$\mathbf{x}_{16}$	0.000	926.777	-48689.114	-39.109	-12.469	-0.792	49358.794
$\mathbf{x}_{17}$	0.000	855.107	-8892.994	-25.814	-20.357	-18.735	92451.977
$\mathbf{x}_{18}$	0.000	830.127	-15326.425	-50.400	-47.961	-45.787	64768.805
$\mathbf{x}_{19}$	0.000	685.514	-9554.601	-53.825	-44.046	-38.881	65470.980
$\mathbf{c}_1$	2.000	0.913	-1.729	1.336	2.290	2.658	5.892
$\mathbf{c}_2$	-1.080	1.282	-5.167	-2.480	-1.031	-0.048	3.525
$\mathbf{c}_3$	-1.788	1.225	-4.929	-3.093	-1.807	-0.834	3.021
$\mathbf{c}_4$	-1.075	1.100	-4.607	-2.134	-1.197	-0.242	3.297
$\mathbf{c}_5$	1.059	1.198	-1.558	-0.237	1.046	1.977	5.950
$\mathbf{c}_6$	-0.868	0.862	-5.507	-1.050	-1.050	-0.357	2.132
$\mathbf{c}_7$	0.397	0.996	-3.915	-0.323	0.664	1.159	4.236
$\mathbf{c}_8$	0.210	0.978	-4.191	-0.484	0.472	0.966	4.272
$\mathbf{c}_9$	1.761	0.781	-1.919	1.412	1.776	2.033	5.943
$\mathbf{c}_{10}$	0.001	1.240	-4.150	-1.010	0.007	1.274	4.196
$\mathbf{c}_{11}$	-0.691	1.052	-5.421	-1.430	-0.870	0.149	2.354
$\mathbf{c}_{12}$	-1.680	1.129	-5.446	-2.781	-1.795	-0.829	3.127
$\mathbf{c}_{13}$	-0.656	0.940	-5.480	-1.247	-0.281	-0.037	2.700
$\mathbf{c}_{14}$	0.282	0.871	-3.901	-0.250	0.692	0.692	5.451
$\mathbf{c}_{15}$	1.080	0.882	-2.700	0.532	0.717	1.648	5.150
$\mathbf{c}_{16}$	-0.464	0.881	-4.592	-1.015	-0.074	0.024	3.230
$\mathbf{f}_1$	0.001	0.037	0.000	0.000	0.000	0.000	2.000
$\mathbf{f}_2$	0.000	0.014	0.000	0.000	0.000	0.000	1.000

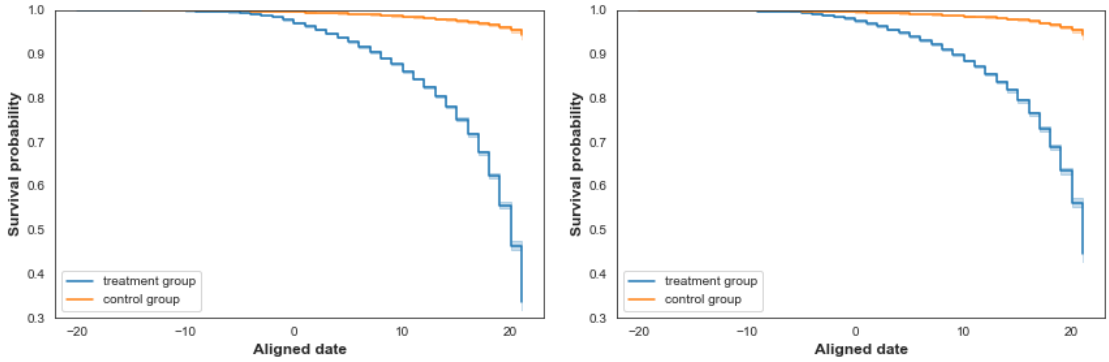
D. Survival rate over time by groups

In Section 3.2.2, we adhere to a standard practice in crafting panel data for diffusion behaviors, as outlined by Belo and Ferreira (2022), by excluding individuals from the panel after they adopt the behavior. Consequently, direct comparison of adoption decisions between the treatment and control groups during the pre-treatment period is not possible. Instead, we compare the likelihood of adoption between the two groups in the pre-treatment period by applying a hazard model (Aral et al. 2009). The hazard model allows us to calculate a survival rate, which represents the likelihood of not adopting the behavior before the treatment, and we compare these rates between the matched samples in the treatment and control groups.

