9-2011

An Empirical Analysis of User Content Generation and Usage Behavior on the Mobile Internet

Anindya Ghose
University of Pennsylvania

Sang P. Han

Recommended Citation

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. http://dx.doi.org/10.1287/mnsc.1110.1350

An Empirical Analysis of User Content Generation and Usage Behavior on the Mobile Internet

Abstract

We quantify how user mobile Internet usage relates to unique characteristics of the mobile Internet. In particular, we focus on examining how the mobile-phone-based content generation behavior of users relates to content usage behavior. The key objective is to analyze whether there is a positive or negative interdependence between the two activities. We use a unique panel data set that consists of individual-level mobile Internet usage data that encompass individual multimedia content generation and usage behavior. We combine this knowledge with data on user calling patterns, such as duration, frequency, and locations from where calls are placed, to construct their social network and to compute their geographical mobility. We build an individual-level simultaneous equation panel data model that controls for the different sources of endogeneity of the social network. We find that there is a negative and statistically significant temporal interdependence between content generation and usage. This finding implies that an increase in content usage in the previous period has a negative impact on content generation in the current period and vice versa. The marginal effect of this interdependence is stronger on content usage (up to 8.7%) than on content generation (up to 4.3%). The extent of geographical mobility of users has a positive effect on their mobile Internet activities. Users more frequently engage in content usage compared to content generation when they are traveling. In addition, the variance of user mobility has a stronger impact on their mobile Internet activities than does the mean. We also find that the social network has a strong positive effect on user behavior in the mobile Internet. These analyses unpack the mechanisms that stimulate user behavior on the mobile Internet. Implications for shaping user mobile Internet usage behavior are discussed.

Keywords

mobile Internet, social networks, content generation, content usage, interdependence, geographical mobility, identification

Disciplines

Industrial Technology | Science and Technology Studies

An Empirical Analysis of User Content Generation and Usage Behavior in the Mobile Internet[1]

Anindya Ghose

Associate Professor, Stern School of Business, New York University

Sang-Pil Han

Postdoctoral Researcher, Stern School of Business, New York University

Abstract

We quantify how users' mobile Internet use is related to some unique characteristics of the mobile Internet space like usage-based data pricing, the extent of their geographical mobility and their social network behavior. We use a unique panel dataset consisting of individual-level mobile Internet usage data encompassing their multimedia content uploading and downloading behavior along with detailed data on their voice and text calls, and demographics. We build and estimate an individual-level simultaneous equations panel data model using three-stage least-squares (3SLS) after controlling for the endogeneity of the social network. We find that there exists negative temporal interdependence between the content generation and usage behavior for a given user. While the extent of geographical mobility of users is positively associated with their mobile Internet activities, users more frequently engage in content downloading compared to content uploading when they travel. We find statistically significant social network effects that are robust across model specifications. We compare the impact of different kinds of social networks – voice call based social networks, text message-based social network and offline location-based spatial network. The impact of the voice call-based social network on is higher than the impact of location-based spatial network. Furthermore, the impact of the overlapping network (that is, combination of the social and spatial networks) is larger than that of either the social network or spatial network. We use our empirical estimates to measure the incremental value to firms from negative temporal interdependence and mobility. Implications for stimulating users' mobile Internet usage, mobile advertising, and targeting of social networks are discussed.

***Keywords*:** Mobile Internet, Uploading, Downloading, Interdependence, Geographical Mobility, Social Networks, Spatial Networks, Econometrics.

[1] We thank participants at the 2009 WIMI-MSI conference, University of Minnesota, University of Texas at Dallas, University of Maryland at College Park, 2009 International Conference in Information Systems, Stanford Internet Economics Conference, SCECR 2009 conference, and the 2009 Marketing Science conference for helpful comments. Anindya Ghose acknowledges the generous financial support from NSF CAREER award (IIS-0643847). Funding for this project was also provided by a grant from the Wharton Interactive Media Institute-Marketing Science Institute (WIMI-MSI) grant competition on user-generated content. The usual disclaimer applies.

