Identification of Peer Effects in Networked Panel Data

Completed Research Paper

Sinan Aral
MIT Sloan School of Management
30 Memorial Drive, Cambridge, MA 02142
sinan@mit.edu

Daniel Rock
MIT Sloan School of Management
30 Memorial Drive, Cambridge, MA 02142
drock@mit.edu

Sean J. Taylor
NYU Stern School of Business
44 W 4th Street, New York, NY 10012
staylor@stern.nyu.edu

Abstract

After product adoption, consumers make decisions about continued use. These choices can be influenced by peer decisions in networks, but identifying causal peer influence effects is challenging. Correlations in peer behavior may be driven by correlated effects, exogenous consumer and peer characteristics, or endogenous peer effects of behavior (Manski 1993). Extending the work of Bramoullé et al. (2009), we apply proofs of peer effect identification in networks under a set of exogeneity assumptions for the panel data case. With engagement data for Yahoo Go, a mobile application, we use the network topology of application users in an instrumental variables setup to estimate usage peer effects, comparing a variety of regression models. We find this type of analysis may be useful for ruling out endogenous peer effects as a driver of behavior. Omitted variables and violation of exogeneity assumptions can bias regression coefficients toward finding statistically significant peer effects.

Keywords: Econometric analyses, Econometrics, IT adoption, Network externalities, Networks, Social influences, Peer Effects
Introduction

The longstanding econometric challenge of accurately measuring peer effects has recently been the subject of renewed interest (Angrist and Lang 2004; Card and Giuliano 2012; Christakis and Fowler 2007; Dupas et al. 2008; Sacerdote 2001). Recent methodological developments, together with the increased availability of granular data on behavior in networks, offer new opportunities to empirically address questions of social cause and effect. However, the increased number of peer effects studies coincides with considerable discussion of their associated methodological challenges. While the idea that individuals may act under the influence of their peers is widely accepted, measuring and interpreting these effects remains difficult.

Current research on identifying peer effects in human behavior is primarily focused on discrete rather than continuous diffusion processes and static rather than dynamic behaviors. Researchers have studied, for example, the diffusion of product adoption in networks (Aral et al. 2013; Bakshy et al. 2011; Bollinger and Gillingham 2012; Manchanda et al. 2008; Muchnik et al. 2013; Watts 2002), which is a discrete process of consumers’ transition from non-adoption to adoption, or static cross sections of networks (Aral and Van Alstyne 2011; Aral and Walker 2012; Currarini et al. 2010). However, a fairly large fraction of human behaviors is dynamic and continuous, in that they occur over time and with some frequency – for example, the process by which we use products over time, after we have adopted them. Our aim is to provide and apply an appropriate model for the estimation of peer effects in a dynamic setting, for example when peers are influencing each other in repetitive, daily behaviors rather than in discrete, one-off decisions.

We discuss some of the econometric issues in using network topology to estimate peer effects. Weak instruments, omitted variables, and violation of the exogeneity assumptions can bias regression coefficients toward finding statistically significant peer effects. Our dataset comes from a Yahoo! mobile browsing application with more than 25 million users. We apply a network-based two-stage least squares estimation procedure to the mobile application’s pageview time series at the individual level. Our results suggest that measuring causal peer effects remains tricky. There are many ways to unwittingly overestimate peer effects. In some cases, a failure to reject the null of no peer effects is therefore more informative.

Manski (1993) characterizes apparent correlations between individual and peer behaviors to come from three coincident sources: correlated effects, “endogenous” peer effects, and “exogenous” peer effects. In the context of a rainy day, the correlated effect might refer to the tendency of those outside to open umbrellas, endogenous peer effects would be the increased tendency for us to open umbrellas given an observation that others are doing the same, and exogenous peer effects would be the change in our propensity to open an umbrella given an observation of what others are wearing. An ordinary least squares regression of individual umbrella use on peer umbrella use with individual controls would pick up all three effects. Angrist (2014) adds that any regression of an individual outcome on group outcomes alone produces a coefficient of 1. Even if we wanted to study umbrella use on a sunny day, we would still face the issue that exogenous and endogenous effects are not individually identified. Angrist reminds us that here we have Manski’s “reflection problem”, meaning that “observed behavior is always consistent with the hypothesis that individual behavior reflects mean reference-group behavior”. This problem is inescapable without a strategy to distinguish between the different types of peer influence. Moffitt (2001) also discusses issues with identification of peer effects and how they might be solved.

Following the model described by Lee (2007), Bramoullé et al. (2009) offer one way to handle the reflection problem in observational data. Relying on spatial econometrics as a guide, they move beyond groups as traditionally defined toward social network connections to develop reference groups on the individual-level. They provide sufficient conditions for identification of endogenous and exogenous peer effects coefficients in static linear-in-means models. In their model the pre-assigned covariates of nth-degree connections (for n>2 in the basic model) in a social network can be used as instruments for peer behavior. Models of this sort have scarcely been tested in the past. The first purpose of our work is to extend this model to the time-varying case and show that the original identification results still hold for panel data. We then elaborate by investigating challenges to the model’s assumptions, offering suggestions for potential applications of this method in business and policy contexts.
Identification of Peer Effects in Networked Panel Data

Business applications for peer effects estimation have already established a foothold via studies of adoption and diffusion (Aral et al. 2009; Manchanda et al. 2008; Miller and Tucker 2009; Tucker 2008). Scholars in many disciplines have been thinking about these ideas for quite a while (Hartmann et al. 2008). Information about the causes of diffusion is valuable. In the next section we discuss research into adoption and engagement in greater depth. While both “diffusive” behaviors, compared to adoption, engagement is relatively understudied given the business value of proper engagement strategy. Extending the model of Bramoullé et al. to a dynamic panel data application in the style of Arellano and Bond (1991) opens the door for more studies of time-dependent behaviors like engagement and product use. This necessitates additional assumptions to maintain identification of the model parameters. As with similar regression structures, the conditions required for the model to accurately estimate peer effects can pose a challenge.

Identification is a second-order concern if a model can be expected to generate spurious estimates of peer effects. Peer effects two-stage least squares (2SLS) setups (like ours in this paper) and ordinary least squares (OLS) regression estimates might diverge for many reasons, some of them quotidian or mechanical (Angrist 2014). Yet in many cases the relevant standard for an approach to be useful is if the exercise provides actionable information to managers or policy-makers. Knowing, for instance, that a product’s engagement is not driven by social processes is vital for a marketer considering a viral distribution channel. We find that 2SLS-based network models might help a manager more reliably conclude that their product lacks a social component given the propensity of unobservables to be positively correlated for peers.

To summarize our contributions: First, we extend a popular and well-known econometric model into the dynamic panel data context—a context that should be relatively common for modern empirical settings. Second, we demonstrate how to carefully apply this model, as well as reasonable alternatives, to a dynamic peer effects estimation problem. Finally, we show how to interpret model estimates in order to make policy decisions, and make the point that our network panel data models can help analysts or policymakers even when we fail to reject the null hypothesis.

This paper proceeds as follows. First we present key parts of the relevant literature on peer effects and engagement. We then review the Bramoullé et al. (2009) model and extend it to the panel data case. Next we describe the data and how our dataset was constructed, and present estimation results and different model specifications. The last section discusses the results and possible extensions before concluding.