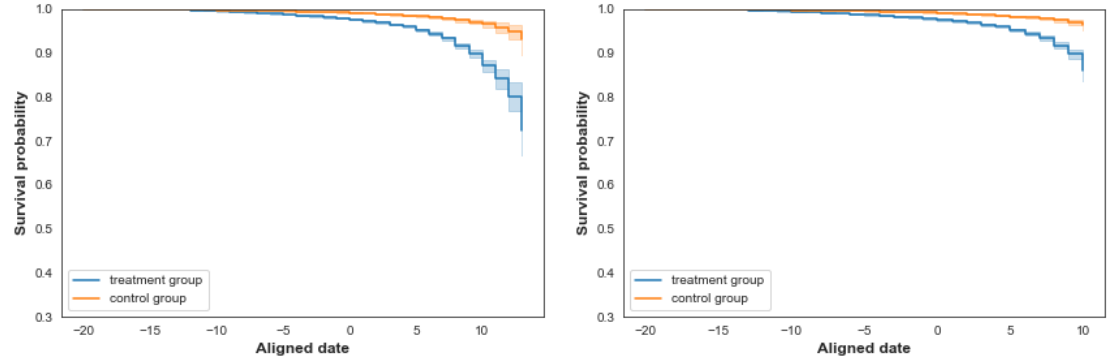
Using the regression model $\eta(t, D) = \eta_0(t)e^{\kappa_0+\kappa_d D}$, we estimate the rate at which individuals attend the offline event. Here, $\eta(t, D)$ denotes the likelihood of adoption; t is the time index; $D \in \{0, 1\}$ pertains to the treatment and control groups; $\eta_0(t)$ symbolizes the baseline adoption likelihood; and $\kappa_0, \kappa_d \in \mathbb{R}$ are the parameters. The change in the likelihood of adoption is measured by κ_d and is linked with switching from the control to the treatment group.

Figure D1 illustrates the survival rate of the control group and each of the treatment groups (within matched pairs), wherein we align the pre-treatment and post-treatment days (represented as t). Here, $t < 0$ denotes the day before the treatment; $t = 0$ represents the treatment day; and $t > 0$ signifies the days in the adoption period.[19] The survival rate at $t < 0$ is relevant to the pre-treatment differences of the two groups as it reflects the pre-treatment differences between the two groups. We observe no systematic visual distinction between the treatment and control groups in terms of the survival rate during the pre-treatment period. However, starting from the treatment day, a growing disparity emerges in the survival rates, with the treated group exhibiting a higher likelihood of attending the event and subsequently being dropped from the sample. The differential survival rate quantitatively captures the cumulative effects of social influence over time.

[19] For example, in reference to Section 3.2 and Figure 3, we can label day one as $t = -2$, day three as $t = 0$, and day five as $t = 2$ for both Bob and Anne.



(a) Hop 1 **(b)** Hop 2



(c) Hop 3 **(d)** Hop 4

Figure D1 **Survival rate by treatment (y-axis) for the aligned treatment date (x-axis). The four figures correspond to the survival rate of individuals in the four treatment groups and the control group.The aligned date < 0 is the pre-treatment period, and the treatment exposure begins at 0.**

E. Estimation table from the DID model on matched samples

We present the estimated results in Tables E1-E2, which include the coefficients for our main variable of interest, $D_{sj}\text{after}_{sjt}$, as well as the coefficients for variables $\mathbf{x}, \mathbf{c}$, and $\mathbf{f}$. Our findings indicate that phone communication can significantly increase the adoption likelihood γ by 0.0086, 0.0067, 0.0052, and 0.0046 from hop 1 to hop 4. For the control group, the adoption likelihood stands at 0.0098. This leads to a comparative percentage increase in the adoption likelihood for each hop group in relation to the control group, amounting to 87.61%, 68.65%, 53.10%, and 46.71% for the first four hops respectively. These results underscore the long-range impact of social influence facilitated by phone communications, highlighting the substantial potential of mobile phones in viral and seeded marketing strategies. It is important to note, however, that due to the binary nature of the outcome variable and our choice of a linear probability model for the analysis, the R^2 measure loses its usual interpretative value (Hanck et al. 2019). This is because a regression line cannot perfectly fit binary dependent variables with continuous regressors.

Table E1 Estimates from the main DID model (part I).