1. Introduction

Spurred by rapid technological advances in the mobile Internet technologies, as smart phones and social networking web sites allow consumers to interact, share, meet, and recommend places based on their physical locations, the ubiquitous access to the mobile Internet provides companies with new marketing opportunities. New marketing strategies may need to be implemented in this environment, in lieu of or in addition to, traditional marketing methods. This requires a deeper understanding of users' behavior in the mobile Internet space. However, little is known about how users' mobile Internet use is related to some unique characteristics of the mobile Internet space like usage-based data pricing, the extent of their geographical mobility and their mobile social network behavior.

There are at least two unique aspects of the user content experience in the mobile Internet space that help distinguish it from other media settings like PC or laptops. *First,* in many countries, users need to explicitly incur expenses (for example, by paying usage-based data transmission charges) during their mobile content experience based on the number of bytes uploaded or downloaded.[2] This is in contrast to the PC web where content experience using a fixed Internet connection or a WiFi access can be done without incurring any monetary costs that are based on usage. *Second*, users can ubiquitously engage in content generation and usage activities using their mobile phones, irrespective of geography as long as they are in an area with cell phone reception.[3] In contrast, PC usage renders much stricter limitations on geographical mobility and access, typically constraining it to office or home.

Usage-based data pricing and ubiquitous access lead to a situation where both monetary constraints and geographical mobility can influence users' mobile Internet behavior over time. Advanced mobile data communication technologies such as 3G services enable people to build and maintain their social relationships through the co-creation and usage of content independent of time and place. Hence people can decide how much content to generate and how much content to use. For example, in a given week higher the amount of time or money spent by a user in using content (e.g., downloading music, games, productivity apps, etc), lower the amount of time or money left for that user to generate content (e.g., uploading photos, reviews, videos, etc). Hence, content generation and usage may not be independent decision-making processes. In contrast to this argument, higher levels of content usage by users in previous periods may motivate them to contribute higher levels of content in subsequent periods

[2] The fee structure in our empirical context is usage-based pricing. Subscribers are charged on a per byte basis for data traffic that they generate through content uploading and downloading.

[3] We use the terms "content generation" and "content uploading" interchangeably in the paper. Similarly, we use the terms "content usage" and "content downloading" interchangeably in the paper.

as they begin to feel an integral part of the community established on these sites and need to engage in reciprocal behavior. Whether the former negative interdependence is stronger than the latter positive one, is an empirical question that we aim to examine in this paper. Also, since users can access the mobile Internet ubiquitously, the extent of their geographical mobility (i.e., their local and national traveling patterns) can also influence their mobile Internet activities. Furthermore, users' mobile Internet usage behavior may be affected by mobile Internet usage behavior of their network neighbors, where the network is drawn from either voice call-based social networks or offline location-based spatial networks.

Keeping these issues in mind, in this paper we focus on the following questions: (i) How is the content generation behavior of users is related to their content usage behavior over time: positive or negative? (ii) How does the extent of user's geographical mobility affect their content generation and usage activities? (iii) How is the behavior of users' social network neighbors associated with their own behavior? We examine these questions using a unique dataset consisting of mobile data across a panel of users encompassing their content uploading and downloading behavior. The dataset consists of 2.34 million individual -level mobile data records across 180,000 users. We also have data on voice and text calls made by the same users that enables to draw the social networks. We have detailed user demographic and geographical data including the location from where a call was placed. This location information helps us impute the extent of their geographical mobility by mapping the different places they visited. We construct two different measures of mobility in order to map both the mean and the variance of their travel patterns. We do this both at the local and national levels, thereby giving us four different mobility metrics. Our analysis utilizes simultaneous equation models as well as generalized method of moments (GMM) -based dynamic panel models and several other models (both linear and non-linear) for robustness checks.