**Diffusion, Peer Influence, and Estimation of Peer Effects**

Social effects can act as an amplifier for policy or business decisions. Human activity is connected; policy-makers or managers should be mindful of the social externalities relevant to their choices. Previous work on diffusion and peer influence has spanned many disciplines, each applying a different bundle of methods (Hartmann et al. 2008). Some of that research has focused on generating contagion, or solving the “influence maximization problem”. Domingos and Richardson (2001) propose algorithms to maximize lift in marketing efforts via social network interactions. Bakshy et al. (2011) suggest that cascade prediction based on prior influence events in the Twitter network is improved by targeting many influential network nodes. Aral et al. (2013) propose strategies to “engineer” contagion in the presence of homophily, i.e. the tendency of “birds of a feather to flock together” in networks (McPherson et al. 2001). These empirical studies of cascades complement theoretical work discussing or modeling how diffusion might occur (Bass 1969; Granovetter 1978; Jackson and Yariv 2007; Schelling 1971; Watts 2002).

Many studies have also directly measured peer effects at the individual level, under both observational and experimental conditions. For observational datasets, a variety of econometric techniques have been used to estimate social multipliers. Tucker (2008) measures peer effects in the diffusion of a video-messaging technology, comparing the relative size of network externalities for managers and other employees at a bank. This work also used an instrumental variables approach, finding that measures of ego-level importance in the network affected the magnitude of observed adoption externalities. Bollinger and Gillingham (2012) model peer effects in adoption of solar panels using a first differences approach, where lags of peer solar panel installation are used to predict adoption. Consistent estimates of peer effect
coefficients in this specification, however, can be problematic because of omitted variables bias even if the “pre-assigned” covariate regressors are truly exogenous to the behavior of interest.

Noting that the individual node characteristics and behaviors tend to be correlated with observed networks, Aral et al. (2009) use covariate matching to distinguish between social influences and homophily in adoption of a mobile application. They find that failure to control for homophily leads to overestimation of behavioral contagion effects in adoption by as much as 300-700%. Even matching methods, however, might not fully account for potential bias in estimation. To assess the magnitude of estimation bias, Eckles and Bakshy (2014) compare the results of a randomized experiment to observational techniques. In an experiment with 67 million users on Facebook, naïve observational estimators overstated the experimental measure of contagion by as much as 300%. They point out that some of the difficulty in measuring peer effects comes from the “implausible assumption that the available covariates are sufficient to make peer behavior unconfounded”. Thus undercontrolled studies are often more likely to find positive peer effects far in excess of the real externalities.

Social contagion over network ties can operate differentially depending on the individual’s “influence” and “susceptibility” (Aral and Walker 2012). This is true beyond the effects of homophily. Network positioning, heterogeneity in tie strength, and dyad-level variation in the sign of social effects combine in real networks to generate the data we observe. Most people can personally relate to the concepts of trendsetters and followers or differences in the strength of social relations. Certainly if peer influence is a fact of life, we have all been either encouraged or discouraged by the actions of others at some point. So-called “chilling effects”, when peer behavior slows down the growth rate of others’ behaviors, were found in a number of contexts by Goldenberg et al. (2010). They point to adopters of CD players, DVD players, and cellular services waiting for early adopters to act first, creating “hockey stick” shaped growth. Of course growth need not rebound if peer use leads to congestion. The wide range of possibilities for different types of influence underscores that average effects may not tell a sufficiently granular story.

Experimental designs get around the typical worries about exogeneity. Experiments are now common in networks and peer effects research, especially as digital tools have made them cheaper to conduct (Bapna and Umyarov 2012). Some studies have focused on experimentally changing the behavior or product of interest instead of networks or group means. Aral and Walker (2011) create treatment and control groups by randomly embedding viral features of different kinds into a Facebook application and tracking the diffusion. With a hazard model specification, they show a 246% increase in social contagion with the addition of passive message broadcasting. Muchnik et al. (2013) demonstrate herding effects in ratings by manipulating the initial rating of posts on a social news aggregation website. These particular designs escape some of the “perils” of peer effects analysis by separating the subjects of analysis from the treatment (whereas a regression of individual behavior on group mean behavior does not). One concern for networked experiments is interference, or the tendency for the stable unit treatment value assumption (SUTVA) to fail when treatments can diffuse into control groups. Athey et al. (2015) discusses p-value calculation in networked contexts.

Other experimental designs focus on changing network exposure conditions and observing differences in diffused behaviors. This might entail changing the connections in a network or experimentally altering the information moving between connected individuals. In the Facebook network, for example, people who were exposed to signals about friends’ information (in the form of different urls) were found more likely to transmit that information themselves (Bakshy et al. 2012). In that study it was the exposure condition that the researchers randomly assigned. In contrast, the treatment in Bapna and Umyarov (2012) was being exposed to a peer who had adopted a paid service. This later design is more analogous to the observational models commonly applied to the study of peer effects.

Many of these studies focus on one-shot behaviors like adoption, ratings, or clicks. But actions that vary over time may often carry more value. Fader et al. (2005) point out that measuring customer lifetime value (CLV), for example, may come from individual-level conditional expectations or aggregate-level sales projections. They offer a taxonomy of probability models for estimating CLV, noting that “recency, frequency, and monetary value” or “RFM” of transactions can be used to estimate “residual lifetime value” (customer value unexplained by covariates). Though one of many possible frameworks, but RFM is conceptually closer to engagement and use than it is to adoption. Peer effects in usage may differ from peer effects in adoption, with large implications for businesses looking to monetize user behavior. Companies can set usage-based pricing structures and subsidize influential individuals (“opinion
leaders”) to generate value or stimulate diffusion (Ghose and Han 2011; Iyengar et al. 2011). These studies on engagement rely on observational data as well. As with Bollinger and Gillingham’s work on solar panel adoption, they use fixed effects in an attempt to handle the reflection problem. Aral and Nicolaides (2016) examine peer effects in exercise over time, relying on exogeneity of the weather conditions for peers in other cities in a continuous IV setup. This research design more convincingly argues for “contagion” in healthy behaviors than observational data would. Future work might compare the parameter estimates from studies of this sort to those retrieved from observational network methods.

Whether an observational or experimental study design, it isn’t clear that “the coefficient on group averages in a multivariate model of endogenous peer effects reveals the action of social forces” (Angrist 2014; Shalizi and Thomas 2011). Consider, as Angrist does, a setup where we estimate the OLS and instrumental variables (IV) coefficients in a regression of individual on group behavior (group average behavior is the instrumented variable). The peer effect is represented as the divergence between the OLS and IV of the relationship between covariates and outcomes, but these estimands can differ for reasons other than peer effects. Many standard econometric problems apply here, plus a few more unique to networks. In a many weak instruments situation, the IV estimates will be biased toward OLS estimates. “Common variance components in outcomes” can produce correlations that resemble social effects, but are driven by omitted factors. Deciding which observation of “the network” is the latent representation of links between units can be a judgment call.

The model we develop is appropriate for observational networked data where outcome behavior and at least one of the covariates vary over time. It is principally based upon the work of Bramoullé et al. (2009), which explicitly leverages network topology to create local group means of outcome and explanatory variables. The presence and linearly independent arrangement of intransitive triads of individuals in the network permits identification results to hold. In other words, a network where all friends of friends are also friends will lead to degenerate estimation outcomes. That kind of model reduces to a regression of outcomes on leave-out group means and covariates.