	Dependent variable: Adoption			
	hop 1	hop 2	hop 3	hop 4
D_{sj} after$_{sjt}$	0.00861414***	0.00674916***	0.00522119***	0.00459264***
	(0.00035388)	(0.00030401)	(0.00060045)	(0.00062162)
$\mathbf{x}_1$	0.00000013**	−0.00000003	−0.00000041	0.00000031
	(0.00000006)	(0.00000005)	(0.00000029)	(0.00000022)
$\mathbf{x}_2$	0.00000010**	0.00000001	−0.00000020	−0.00000052
	(0.00000005)	(0.00000005)	(0.00000037)	(0.00000050)
$\mathbf{x}_3$	0.00000002	0.00000003	−0.00000008	−0.00000012
	(0.00000004)	(0.00000004)	(0.00000020)	(0.00000022)
$\mathbf{x}_4$	0.00000016**	−0.00000005	−0.00000020	0.00000003
	(0.00000006)	(0.00000005)	(0.00000045)	(0.00000011)
$\mathbf{x}_5$	0.00000008	−0.00000004	−0.00000008	−0.00000002
	(0.00000010)	(0.00000005)	(0.00000066)	(0.00000032)
$\mathbf{x}_6$	0.00000027**	−0.00000002	−0.00000106	−0.00000023
	(0.00000011)	(0.00000009)	(0.00000089)	(0.00000035)
$\mathbf{x}_7$	−0.00000011*	−0.00000009	−0.00000037	0.00000078**
	(0.00000006)	(0.00000007)	(0.00000027)	(0.00000031)
$\mathbf{x}_8$	0.00000016	−0.00000008	−0.00000066	0.00000009
	(0.00000012)	(0.00000007)	(0.00000079)	(0.00000029)
$\mathbf{x}_9$	−0.00000011	−0.00000005	−0.00000017	0.00000015
	(0.00000011)	(0.00000008)	(0.00000031)	(0.00000035)
$\mathbf{x}_{10}$	−0.00000031**	0.00000020*	−0.00000031	−0.00000000
	(0.00000014)	(0.00000011)	(0.00000112)	(0.00000040)
$\mathbf{x}_{11}$	0.00000017	−0.00000029	−0.00000266	0.00000009
	(0.00000022)	(0.00000022)	(0.00000313)	(0.00000031)
$\mathbf{x}_{12}$	−0.00000012	0.00000008	0.00000034	−0.00000002
	(0.00000015)	(0.00000012)	(0.00000082)	(0.00000019)
$\mathbf{x}_{13}$	−0.00000010	−0.00000004	0.00000001	−0.00000029
	(0.00000009)	(0.00000010)	(0.00000022)	(0.00000037)
$\mathbf{x}_{14}$	0.00000015	0.00000030***	0.00000204***	0.00000008
	(0.00000021)	(0.00000011)	(0.00000066)	(0.00000031)
$\mathbf{x}_{15}$	−0.00000010	0.00000023*	−0.00000161*	−0.00000078
	(0.00000009)	(0.00000012)	(0.00000094)	(0.00000079)
$\mathbf{x}_{16}$	−0.00000007	0.00000014	−0.00000012	−0.00000115
	(0.00000012)	(0.00000009)	(0.00000029)	(0.00000090)
$\mathbf{x}_{17}$	−0.00000005	−0.00000002	0.00000002	−0.00000985
	(0.00000014)	(0.00000010)	(0.00000021)	(0.00000999)
$\mathbf{x}_{18}$	0.00000027*	−0.00000006	−0.00000019	0.00000013
	(0.00000015)	(0.00000011)	(0.00000092)	(0.00000062)
$\mathbf{x}_{19}$	0.00000012	0.00000021	−0.00000024	0.00000068
	(0.00000019)	(0.00000016)	(0.00000086)	(0.00000078)

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

Table E2 Estimates from the main DID model (part II).

	Dependent variable: Adoption			
	hop 1	hop 2	hop 3	hop 4
$\mathbf{c}_1$	−0.00065460***	0.00043623***	−0.00025075	0.00066402
	(0.00016792)	(0.00014262)	(0.00063764)	(0.00044803)
$\mathbf{c}_2$	0.00002068	0.00015440	0.00037303	0.00039341
	(0.00014550)	(0.00012101)	(0.00035397)	(0.00036414)
$\mathbf{c}_3$	−0.00123777***	0.00010076	0.00026631	−0.00029249
	(0.00028330)	(0.00024246)	(0.00308679)	(0.00116076)
$\mathbf{c}_4$	−0.00059973***	0.00022627	0.00008107	0.00015001
	(0.00020343)	(0.00018018)	(0.00234160)	(0.00091419)
$\mathbf{c}_5$	−0.00018366	−0.00003268	−0.00023204	0.00089503**
	(0.00016264)	(0.00014281)	(0.00048211)	(0.00039013)
$\mathbf{c}_6$	0.00168100***	−0.00073436**	0.00006931	−0.00100788
	(0.00042476)	(0.00037069)	(0.00529100)	(0.00205883)
$\mathbf{c}_7$	−0.00057674***	−0.00021529	0.00005045	0.00087626
	(0.00019040)	(0.00017731)	(0.00205287)	(0.00084908)
$\mathbf{c}_8$	0.00035249**	−0.00021696*	−0.00073306*	−0.00029406
	(0.00014365)	(0.00012257)	(0.00037555)	(0.00035734)
$\mathbf{c}_9$	0.00047624***	0.00004060	0.00055956	−0.00005796
	(0.00017922)	(0.00016301)	(0.00137390)	(0.00059330)
$\mathbf{c}_{10}$	−0.00150325***	0.00038904	0.00017426	0.00030160
	(0.00034242)	(0.00028399)	(0.00353222)	(0.00151574)
$\mathbf{c}_{11}$	0.00219535***	−0.00031302	−0.00092488	−0.00018391
	(0.00061206)	(0.00053461)	(0.00841970)	(0.00334316)
$\mathbf{c}_{12}$	−0.00120020***	−0.00008805	0.00056117	0.00000739
	(0.00029552)	(0.00025087)	(0.00328136)	(0.00114837)
$\mathbf{c}_{13}$	0.00064076***	−0.00010673	−0.00035169	−0.00013720
	(0.00018354)	(0.00014412)	(0.00136448)	(0.00053599)
$\mathbf{c}_{14}$	−0.00043075**	−0.00028706**	0.00036648	0.00029569
	(0.00016806)	(0.00014026)	(0.00162522)	(0.00062522)
$\mathbf{c}_{15}$	0.00007766	0.00001824	−0.00096334	0.00022993
	(0.00016787)	(0.00014789)	(0.00067258)	(0.00037780)
$\mathbf{c}_{16}$	−0.00009547	0.00016336	0.00051048	−0.00051338
	(0.00017996)	(0.00015555)	(0.00127812)	(0.00067177)
$\mathbf{f}_1$	0.00064243	0.00103946	−0.00573922	−0.01427646
	(0.00302893)	(0.00364255)	(0.01284173)	(0.01010160)
$\mathbf{f}_2$	−0.00172523	0.01603047	0.00846063	0.01756889
	(0.00770979)	(0.01110770)	(0.03698765)	(0.01226223)
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
R^2	0.03201922	0.03190918	0.03895760	0.04105386
Residual Std. Error	0.03860605 (df = 348,243)	0.03405705 (df = 355,980)	0.02265690 (df = 58,007)	0.02054692 (df = 47,680)

