HARVARD | BUSINESS | SCHOOL

Separating homophily and peer influence with latent space

Joseph P. Davin
Sunil Gupta
Mikolaj Jan Piskorski

Working Paper

14-053

January 7, 2014

Separating homophily and peer influence with latent space

Joseph P. Davin, Sunil Gupta, Mikolaj Jan Piskorski*

December 30, 2013

Abstract

We study the impact of peer behavior on the adoption of mobile apps in a social network. To identify social influence properly, we introduce latent space as an approach to control for latent homophily, the idea that "birds of a feather flock together." In a series of simulations, we show that latent space coordinates significantly reduce bias in the estimate of social influence. The intuition is that latent coordinates act as proxy variables for hidden traits that give rise to latent homophily. The approach outperforms existing methods such as including observed covariates, random effects, or fixed effects. We then apply the latent space approach to identify social influence on installation of mobile apps in a social network. We find that peer influence account for 27% of mobile app adoptions, and that latent homophily inflates this estimate by 40% (to 38%). In some samples, ignoring latent homophily can result in overestimation of social effects by over 100%.

*Joseph Davin (jdavin@hbs.edu) is a doctoral student at Harvard Business School, Soldiers Field, Boston MA; Sunil Gupta is the Edward W. Carter Professor of Business Administration at the Harvard Business School; Mikolaj Jan Piskorski is an Associate Professor of Business Administration and Richard Hodgson Fellow in the Strategy Unit at the Harvard Business School.

Keywords: Social influence, social network, mobile apps, peer effects, latent homophily, latent space, proxy variables

1 Introduction

What if your friend becoming obese caused you to become obese as well? That was the main finding of Christakis & Fowler (2007) who claimed that obesity was contagious through social networks. Although the study was picked up widely by the national press and hailed as “one of the most exciting studies in medical sociology” by the National Institute on Aging (Kolata 2007), the study also drew critiques that social influence was not properly separated from homophily. Homophily is the concept that people who share things in common are more likely to be friends. Because of this, people who are friends are also expected to act in similar ways because they possess similar traits. So, when an analyst observes clusters of friends behaving in similar ways, it is hard to identify whether the underlying cause is social influence or homophily. Even more troublesome is latent homophily: when people become friends based on unseen traits. Returning to the obesity study, friends could become obese together not because of social contagion effects but because of similar underlying traits that led them to be friends as well as increased their odds of becoming obese. This does not mean that social influence did not cause obesity, but that more work is needed to isolate social effects from homophily.

Social influence potentially affects not just health outcomes but many areas of consumer decision making. McKinsey & Company estimates that a third of all consumer spending, c. $940 billion in annual consumption in the US and Europe alone, will be influenced by social interactions (Chui et al 2012). The ubiquitousness of social influence in consumption decisions makes it an attractive marketing opportunity for firms. As a consequence, researchers want to quantify contagion effects in a myriad of consumer settings (Godes et al 2005, Hartmann et al 2008).

The inability to separate social influence from homophily is troubling for both firms and researchers alike. To be effective in social networks, marketers have to take actions congruent with the dominant process that guides consumer decision making among groups of connected individuals. If homophily was the dominant force, then there are latent traits responsible for purchase. This means marketers should uncover latent traits for segmentation and targeted marketing actions (Hill et al 2006). However, if purchase decisions are mainly driven by social influence, then marketers need to augment their marketing strategy to target the most influential individuals in the social network or build product features that promote peer influence (Tucker 2008, Trusov et al 2010, Hartmann 2010, Aral & Walker 2011a). Therefore, to have greater impact on social networks, it is important to separate homophily and social influence.

Motivated by this urgent and critical problem, a number of researchers have sought to isolate the effect of social influence from homophily (Tucker 2008, Aral et al 2009, Bramoulle et al 2009, Nair et al 2010, Hartman 2010, Shalizi & Thomas 2011, Goldsmith-Pinkham & Imbens 2013). Existing approaches to separate social influence and homophily ignore latent homophily, require panel data, are unable to accommodate lagged independent variables, or model homophily as homophily as group effects. We extend current methods by introducing latent space which can control for observed as well as latent homophily, and is applicable to single cross-section or panel data, lagged independent variables, and estimates homophilous traits on the individual level.

Although latent space is a well-accepted model in social network analysis, it has primarily been used to visualize individuals and identify communities within a network. Prior work in this area focused on latent space distance between pairs of individuals. Our contribution is to use latent coordinates as proxy variables to control for unobserved homophily when estimating social effects. We empirically illustrate the importance of social influence in app adoption in a social network. We find that peer usage accounts for more than a quarter of all mobile app adoptions, even after controlling for homophily. Failing to account for homophily would inflate the impact of social influence by 40% on average, even up to 100% in certain samples.

The rest of the paper is laid out as follows: In section 2, we review the extant literature on the confluence of homophily and social influence and outline the intuition of using latent space to control for latent homophily. In section 3, we describe the latent space model. We use a series of simulations to show that latent homophily can create bias, and that latent space can reduce this bias. We compare the approach with other existing methods in section 4. In section 5, we investigate whether peers influence mobile app adoption and conclude in section 6.

2 Social influence & homophily

Social influence has profound impact on many aspects of customer decision making and marketing, including the adoption of new technologies (Tucker 2008, Aral et al 2009, Aral & Walker 2011a, Aral & Walker 2012), social network usage and adoption (Trusov et al 2010, Aral et al 2011b, Ghose and Han 2011, Katona et al 2011), eCommerce (Stephen & Toubia 2010), new prescriptions of pharmaceutical drugs (Manchanda et al 2008, Nair et al 2010, Iyengar et al 2011), ad effectiveness (Bakshy et al 2012), group decision making (Hartmann 2010), and customer retention (Nitzan & Libai 2011). Effect size varies by the empirical setting. For example, Hartmann (2010) finds that 35% of customer value is attributable to peer effects while Trusov et al (2010) found that only 1 in 5 customers actually influence their peers.

Tempering the impact of these studies is that peer effects studies in social science research have been called into question. Initially, a series of high profile studies suggest that medical and human conditions such as obesity, smoking cessation, happiness, and loneliness that are traditionally understood to be non-contagious are actually socially transmitted along network ties (Christakis & Fowler 2007, Christakis & Fowler 2008, Fowler & Christakis 2008, Cacioppio et al 2009). In response, critics question their methodology and by extension other research in social influence (Cohen-Cole & Fletcher 2008, Lyons 2011, Shalizi & Thomas 2011). The main criticism surrounds the confounding of homophily and social influence. Homophily, the phenomenon that "birds of a feather flock together", has been widely documented in sociology and organizational behavior (Kandel 1978, Ibarra 1992, McPherson et al 2001, Kossinets & Watts 2009). Homophily arise from correlated unobservable traits among friends (sometimes called endogenous tie formation) and can inflate estimates of social influence if those latent traits also cause the outcome of interest. This is analogous to omitted variable bias or confounding which induces correlation between friendship and the outcome of interest. The critique is not new; Manski (1993) discusses identification problems in endogenous social effects for linear-in-means models. Figure 1 illustrates a simplified diagram of how an unobserved common trait among friends can induce bias in the estimate of social influence effect. If we knew the omitted traits involved, we would control for them when estimating social influence and thereby eliminate any bias potentially induced by latent homophily. Alas, we do not observe these traits, though we use this exact intuition with latent space.[1]

[Insert Figure 1 here]

Marketing literature dealing with social influence has shed some light on identifying social influence effects. Van den Bulte & Lilien (2001) show that marketing efforts, which is similar among doctors who know each other in a network, can explain the pattern of new drug adoption. Once they control for these marketing actions, they find no evidence of social influence among doctors in adoption of a new pharmaceutical drug. Tucker (2008) suggest that not properly controlling for endogenous group ties could bias peer effects by 50% in technology adoption among employees in a firm.

To identify peer effects, researchers increasingly resort to experimental methods to make causal statements about social influence (Fowler & Christakis 2010, Aral & Walker 2011a, 2011b, 2012). However, experiments in social networks can be wrought with statistical concerns. There is active research on how to conduct experiments in the presence of interference that naturally exists between observations in a social network (Aronow & Samii 2012, Bow-

[1] For a full causal graph on how the omission of unobserved traits can create a “backdoor” for bias when estimating social influence, see Shalizi & Thomas (2011)

ers et al 2013, Mebane & Poast 2013). More practically, experiments ignore petabytes of secondary "Big Data" which firms are anxious to use.

In order to make use of observational data, one common attempt to separate the confounding of homophily and social influence is to control for covariates (e.g., Aral et al 2009, Nitzan & Libai 2011, Iyengar et al 2011). Covariates could contribute to social ties as well as the outcome, and so controlling for them should remove bias arising from observed homophily. In the adoption of a new instant messaging system, Aral et al (2009) show that naive models that do not control for covariates could bias peer effect estimates by 300-700%.