**Extension of Bramoullé, Djebbari, and Fortin’s Social Network Model (BDF)**

**A Review of the Static Network Model**

The baseline structural model used describes the (linear) relationship between the individual behavior vector $y$, the vector of connected peers’ behaviors $Gy$, the vectors of exogenous (i.e. pre-assigned in this case) covariates $x$, and the vectors of pre-assigned covariates for connected peers $Gx$. $G$ is the row-normalized adjacency matrix representing the social network of peers. Entries in $G$ are zero if two individuals are not connected ($G_{ij} = 0$ if there is no $ij$ dyad). For simple averaging of peer behavior, entries in each row (representing the peers of individual $i$) will sum to 1 and will have the value $1/n$ where $n$ is the count of nonzero entries in the row. This is equivalent to a simple leave-out mean. To review, following the simplest version of Bramoullé et al. (2009) (or “BDF” for the rest of the paper) we have the equation:

$$y = \alpha + \beta Gy + \gamma x + \delta Gx + \epsilon; \quad E[\epsilon | x] = 0 \quad (1)$$

With $|\beta| < 1$ and $(I - \beta G)^{-1} = \sum_{k=0}^{\infty} \beta^k G^k$, we have that:

$$y = \alpha(I - \beta G)^{-1} + (I - \beta G)^{-1}(\gamma I + \delta G)x + (I - \beta G)^{-1} \epsilon \quad (2)$$

$$y = \left(\frac{\alpha}{1-\beta}\right) + \gamma x + (\gamma \beta + \delta) \sum_{k=0}^{\infty} \beta^k G^{k+1}x + \sum_{k=0}^{\infty} \beta^k G^k \epsilon \quad (3)$$

Equations (2) and (3) describe the reduced form structure of the model where $y$ is purely a function of observables $x$ and network $G$. We then can write the expected average peer behavior as:

$$E(Gy|x) = \left(\frac{\alpha}{1-\beta}\right) + \gamma Gx + (\gamma \beta + \delta) \sum_{k=0}^{\infty} \beta^k G^{k+2}x \quad (4)$$

BDF goes on to show in their proposition 1 that if $\gamma \beta + \delta \neq 0$ and the matrices $I$, $G$, and $G^2$ are linearly independent, then social effects are identified. Additionally, if those 3 matrices are linearly dependent and
no individual is isolated, social effects are not identified (the paper has more detailed results). Without linear independence, then it is not possible to find an identifying instrument for $G_y$ in equation (1) above. Otherwise we can use network structure to find exogenous covariates that instrument for $G_y$. This is to say that friends-of-friends’ covariates ($G^2x$) or even deeper network connections’ covariates ($G^nx$ for n>2) can be used to predict friends’ behavior in an IV setup. While for our purposes we will restrict the model in this analysis to one connected component of a network, writing a block diagonal matrix $\Gamma$ where the diagonal is composed of disjoint subnetworks $G_1,...,G_n$ is a natural extension of this model. Including network fixed effects, BDF generalizes the model to partially deal with correlated effects. Note that this does not deal with the tendency for networks to be formed because of behaviors. We still must assume that the network formation process is strictly exogenous to vector $y$. With the correlated effects, for a network $k$:

$$y_k = \alpha_k + \beta G_k y_k + \gamma x_k + \delta G_k x_k + \epsilon_k; \quad \mathbb{E}[\epsilon_k | \alpha_k, G_k, x_k] = 0 \quad (5)$$

Subtracting the local network average from the individual, BDF shows the result that $I, G, G^2,$ and $G^3$ must be linearly independent in addition to the assumption that there are social effects, i.e. $\gamma \beta + \delta \neq 0$. In this case the valid identifying instrument set will be $(I - G_k)G_k^2x, (I - G_k)G_k^3x, ..., (I - G_k)G_k^n x$. This “within local transformation” is analogous to a first difference estimator where the fixed effect for network $k$ is differenced out across the network. It is this model and Arellano and Bond (1991) which inspire the panel model. We present the setup for a single connected network, though the results can be extended to include multiple networks with correlated effects in a stacked matrix.

**The Simple Panel Model**

For a single connected network, if $t$ indexes time, we have:

$$y_t = \alpha + \beta G_t y_t + \gamma x_t + \delta G_t x_t + \epsilon_t; \quad \mathbb{E}[\epsilon_t | x_t, G_t] = 0 \forall t \quad (5)$$

And analogously to (2) and (3), with $|\beta| < 1, I - \beta G_t$ and $(I - \beta G_t)^{-1} = \sum_{k=0}^{\infty} \beta^k G_t^k$, we have that:

$$y_t = \alpha (I - \beta G_t)^{-1} + (I - \beta G_t)^{-1}(\gamma I + \delta G_t)x_t + (I - \beta G_t)^{-1}\epsilon_t \quad (6)$$

$$y_t = \frac{\alpha}{1 - \beta} I + \gamma G_t x_t + (\gamma \beta + \delta) \sum_{k=0}^{\infty} \beta^k G_t^{k+1}x_t + \sum_{k=0}^{\infty} \beta^k G_t^{k+2} x_t \quad (7)$$

$$\mathbb{E}(G_t y_t | x_t) = \frac{\alpha}{1 - \beta} I + \gamma G_t x_t + (\gamma \beta + \delta) \sum_{k=0}^{\infty} \beta^k G_t^{k+2} x_t \quad (8)$$

(see Appendix A.1 in Bramoullé et al. (2009) for a proof that proves results for all models once they have been reduced to equations analogous to (5)-(8)).

**The Panel Model with Fixed Effects and Static Networks**

This setup has no means of handling the propensity for “common shocks” to affect networks over time or within connected components. If we want time-specific fixed effects, we have a structural equation:

$$y_t = \lambda_t + \alpha t + \beta G_t y_t + \gamma x_t + \delta G_t x_t + \epsilon_t \quad (9)$$

Using first differences, this setup too can be used to derive the reduced form in analogous way. Let

$$\Delta y_t = y_{t+1} - y_t, \quad \Delta x_t = x_{t+1} - x_t, \quad \Delta \epsilon_t = \epsilon_{t+1} - \epsilon_t, \quad \Delta \lambda_t = \lambda_{t+1} - \lambda_t$$

By differencing two equations at time $t$ and time $t+1$, we have:

$$\Delta y_t = \Delta \lambda_t + \beta (G_{t+1} y_{t+1} - G_t y_t) + \gamma \Delta x_t + \delta (G_{t+1} x_{t+1} - G_t x_t) + \Delta \epsilon_t; \quad \mathbb{E}[\Delta \epsilon_t | \Delta x_t, G_{t+1}, G_t] = 0 \forall t \quad (10)$$

Assume for now that the latent network is stable and $G_t$ is isomorphic for all time $t$. We’ll denote this graph $G_t$ for “final graph”. In this case the problem reduces cleanly as we can factor the final graph out. We get:

$$\Delta y_t = \Delta \lambda_t + \beta G_f \Delta y_t + \gamma \Delta x_t + \delta G_f \Delta x_t + \Delta \epsilon_t \quad (11)$$