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

F. Robustness checks

F1. Balance between the treatment and control groups To mimic the random assignment of treatment in a randomized controlled experiment, it is crucial to ensure that the treatment and control groups are comparable and that any observed differences between them are due to chance. One way to achieve this is by checking for post-matching covariate and propensity score imbalances. Analyzing both the covariates for observed and latent homophily, as well as the propensity score, is a critical step in this process. By doing so, we can determine whether pairs in the treatment and control groups are sufficiently similar and whether the treatment effect estimates are reliable.

Checking for overlap in covariates between the treatment and the control group. To effectively remove confounding effects, we need to balance the covariates ($\boldsymbol{x}, \mathbf{c}$, and $\mathbf{f}$) between the matched pairs. We use the standardized mean difference (SMD) to evaluate if the covariates in the treatment and control groups have sufficient overlap (Cohen 1988). The SMD measures the difference in means in the unit of pooled standard deviation for a specific covariate. Following this formula:

$$\text{SMD} = \frac{\overline{\boldsymbol{x}}_{j,h} - \overline{\boldsymbol{x}}_{j,c}}{\sqrt{(\sigma_{j,h}^2 + \sigma_{j,c}^2)/2}},$$

where $\overline{\boldsymbol{x}}_{j,h}$ and $\overline{\boldsymbol{x}}_{j,c}$ are the means of the covariate j for the treatment group on hop h and the control group c, respectively, and $\sigma_{j,h}$ and $\sigma_{j,c}$ are the standard deviations of covariate j for the treatment group on hop h and the control group c, respectively. We perform a similar analysis on $\mathbf{c}$ and $\mathbf{f}$. Guidelines suggest that an SMD below 0.25 or 0.1 for a particular covariate indicates sufficient overlap between the treatment and control groups (Stuart et al. 2013). Figure F1 shows that all the variables we choose pass this robustness check.

We also analyze the differences in the control variables of the treatment and the control groups before and after implementing PSM (Table F1). This table shows the differences in the sample means of the treatment and control groups on the different conditioning variables before and after matching was performed. The standardized difference, short for Std. dif., is computed as the absolute difference normalized by the standard deviation of the treatment group. The percentage reduction in standardized bias after matching is shown in the last column. We observe that after matching, most variables achieve substantial bias reduction, with a few exceptions. Notably, the average standardized bias reduction is 36.15%, indicating that the matching achieved good quality and adequate balance between the two groups. The patterns for all hop groups are similar, so we only show hop 1 in this appendix.

Checking for balance in propensity scores between the treatment and control groups To assess if the propensity scores of the matched pairs in the treatment and control groups are balanced, we plot their distributions (Figure F2). The graphs indicate that the distributions of propensity

Table F1 Differences in the treatment and the control groups before and after PSM for hop 1.