Results show that there exist negative temporal interdependence between the content generation and usage behavior for a given user. This finding implies that the resource constraint (e.g., time and money constraint) binds at least for some people. The effect is asymmetric such that the negative impact of previous period's content generation on current period's content usage is much higher than vice-versa. While the extent of geographical mobility of users is positively associated with their content generation and usage activities, its impact varies by specific mobility metrics that capture the mean and the variance of their travel patterns, at both the local and national level. This finding suggests that when travelling and visiting different places, users more frequently engage in content downloading compared to content uploading. When traveling to a location that is different from their regular travel (i.e., away from home or office), users more frequently engage in content downloading compared to content uploading.

Furthermore, we find strong suggestive evidence that content generation and usage behavior of network neighbors positively influences user behavior. The marginal effects are also asymmetric with social network behavior having a stronger association with content downloading compared to content uploading. Furthermore, we discuss the economic impact of the estimates by converting them into monetary values.

Our paper aims to makes three key contributions to the literature. We are the first to simultaneously model and estimate the drivers of users' content generation and usage behavior in the mobile Internet space and the nature of temporal interdependence between these processes. We build and estimate an individual-level simultaneous equations panel data model using three-stage least-square (3SLS) estimation. We further demonstrate the robustness of this analysis by conducting GMM-based dynamic panel data analyses and other analyses that include models with a random coefficient in a constant term, count data models, content share models, content size models, models with different time lags, and models using social network variables based on different kinds of social networks and different specifications. We use our empirical estimates to measure the incremental annual value to firms from negative temporal interdependence and various mobility-based measures.

Second, the unique nature of our data allows us to map how the mean and variance of users' geographical mobility, at both the local and national levels, drives their content generation and usage behavior. No prior work has linked the content generation and usage process to the extent of their geographical mobility. Our paper thus contributes to the literature on the economic and social impact of the mobile Internet.

Third, our data allows us to observe the mobile Internet behavior by both mobile communication-based social networks and offline location-based spatial networks. Our empirical estimates provide an upper bound on the causal effect of the social network on user behavior in mobile Internet usage and consumption. We compare and contrast different kinds of social networks – voice call-based social networks, text message-based social network and offline location-based spatial network. The impact of the voice call-based social network on users' mobile Internet usage behavior is higher than the impact of location-based spatial network. Furthermore, the impact of the overlapping network (that is, the combination of the social and spatial networks) is much larger than that of either the social network or spatial network.

The rest of this paper is organized as follows. In Section 2, we provide prior literature relevant to our paper to build the theoretical framework. Section 3 describes the data that we employ and provides some individual-level usage patterns to motivate the need for incorporate various econometric issues involved in the model. In Section 4, we describe the individual-level econometric model. Section 5 presents estimation results, and several robustness tests. Section 6 discusses implications of the result and concludes.

2. Literature Review

Our paper is related to a small literature that discusses the interplay among user Internet usages, mobility, and social networks.

First, we relate the dynamic interdependence between user content generation and usage behaviors to two relevant streams of literature – economic behavior under resource constraints and reciprocity stemming from social exchange theory. Researchers have long recognized that time also acts as a constraint (Becker 1965, Jacoby et al. 1976). In online settings, users need to allocate resources between content generation and content usage activities since they can take on the dual role of creators as well as consumers (Trusov et al. 2006). In the mobile Internet space, not only do users need to invest time but also incur explicit transmission charges to generate and use content in some countries. This suggests that there would be a negative temporal interdependence between content generation and content usage activities over time. In contrast, prior work in the online content sharing literature (Asvanund et al. 2004; Xia et al. 2007) draws on reciprocity stemming from social exchange theory (Homans 1958) and suggests that the more a user benefits from the contribution of other users, the more the user is willing to create content (Xia et al. 2007). This behavior suggests that there is a positive temporal interdependence between content generation and content usage activities over time. Since the extent of reciprocal interactions with each other in the mobile Internet setting is unknown, the overall extent and directional interdependence of the temporal effect between content generation and usage remains an empirical question.