With reduced form, by an argument similar to the ones above:
\[
\Delta y_t = \frac{\Delta \lambda_i}{1 - \beta} t + \gamma \Delta x_t + (\gamma \beta + \delta) \sum_{k=0}^{\infty} \beta^k G_{Y}^{k+1} \Delta x_t + \sum_{k=0}^{\infty} \beta^k G_{Y}^{k} \Delta \epsilon_t \quad (12)
\]

\[
E(G_{Y} \Delta y_t | \Delta x_t) = \frac{\Delta \lambda_i}{1 - \beta} t + \gamma G_{Y} x_t + (\gamma \beta + \delta) \sum_{k=0}^{\infty} \beta^k G_{Y}^{k+2} \Delta x_t \quad (13)
\]

Incidentally this will be the reduced form equation that applies to the Yahoo! Go network. We can only observe the network on the final day of a 28-day period, and therefore take the network at the end of the period as the latent "true" network (this assumes that if two people were friends by the end of the month, they were likely friends at the beginning of the month too). Where these results fail, however, is in the case that \( G_i \neq G_f \) for all time \( t \). In that case equation 10 does not reduce as cleanly. It may be that the network is measured with error; some observed connections might not be active while other unobserved connections are activated. We assume we can observe the true latent network, though previous work has developed methods to more concretely describe links between nodes (A. Goldenberg et al. 2010).

**The Panel Model with Fixed Effects and Dynamic Networks**

Returning to (10), to use the series expansion shortcut to the reduced form we must find some matrix which represents the change in the network over time. Let us now modify \( G \) slightly to use matrix \( F \), which will represent a version of \( G \) that is not yet row-normalized. \( F_{ij} = 1 \) if individual \( i \) and individual \( j \) are connected. \( \text{F}_y \) is then the vector of sums of peer behaviors. This comes from the fact that for a given individual \( i \) we have \( F_{xy} = 1 \) whenever \( i \) is connected to \( j \). Taking the inner product of a row vector from \( F \) (row \( i \)) and \( y \) (the column vector of behaviors), the output will be the sum of peer behaviors for the individual corresponding to individual \( i \). We have:

\[
\Delta y_t = \Delta \lambda_i t + \beta (F_{t+1} y_{t+1} - F_t y_t) + \gamma \Delta x_t + \delta (F_{t+1} x_{t+1} - F_t x_t) + \Delta \epsilon_t;
\]

\[
E[\Delta \epsilon_t | \Delta x_t, F_t, F_{t+1}] = 0 \forall t \quad (14)
\]

so it follows that:

\[
\Delta y_t - \beta (F_{t+1} y_{t+1} - F_t y_t) = \Delta \lambda_i t + \gamma \Delta x_t + \delta (F_{t+1} x_{t+1} - F_t x_t) + \Delta \epsilon_t
\]

The researcher must decide on the precise timing of peer effects to allow for separation of changing network effects from changing behavior effects. We have two boundary conditions. In the “early” condition, former friends are discarded and the relevant adjacency matrix for the next period will be \( F_{t+1} \). In the “late” condition, new friends have no peer influence and the relevant adjacency matrix is \( F_t \). In general, we can define a parameter (or potentially a vector of parameters) \( \rho \in [0,1] \) to determine relative influence levels of new friends and old friends. This weighting value determines which convex combination of friends to use.

\[
F_{\rho,t} = \rho F_{t+1} + (1 - \rho) F_t \quad (15)
\]

With \( \rho \) selected, we can renormalize \( F_{\rho,t} \) row-wise to get \( G_{\rho,t} \), and then we have:

\[
\Delta y_t - \beta G_{\rho,t} \Delta y_t = \Delta \lambda_i t + \gamma \Delta x_t + \delta G_{\rho,t} \Delta x_t + \Delta \epsilon_t \quad (16)
\]

And then using the representation of \( |\beta| < 1, (I - \beta G_{\rho})^{-1} = \sum_{k=0}^{\infty} \beta^k G_{\rho}^k \), equation (12) will hold subbing in \( G_{\rho,t} \) for \( G_t \). Below we briefly discuss a few ways to calculate \( G_{\rho,t} \). For the “between” estimator, where we have individual specific fixed effects, there is no problem factoring out the network after de-meaning the vectors in the equation. The network is stable in cross-section after fixing the time period.

**Random Effects Models**

Random effects (RE) models take a similar form to (9). In vector form:

\[
y_t = \alpha + \beta G_y y_t + \gamma x_t + \delta G_x x_t + \nu_t; \quad \nu_t = \epsilon_t + \lambda \quad (17)
\]

Now we assume that the error term contains a random effect \( \lambda \) that generates serial correlation. This random effect comes with a strong assumption: it must be orthogonal to all of the included explanatory variables. Formally,
Identification of Peer Effects in Networked Panel Data

A similar logic as we have demonstrated for fixed effects models in (10-13) applies to recover identification results for random effects models. Yet the random effect specification will often be inappropriate in social contexts. To see why, note that the random effect generating individual or time-specific heterogeneity is assumed to be orthogonal to all of the included explanatory variables as well as the network. RE models require that sources of unobserved heterogeneity are unrelated to all explanatory variables, which are often included because they are behaviorally important. In the presence of homophily, it is rather heroic to assume that all covariates in addition to network formation processes are orthogonal to individual or time-specific unobserved heterogeneity. This poses an opportunity to use a Hausman specification test to check for the presence of homophily, a proposition we will explore in future work.

Calculation Strategies for $G_{\rho}$

The simplest way to start is to choose either the new network or the old network as the relevant period-level isomorphic graph. We might also define a scalar to indicate the relative “socialness” of each period as well. One way to determine a weighting value between 0 and 1 is to define a row vector of weights for each individual, where the $i^{th}$ element of the row vector is the proportion of the total count of connections belonging to the new network for individual $i$. This is to say that:

$$\rho_{i,t} = \frac{n_{i,t+1}}{n_{i,t} + n_{i,t+1}}; \; n_{i,t} \text{ is the connection count of person } i \text{ in time } t$$

We could treat $\rho$ as another parameter to optimize in the estimation procedure, choosing its value in each period (as a scalar) to minimize the sum of squared residuals. This approach is not discussed in this paper. Lastly we might abandon the convex combination idea entirely, focusing instead on finding the graph union or graph intersection of $F_t$ and $F_{t+1}$. The graph union has the advantage of including all connections as important in peer effects. The graph intersection represents the part of the graph that is stable over time. Most network software packages contain algorithms for calculating these matrices. A combination of both approaches may be worthwhile, taking the “true” network to be some convex combination of the intersection and union graphs.

General Features of the Models

All of these models struggle to handle some of the key problems with measuring peer effects. Nowhere do we fix the problem of distinguishing genuine social contagion from homophily. Correlated effects can be differenced out at the network-level, but subgraphs may contain their own localized correlated effects that further influence connected peers. To the extent that pre-assigned characteristics of peers influence behaviors (without, somehow, influencing formation of ties), we can calculate an estimate of an average peer effect of behavior contingent on observable types of homophily. Yet rarely do researchers have sufficient data on covariates to control for every type of correlation in tie formation. Additionally, there is often good reason to think that a given behavior of interest generates peer connections from period to period. In that case the moment restrictions do not hold. It is therefore difficult to fully trust a causal interpretation of the peer effects coefficients generated by these models. The identification of these parameters is subordinate to the concern that they are fragile. In other words, we should maintain skepticism of the estimated parameters because previous research on behavior in networks suggests these models will often fail to fully describe social dynamics.