However, merely including covariates does not control for latent homophily (Shalizi & Thomas 2011). To control for latent homophily effectively, we introduce the latent space model which infers latent coordinates for individuals based on their connections with each other in a social network. Latent space coordinates act as proxy variables for unseen traits that drive homophily. Proxy variables have a long history in econometrics for controlling and significantly reducing the bias stemming from omitted variables (McCallum 1972, Aigner 1974). For example, IQ is used as a proxy for innate ability; R&D and patents are used as proxy for innovative activity of a firm. In our proposed approach, we use the latent space model to approximate those hidden traits which underlie homophily and reduce the bias on the estimate of social influence.

Our use of latent space to control for homophily is motivated by Shalizi & Thomas (2011) who suggest that to get identification of social influence, the parts of the unobserved trait that influence tie formation must be made observable. We achieve this goal by using latent space to model unobserved traits that underlie social tie formations. Latent space provides a model of tie formation in a network by co-locating friends together and separating apart strangers in a k-dimensional space (Hoff et al 2002, Hoff 2005, Krivitsky et al 2007). The approach is also no stranger to marketing. Ansari et al (2010) use latent space to model multiple relationship formation within networks of managers and musicians. Braun & Bonfrer (2011) speed up estimation of latent space for large datasets using Dirichlet processes and apply the approach to a telecommunications network.

The reason latent space has not been used as bias control in social influence studies is because latent space has traditionally been concerned with the distances between actors not the actual positions that these actors take in latent space. Braun et al (2011) articulate this view by cautioning researchers that it is “the relative distances among individual latent coordinates, and *not the absolute positioning in the latent space*, that matter” (emphasis added).

However, when controlling for homophily, we are interested in the actual latent space coordinates because these capture unseen traits that drive homophily. If we can infer the forces that underlie homophily, then we can use them to control for homophily in social influence studies. In the econometrics literature, Goldsmith-Pinkham & Imbens (2013) propose the use of a binary latent space for addressing homophily arising from clusters but do not discuss more general formulations of latent space or explore how effective the approach is in reducing bias. Our contribution is to introduce latent space as proxy variables that control for homophily in social influence studies. We show through extensive simulations that latent space reduces a significant proportion of the bias induced by latent homophily.

Fixed and random effects attempt are alternative methods but possess shortcomings which may not be attractive for some empirical settings. Hartmann (2010) shows in a simulation that introducing heterogeneity in the model via random effects addresses correlated unobservables. Stephen & Toubia (2010) also use random “ability” effects to control for network endogeneity. Fixed effects could potentially be used at the network level (Bramoulle et al 2009) or at the individual level (Nair et al 2010). However, these approaches do not leverage the social network data of who is connected to whom. More importantly, random and fixed effects models do not work if there are dynamics (i.e., lagged independent variable) in the underlying data generating process (Angrist & Pischke 2009). It is known that using fixed effects in dynamic models creates bias in coefficient estimates (Nickell 1981), and can even change the sign of estimated coefficients (Phillips & Sul 2007) although this is an ongoing area of research (Bai 2013).

3 Latent space model

A latent space model uses information on dyadic ties between individuals in the network to estimate a set of coordinates for each individual in the network, such that the resulting distances between the individuals in latent space corresponds (via some link function) to the presence or absence of dyadic ties.

3.1 Dyadic ties

We start with an $N \times N$ symmetric matrix A containing information of dyadic ties between pairs of actors among N actors within the network, with $A_{ij} = 1$ if actors i and j are friends and 0 otherwise.[2]. This matrix will serve as the data to estimate a latent space model.

3.2 Latent space coordinates

We assume conditional independence of dyads so that the probability of forming a link between two actors in the network depend only on their positions in the latent space:

$P\left(A|X, Z, \theta\right) = \prod_{i \neq j} P\left(A_{ij} = 1|X_i, X_j, Z_i, Z_j, \theta\right)$

where $Z = \{Z_1, \cdots, Z_N\}$ are observed covariates and $X = \{X_1, \cdots X_N\}$ are latent traits for each individual $i = 1, \cdots, N$. $\theta = \{\alpha_0, \alpha_1\}$ is the set of parameters. The model can accommodate multiple covariates: each X_i, Z_i can be vectors capturing k and l traits of an individual, which means that α_1, α_2 are also vectors of length k, l. Under homophily, we can model dyadic ties as a function of distance in this latent space:

$logit\left[P\left(A_{ij} = 1|X_i, X_j, Z_i, Z_j, \theta\right)\right] = \alpha_0 - \alpha_1 \cdot \|Z_i - Z_j\| - \|X_i - X_j\|$

where $\alpha_1 > 0$ so that actors who are far apart have smaller chance of being friends. There is no coefficient on distances of X because we cannot separately identify the coefficient from

[2] The dyadic ties can be asymmetric or non-binary (e.g., Ansari et al 2010); these are simple extensions of the base latent space model

the variance of X. We use Euclidean distance although any other distance metric satisfying the triangle inequality can be used (e.g., cosine distance). The number of dimensions for the latent space can be determined by BIC or other likelihood-based criteria (Krivitsky et al 2009).

3.3 Capturing homophily with latent space

An important feature of a latent space model is that although the likelihood depends on information between dyads, the model inherently captures more complex network relationships, for example, relationships that are reciprocal and transitive. This is because the location of an individual depends not only on its direct ties but also on how everyone else in the network is located.

As an example, consider a six person network where two actors have ties with the other four actors (see Figure 2). Look at the two actors in the middle (dark grey) – they have a tie to each other which can convey influence. However, these two actors in the middle are also connected to the same actors (light grey) in the same way, so they would actually show up as very close to each other in the latent space. If the two actors did not share common friends, they would be further apart in latent space. Since they do share common friends, then we can infer something about their similarity on many characteristics. Using latent space to infer their unobserved similarity, and then controlling for it, we can then get at the effect of social influence above and beyond homophily.

[Insert Figure 2 here]

3.4 Estimation

The likelihood function follows from the logistic form:

$$L\left(\theta, X_1, ..., X_N | A\right) = \prod_{i,j} P\left(A_{ij} = 1\right)^{I(A_{ij}=1)} \left[1 - P\left(A_{ij} = 1\right)\right]^{1-I(A_{ij}=1)}$$

The latent coordinates could be estimated through MLE but Hoff et al (2002) warns that the likelihood function is not convex. Instead, they suggest a Bayesian estimation approach which draws latent positions X and other parameters in succession:

1. Start with $X_0 = \tilde{X}$, MLE solution through direct optimization of the likelihood function. This is feasible for small N but difficult for large N. An alternative is to start with the MDS solution to geodesic distances based on the sociomatrix.

2. At iteration t, draw a candidate X^p from a symmetric proposal distribution centered around the previous draw $X^{(t-1)}$ (e.g., multivariate normal)

3. Accept $X^{(t)} = X^p$ with probability $max\left(1, \frac{p\left(A|X^p,\theta^{(t-1)}\right)\pi(X^p)}{p\left(A|X^{(t-1)}\theta^{(t-1)}\right)\pi\left(X^{(t-1)}\right)}\right)$; else $X^{(t)} = X^{(t-1)}$

4. One identification problem with latent space is that translations, reflections, and rotations of a position still generates the same likelihood function albeit different positions. A simple example is a star network – just rotate the star by 90 degrees and you get different positions for each actor. To group together these trivial transformations, the standard approach is to "procruste" a set of positions towards a fixed orientation. We do this by minimizing the distance of a network X to some fixed point X_p (usually the multidimensional scaling solution). We also cannot identify the center of the locations, so we set the mean of each dimension of X to be centered at 0.

5. Update θ with Metropolis-Hastings

We use diffuse prior distribution $\theta \sim MVN(\xi, \Psi)$, $\pi(X_i)$ is $MVN(0, I_k)$ where ξ, Ψ are hyperparameters to be specified by the analyst. We set $\xi = 0, \Psi = 9 \cdot I$ to allow for a wide range of parameter values. The R package `latentnet` estimates latent space models (Krivitsky & Handcock 2008, Krivitsky & Handcock 2009).

Not all networks are identified. For example, a complete network where individuals are connected to every other individual provides no variance to infer the relative distances of any pair of actors. Similarly, a network made of disconnected subnetworks does not shed light on the distances between individuals in separate subnetworks. For our simulations and analysis, we keep the largest connected network to get around this problem.

4 Simulation

To show latent space models can reduce the homophily bias in social influence studies, we set up a series of simulations to explore the approach across different scenarios. First, we describe the basic data generating process and estimating equation. Then, we describe the variations on the basic simulation that we tested. We ran 65,000 independent scenarios and 520 million MCMC iterations in total.

4.1 Data generating process

We generate network and outcome data loosely based on Shalizi & Thomas (2011):

1. We simulate N actors in a social network. Each actor i has a latent characteristic $X_i \sim F_X$ and observed characteristic $Z_i \sim F_Z$.