Second, we examine the impact of the extent of the geographical mobility of a user on the mobile Internet activity of that user. There are two possible scenarios here. The first scenario is that the more a user travels, the more travel-related discretionary time the user is likely to have. Shim et al. (2008) analyze mobile usage patterns where people view TV programs on their phone screens. They find that the highest usage is between 6 AM and 9 AM in the morning and between 6 PM and 8 PM in the evening, which is consistent with the notion that most users view content using mobile phones while commuting from home to work place and back. O'hara et al. (2007) find that people use mobile video content to pass time, manage solitude, and disengage from others. In contrast, it is also possible that mobile internet usage can occur at geographically fixed places as well. Hence, the overall extent of the impact of a user's geographical mobility on the same user's propensity to engage in mobile Internet activity remains an empirical question.

Third, users can be influenced by others with whom they communicate. There is some evidence of this in the business world – such as the adoption of new services and products (Hill et al. 2006, Tucker 2008, Aral et al. 2009, Nair et al. 2010, Iyengar et al. 2010), the switching from an existing service provider (Dasgupta et al. 2008), and diffusion of user-generated content in the online space (Susarla et al. 2010). Further, people who live in the same geographical vicinity can have a higher probability of communicating with each other offline. Prior studies use physical proximity between people as a proxy for the extent of social influence with each other through direct social interaction or observation (Yang and Allenby 2003 and Bell and Song 2007). Therefore, users' mobile Internet activities can be associated with the mobile Internet activity of both their social network and spatial network neighbors.

In addition, there is another kind of network effect that is worth discussing here – that of two-sided networks even though the mobile Internet is also sufficiently distinct from existing two-sided markets. In traditional two-sided markets, there are network effects created by two distinct user groups or entities. For example, in two-sided offline media such as TV or magazines, consumers typically only buy and consume media content while professional editors produce and sell content. A wide body of empirical research exists in two-sided markets (see for example, Bucklin and Sismeiro 2003, Nair et al. 2004, Ellison and Ellison 200, Tucker and Zhang 2010). However, mobile Internet settings are distinct in that people can take on the dual role of content creators as well as content consumers.

3. Data Description and Basic Patterns

In this section, we provide a short overview of the mobile Internet service in our data, describe the data that we obtained from a large telecommunications service company in South Korea, and finally provide basic patterns in the data to motivate our subsequent model development.

3.1 Data Source

Our sample consists of 2.34 million mobile data records from 180,000 3G mobile users who used the services of the company between March 15, 2008 and June 15, 2008.The South Korea 3G mobile market witnessed over 10 million subscribers in June 2008. 3G mobile services enable users to upload and download their content faster than conventional mobile services. Further, these services are more commonly available on the larger screen handsets that facilitate relatively more user-friendly content generation and usage.

There are two broad categories of websites users can access through their mobile phones in our data. The first category is one consisting of regular social networking and community websites. Examples of such websites in our data include Cyworld and Facebook. The second category of websites includes portal sites specifically created by the mobile phone company. Examples of such websites in our data include Nate Portal and KTF Portal, which are the Asian equivalent of US sites like Vodafone live and T-Mobile's Web 'n' Walk. The content on these sites can be accessed through a mobile phone by users who subscribe to the services of the mobile operator. Like social networking sites, these mobile portals are also community-oriented sites that allow users to download and upload (in order to share with others) multimedia content like photos, music, videos, apps, podcasts, etc. The transmission charges are the same in our data, irrespective of whether users upload or download content and whether users access to social networking community sites or mobile portal sites. Hence, we measure the level of users' mobile Internet activities based on the frequency of content uploading and downloading.[4]

3.2 Variable Description

Our mobile Internet data include individual-level information on users' content generation and usage activities over time. The temporal unit in our analysis is a "week." Technically, a unique mobile Internet session initiates when a user pushes a button in a keypad on the handset or by clicking an icon on a touchpad, and it ends when the user deactivates the mobile Internet session. Only when a user initiates a mobile Internet session can the user either download content or upload content, or both. Such an activity involving more than zero bytes of data transmission is an event. One mobile Internet session usually consists of multiple events. As shown in Table 2, a user's content generation occurs far less frequently than content usage. We construct four metrics with respect to users' geographical mobility: 1) local location mobility, 2) national location mobility, 3) local mobility dispersion, and 4) national mobility dispersion.