Pessimism aside, the structure of the reduced form of the type in equations (3), (7), and (12) permits a robustness check on the assumptions of the model. We could run a kind of truncated reduced form regression of outcomes on covariates, friends’ covariates, friends-of-friends’ covariates, ..., n-distant friends’ covariates, etc. Since $|\beta| < 1$, we would expect to see coefficients on friends’ covariates declining exponentially in network distance. These coefficients would of course be biased (likely upward) by the omitted variables calculated by letting $n$ go to infinity. Nevertheless, normalizing the coefficient on the immediate neighbors’ covariates to be 1, we should see coefficients rapidly shrink as a function of network distance. This kind of check would be most useful in the case that the model is just identified. Otherwise a test of overidentification suffices. So far we have also assumed that social effects exist. In the case that
they do not exist, we expect that the first stage of the IV regression collapses. This condition of no social effects implies that the 2SLS estimates will not differ greatly from OLS estimates in expectation. That weak instrument scenario presents an asset: either the included covariates have no social content, the behavior of interest is unlikely to be contagious, or both. These are useful facts for managers to know.

The Yahoo! Go Data and Network

Our network, engagement, and covariate data come from Yahoo! Go, a mobile application designed for browsing behavior. Yahoo! designed Go as a way for online users to access Yahoo! services on their mobile phones and/or PDAs. Launched in July 2007 and discontinued in January 2010, the network of users had over 27 million members. Go services enabled users to check sports scores, look up stock quotes, send and receive email, search, and read news (see Figure 1 for an image of one of the application screens). Each user has a unique Yahoo! ID across all Yahoo! services (mobile or otherwise). These IDs have been anonymized by the company and all users under age 18 have been removed from the dataset.

Peer effects in the Yahoo! Go adoption were previously studied in Muchnik et al. (2009), with propensity score matching estimators suggesting that the social network played a role in app adoption decisions. However, Go does not support instant messaging and social activities like link sharing, commenting, or liking, so Go activities should not be expected to exhibit network externalities in usage. This is to say that though adoption might spread via word-of-mouth or similar channels, use is not likely to be social on the basis that an observable network exists. As a result, we feel that the Go empirical context is an excellent one to test whether a method for estimating dynamic peer effects is biased since we believe that ground truth is very likely to be close to zero, but there likely to be unobserved confounders related to homophily.

The exogeneity of the network with respect to the behavior of interest is assumed in the model setup we have described. In practice these types of networks are hard to come by, especially with the proliferation of social or “viral” application features in digital products. The separation of network formation from the behavior of interest makes the Go network an attractive context in which to apply the model. We wish to rule out the criticism that our results might be driven by network endogeneity in discussing the performance and estimation results of our model specifications. Go serves that purpose, facilitating a more nuanced discussion of the research value of network topology-based panel models. Since Go has no explicitly social features, we can more reasonably make the assumptions necessary for the model identification results to hold. Otherwise even exogenous covariates will be polluted with endogenous information from the network. While we are specifying models in which individual behavior depends on peer behavior, we effectively “select” a peer group which is not created by the behavior of interest. In contrast to many other mobile applications, Go pageview behavior is not by default communicated to other connected users. There is no social signal, nor algorithm and/or collaborative filter to serve friends similar content. Whereas networks are likely formed by the behavior of interest, the network cannot be assumed to be exogenous.

The network of interest is formed by instant messaging (IM) behavior on Yahoo! Messenger and we code at least one message sent and received between two users in the 30 days prior to the launch of Yahoo! Go as a “friend” or edge in the network. Friend-of-friends were coded as Go users within two hops of the focal user, not including the user’s one hop friends. Note that Yahoo! Messenger is an entirely separate product, but that the majority of Yahoo! Go adopters were active users of Messenger. In this way we can observe the network without Go having any explicitly social features like sharing or mail.

Our data set includes daily use data in the form of mobile page views for October 2007, as well as covariate data on user demographics, including age, gender, and friend counts (degree). Using the Messenger network, we also build a series of network covariates which are static over the course of our panel (since we use the end-of-month network to represent the latent network). We use the total mobile pageviews because the desktop behaviors might also be linked to network formation processes. Friends’ usage and Friends-of-Friends’ usage are calculated as the average total pageviews for user friends connected to a given user and average total pageviews for friends-of-friends of a given user (respectively). Starting with the 27 million Yahoo! Messenger users, we narrow the user set to adopters of Go who also have friends using Go (21,896 users), and then perform the analysis on the largest connected component to satisfy the condition that $I$, $G$, and $G^2$ are linearly independent. This leaves us with 2,203 users with use behavior observed over 28 days. Figure 2 depicts the network of users, while Figure 3 zooms in on the
largest connected component. Figure 4 shows the degree distribution of nodes in the largest connected component and Figure 5 shows the distribution of node counts in all connected components of users except the largest one. Our largest connected component is exceptionally large in comparison to the others.

In our largest connected component, the degree distribution closely follows an exponential density and the network diameter is 21. This “scale-free” characteristic is common in networks (Barabási and Albert 1999). Aside from the largest connected component, there are 11,108 smaller networks of users which
either fail to have a large enough user pool to develop precise estimates, fail to meet the identification conditions of the model, or both. Summary statistics for the largest connected component are reported below (Table 1).

<table>
<thead>
<tr>
<th>Table 1. Summary Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Age</td>
</tr>
<tr>
<td>Friend Age</td>
</tr>
<tr>
<td>Friend-of-Friend Age</td>
</tr>
<tr>
<td>Gender</td>
</tr>
<tr>
<td>Friend Gender</td>
</tr>
<tr>
<td>Friend-of-Friend Gender</td>
</tr>
<tr>
<td>Degree</td>
</tr>
<tr>
<td>Friend Degree</td>
</tr>
<tr>
<td>Friend-of-Friend Degree</td>
</tr>
<tr>
<td>PageRank</td>
</tr>
<tr>
<td>Friend PageRank</td>
</tr>
<tr>
<td>Friend-of-Friend PageRank</td>
</tr>
<tr>
<td>Use (pageviews)</td>
</tr>
<tr>
<td>Lagged Use (1 Day)</td>
</tr>
<tr>
<td>Lagged Use (2 Days)</td>
</tr>
<tr>
<td>Friend Use</td>
</tr>
<tr>
<td>Lagged Friend Use (1 Day)</td>
</tr>
<tr>
<td>Lagged Friend Use (2 Day)</td>
</tr>
<tr>
<td>Friend-of-Friend Use</td>
</tr>
<tr>
<td>Lagged Friend-of-Friend Use (1 Day)</td>
</tr>
<tr>
<td>Lagged Friend-of-Friend Use (2 Day)</td>
</tr>
</tbody>
</table>

Table 1. Summary Statistics

It is not surprising that friend and friend-of-friend summary statistics are so similar to individual statistics (everyone is someone's friend). We do see a common feature of human networks that the average degree of friends is higher than the individual's degree. This is because the most connected individuals are more likely to be friends with many people. We also calculate the PageRank score for all nodes in the network as a means of measuring network centrality (Page et al. 1998). This will be used as a control in our peer effects regressions. The average age is 35.53 with a standard deviation of 7.93. Gender is coded as 1 for men and 0 for women. Since we only have use (mobile pageviews) as a time-varying characteristic of individuals in our network, we will use lags of use as our pre-assigned covariate for individuals, friends, and friends-of-friends. In the next section we present the regression results from estimation of our panel specifications.