2. Characteristics shape the friendship network. Let A be a sociomatrix where A_{ij} means that persons i and j are friends. Friendships between actors are undirected, i.e., A is symmetric. We assume homophily, that is, people with similar traits are closer in social distance and therefore are more likely to be friends:

$$\begin{aligned} P(A_{ij}=1) &\propto logit^{-1}\left(\alpha_0-\alpha_1\|Z_i-Z_j\|-\alpha_2\|X_i-X_j\|\right) \\ &= \frac{exp\left(\alpha_0-\alpha_1\|Z_i-Z_j\|-\alpha_2\|X_i-X_j\|\right)}{1+exp\left(\alpha_0-\alpha_1\|Z_i-Z_j\|-\alpha_2\|X_i-X_j\|\right)} \end{aligned}$$

3. The traits also affect an outcome of interest. There is a time series of outcomes:

   - Initialize random starting points $Y_{i0} \sim F_Y^0$
   - $Y_{it} = \beta_0 + \beta_1 Y_{i,t-1} + \beta_2 X_i + \beta_3 Z_i + \beta_4 \frac{\sum_j A_{ij} Y_{j,t-1}}{\sum_j A_{ij}} + \epsilon_{it}$

Peer activity is calculated as $\frac{\sum_j A_{ij} Y_{j,t-1}}{\sum_j A_{ij}}$ which is the average lagged Y among actor i's friends (Aral et al 2009, Shalizi & Thomas 2011). β_4 is the effect of social influence on the outcome variable. The problem with latent homophily is that if β_2 is nonzero, then latent trait X affects the outcome variable, and therefore the omission of X in the estimating equation will generate bias in the estimate of social influence, β_4.

4.2 Variations on the simulation

4.2.1 Basic simulation

The basic simulation has no observed covariates (no Z) and sets parameter values at $\alpha = \{\alpha_0, \alpha_1, \alpha_2\} = \{0, 0, 3\}$ and $\beta = \{\beta_0, \beta_1, \beta_2, \beta_3, \beta_4\} = \{0, 0, 0.4, 0, 0\}$, i.e., no lagged effect and no social influence effect[3]. We assume distributions on $X_i \sim U(0,1)$, $Y_{i0} \sim N(0, 0.25)$, and $\epsilon_{it} \sim N(0, 0.25)$.[4] We run the data generating process for two periods and estimate Y_2 with three models:

- The **naive** estimate attempts to use lagged Y to control for homophily, using independent variables from the prior period to avoid simultaneity: $\hat{Y}_{it} = b_0 + b_1 Y_{i,t-1} + b_4 \frac{\sum_j A_{ij} Y_{j,t-1}}{\sum_j A_{ij}}$
- **Social network statistics** such as centrality play a major role in social network analysis. We test to see whether betweenness centrality, i.e., the number of geodesic paths that run through an actor, and degree, i.e., the number of ties, help control for homophily: $\hat{Y}_{it} = b_0 + b_1 Y_{i,t-1} + b_4 \frac{\sum_j A_{ij} Y_{j,t-1}}{\sum_j A_{ij}} + c_1 Betweenness_i + c_2 degree_i$
- **Latent space** estimates latent coordinates $\hat{X}$ from the sociomatrix A. We input these latent coordinates in the regression equation to control for latent traits X: $\hat{Y}_{it} = b_0 + b_1 Y_{i,t-1} + b_2 \hat{X}_i + b_4 \frac{\sum_j A_{ij} Y_{j,t-1}}{\sum_j A_{ij}}$

[3] We tested a lagged variable in the data generating process with the same results

[4] Values from prior studies that use simulation studies to demonstrate bias in social influence studies (Shalizi & Thomas 2011).

In the basic simulation and its variations, we simulate and use data from time period $t = 1$ to predict Y_{i2}. Since there is only one observation per individual, we cannot use random or fixed effects to control for bias stemming from the omission of X.

4.2.2 No latent trait effect on Y

If the underlying traits that drive social ties do not affect Y, then latent homophily should not create any bias in the effect of social influence. We change $\beta_2 = 0$ to remove the relationship between latent traits X and the outcome variable Y. We expect the bias to be close to zero in all three estimating equations: naive, social network statistics, and latent space.

4.2.3 Social influence effect > 0

To provide assurance that latent space does not merely bias the estimate towards zero, we set the effect of social influence to be greater than 0 to see if latent space will remove the bias in the presence of social influence. Therefore, we set $\beta_4 = 0.3$.

4.2.4 Y is binary

We test latent space when Y is binary. The data generating process is amended as follows:

- $Y_{it} = \beta_0 + \beta_1 Y^{bin}_{i,t-1} + \beta_2 X_i + \beta_3 Z_i + \beta_4 \frac{\sum_j A_{ij} Y_{j,t-1}}{\sum_j A_{ij}} + \epsilon_{it}$
- $Y^{bin}_{it} = 1$ if $Y_{it} > 0$
- $Y^{bin}_{it} = 0$ otherwise; and

The estimating equations specifications are as in the basic simulation, but we estimate Y^{bin}_{i2} with probit regression instead of linear regression.

4.2.5 Alternate distributions of X

To check that the results do not depend on distributional assumptions, we relaxed the assumption that latent traits X are drawn from a uniform distribution. We test two additional distributions: $X_i \sim Beta(3,3)$, a **hump-shaped** density function where traits are concentrated towards the middle of the distribution, and $X_i \sim Beta(0.4, 0.4)$, a **U-shaped** density function where traits are polarized towards the extremes. We expect that when traits are concentrated towards the middle of the distribution, there is more homogeneity and therefore less information to distinguish one actor from another. So, the estimated latent coordinates will do a significantly poorer job of reducing bias.

4.2.6 Observed covariates

Existing methods use observed covariates to separate social influence from homophily. However the performance of observed covariates would depend on whether it accounts for a small or large portion of total homophily. If the observed covariate is uncorrelated with latent traits which drive latent homophily, then controlling for observed covariates should remove a small proportion of the bias. On the other hand, if the covariate is highly correlated with the latent trait, then merely controlling for the covariate could reduce a larger proportion of the bias.

To operationalize covariates in the simulation, we draw $Z_i \sim \tau_1 U_i + \tau_2 \left(X_i - 0.5\right) + 0.5$ where $U_i \sim U\left(-0.5, 0.5\right)$; the values of $\{\tau_1, \tau_2\}$ control the correlation between X and Z. We test three correlation levels:

- Correlation $= 0$: $\{\tau_1, \tau_2\} = \{1, 0\}$
- Correlation ≈ 0.5: $\{\tau_1, \tau_2\} = \{0.85, 0.5\}$
- Correlation ≈ 0.8: $\{\tau_1, \tau_2\} = \{0.5, 0.85\}$

We allow Z to have the same effect as X on Y, i.e., $\alpha_1 = 3$ and $\beta_3 = 0.4$.

4.2.7 Panel: random & fixed effects

In panel data, estimating random or fixed effects are standard econometric methods to accommodate omitted variables. However, they have limited value in social influence studies.

Random effects are expected to be biased because the omitted variable (latent trait) is correlated with regressors (social influence). The bias in the estimate of the regressor social influence is induced because of this correlation, and the only way that random effects is unbiased is if we impose the restriction that there is no latent homophily. Fixed effects allows for correlation between the omitted variable and regressor, but does not perform well when the data is autoregressive.

To compare the performance of latent space with random and fixed effects estimates, we perform a series of simulations with short (3 time periods) and long (20 time periods) panel data. We test two cases: when the data is not autoregressive: $\beta_1 = 0$ and when the data is autoregressive: $\beta_1 = 0.7$. For each case, we estimate random and fixed effects including and excluding lagged dependent variables as regressors in the estimating equation:

- Without lagged dependent variable: $\hat{Y}_{it} = b_{0i} + b_4 \frac{\sum_j A_{ij} Y_{j,t-1}}{\sum_j A_{ij}}$
- With lagged dependent variable $\hat{Y}_{it} = b_{0i} + b_1 Y_{i,t-1} + b_4 \frac{\sum_j A_{ij} Y_{j,t-1}}{\sum_j A_{ij}}$

4.2.8 Additional simulation details

We simulate 100 individuals per simulation and estimated regression coefficients using OLS (except in binary Y where we use probit regression)[5]. Fixed effects regressions were operationalized with indicator variables and random effects regressions were estimated using the *lme*4 package in *R*. We tested 4 (no social influence, no homophily, social influence $>$ 0, binary Y) + 2 (hump-shaped and U-shape distribution on X) + 3 (observed covariate correlation with latent trait = 0, 0.5, 0.8) + 2 x 2 (short vs. long panel and data generated with & without lagged dependent variable) = 13 scenarios in total. We ran 5,000 separate independent repetitions of each scenario for a total of 5,000 x 13 = 65,000 runs. For each sociomatrix in each repetition, we estimate latent space by running the MCMC chain for 8,000 iterations with 3,000 burn-in, so we ran MCMC for 520 million iterations.

[5] We find that the more observations in the network, the better latent space performs in reducing bias

4.3 Simulation results

The main statistic of interest is the bias in the coefficient of social influence with each approach. We calculate the bias as the difference between the mean of the estimated coefficient across 5,000 runs of the scenario and the true coefficient value in the simulation. We report (1) how much bias is induced by homophily and more importantly (2) what proportion of bias is reduced. Additional statistics for the simulations can be found in the online appendix.