The first two metrics capture the *mean* of users' travel patterns, and the last two metrics capture the *variance* of their travel patterns. Users are often engaged in mobile content activities when they are outdoors and when they are traveling. In such circumstances, the geographical locations from where a call was placed will change over time. We refer to the number of unique locations from where calls were placed by a user as a measure of their "*location mobility*." In a sense, this variable captures the *mean* travel pattern of a user. There are two different levels of granularity with respect to the extent of location

[4] In our robustness checks, we also use the amount of bytes transmitted via users' content uploading and downloading as an alternative measure for the level of users' mobile Internet activity.

mobility of each user: local and national. Both variables measure the number of distinct locations from where a user made calls at the zip code level and the province/state level, respectively. Since the total number of unique zip codes is much higher than the number of provinces or states in our data (i.e., 30,116 vs. 16), the number of distinct zip code-level locations corresponds to a user's "local location mobility" and the number of distinct province-level locations corresponds to a user's "national location mobility."

The "*mobility dispersion*" measures the extent of geographical deviation from one's commonly visited places (i.e., home and office) and in that sense, it captures the *variance* of a user's travel patterns. A user's mobility dispersion refers to a fraction of uncommonly visited places with respect to the total number of places visited during a given week.[5] The key notion behind the use of this variable is that deviations from routine patterns of travel might indicate a business trip or a leisure trip to a location that is different from their regular travel. Such unique travel occasions could lead to a higher propensity (compared to the mean) to upload and download content in order to share travel experiences (i.e., sharing photos taken at tourist attractions). As before, we have both local and national levels of granularity for the mobility dispersion metric.

We also have data on voice calls and text messages made by the same users that enables to draw social networks.[6] For example, voice and text call records contain both callers' telephone numbers from the initial 180,000 users and call receivers' telephone numbers, and their call duration. Therefore, our data help us identify an exogenously defined network of social neighbors because we do not use mobile Internet activity itself to construct the network or geographical proxies for reference groups.[7] Social network variables in our sample correspond to the level of content activities (i.e., frequencies of content uploads and downloads) of network neighbors of each user in the sample.

Note that we do not assume that mobility drives and determines the social network variable in our paper. Instead, we aim to utilize the extent of the geographical mobility of a user to measure the amount of travel-related discretionary time of the user. In addition, we use the individual-level interactions data among users and their social network to capture a source of learning identified in prior work. There are four possible types of user behavior that accrue as a consequence of interplay between the two variables – (a) low mobility and infrequent calling to people at travel destinations, (b) high mobility but infrequent

[5] We define *commonly* visited places as those places where a user visited at least once every week. Hence, we define an *uncommonly* visited place as a place a user did not visit every week.

[6] To be clear, we have voice call records only for 5 weeks. Thus, for the analysis in which we use social network data and mobility metrics based on call records, the time-period is reduced from 13 weeks to 5 weeks.

[7] For robustness checks, we also use text messaging-based communication records and geographical location of users to identify alternative social network neighbors and offline spatial network neighbors, respectively. We find qualitatively the same results as in the main results using voice records data.

calling to people at travel destinations, (c) low mobility but frequent calling to people at travel destinations, and (d) high mobility and frequent calling to people at travel destinations. The inclusion of both, the social network variable and the mobility variable helps capture all four types of user behavior.

Finally, we also have data on demographics like age, gender and product characteristics such as handset age. The summary statistics of the key variables used is provided in Table 2.

3.3 Basic Patterns in the Data

We now discuss stylized patterns in the data to motivate our subsequent econometric model development. First, we describe the interrelationships in the content generation and content usage both at the aggregate-level and the individual-level, which is linked to the key objective of this paper. Second, we describe the interrelationship of mobile Internet session initiation and content downloading and uploading activities. We discuss some evidence supporting the need for incorporating various econometric issues we address in our model.