Regression Results

The tables are organized as follows: Table 2 shows estimates from OLS regressions of individual use (in mobile pageviews) on the average friend use and other covariates. Table 3 shows the IV estimates for the same equations, using all available friend-of-friend covariates as instruments for the endogenous average
friend use. For example, we have in columns 3 and 4 of Table 2 pooled and fixed effect versions (respectively) of a regression of individual use on friends’ use, lagged individual use (1 and 2 days lagged) and lagged friends’ use. We use lagged friends-of-friends’ use (1 and 2 day) to instrument for friends’ use in columns 1 and 2 of Table 3. When we include more demographic covariates to predict individual use, we similarly use friends-of-friends’ covariates of the same type as instruments. First stage results are presented in Table 4. Table 5 presents Sargan J statistics and Cragg-Donald Wald F statistics and Stock and Yogo (2005) weak identification critical values for the 6 IV specifications. The “full model” containing all covariates and lagged usage behaviors for the individual and their friends has the following equation (similar to equation 9 in section 2):

\[ \text{Use}_t = \alpha + \beta G_t \text{Use}_t + \gamma_1 \text{Lag Use}_{1,t} + \gamma_2 \text{Lag Use}_{2,t} + \gamma_3 \text{Age} + \gamma_4 \text{Gender} + \gamma_5 \text{Degree} + \gamma_6 \text{PageRank} + \delta_1 G_t \text{Lag Use}_{1,t} + \delta_2 G_t \text{Lag Use}_{2,t} + \delta_3 G_t \text{Age} + \delta_4 G_t \text{Gender} + \delta_5 G_t \text{Degree} + \delta_6 G_t \text{PageRank} + \lambda_t + \epsilon_{i,t} \]

We substitute in the network we observe at the end of the period of analysis \( G_t \) (in row-normalized form) for \( G_t \) for each day in the 28-day period we examined. The other models use a subset of the variables included in this larger equation. Multiplying a vector by \( G \) therefore returns a vector containing the average for an individual’s friends’ values (use or covariate). We use 1 and 2 day lags in these specifications.

### Table 2. OLS Regression Coefficients

<table>
<thead>
<tr>
<th>Estimates (Dependent Variable is Pageviews)</th>
<th>OLS - Pooled (1)</th>
<th>OLS - Fixed Effects (2)</th>
<th>OLS - Pooled with Lags (3)</th>
<th>OLS - Fixed Effects with Lags (4)</th>
<th>OLS - Pooled with Demo. (5)</th>
<th>OLS - FE with Demo. (6)</th>
<th>OLS - Pooled Full Model (7)</th>
<th>OLS - FE Full Model (8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friend Use</td>
<td>0.0665***</td>
<td>0.0560***</td>
<td>0.0444***</td>
<td>0.0299***</td>
<td>0.0685***</td>
<td>0.0580***</td>
<td>0.0431***</td>
<td>0.0286***</td>
</tr>
<tr>
<td></td>
<td>(0.0064)</td>
<td>(0.0064)</td>
<td>(0.0070)</td>
<td>(0.0070)</td>
<td>(0.0064)</td>
<td>(0.0064)</td>
<td>(0.0070)</td>
<td>(0.0070)</td>
</tr>
<tr>
<td>Lag Use (1 Day)</td>
<td>0.299***</td>
<td>0.298***</td>
<td>0.297***</td>
<td>0.297***</td>
<td>0.297***</td>
<td>0.296***</td>
<td>0.297***</td>
<td>0.296***</td>
</tr>
<tr>
<td></td>
<td>(0.0122)</td>
<td>(0.0121)</td>
<td>(0.0121)</td>
<td>(0.0121)</td>
<td>(0.0121)</td>
<td>(0.0121)</td>
<td>(0.0121)</td>
<td>(0.0121)</td>
</tr>
<tr>
<td>Lag Use (2 Day)</td>
<td>0.189***</td>
<td>0.191***</td>
<td>0.187***</td>
<td>0.187***</td>
<td>0.189***</td>
<td>0.189***</td>
<td>0.189***</td>
<td>0.189***</td>
</tr>
<tr>
<td></td>
<td>(0.0099)</td>
<td>(0.0099)</td>
<td>(0.0099)</td>
<td>(0.0099)</td>
<td>(0.0099)</td>
<td>(0.0099)</td>
<td>(0.0099)</td>
<td>(0.0099)</td>
</tr>
<tr>
<td>Friend Lag Use (1 Day)</td>
<td>0.00295</td>
<td>0.00363</td>
<td>0.00208</td>
<td>0.00208</td>
<td>0.00268</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0077)</td>
<td>(0.0077)</td>
<td>(0.0077)</td>
<td>(0.0077)</td>
<td>(0.0077)</td>
<td>(0.0077)</td>
<td>(0.0077)</td>
<td>(0.0077)</td>
</tr>
<tr>
<td>Friend Lag Use (2 Day)</td>
<td>-0.00029</td>
<td>0.00975</td>
<td>-0.00134</td>
<td>0.00863</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0065)</td>
<td>(0.0065)</td>
<td>(0.0065)</td>
<td>(0.0065)</td>
<td>(0.0065)</td>
<td>(0.0065)</td>
<td>(0.0065)</td>
<td>(0.0065)</td>
</tr>
<tr>
<td>Age</td>
<td>0.00735</td>
<td>0.00828</td>
<td>-0.00574</td>
<td>-0.00554</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0218)</td>
<td>(0.0218)</td>
<td>(0.0192)</td>
<td>(0.0192)</td>
<td>(0.0192)</td>
<td>(0.0192)</td>
<td>(0.0192)</td>
<td>(0.0192)</td>
</tr>
<tr>
<td>Gender</td>
<td>3.125***</td>
<td>3.087***</td>
<td>1.784***</td>
<td>1.769***</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.453)</td>
<td>(0.453)</td>
<td>(0.428)</td>
<td>(0.428)</td>
<td>(0.428)</td>
<td>(0.428)</td>
<td>(0.428)</td>
<td>(0.428)</td>
</tr>
<tr>
<td>Friend Age</td>
<td>0.116***</td>
<td>0.119***</td>
<td>0.0522</td>
<td>0.0525</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0411)</td>
<td>(0.0411)</td>
<td>(0.0382)</td>
<td>(0.0382)</td>
<td>(0.0382)</td>
<td>(0.0382)</td>
<td>(0.0382)</td>
<td>(0.0382)</td>
</tr>
<tr>
<td>Friend Gender</td>
<td>-3.558***</td>
<td>-3.535***</td>
<td>-1.243**</td>
<td>-1.230**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.580)</td>
<td>(0.580)</td>
<td>(0.560)</td>
<td>(0.559)</td>
<td>(0.560)</td>
<td>(0.559)</td>
<td>(0.559)</td>
<td>(0.559)</td>
</tr>
<tr>
<td>Degree</td>
<td>0.344***</td>
<td>0.343***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Thirty Seventh International Conference on Information Systems, Dublin 2016 12
### Table 2. OLS Regression Coefficients