Figure 3 shows results from the basic simulation and its variants. In the basic simulation, the true value of social influence zero, but the estimated coefficient is centered around 0.6, so the bias is around 0.6. Social network statistics reduces the estimated bias by 7%, not a very significant amount. This pattern continues across all scenarios. When we use latent space as control for homophily, we reduce the bias by 93%. In the scenario with no latent homophily, we remove the link between the latent trait and the outcome variable. As a consequence, there is minimal bias on the social influence coefficient. In the next two scenarios, latent space continues to reduce bias when the true effect of social influence is 0.3 or when the outcome variable Y is binary. When we perturb the distribution of the latent trait to be U-shaped, latent space continues to be effective in reducing the bias by about 90%. When the distribution of the latent trait is hump-shaped, there is less variance in the latent trait across actors in the network. Therefore, there is a weaker relationship between the latent trait and the outcome variable, meaning there is less bias in the social influence coefficient (around 0.2). There is also less information to estimate latent space since there is less variance in the latent trait, and so the approach reduces 76% of the bias. Across these simulations, latent space removes most of the bias that is generated by homophily.

[Insert Figure 3 here]

In the next set of simulations, we consider scenarios with observed covariates (Figure 4). In the scenario with zero correlation between the covariate and latent trait, controlling for the covariate removes 29% of the bias. As the correlation between the covariate and the latent trait grows, including the covariate in the regression helps remove more bias from the estimate of social influence effect. This pattern is articulated by Shalizi & Thomas (2011) who state that methods that use covariates (e.g., matching on observed covariates) only work if there is high correlation between the observed and latent traits.

[Insert Figure 4 here]

Using latent space coordinates only, approximately 60-70% of the bias is reduced. Even without information on observed covariates, latent space is able to pick up both observed and latent covariates that drive homophily. In prior scenarios where there is only one latent trait driving homophily, latent space reduced bias by about 90-95%. Here, homophily is driven by two traits, so there is greater dimensionality to locate actors in latent space and therefore greater sparsity and higher uncertainty, resulting in lower bias reduction. We do a fair job in reducing bias even if we did not have information on the observed covariates.

The results for panel data scenarios are summarized in Figure 5 and figure 6. In figure 5, the data generating process has no autoregressive term. The fixed effect model performs very well to remove almost all the bias. Random effects can be thought of as a weighted mean between the pooled (naive) estimate and fixed effects. When there is a short panel (3 periods), the pooled estimate dominates and random effects do not shift the estimated coefficient very much. With a longer panel (20 periods), random effect moves towards the fixed effect solution and removes some of the bias (37%) but does not perform as well as the fixed effect or the latent space solution.

[Insert Figure 5 here]

[Insert Figure 6 here]

When the data is autoregressive (figure 6), fixed effects does poorly, increasing the bias from the naive model, especially in a short panel scenario, reflecting the phenomenon documented by Nickell (1981). When the panel is longer (20 periods), adding a lag to fixed effects estimation reduces 80% of the bias (highlighted in light green), but not as much as latent space which reduces 100% of the bias. Random effects also perform poorly when applied to autoregressive data.

4.4 Summary of simulations

In this section, we describe the latent space model and compare its performance with competing methods in an extensive set of simulations. Across all the scenarios, latent space reduces a significant proportion of bias and performs as well or better than existing methods. Latent space reduces bias whether the true effect of social influence is zero or nonzero, whether the outcome variable Y is binary or continuous, and under different distributions of the latent trait.

5 Impact of social influence on mobile apps adoption

Having established an approach that reduces bias in social influence studies, we can now turn to explore our main substantive question: Do peers influence mobile app installations in a social network?

Widespread smartphone penetration provides increasing clarity that consumers are moving from PC to mobile and more notably to mobile apps on smartphones. Mobile apps make up 82% of time on the mobile phone vs. browsers, yet four in five consumers dislike advertisements on mobile phones (Gupta 2013). Because of these findings, many companies desire to engage and connect with consumers via mobile apps. Since mobile apps are often social in nature, especially those that are linked to a social network platform, it is natural to think that peer influence plays a role in product adoption.

However, identifying peer effects in mobile app adoption is not straightforward because social influence can be confounded with homophily. When we see one friend uses an app and another installs a new app, is it their latent similarity that drives adoption or is it because of peer influence? We use latent space to untangle homophily from social influence. First, we use the social network structure to extract latent traits that govern friendship formation. Then, we use the estimated latent space coordinates to control for homophily when estimating the effect of peer influence on adoption. We find that homophily can inflate the effect of peer influence by 40%, but even after controlling for homophily, peers still drive more than a quarter of all mobile app installations.

5.1 Data

We use data from mixi, a leading social network in Japan. The data was shared by the company using a password protected medium. User IDs were scrambled to prevent identification of users on the platform. Other personally identifiable information were also scrambled. The data consists of the complete social graph of all its members, a total of 600 million connections among its 22 million users in October 2010. The social network provides a platform for apps made by application developers. Most of the apps are games (similar to Zynga games on Facebook) that are also linked with the social network.

The dataset records the time when an individual uses mobile apps or installs apps over seven days. We use app usage and installation information from the first six days to predict mobile app installation in the seventh day, avoiding simultaneity issues. Because adoption for individual apps is low in the observation window of the data, we could limit the analysis only to those with a large number of installations. However, this would ignore a majority of apps. In order to include information of all apps, especially those in the long tail, and to focus on the installation of apps on a social network not on individual apps, we aggregate usage and installation information across all apps.

We use snowball samples from the network because current latent space approaches have sample size limitations and we are not able to use the entire network (22 million users). Because mobile app adoption is sparse on a single day, we sample from a seed who adopts a mobile app in day seven to ensure that at least one actor in the network adopts, then include all actors two degrees from the seed in the sample. We keep only samples that contain 300 and 600 actors to simplify estimation of latent space, for a total of 28,007 actors across 62 snowball samples. Data is summarized in Table 1 below.

[Insert Table 1 here]

5.2 Model & estimation

The dependent variable of the model is whether an individual installs a mobile app or not. This latent utility of installing an app is allowed to vary by peer usage, own usage, and personal traits, with heterogeneous coefficients across samples:

$$u_{sit} = \beta_{s0} + \beta_{s1} Peer_{i,t-1} + \beta_{s2} M_{i,t-1} + \beta_{s3} Deg_i + \beta_{s4} X_i + \beta_{s5} LS_{si} + \epsilon_{sit}$$

where

- s denotes the sample, i denotes the individual, j denotes friends, and t denotes the time period.

- $Peer_{i,t-1} = \frac{\sum_j A_{ij} M_{j,t-1}}{\sum_j A_{ij}}$ is the proportion of friends who use mobile apps in days 1-6, where $A_{ij} = 1$ if persons i and j are friends, and $M_{j,t-1} = 1$ if person j used mobile apps in days 1-6, and 0 otherwise.

- $M_{i,t-1}$ captures whether person i accessed any mobile apps in the prior six days or not. This variable takes on 1 if the person used mobile apps, and 0 otherwise.

- $Deg_i = log \sum_j A_{ij}$ is log number of degrees, or the number of friends person i has.

- X_i are time-invariant covariates for person i, namely, gender, age, log number of photos and log number of comments. Although the latter two may vary over time, this is unlikely to change significantly over the observation period of a week.

- LS_i are the latent space coordinates for individual i.

- ϵ_{sit} are independent type I extreme value random variables.

- $\beta'_s = \left[\begin{array}{cccccc} \beta_{s0} & \beta_{s1} & \beta_{s2} & \beta_{s3} & \beta_{s4} & \beta_{s5} \end{array} \right]$ are the coefficients to be estimated in the model.

We formulate this as a logistic regression problem where individual i from sample s installs ($Y_{sit} = 1$) if the latent utility of installing is greater than 0 at time t:

$$\begin{aligned} Y_{sit} &= 1 \text{ if } u_{sit} > 0 \\ &= 0 \text{ otherwise} \end{aligned}$$

Latent space coordinates are estimated on each snowball sample and used as input in the regression. We impose a hierarchy to estimate the coefficients in the model since the effects of variables are expected to vary by sample with shrinkage towards a set of population hyperparameters:

$$\begin{aligned} \beta_s &= \delta + \omega_s \\ \omega_s &\sim N(0, V_\beta) \end{aligned}$$

with prior distributions

$$\begin{aligned} \delta &\sim N\left(\bar{\delta}, A_{\delta}^{-1}\right) \\ (V_\beta)^{-1} &\sim W(\nu, V) \end{aligned}$$

where $W(\nu, V)$ is the Wishart distribution with ν degrees of freedom and V location parameter.

The hierarchy allows us to make population level statements about social influence while allowing the patterns of homophily and social influence to vary across samples. This is important because while we might expect that homophily plays a big role in mobile app adoption and therefore generate a large bias in social influence if ignored, the phenomenon may vary by consumers in the population, so the bias may be large or small. The hierarchy will help uncover which groups of consumers are most affected by homophily bias and which are not.