3.3.1 Interrelationship between User Content Generation and Content Usage

In order to have a sense for how content generation and usage patterns are associated with each other, we plotted the total number of content uploads and content downloads in our sample over 13 weeks (i.e., 91 days). Figure 1 shows that there exists a strong weekly cycle for both content uploads and downloads. That is, mobile internet activity during weekdays is generally higher than that during weekends except the national holidays. This finding is consistent with prior work that has shown that consumers surf on the internet more during weekdays and regular working hours than during weekends and off-peak hours (Baye et al. 2009). The plot also shows similar patterns between the two time-series, implying that content generation and usage are indeed associated with each other at the aggregate level.

Further, we plotted the individual-level mobile Internet use patterns of some users in our data. Plot 1 in Figure 2 shows the weekly frequency patterns of content uploading and downloading of user 9116, and provides suggestive evidence for the need to incorporate “simultaneity” into our model. This plot shows that the uploading frequency is highly correlated with the downloading frequency in the case of this user. This kind of correlated behavior can be driven by the following factors – observed user heterogeneity (i.e., young users like both content uploading and downloading), correlated unobservables (i.e., users who like to upload also like to download content, and vice versa), etc. Although a single-equation panel data model can accommodate observed user heterogeneity (e.g., by including variables such as age and gender in the random effect specification and by differencing-out any user-specific time-invariant characteristics), it cannot address the correlated unobservables mentioned above. This motivated the use of a "simultaneous equation model" for content uploading and downloading equations. Furthermore, Plot 1 provides suggestive evidence of negative interrelationship between content generation and usage, especially during the first 9 weeks.

3.3.2 User Mobile Internet Session Initiation and Content Downloading and Uploading

Plot 2a and 2b in Figure 2 show the weekly frequency patterns of content uploading and downloading of user 1927 and user 3186, respectively. Although these two users have the same demographic characteristics (i.e., age and gender), they are very different in terms of their propensity to initiate a mobile Internet session as well as in terms of the number of content uploads and downloads. Note that user 1927 in Plot 2a with a higher frequency of mobile Internet session initiation (13 times) has engaged more frequently in content usage in comparison to user 3186 in Plot 2b who has a lower propensity to initiate a mobile Internet session initiation (2 times). In other words, even after controlling for observed demographics like age and gender, some unobserved user characteristics like inherent interest in initiating a mobile Internet session can explain such behavioral differences between users. This motivates the need to incorporate a "selection constraint" in our model in order to control for the non-randomness in users' mobile Internet session initiation stage.[8]

Lastly, Plots 1, 2a, and 2b indicate that the frequency of uploading and downloading at week 1 vary greatly by user. For example, user 1927 (Plot 2a) had 56 instances of downloading, whereas user 9116 (Plot 1) had about 18 instances of downloading. This implies that users could have differences in the extent of prior experience with respect to content uploading and downloading at the beginning of the sample (i.e., week 1). We model this "initial condition" issue by specifying selection equations for week 1 and for week 2 – 13, separately.

4. Econometric Model

To analyze the underlying process of user content generation and content usage, we build and estimate an individual-level simultaneous equations panel data model using three-stage least-square (3SLS)

[8] Descriptive statistics from our data suggests that young, male users tend to actively engage in mobile content generation and usage more than others. Thus, there could be disproportionately higher number of young, male users in the sample for mobile-session initiated users (i.e., selected sample) compared to in the total sample. Indeed results from a random effect dynamic probit model for users' session initiation equation supports this argument (see Appendix B). Further, if we were to group those users who have initiated mobile Internet sessions by the amount of their content activities (both uploads and downloads) and divide them into heavy-users vs. light-users, we see a disproportionately higher number of young, male users in the heavy-user sample compared to the light-user sample.

estimation. We further demonstrate the robustness of this analysis by conducting GMM-based dynamic panel data analyses. We also present and discuss a series of other robustness check analyses that include models with a random coefficient in a constant term, count data models, content share models, content size models, models with different time lags, and models using network variables based on different kinds of networks and different specifications.