<table>
<thead>
<tr>
<th>Estimates (Dependent Variable is Pageviews, Friend Use is Instrumented)</th>
<th>IV - Pooled with Lags (1)</th>
<th>IV - FE with Lags (2)</th>
<th>IV - Pooled with Demographics (3)</th>
<th>IV - FE with Demographics (4)</th>
<th>IV - Pooled Full Model (5)</th>
<th>IV - FE Full Model (6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PageRank</td>
<td>(0.106)</td>
<td>(0.106)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Friend Degree</td>
<td>0.0745</td>
<td>0.0774</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Friend PageRank</td>
<td>(0.0851)</td>
<td>(0.0850)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>15.65***</td>
<td>16.04***</td>
<td>7.557***</td>
<td>2.844***</td>
<td>11.57***</td>
<td>11.84***</td>
</tr>
<tr>
<td></td>
<td>(0.205)</td>
<td>(0.941)</td>
<td>(0.226)</td>
<td>(0.672)</td>
<td>(1.513)</td>
<td>(1.773)</td>
</tr>
<tr>
<td>Obs.</td>
<td>61,684</td>
<td>61,684</td>
<td>57,278</td>
<td>61,684</td>
<td>61,684</td>
<td>57,278</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.002</td>
<td>0.005</td>
<td>0.180</td>
<td>0.184</td>
<td>0.003</td>
<td>0.006</td>
</tr>
<tr>
<td>Time FE</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

### Table 3. IV Regression Coefficients

<table>
<thead>
<tr>
<th>Estimates (Dependent Variable is Pageviews, Friend Use is Instrumented)</th>
<th>IV - Pooled with Lags (1)</th>
<th>IV - FE with Lags (2)</th>
<th>IV - Pooled with Demographics (3)</th>
<th>IV - FE with Demographics (4)</th>
<th>IV - Pooled Full Model (5)</th>
<th>IV - FE Full Model (6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PageRank</td>
<td>(0.106)</td>
<td>(0.106)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Friend Degree</td>
<td>0.0745</td>
<td>0.0774</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Friend PageRank</td>
<td>(0.0851)</td>
<td>(0.0850)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>15.65***</td>
<td>16.04***</td>
<td>7.557***</td>
<td>2.844***</td>
<td>11.57***</td>
<td>11.84***</td>
</tr>
<tr>
<td></td>
<td>(0.205)</td>
<td>(0.941)</td>
<td>(0.226)</td>
<td>(0.672)</td>
<td>(1.513)</td>
<td>(1.773)</td>
</tr>
<tr>
<td>Obs.</td>
<td>61,684</td>
<td>61,684</td>
<td>57,278</td>
<td>61,684</td>
<td>61,684</td>
<td>57,278</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.002</td>
<td>0.005</td>
<td>0.180</td>
<td>0.184</td>
<td>0.003</td>
<td>0.006</td>
</tr>
<tr>
<td>Time FE</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
### Table 3. IV Regression Coefficients

<table>
<thead>
<tr>
<th>First Stage Estimates (Friend Use is Dependent Variable)</th>
<th>First Stage - Pooled with Lags</th>
<th>First Stage - FE with Lags</th>
<th>First Stage - FE with Demographics</th>
<th>First Stage - FE with Demographics</th>
<th>First Stage - Pooled Full Model</th>
<th>First Stage - FE Full Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lag Friend-of-Friend Use (1 Day)</td>
<td>0.0476***</td>
<td>0.0306***</td>
<td>0.0451***</td>
<td>0.0282***</td>
<td>0.0327***</td>
<td>0.0156*</td>
</tr>
<tr>
<td></td>
<td>(0.00922)</td>
<td>(0.00932)</td>
<td>(0.00923)</td>
<td>(0.00933)</td>
<td>(0.00926)</td>
<td>(0.00937)</td>
</tr>
<tr>
<td>Lag Friend-of-Friend Use (2 Day)</td>
<td>0.00325</td>
<td>0.0428***</td>
<td>0.000941</td>
<td>0.0405***</td>
<td>-0.0122</td>
<td>0.0272***</td>
</tr>
<tr>
<td></td>
<td>(0.00920)</td>
<td>(0.00930)</td>
<td>(0.00922)</td>
<td>(0.00932)</td>
<td>(0.00925)</td>
<td>(0.00936)</td>
</tr>
<tr>
<td>Friend-of-Friend Age</td>
<td>-0.0560</td>
<td>-0.0644</td>
<td>-0.0722</td>
<td>-0.0722</td>
<td>-0.0803</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0496)</td>
<td>(0.0492)</td>
<td>(0.0499)</td>
<td>(0.0495)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.767)</td>
<td>(0.760)</td>
<td>(0.790)</td>
<td>(0.782)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Friend-of-Friend Degree</td>
<td>0.0796</td>
<td></td>
<td></td>
<td>0.0543</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.158)</td>
<td></td>
<td></td>
<td>(0.156)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Friend-of-Friend PageRank</td>
<td></td>
<td></td>
<td></td>
<td>618.1</td>
<td>2,232</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(29,045)</td>
<td>(28,776)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.193)</td>
<td>(0.524)</td>
<td>(1.330)</td>
<td>(1.407)</td>
<td>(1.435)</td>
<td>(1.506)</td>
</tr>
<tr>
<td>Obs.</td>
<td>57,278</td>
<td>57,278</td>
<td>57,278</td>
<td>57,278</td>
<td>57,278</td>
<td>57,278</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.199</td>
<td>0.215</td>
<td>0.201</td>
<td>0.216</td>
<td>0.205</td>
<td>0.220</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1
Table 5. Overidentification and Weak Identification Tests

<table>
<thead>
<tr>
<th></th>
<th>Lags Models</th>
<th>Demographics Models</th>
<th>Full Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>P-Value</td>
<td>0.3237</td>
<td>0.0006</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

Table 6. Specification Tests (RE vs. FE)

The OLS regression of individual pageviews on peer average pageviews would suggest there are peer effects in our network. A simple regression of individual on peer use without controls suggests a 1 view increase in the peer average corresponds to a 0.0665 view increase for individuals (Table 2 column 1). With more controls, this increase remains statistically significant but declines to 0.0299 once we include fixed effects for the day of the month and lagged use for the individual and peers. Unsurprisingly, lagged individual use is the best predictor of future individual use. An extra pageview one day prior and two days prior is associated with a respective (statistically significant at 5%) increases of around .3 and .19 pageviews in nearly all models. The full OLS model (Table 2 column 8) with fixed effects predicts a statistically significant 0.0286 extra pageviews for an increase in mean peer pageviews of 1. Gender is significant and also predicts more engagement (1.77 additional pageviews if the individual is male), though having more male friends predicts a decline in individual engagement. Individuals of higher degree also tend to have an additional 0.34 pageviews for every additional friend they have. The purely correlational analysis appears consistent with peer effects, though these correlations are misleading.

We should note here that aside from use and gender of friends, there are no statistically significant friend-based measures in the full model. This is an indication that some peer covariates may not have exogenous peer effects (to use Manski’s framing). It is also worth pointing out that model fit here (and in the IV regressions in the following section) is quite poor, explaining at most 20% of the variance in usage behavior even after including lags of the behavior of interest. Although this might seem like a limitation, we are not conducting a predictive exercise—the estimate of interest is the local, linear slope of the conditional expectation function of usage with respect to peer usage which is measured by the model.