Variable	Treatment	Control (before)	Control (after)	Std. dif. (after)	Std. dif. (before)	Bias reduction (%)
$\mathbf{c}_1$	1.6626	2.0779	1.6999	0.0417	0.4653	91.038
$\mathbf{c}_2$	-0.3495	-1.3051	-0.4903	0.1561	1.0591	85.2611
$\mathbf{c}_3$	-1.2529	-1.8771	-1.3875	0.1536	0.7125	78.4421
$\mathbf{c}_4$	-0.552	-1.1813	-0.6468	0.1067	0.7087	84.9443
$\mathbf{c}_5$	1.7363	0.8259	1.5676	0.1914	1.0323	81.4589
$\mathbf{c}_6$	-0.5001	-1.0782	-0.5423	0.0516	0.7067	92.6985
$\mathbf{c}_7$	-0.0647	0.5851	0.0089	0.0796	0.7024	88.6674
$\mathbf{c}_8$	-0.2329	0.3381	-0.1535	0.0874	0.6286	86.0961
$\mathbf{c}_9$	1.7752	1.7176	1.7714	0.0044	0.0666	93.3934
$\mathbf{c}_{10}$	-0.9493	0.3611	-0.8054	0.1619	1.474	89.0163
$\mathbf{c}_{11}$	0.0895	-1.102	-0.0095	0.1338	1.6112	91.6956
$\mathbf{c}_{12}$	-1.2336	-1.7442	-1.3458	0.1263	0.5747	78.0233
$\mathbf{c}_{13}$	-0.8653	-0.6	-0.845	0.0215	0.2812	92.3542
$\mathbf{c}_{14}$	0.0087	0.3929	0.0514	0.0475	0.4286	88.9174
$\mathbf{c}_{15}$	1.4219	0.9551	1.3499	0.0808	0.5241	84.5831
$\mathbf{c}_{16}$	-0.7781	-0.3269	-0.7489	0.0332	0.5116	93.5106
$\mathbf{x}_1$	516.8323	-534.716	168.6711	0.1064	0.3214	66.8948
$\mathbf{x}_2$	65.1144	77.4631	58.1673	0.0031	0.0055	43.6364
$\mathbf{x}_3$	-59.7531	-179.0862	-156.8535	0.0362	0.0445	18.6517
$\mathbf{x}_4$	12.7363	-166.4732	-8.697	0.0142	0.1189	88.0572
$\mathbf{x}_5$	-7.8875	-131.1355	-9.233	0.0011	0.1	98.9
$\mathbf{x}_6$	184.0699	-191.6389	-95.3678	0.0757	0.1018	25.6385
$\mathbf{x}_7$	37.8163	-36.416	-64.1527	0.0378	0.0275	-37.4545
$\mathbf{x}_8$	4.022	-76.3275	-13.4865	0.0151	0.0695	78.2734
$\mathbf{x}_9$	-25.9622	-10.8598	-22.6436	0.006	0.0273	78.022
$\mathbf{x}_{10}$	-16.264	-96.312	83.326	0.0438	0.0352	-24.4318
$\mathbf{x}_{11}$	8.9739	-9.3456	-36.9819	0.0637	0.0254	-150.7874
$\mathbf{x}_{12}$	-6.3109	-31.9376	-0.0512	0.0069	0.0284	75.7042
$\mathbf{x}_{13}$	48.4492	-29.8367	-45.8742	0.0715	0.0593	-20.5734
$\mathbf{x}_{14}$	22.5637	-17.9400	-37.5643	0.0788	0.0531	-48.3992
$\mathbf{x}_{15}$	21.1188	-31.6315	10.9908	0.0061	0.0315	80.6349
$\mathbf{x}_{16}$	1.6597	-1.4429	23.5587	0.0201	0.0028	-617.8571
$\mathbf{x}_{17}$	-4.0051	-9.0715	-8.3151	0.0055	0.0064	14.0625
$\mathbf{x}_{18}$	18.7481	-31.3464	-44.0867	0.0779	0.0621	-25.4428
$\mathbf{x}_{19}$	5.9943	-24.3380	-12.2617	0.0198	0.0328	39.6341
f (number)	0.0018	0.0006	0.0017	0.0022	0.0281	92.1708
f (percentage)	0.0006	0.0001	0.0005	0.0088	0.0247	64.3725