In the mobile Internet space where users generate and use content using their phones, they face a two-step decision-making process. In step 1, they decide whether to initiate a mobile Internet session by clicking a button on the mobile phone. In step 2, if they have initiated a mobile Internet session, they determine how much to upload (if any) and how much to download (if any). They can engage in both uploading and downloading activities multiple times in a given mobile Internet session. Hence, there are three related user decisions – (a) mobile Internet session initiation, (b) content generation, and (c) content usage. Conditional on (a), the user could do (b), (c), or both.

There are five econometric issues to be considered here: (i) sample selection bias, (ii) social network endogeneity, (iii) initial conditions problem, (iv) unobserved user heterogeneity, and (v) simultaneity.

First, recall that we (researchers) can observe the content upload and download frequency of users only if they initiate their mobile internet sessions. Sample selection can arise in this setting because some people who initiate their mobile internet sessions more frequently can be also more prone to uploading (or downloading) content as opposed to those who less frequently initiate their mobile Internet sessions. If uncorrected, the estimates in the main equations are biased and inconsistent, leading to misleading inference (Heckman 1979). Further, it could undermine the external validity of estimates in that the estimates are only relevant to the selected sample (i.e., people who initiated mobile Internet sessions), thereby limiting the generalizability of results. In other words, the main source of the selection bias here is that users can decide whether to initiate a mobile internet session based on their discretion and intrinsic preferences as opposed to being randomly chosen. Hence, we controlled for this sample selection bias by including a selection correction term in main equations (i.e., content upload and download frequency).

Second, mobile Internet behaviors of a user and social network of that user can seem to be correlated regardless of whether it was because causal influence or not. To address the endogeneity issue of the network variable, we adopted the identification strategy and modeling approach of Nair et al. (2010). That is, we use the lagged social network variable in accordance with Manski (1993) and other work that has studied the impact of the social network effect on user behavior (Van den Bulte and Lilien 2001, Manchanda et al. 2008, Iyengar et al. 2010).

Third, we account for the well-known initial conditions problem in our model because for each user the first observation in our sample may not be the true initial outcome of his/her mobile content generation and usage process. *Fourth*, users are, in general, different in terms of their propensities and preferences towards content generation and usage. While some consumers tend to be users of content created by others, some others contribute by creating and uploading content to web portals and social networking sites. Thus, we account for this phenomenon by incorporating both observed and unobserved user heterogeneity in our model. *Finally*, as several plots have shown before, there could be simultaneity between content generation and usage, and we need to account for that too. We address each of five econometric issues above in the following Sections 4.1 and 4.2. Notations and variable descriptions are provided in Table 1 (see Appendix A).

4.1 Selection Equations: Mobile Internet Session Initiation

To statistically address the sample selection bias, we explicitly specify our econometric model by extending Verbeek and Nijman's (1996) two-step method.[9] In step 1, related to user's decision in step 1, we run a random effect (hereafter RE) dynamic probit model for user's binary decision of whether to initiate a mobile Internet session or not during a given week. Estimates from step 1 are used to obtain a Heckman (1979)'s selection correction term (see Appendix B for a detailed procedure and formula to compute the selection correction term.) In step 2, we insert the correction term into content generation and content usage equations, respectively, and estimate the two equations simultaneously using a three stage-least-square (3SLS) method. As mentioned before, to account for social network effects, we use a *lagged* social network variable. To control for observed user heterogeneity, we include time-invariant user-specific variables like age, sex, and handset age.