We estimate the model with MCMC using a Metropolis algorithm. At iteration t we accept a candidate draw for each sample with probability proportional to the ratio of posterior density:

$P\left(\text{accept } \beta_s^{new}\right) = min\left[1, \frac{\pi\left(\beta_s^{new}|\delta^{(t-1)}, V_\beta^{(t-1)}\right)p(Y|\beta_s^{new}, X)}{\pi\left(\beta_s^{(t-1)}|\delta^{(t-1)}, V_\beta^{(t-1)}\right)p\left(Y|\beta_s^{(t-1)}, X\right)}\right]$

otherwise, keep $\beta_s^{(t)} = \beta_s^{(t-1)}$

Then, given the draws for all samples $\left\{\beta_s^{(t)}\right\}$, we draw parameters from distributions centered at the weighted average of the priors and the mean and variance of these samples:

$$\begin{aligned}
\delta^{(t)} &\sim N\left[\bar{d}, V_{beta}^{(t-1)} \otimes \left(S + A_\delta\right)^{-1}\right] \\
\left(V_\beta^{(t)}\right)^{-1} &\sim W\left(\nu + S, V + O\right) \\
\text{where } \bar{d} &= \left(S + A_\delta\right)^{-1}\left(\sum_s \beta_s + A_\delta \bar{d}\right) \\
O &= \sum_s \left(\beta_s - \delta^{(t)}\right)\left(\beta_s - \delta^{(t)}\right)' \\
S &= \text{Number of snowball samples}
\end{aligned}$$

We estimate four models for mobile app adoption. The first model is a naive model, including only peer usage, own usage, and degrees. The second model adds covariates but no latent space, similar to existing methods that control for observed homophily but not latent homophily. The third model adds latent space coordinates but no covariates (gender, age, photos, comments), reflecting a scenario where latent space is estimated but no variables are available to control for observed homophily. The last model controls for both observed and latent homophily using covariates and latent space coordinates. When latent space coordinates are involved, we use 1-10 dimensions because we do not know a priori how many latent dimensions may be appropriate for the network.

We demean the data, divide each variable by its standard deviation and estimate the model with MCMC. The MCMC chain is run for 400,000 iterations, keeping every 5 draws. Convergence is checked visually with plots of coefficients after a burn in of 200,000 draws. We check model fit with log marginal likelihood (Chib 1995, Rossi et al 1996; for calculating LML, Rossi et al 1996 electronic companion appendix 3 has a clear step-by-step walkthrough). We set $\bar{\delta} = 0, A_\delta = 0.1, \nu =$ number of X variables $+ 3, V = \nu I$.

5.3 Results

In this section, we report the model fit and parameter estimates from our MCMC draws from the posterior distribution. We quantify the impact of social influence and show how far off estimates would be if we ignored latent homophily. We finish with some managerial implications.

5.3.1 Model fit

We report the log marginal likelihood of different model specifications in a plot in figure 7. The full model (Model 4) with both covariates and latent space of 8 dimensions fits the data best. The improvement is significant over either the covariates only model or the latent space only model (Models 2 and 3). This suggests latent space captures an aspect of the data that is separate and additional to observed homophily.

[Insert Figure 7 here]

5.3.2 Parameter estimates

Table 2 shows the estimated posterior mean and 95% posterior interval for population parameters across different specifications of the model. In all models, own mobile app activity from the past six days is positively correlated with new app installation. Female consumers are more likely to install apps than their male counterparts.

The peer activity coefficient, a measure of peer influence, is positive and significant. The parameter estimate decreases when we control for latent homophily with latent space (from Model 2 to Model 4).

As expected, the coefficients for latent space only make sense at the sample level and therefore are not significantly different from zero at the population level. This is because latent space captures different aspects of latent homophily across different samples samples from the population. To check the face validity of the latent space model, we compare the estimated latent distances between friends and non-friends in the sample. On average, distances between non-friends are twice as far apart as friends. This supports the idea that there is significant latent homophily in the data.

[Insert Table 2 here]

5.3.3 Peer impact on mobile app adoption

To quantify the impact of social influence, we simulate two worlds: one with social influence and another without social influence. In the current dataset, individuals can be influenced by peers who use mobile apps. For the world without social influence, we predict the level of mobile app adoption when peer influence is removed.

When we remove peer influence, the average mobile app adoption likelihood falls from 6.80% to 4.98% (under model 4). The decline of 1.82% is significantly different from zero. From these numbers, we can attribute 1.82% / 6.80% = 27% of all mobile app installations to social influence (95% posterior interval: [25%, 28%]).

Next, we plot the change in probability of adoption with and without social influence for individuals (Figure 8). For 5% of individuals, peers play no role at all because none of their peers use mobile apps. For the middle half of individuals, peer influence accounts for 23% to 43% of the motivation for app adoptions.

[Insert Figure 8 here]

5.3.4 Bias from latent homophily

To assess the importance of accounting for latent homophily, we repeat the simulation in the previous section but use model 2 which does not control for latent homophily. With this model, we find that social influence accounts for 38% of all installations (see figure 9). Thus, ignoring homophily resulted in a $\frac{38\%-27\%}{27\%} = 40\%$ inflation of social influence effects (p < 0.001, based on posterior distributions of difference in peer effects). The size of the bias due to latent homophily is in line with prior research which find that latent homophily can inflate social influence by 10-50% (Tucker 2008, Hartmann 2010).

[Insert Figure 9 here]

Our modeling approach allows us to estimate the bias distribution across different samples from the network. We plot the histogram of biases as a percent of effect size in Figure 10. This is how much social influence is overestimated if we do not account for latent homophily. We find positive bias in every sample. On the low end, ignoring latent homophily would bias social influence estimates by 15-20%. On the high end, the bias could inflate the effect by more than 100%. The variability across samples could be due to sampling error or because homophily and social influence have heterogeneous effects in the network. At this point, we are not able to tease these two effects apart.

[Insert Figure 10 here]

5.3.5 Managerial targeting

Firms may be interested in increasing mobile app adoption by motivating peer app usage. To illustrate the power of peer influence, we simulate a scenario where five additional friends in each individual's network use mobile apps. For example, if an individual currently has two out of twenty friends use mobile apps, we predict the probability of adoption for the individual when seven out of twenty friends use mobile apps (capped at 100% of friends). With targeted efforts to stimulate app usage, we find that overall weekly adoption rates jump from 6.80% to 8.58%, a 26% increase in installations.

6 Conclusion

We propose using latent space to model homophily to reduce bias in social influence studies. Our extensive set of simulations shows that inclusion of latent coordinates in regressions reduces bias when estimating social influence effects. We apply the method to investigate whether mobile app adoption is influenced by peers and find that peers drive more than a quarter of all mobile app adoptions, even after controlling for homophily.

Ignoring homophily could result in extremely biased estimates. In our setting, we find that latent homophily could inflate the proportion of adoption attributed to social influence by 40%. Alarmingly, there is large variation across samples; the bias in some samples was as large as the effect size itself. Based on our simulations and empirical findings, we conclude that it is necessary to control for latent homophily when estimating social effects. Latent space presents an appealing and intuitive way to do so.

There are some methodological gaps for further research. The approach we use involves two step estimation which may have implications for standard errors. We do not explore this issue because the primary focus of this paper is to address a more fundamental social science question: whether we can detect and reduce the bias induced by homophily on the coefficient of social influence.

Separating social influence from homophily continues to be a challenging area for research. Providing bounds for the effect of social influence is of great interest to the social influence research community but require additional work (Aral et al 2009, Shalizi & Thomas 2011). If certain assumptions are satisfied, bounds for the regression coefficients in the presence of proxy variables can be stated (Klepper & Learner 1984, Bollinger 2003, Bollinger & Minier 2012). Since latent space coordinates act as proxy variables for latent traits that drive homophily, there may be potential to provide bounds for social influence effects. We leave this for future research.

Tables & Figures

Figure 1: Difference between social influence and homophily

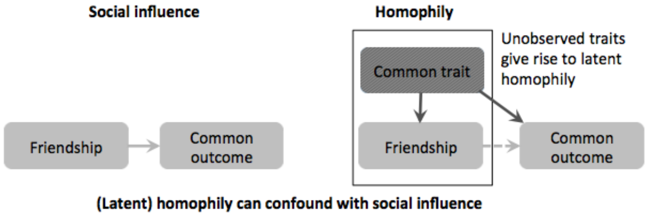

Social influence
Homophily
Unobserved traits give rise to latent homophily
Common trait
Friendship
Common outcome
Friendship
Common outcome
(Latent) homophily can confound with social influence