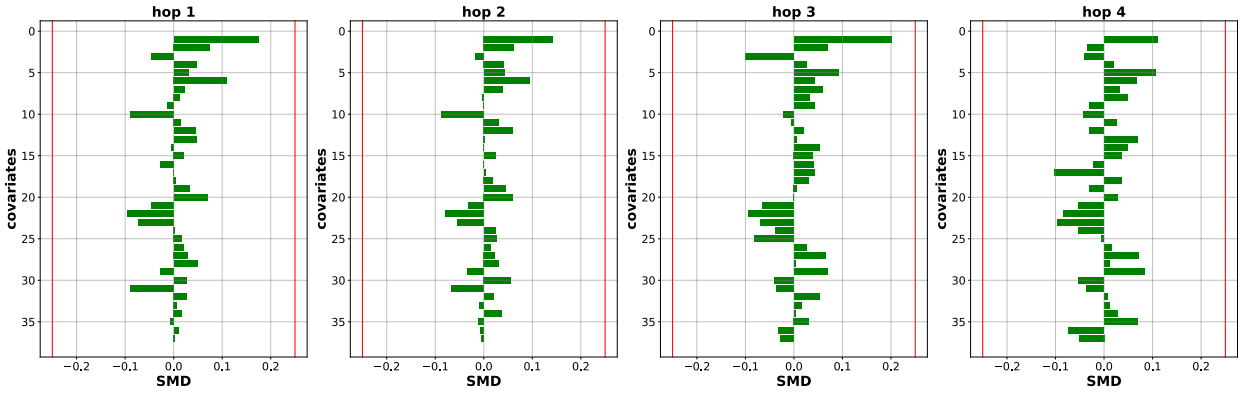
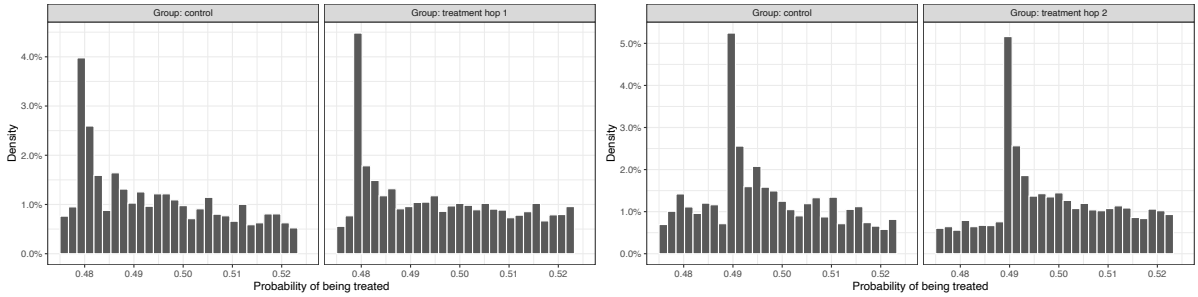


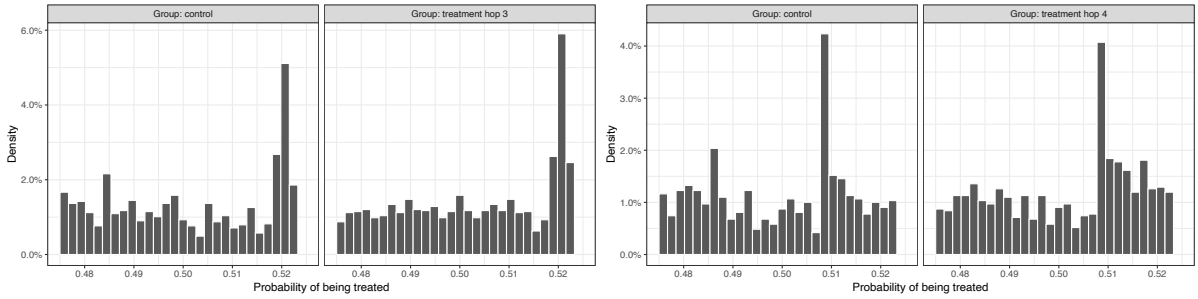
Figure F1 SMD for the matched samples of each treatment group.

scores for the two groups are similar and significantly overlap after matching. We also calculate the pairwise differences in propensity scores between the two groups and present them in Table F2. The differences are on the order of 10^{-2} for the four hops, where propensity scores range from 0 to 1. These results suggest that the treatment and control groups are well-balanced regarding propensity scores.



(a) Hop 1

(b) Hop 2



(c) Hop 3

(d) Hop 4

Figure F2 Distribution for propensity scores of the treatment and control groups for different hop indices h.

Table F2 Pairwise difference in the propensity scores between the treatment and the control group.

Hop	Difference in mean	95% confidence interval
1	0.04099197	(0.03893030, 0.04305363)
2	0.02779980	(0.02602115, 0.02957844)
3	0.01602217	(0.01326176, 0.01878257)
4	0.03068643	(0.02702569, 0.03434718)

F2. Rosenbaum sensitivity test Since the treatments in our study, namely, the phone calls, are not randomized, a certain degree of bias may persist in our analysis, despite our attempts to control for observed and latent homophily. We examine the sensitivity to selection on unobservables utilizing the Rosenbaum bounds approach (Rosenbaum 2005). This approach gauges the potential impact of unobserved variables on an individual's assignment to the treatment group or a control group and consequently on inference.