The IV regressions offer mixed evidence on the existence of peer effects. In Table 3 column 1, we see the results for the pooled IV with only lagged individual and friend behavior (no fixed effects for the day). Friend Use is instrumented by friend-of-friend lagged use of 1 and 2 days. Once again we have the statistically significant lagged individual coefficients of about 0.3 and 0.19 for 1 and 2 day lags respectively. But the average friend pageviews are not significantly different from 0. Adding time fixed effects (column 2) gets friend use to marginal but not convincing significance (10%). The specifications in columns 3 and 4 add age and gender covariates for the individual and friends, instrumenting for friend use with the lagged friend-of-friend use as before in addition to friend-of-friend “average” gender and age. Now peer effects appear to be statistically significant at the 1% level and quite strong! In column 4, a 1 pageview increase in peer averages appears to cause a 0.633 pageview increase in the individual’s use. The full specification (Table 3 column 6) tempers the apparent effect: here an increase in peer average pageviews of 1 corresponds to a 0.555 increase in individual views at a 10% significance level. The coefficients on friends’ covariates are quite close to zero. This is an important red flag to which we will return soon.

Table 6 describes the Hausman Specification Test p-values to determine whether fixed or random effects models would be more appropriate. The Hausman Tests reject in favor of fixed effects models (with exceptionally low p-values for the probability that the random effects specifications satisfy the orthogonality conditions). Though we did try random effects models, with a dataset of this size we have the opportunity to put fewer constraints on the variance structure of the residuals. Since random effects versions of each of these models had similar coefficients to the pooled versions and each Hausman Test rejected the null, we do not report the random effects model coefficients here. Broadly it seems challenging to satisfy the assumption in random effects models that unobserved heterogeneity is orthogonal to the other included regressors. This does yield a possible insight into the degree of homophily though. The model with only lags fails to reject in favor of fixed effects, likely because the random effect is orthogonal to previous use. The coefficients for the random effects model with only lags
are nearly identical to the pooled model with only lags. It is only once we include covariates that describe individuals’ demographics and networks that we begin to reject in favor of fixed effects (as expected). In future work we will explore whether a failure to reject random effects specifications offers evidence that homophily on the basis of included regressors is weak.

Discussion of Results

How can we reconcile the mixed indications for evidence of peer effects? A combination of the overidentification and weak instruments tests is informative. Looking at Table 5, the J statistics are small enough that we fail to reject at even a 10% significance level the null hypothesis that the instruments are exogenous. Under the assumption that contemporaneous peer pageviews are indeed the only endogenous variable, lags of friends-of-friends’ use behavior appear to be valid instruments. Yet once we add in instruments that reflect demographics (age and gender), Table 5 (columns 3 and 4) tell us that at least one of the instruments is not exogenous. We reject the null hypothesis for the Sargan Overidentification Test at nearly any reasonable significance level (this is true for the full model in Table 5 columns 5 and 6 as well). The exclusion restriction fails for at least one covariate. The correlation with the error term of friend-of-friend demographics suggests an unobserved covariate related to the individual’s outcomes. It would appear that the evidence for peer effects is relatively weak. Either we fail to include covariates which are likely related to the data-generating process, or our peer effects estimates are too imprecise to be convincing. We assumed when applied this model BDF’s “natural” condition that \( \gamma \beta + \delta \neq 0 \), i.e. there are social effects in this context. Maybe there are social effects, but evidently the friend-of-friend covariate instruments which make that condition true are in fact endogenous. These are the unobservable “shared components of variance in outcomes” which make precisely measuring endogenous peer effects challenging. Models relying on network structure to generate exogenous network instruments fall victim to the likely condition of widespread homophily. Setups such as the one we have presented and the original static version in BDF are especially prone to find positive peer effects whenever positive sorting on unobserved covariates is correlated with the behavior of interest.

If “birds of a feather flock together”, then they likely roost together as well (as it were). Indeed, models like these effectively stack the deck in favor of finding peer effects by first isolating connected individuals and only then applying an estimation technique within similar groups. Regressions on group leave-out average outcomes are also problematic, but might have additional noise in the definition of “peer” so as to avoid intensified homophily. Furthermore, in some applications it may be a problem that the matrix \( G^n \) represents paths of length \( n \), including paths that return to the individual (though not if assumptions are strictly satisfied). This can create mechanical nonzero correlations between instrumented peer behavior and outcomes that do not exist for “leave-out” matrices. It is also possible that there are no exogenous or endogenous peer effects, in which case any coefficients on peers we have found is due to statistical noise. We might have weak instruments.

Our first stage estimates have predictive power though. So why isn’t there substantial evidence for peer effects? Yahoo! Go lacks social features, but the network exhibits homophily. It would be surprising if we found endogenous peer effects. The Yahoo! Go product engagement is driven primarily by previous engagement. Marketing strategies to leverage “virality” in usage behaviors for Go would likely fail. An important implication of our work is that failing to find significant peer effects can be useful by ruling out particular types of interventions as unpromising.

Conclusion

Given the multitude of reasons that a 2SLS IV estimator and an OLS estimator might diverge, of which peer effects is one of many, what is the place for network structure-based IV setups like the one we have used and BDF? Because the model setup is so fragile, estimation results reflecting social multipliers are unlikely to be trustworthy. But arguably there is a meaningful practical application of these models. A null result, however disappointing in some cases, can be valuable and informative by placing reasonable bounds of effect sizes. Marketers, project managers, and business analysts can use this setup to pick out products without peer influence. In any case where homophily is likely to lead to positive correlation in behaviors, a failure to reject the null of zero peer effects is a strong statement. Null results in the presence of a bias of magnitude are more reliable null results. With the results we have presented, there is now a
dynamic framework to apply as well (if only as a robustness check). There are mitigating factors which might make a null result less reliable too. Heterogeneity, congestion effects, and omitted variables could make a null result more likely. Using an instrumental variables setup can also widen out standard errors; lower precision estimates make it harder to trust a failure to reject. An important implication then of this study is an emphasis on the fragility of networked panel data models which use distant network peer pre-assigned covariates. Homophily in social networks abounds, perhaps in varieties for which we have no explicit terms. Putative peer effects are, as Angrist (2014) reminds us, possibly due to unobserved confounders.

Future work might extend to matching estimators or simulation techniques to handle these challenges, as well as to modeling the case where the network can co-evolve dynamically alongside usage behavior as in many social applications. Yet in the simplest of cases homophily will often lead us to find a putative peer effect that isn’t really there. In our case, knowing that Yahoo! Go lacked a positive social externality might have helped Yahoo! end support for the product earlier, or to focus their efforts on the most engaged users. Engineering and exploiting contagion in marketing can be a worthy pursuit, but often an expensive strategy to pursue. These models can help decide if resources are better deployed elsewhere. With the right data though, marketers have an additional tool for checking the suitability of social marketing strategies for their products.

Acknowledgements

We are grateful to Yahoo, Inc. for providing data access. We also thank Erik Brynjolfsson, Lev Muchnik, Dean Eckles, Guillaume Saint-Jacques, Hossein Ghasemkhani, and participants at the MIT Labi Lunch and TIES Student Seminar at MIT Sloan for helpful comments and discussion. Generous financial support was provided by the MIT Initiative on the Digital Economy (IDE).

References


Identification of Peer Effects in Networked Panel Data