Figure 2: Illustrative network of six actors

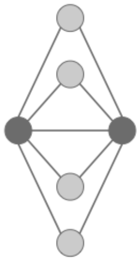


Figure 3: Bias reduction with latent space across different simulation scenarios

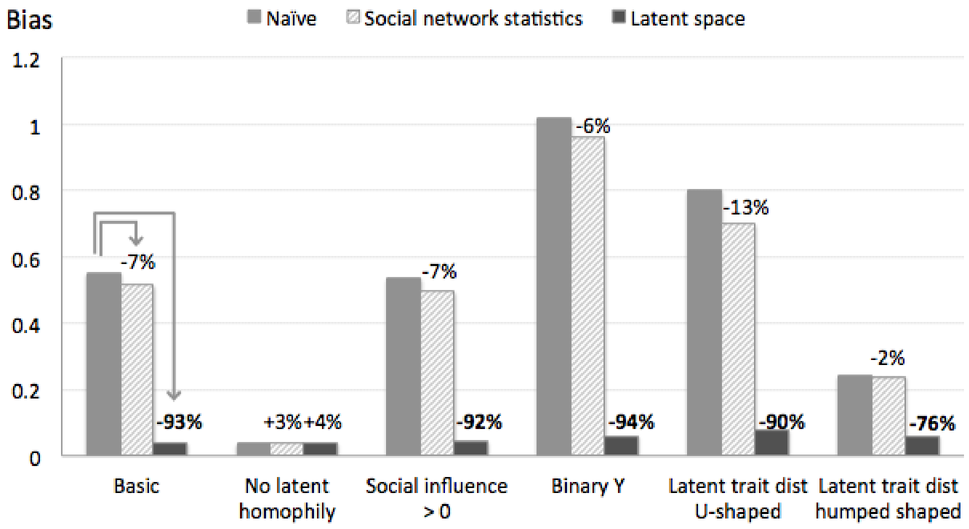


Figure 4: Bias reduction of observed covariates vs. latent space

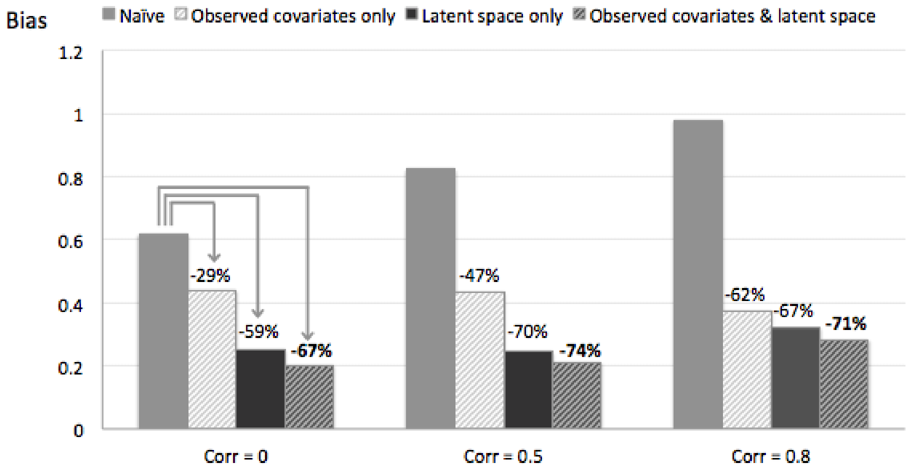


Figure 5: Bias reduction of fixed & random effects vs. latent space (panel data)

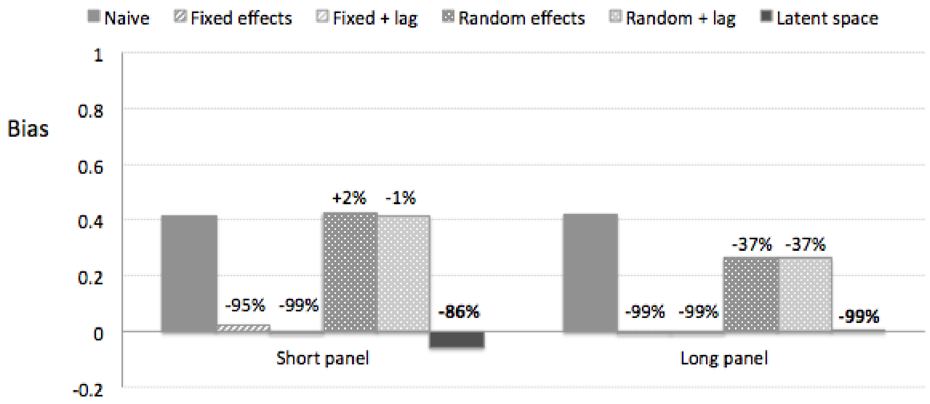


Figure 6: Bias reduction of fixed & random effects vs. latent space (autoregressive data)

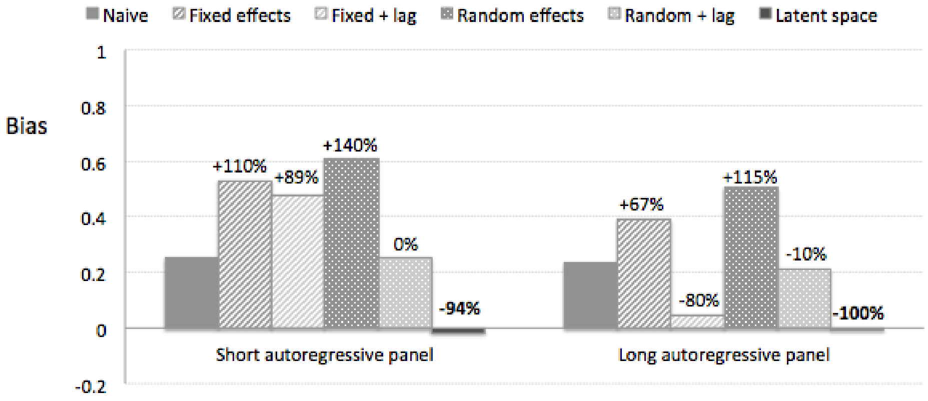


Table 1: Data summary statistics (of means across 62 samples)

Variable	mean	SD	min	max
% install app in day 7	6.4%	4.0%	2.8%	3.3%
Peer usage (% friends who used mobile apps in days 1-6)	14.1%	4.3%	8.4%	9.4%
Own usage (% who used mobile apps in days 1-6)	14.8%	61.3%	6.8%	52.8%
Log degrees (No. of connections)	4.025	0.313	3.046	4.819
Female	0.541	0.119	0.257	0.820
Age	27.98	4.187	19.90	37.45
Log number of photos	2.638	0.449	1.463	3.759
Log number of comments	3.012	0.442	0.934	3.648

Figure 7: Log marginal likelihood (LML) fit by dimensions of latent space

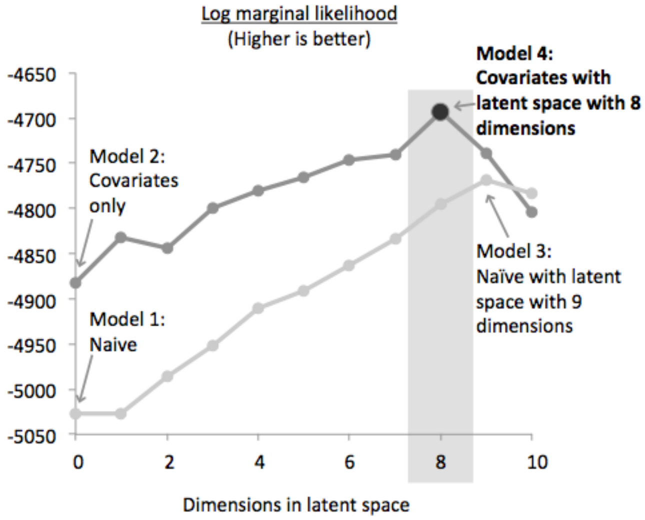


Table 2: Regression coefficients estimates (Logistic regression population parameters)

Model	1	2	3	4
	Naive	Covariates	Naive, LS	Covariates, LS
Intercept	**-4.48** [-4.82, -4.15]	**-4.59** [-5.05, -4.15]	**-4.77** [-5.07, -4.45]	**-5.28** [-5.67, -4.85]
Peer activity	**3.21** [2.66, 3.79]	**3.48** [2.89, 4.15]	**2.74** [2.15, 3.23]	**2.53** [2.17, 2.88]
Own activity	**2.73** [2.53, 2.93]	**2.76** [2.54, 2.98]	**3.08** [2.83, 3.34]	**3.13** [2.87, 3.40]
Degree	0.07 [-0.04, 0.18]	0.01 [-0.13, 0.16]	0.01 [-0.14, 0.15]	0.04 [-0.13, 0.21]
Female	-	**0.21** [0.02, 0.41]	-	**0.28** [0.05, 0.50]
Age	-	-0.01 [-0.011, 0.10]	-	-0.01 [-0.14, 0.13]
Photos	-	0.06 [-0.05, 0.18]	-	0.08 [-0.07, 0.23]
Comments	-	0.02 [-0.11, 0.15]	-	0.03 [-0.14, 0.19]
LS 1	-	-	0.01 [-0.12, 0.14]	-0.01 [-0.15, 0.13]
LS 2	-	-	0.01 [-0.14, 0.12]	0.01 [-0.13, 0.15]
LS 3	-	-	0.00 [-0.12, 0.13]	0.00 [-0.14, 0.14]
LS 4	-	-	0.02 [-0.11, 0.15]	-0.01 [-0.15, 0.13]
LS 5	-	-	-0.01 [-0.13, 0.12]	0.00 [-0.14, 0.14]
LS 6	-	-	0.00 [-0.13, 0.13]	0.00 [-0.13, 0.14]
LS 7	-	-	0.01 [-0.12, 0.13]	-0.01 [-0.14, 0.14]
LS 8	-	-	0.00 [-0.13, 0.12]	0.00 [-0.14, 0.14]

Table reports mean of posterior distribution; [] contain 2.5% and 97.5% posterior interval; bold = 95% posterior interval does not cover 0

Note: Model 3 LS 9: 0.00 [-0.13, 0.12]