To quantify the amount of bias from unobserved variables that could qualitatively alter the results, we use the odds ratio of treatment assignment (Γ). As illustrated in Figure F3 in Appendix F2, our findings indicate that the critical level of Γ at which we would question the validity of the PSM is 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 exceeds 0.05. These results suggest that if an unobserved confounder were to cause the odds ratio of treatment assignment to differ by less than to 8.5

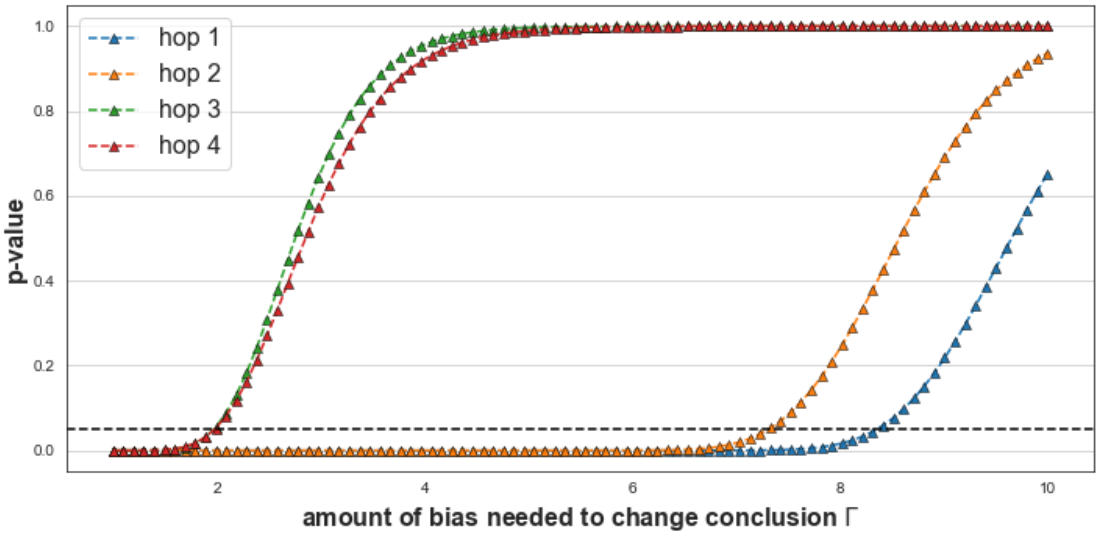


Figure F3 **Rosenbaum sensitivity test. The x-axis and y-axis correspond to the Γ value and the upper bound significance level, respectively. The dotted horizontal line corresponds to an upper bound p-value of** 0.05.

between the treatment and control groups, the confidence interval for the social influence effect would not include zero.

While there is no universally agreed upon rule-of-thumb value for Γ, some studies propose that any value above $\Gamma = 1.5$ signals substantial insensitivity to unobserved confounders (Sen 2014, Ransbotham et al. 2019). Our Γ values substantially exceed this threshold across four hops, indicating that our results are considerably insensitive to hidden bias and provide strong support for the existence of social influence through phone communications up to four degrees of separation in our data. However, beyond the fourth hop, our results become more susceptible to unobserved confounders, and hence, we omit them from our analysis.

G. Other analysis methods based on observational data

We have expanded upon the main analysis in Section 4 by incorporating various alternative observational methods. Alongside PSM, we have utilized other methods such as coarsened exact matching, subclassification, Mahalanobis distance matching, and Post-Lasso estimation. These methods are explained below:

1. Subclassification involves stratifying the samples based on propensity scores (Imbens and Rubin 2015). We divide the samples into strata comprising units with similar propensity scores and calculate the treatment effect estimate of each stratum. This method helps control for covariates within each stratum. We then estimate the average treatment effects on the treated units using regression within the strata.

2. Coarsened Exact Matching (CEM) is a matching method that reduces reliance on functional forms (Iacus et al. 2012). Unlike PSM, which necessitates defining the functional form of the covariates to estimate the propensity score, CEM coarsens each covariate of the treated and control units into predefined strata and executes exact matching on these coarsened covariates. This method helps avoid potential complications arising from functional form misidentification and reduces bias due to model misrepresentation.

3. Mahalanobis Distance Matching (MDM) is akin to propensity matching but uses a different distance function, the Mahalanobis distance, between data pairs instead of differences in propensity scores. Mahalanobis distance is a scale-free Euclidean distance where the Euclidean distances between two units are normalized by the covariance matrix.

4. Post-Lasso estimation estimates the effects of treatments with a data-driven penalty (Belloni et al. 2013). Specifically, we calculate the difference in adoption likelihood of the treatment and control groups by applying ordinary least squares to the model selected by Lasso regression.

Figure G1 illustrates the effect of social influence, represented by the difference in adoption likelihood (on the y-axis), for different hop indices (on the x-axis) using the above-mentioned alternative methods. This analysis reveals a consistent decaying pattern of social influence effect across all methods, demonstrating the robustness of this decay pattern.