Figure 8: Distribution of peer effects as a driver of mobile app adoption

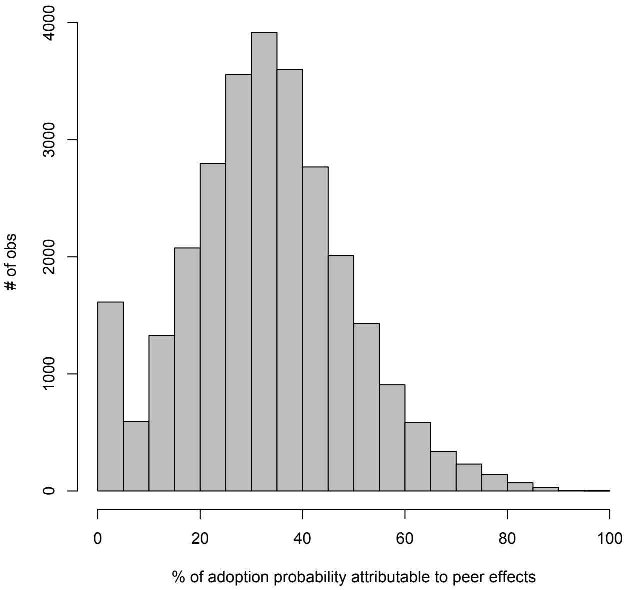


Figure 9: Drivers of mobile app adoption (% of total adoptions)

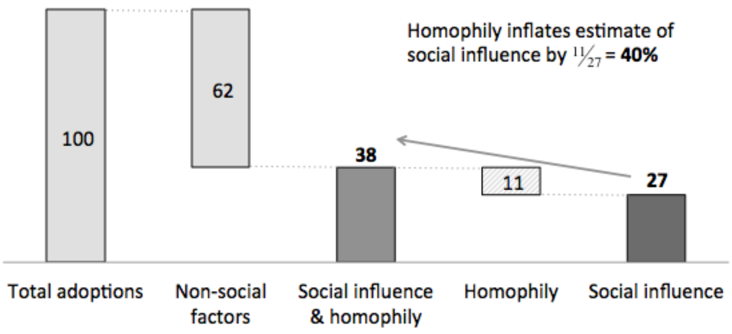
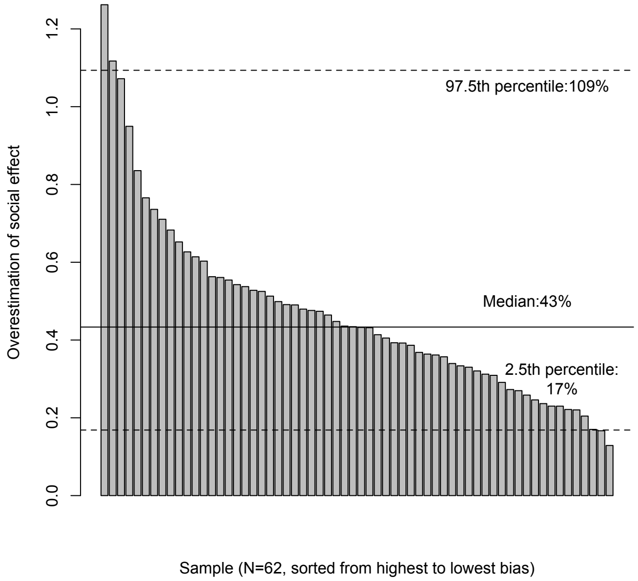


Figure 10: Bias to effect size ratio in each sample



Online appendix: Detailed simulation statistics

In the text, we summarized the simulations by bias in the naive scenario and the bias reduction of various methods as a percent of the bias in the naive scenario. Here, we show the average bias with each method, and also present the standard deviation, minimum, and maximum. One measure of interest is whether the estimated values is centered around the true value, captured by the percent of coefficients larger than the true value. The closer this statistic is to 50%, the more centered the sampling distribution of the coefficient. Latent space reduces a significant amount of bias and centers the distribution of the estimated coefficients across all scenarios. In panel scenarios, latent space helps guard against autoregressive data generating processes, which can severely bias random and fixed effects estimates (tables A8a and A8b).

The true value of social is 0 except in Table A3 where the true value is 0.3.

Table A1: Basic simulation results

Statistic	Naive	Network	Latent space (LS)
Mean (true val = 0)	0.552	0.515	0.041
SD	0.602	0.620	0.665
Min	-1.793	-1.799	-2.568
Max	2.779	3.178	3.215
Bias reduction	-	-7%	-93%
% above true value	83%	80%	52%

Table A2: When there is no latent homophily bias

Statistic	Naive	Network	LS
Mean(true val = 0)	0.038	0.039	0.040
SD	0.636	0.645	0.662
Min	-2.174	2.317	-2.415
Max	3.194	3.535	3.116
Bias reduction	-	+3%	+4%
% above true value	53%	53%	53%

Table A3: Social influence > 0, true value = 0.3

Statistic	Naive	Network	LS
Mean (true val = 0.3)	0.835	0.795	0.343
SD	0.601	0.611	0.672
Min	1.258	-1.385	-1.568
Max	2.781	2.816	2.729
Bias reduction	-	-7%	-92%
% above true value	83%	80%	54%

Table A4: Binary Y

Statistic	Naive	Network	LS
Mean (true val = 0)	1.019	0.958	0.060
SD	1.791	1.878	1.972
Min	-5.882	-6.023	-8.191
Max	7.933	8.819	7.623
Bias reduction	-	-6%	-94%
% above true value	73%	70%	51%

Table A5a: Latent traits are distributed U-shaped

Statistic	Naive	Network	LS
Mean (true val = 0)	0.800	0.698	0.079
SD	0.464	0.505	0.608
Min	-1.455	-1.568	1.848
Max	3.044	3.078	2.655
Bias reduction	-	-13%	-90%
% above true value	96%	91%	55%

Table A5b: Latent traits are distributed hump-shaped

Statistic	Naive	Network	LS
Mean (true val = 0)	0.243	0.238	0.059
SD	0.733	0.743	0.749
Min	2.492	-2.548	-2.621
Max	3.091	3.024	3.238
Bias reduction	-	-2%	-76%
% above true value	63%	63%	53%

Table A6a: Observed covariate and latent trait has 0 correlation

Statistic	Naive	Obs only	LS without observed	LS with observed
Mean (true val = 0)	0.619	0.440	0.252	0.202
SD	0.304	0.302	0.351	0.352
Min	-0.397	-0.424	-1.034	-1.040
Max	2.019	1.992	1.994	1.991
Bias reduction	-	-29%	-59%	-67%
% above true value	99%	98%	77%	71%

Table A6b: Observed covariate and latent trait has 0.5 correlation

Statistic	Naive	Obs only	LS without observed	LS with observed
Mean (true val = 0)	0.823	0.434	0.245	0.212
SD	0.291	0.298	0.369	0.364
Min	-0.118	-0.294	-1.009	-0.982
Max	2.585	1.918	1.847	1.953
Bias reduction	-	-47%	-70%	-74%
% above true value	100%	97%	75%	72%

Table A6c: Observed covariate and latent trait has 0.8 correlation

Statistic	Naive	Obs only	LS without observed	LS with observed
Mean (true val = 0)	0.980	0.374	0.321	0.284
SD	0.292	0.313	0.360	0.356
Min	0.091	-0.603	-0.826	-0.815
Max	2.270	1.679	1.832	1.849
Bias reduction	-	-62%	-67%	-71%
% above true value	100%	93%	84%	79%

Table A7a: Not autoregressive, 3 periods, FE = Fixed, RE = Random effects

Statistic	Naive	FE	FE + lag	RE effects	RE + lag	LS
Mean (true val = 0)	0.416	0.020	-0.003	0.426	0.413	-0.057
SD	0.332	0.353	0.308	0.338	0.333	0.320
Min	-0.776	-1.210	-1.099	-0.704	-0.766	-1.112
Max	1.658	1.473	1.199	1.691	1.658	1.069
Bias reduction	-	-95%	-99%	+2%	-1%	-86%
% above true value	89%	50%	48%	90%	89%	42%

Table A7b: Not autoregressive, 20 periods, FE = Fixed, RE = Random effects

Statistic	Naive	FE	FE + lag	RE effects	RE + lag	LS
Mean (true val = 0)	0.419	-0.004	-0.004	0.263	0.264	0.005
SD	0.135	0.118	0.118	0.129	0.130	0.118
Min	-0.064	-0.404	-0.397	-0.194	-0.194	-0.392
Max	0.870	0.378	0.372	0.753	0.767	0.368
Bias reduction	-	-99%	-99%	-37%	-37%	-99%
% above true value	100%	48%	48%	98%	98%	51%

Table A8a: Autoregressive, 3 periods, FE = Fixed, RE = Random effects

Statistic	Naive	FE	FE + lag	RE effects	RE + lag	LS
Mean (true val = 0)	0.253	0.531	0.479	0.608	0.254	-0.016
SD	0.213	0.250	0.252	0.241	0.214	0.204
Min	-0.521	-0.508	-0.626	-0.641	-0.521	-0.859
Max	1.519	1.648	1.567	1.739	1.519	0.985
Bias reduction	-	+110%	+89%	+140%	0%	-94%
% above true value	89%	98%	97%	99%	89%	46%

Table A8b: Autoregressive, 20 periods, FE = Fixed, RE = Random effects

Statistic	Naive	FE	FE + lag	RE effects	RE + lag	LS
Mean (true val = 0)	0.235	0.391	0.048	0.505	0.211	-0.001
SD	0.061	0.144	0.071	0.130	0.066	0.061
Min	-0.013	-0.322	-0.254	-0.128	-0.028	-0.266
Max	0.450	0.856	0.326	0.957	0.446	0.248
Bias reduction	-	+67%	-80%	+115%	-10%	-100%
% above true value	100%	99%	76%	100%	100%	50%

References

Aigner, D. J. (1974). MSE dominance of least squares with errors-of-observation. Journal of Econometrics, 2(4), 365-372.