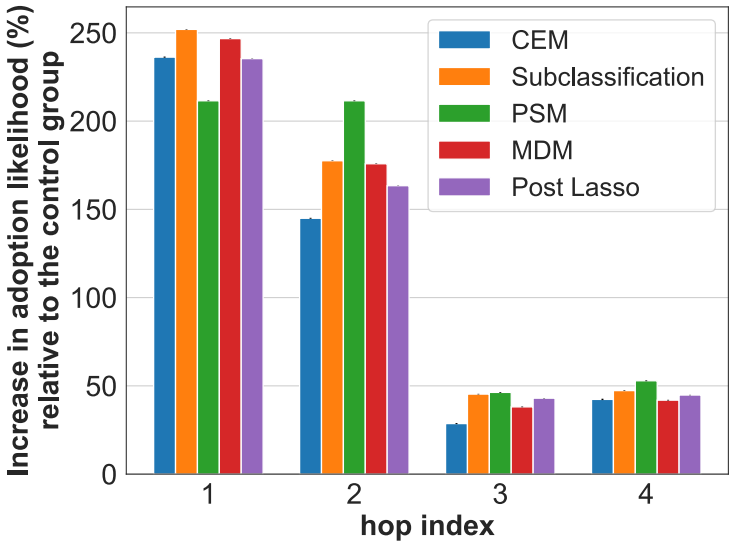


Figure G1 **Difference in increased adoption likelihood for treatment group (relative to the control group). The vertical bars cover 95% confidence intervals.**

H. Reference prices from a marketing website

In our simulation detailed in Section 5.2.2, we utilized reference marketing prices obtained from the content marketing platform Zerys, accessible at `https://z3i.zerys.com`. Figure H1 illustrates the pricing structure for three tiers of content marketing services aimed at creating viral content.

It is important to note that these prices serve as a basis for our analysis and are used purely for illustration. In a practical scenario, marketing firms can estimate their own costs based on their specific contexts and requirements.

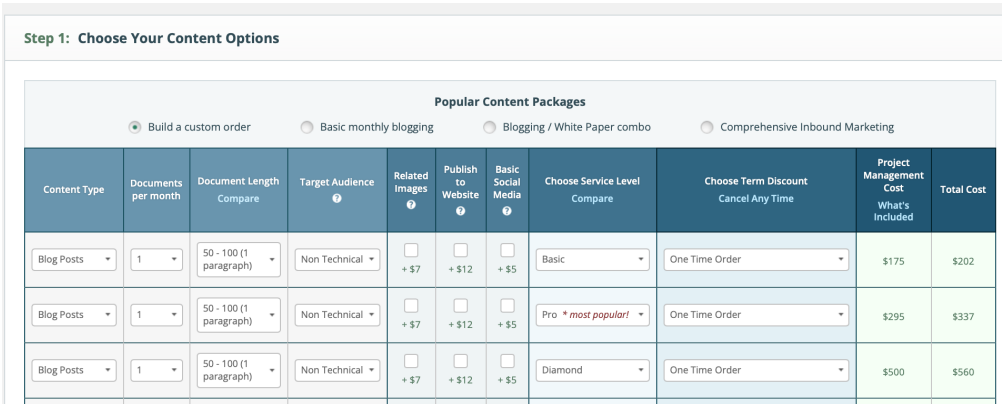


Figure H1 **Reference prices for three marketing effort tiers (low, medium, and high) used in the cost-benefit analysis simulation presented in Section 5.2.2.**

I. Practical accessibility of phone communication data

The increasing availability of call detail records (CDRs) to researchers, governments, and commercial firms highlights the potential of our work to yield broad theoretical and practical implications. First, CDRs have been increasingly utilized in a range of academic disciplines, including sustainable urban development and mobility analysis (Barbour et al. 2019, Leng et al. 2021c), tourism management (Leng et al. 2021b), healthcare (Jones et al. 2018), inequality analysis (Ucar et al. 2021, Leng et al. 2021a), and migration analysis (Salah et al. 2019). Second, governments and public authorities have turned to CDRs for data-driven policy-making, such as tracking population movements and modeling epidemics during the Covid-19 pandemic in countries like China, South Korea, Israel, and several European nations (Oliver et al. 2020). Third, CDRs are becoming commercially available through partnerships with network providers and collaborations with companies such as Flowminder[20] for business insights and policy-making. Marketing companies and event managers can partner with network providers to develop joint marketing strategies that benefit both parties. Privacy-preserving analysis frameworks like Open Algorithms[21] make these collaborations more practical by enabling privacy-preserving analysis of mobile phone data.

In summation, the expanded accessibility of phone communication data to a range of stakeholders underlines the necessity for customized technical frameworks that can analyze and extract meaningful insights from this data. Our framework, centered around mobile phone call data, serves as a valuable instrument to investigate emerging and critical research questions within this domain.

[20] `https://www.flowminder.org`

[21] `https://www.opalproject.org`