Angrist, J. D., & Pischke, J. S. (2008). Mostly harmless econometrics: An empiricist's companion. Princeton University Press.

Ansari, A., Koenigsberg, O., & Stahl, F. (2011). Modeling multiple relationships in social networks. Journal of Marketing Research, 48(4), 713-728.

Aral, S., & Walker, D. (2011a). 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. (2011b). Identifying social influence in networks using randomized experiments. Intelligent Systems, IEEE, 26(5), 91-96.

Aral, S., & Walker, D. (2012). Identifying influential and susceptible members of social networks. Science, 337(6092), 337-341.

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.

Aronow, P. M., & Samii, C. (2012). Estimating Average Causal Effects Under General Interference.

Bai, J. (2009). Panel data models with interactive fixed effects. Econometrica, 77(4), 1229-1279.

Bakshy, E., Eckles, D., Yan, R., & Rosenn, I. (2012, June). Social influence in social advertising: evidence from field experiments. In Proceedings of the 13th ACM Conference on Electronic Commerce (pp. 146-161). ACM.

Bollinger, C. R. (2003). Measurement error in human capital and the black-white wage gap. Review of Economics and Statistics, 85(3), 578-585.

Bollinger, C. R., & Minier, J. (2012). On the robustness of coefficient estimates to the inclusion of proxy variables. Working paper. Lexington, KY: University of Kentucky.

Bowers, J., Fredrickson, M. M., & Panagopoulos, C. (2013). Reasoning about Interference Between Units: A General Framework. Political Analysis, 21(1), 97-124.

Bramoullé, Y., Djebbari, H., & Fortin, B. (2009). Identification of peer effects through social networks. Journal of econometrics, 150(1), 41-55.

Braun, M., & Bonfrer, A. (2011). Scalable inference of customer similarities from interactions data using Dirichlet processes. Marketing Science, 30(3), 513-531.

Cacioppo, J. T., Fowler, J. H., & Christakis, N. A. (2009). Alone in the crowd: the structure and spread of loneliness in a large social network. Journal of personality and social psychology, 97(6), 977.

Chib, S. (1995). Marginal likelihood from the Gibbs output. Journal of the American Statistical Association, 90(432), 1313-1321.

Christakis, N. A., & Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. New England journal of medicine, 357(4), 370-379.

Christakis, N. A., & Fowler, J. H. (2008). The collective dynamics of smoking in a large social network. New England journal of medicine, 358(21), 2249-2258.

Chui, M., Manyika, J., Bughin, J., Dobbs, R., Roxburgh, C., Sarrazin, H., Sands, G., & Westergren, M. (2012). The social economy: Unlocking value and productivity through social technologies.

Cohen-Cole, E., & Fletcher, J. M. (2008). Is obesity contagious? Social networks vs.

environmental factors in the obesity epidemic. Journal of Health Economics, 27(5), 1382-1387.

Fowler, J. H., & Christakis, N. A. (2008). The dynamic spread of happiness in a large social network. BMJ: British medical journal, 337, a2338.

Ghose, A., & Han, S. P. (2011). An empirical analysis of user content generation and usage behavior on the mobile Internet. Management Science, 57(9), 1671-1691.

Godes, D., Mayzlin, D., Chen, Y., Das, S., Dellarocas, C., Pfeiffer, B., Libai, B., Sen, S., Shi, M., & Verlegh, P. (2005). The firm's management of social interactions. Marketing Letters, 16(3), 415-428.

Gupta, S. (2013). For Mobile Devices, Think Apps, Not Ads. Harvard Business Review, 71-75.

Hartmann, W. R. (2010). Demand estimation with social interactions and the implications for targeted marketing. Marketing Science, 29(4), 585-601.

Hartmann, W. R., Manchanda, P., Nair, H., Bothner, M., Dodds, P., Godes, D., Hosanagar, K., & Tucker, C. (2008). Modeling social interactions: Identification, empirical methods and policy implications. Marketing letters, 19(3), 287-304.

Hill, S., Provost, F., & Volinsky, C. (2006). Network-based marketing: Identifying likely adopters via consumer networks. Statistical Science, 256-276.

Hoff, P. D. (2005). Bilinear mixed-effects models for dyadic data. Journal of the American Statistical Association, 100(469), 286-295.

Hoff, P. D., Raftery, A. E., & Handcock, M. S. (2002). Latent space approaches to social network analysis. Journal of the american Statistical association, 97(460), 1090-1098.

Goldsmith-Pinkham, P. & Imbens, G. (2013) Social networks and the identification of peer effects. Journal of Business and Economic Statistics.

Ibarra, H. (1992). Homophily and differential returns: Sex differences in network structure and access in an advertising firm. Administrative science quarterly, 422-447.

Iyengar, R., Van den Bulte, C., & Valente, T. W. (2011). Opinion leadership and social contagion in new product diffusion. Marketing Science, 30(2), 195-212.

Kandel, D. B. (1978). Homophily, selection, and socialization in adolescent friendships. American Journal of Sociology, 427-436.

Katona, Z., Zubcsek, P. P., & Sarvary, M. (2011). Network effects and personal influences: The diffusion of an online social network. Journal of Marketing Research, 48(3), 425-443.

Klepper, S., & Leamer, E. E. (1984). Consistent sets of estimates for regressions with errors in all variables. Econometrica: Journal of the Econometric Society, 163-183.

Kolata, G. (2007, July 25th). Study Says Obesity Can Be Contagious. New York Times. Retrieved from http://www.nytimes.com/2007/07/25/health/25cnd-fat.html

Kossinets, G., & Watts, D. J. (2009). Origins of homophily in an evolving social network1. American Journal of Sociology, 115(2), 405-450.

Krivitsky, P. N., & Handcock, M. S. (2008) Fitting position latent cluster models for social networks with latentnet. Journal of Statistical Software, 25(5).

Krivitsky, P. N., & Handcock, M. S. (2009) latentnet: Latent position and cluster models for statistical networks. R package version 2.2-2. URL http://statnetproject.org.

Krivitsky, P. N., Handcock, M. S., Raftery, A. E., & Hoff, P. D. (2009). Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. Social Networks, 31, 204-213.

Lyons, R. (2011). The spread of evidence-poor medicine via flawed social-network analysis. Statistics, Politics, and Policy, 2(1).

Manchanda, P., Xie, Y., & Youn, N. (2008). The role of targeted communication and contagion in product adoption. Marketing Science, 27(6), 961-976.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. The review of economic studies, 60(3), 531-542.

McCallum, B. T. (1972). Relative asymptotic bias from errors of omission and measurement. Econometrica, 40(4), 757-758.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. Annual review of sociology, 415-444.

Mebane, W. R., & Poast, P. (2013). Causal Inference without Ignorability: Identification with Nonrandom Assignment and Missing Treatment Data. Political Analysis.

Nair, H. S., Manchanda, P., & Bhatia, T. (2010). Asymmetric social interactions in physician prescription behavior: The role of opinion leaders. Journal of Marketing Research, 47(5), 883-895.

Nickell, S. (1981). Biases in dynamic models with fixed effects. Econometrica: Journal of the Econometric Society, 1417-1426.

Nitzan, I., & Libai, B. (2011). Social effects on customer retention. Journal of Marketing, 75(6), 24-38.

Phillips, P. C., & Sul, D. (2007). Bias in dynamic panel estimation with fixed effects, incidental trends and cross section dependence. Journal of Econometrics, 137(1), 162-188.

Rossi, P. E., McCulloch, R. E., & Allenby, G. M. (1996). The value of purchase history data in target marketing. Marketing Science, 15(4), 321-340.

Shalizi, C. R., & Thomas, A. C. (2011). Homophily and contagion are generically confounded in observational social network studies. Sociological Methods & Research, 40(2), 211-239.

Stephen, A. T., & Toubia, O. (2010). Deriving Value from Social Commerce Networks. Journal of Marketing Research, 47(2).

Trusov, M., Bodapati, A. V., & Bucklin, R. E. (2010). Determining Influential Users in Internet Social Networks. Journal of Marketing Research, 47(4), 643-658.

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

Van den Bulte, C., & Lilien, G. L. (2001). Medical Innovation Revisited: Social Contagion versus Marketing Effort1. American Journal of Sociology, 106(5), 1409-1435.