In addition, to account for the initial condition problem, we specify two separate equations for a user's mobile Internet session initiation decision: one for the first time period only (i.e., $t = 1$) and one for the remaining time periods (i.e., $t \geq 2$). We also include a lagged dependent variable. This also allows us

[9] A single-shot estimation model cannot address all econometric issues involved in our context. This has been documented in prior work. For example, Verbeek and Nijman (1996) and Stewart (2006) have pointed out that a single-shot estimation is not able to control for simultaneity. Similarly, BiØrn (2004) has pointed out that a single shot estimation is unable to handle the selection issue. Furthermore, a single shot estimation is also computationally very intensive, given the size of our data (2.3 million observations). Hence, we utilize the computationally less stringent two-step estimator for our model while explicitly incorporating all economic issues involved in our context.

to identify state dependence.[10] In addition, to account for the individual-level social network effect, we included *lagged* social network variables to alleviate the endogeneity bias that can arise with the use of the social network variable.[11] Furthermore, an orthogonality problem may arise in our model. Because our selection equation is based on a RE model, the estimator will be inconsistent if the unobserved, user-specific, time-invariant factor is correlated with the regressors therein. Hence, we follow the Mundlak (1978) and Zabel (1992) approach and add the mean values of time-varying regressors (in our case, the mean value of session initiation by social network neighbors) in the selection equation.

To be specific, we specify that user i decides whether to initiate mobile Internet sessions using an indicator function (i.e., 1 = Yes and 0 = No). We specify a model for the initial period (t = 1) as follows:

$$\mathrm{Session}^{*}_{i,1} = \pi_0 + \pi_1 \mathrm{Age}_i + \pi_2 {\mathrm{Age}_i}^2 + \pi_3 \mathrm{Sex}_i + \pi_4 \mathrm{Handset\ Age}_i + \pi_5 z_{-i,t} + u_i \qquad (1)$$

$$\mathrm{Session}_{i,1} = 1\left(\mathrm{Session}^{*}_{i,1} > 0\right).$$

For the remaining periods (t ≥ 2), we specify a model as follows:

$$\mathrm{Session}^{*}_{i,t} = \alpha_0 + \alpha_1 \mathrm{Session}_{i,t-1} + \alpha_2 \mathrm{Social\ Network\ Session}_{i,t-1} + \alpha_3 \mathrm{Age}_i + \alpha_4 {\mathrm{Age}_i}^2 + \alpha_5 \mathrm{Sex}_i$$

$$+\alpha_6 \overline{\mathrm{Social\ Network\ Session}_i} + \delta_i + \lambda_t + \alpha_7 z_{-i,t} + \eta_{i,t} \qquad (2)$$

$$\mathrm{Session}_{i,t} = 1\left(\mathrm{Session}^{*}_{i,t} > 0\right).$$

where δ_i is a user-specific random coefficient, λ_t is a time-period dummy, $z_{-i,t}$ is a mean mobile Internet session initiation of all other users in user i’s billing zip code, and $\eta_{i,t}$ is a user- and time-specific error term.

[10] There are two forms of dynamics in consumer decisions that can be modeled in a reduced form model like ours – (i) habit persistence due to switching costs and (ii) variety-seeking behavior that have been identified in prior work (see for example, Osborne 2007). By including the lagged dependent variable in the selection equation, we aim to control for these two possibilities. First, switching costs could arise in experience goods settings due to the additional time and psychological effort involved in facing the uncertainty of consuming new products. Hence, consumers with switching costs will continue to engage in the same activity in the current period as in the previous period. Second, some consumers may be variety-seeking with a preference for consuming new products or services (McAlister and Pessemier 1982). We can also expect this kind of switching behavior across users in mobile Internet settings. Hence, by including a lagged dependent variable in the selection equation, we can examine if there are switching costs experienced by users in this context (if the sign of the lagged dependent variable is positive and statistically significant). Likewise, we can also examine if there is a variety-seeking behavior among users in this context (if the sign of the lagged dependent variable is negative and statistically significant).

[11] We employed an individual-level word-of-mouth interactions data via mobile communications among users to capture a source of social influence based learning in consumer behavior (Ghose and Han 2009). Existing work in the consumer learning literature (for example, Erdem et al. 2008) has considered three sources for consumer learning – 1) consumers’ own experience, 2) firms’ marketing activities, and 3) social network effects. In particular, the learning can occur in the mobile media context through indirect word-of-mouth (WOM) signals such as the content creation and usage behavior of their social network neighbors. We capture this by including the social network variable in our study.

If the initial conditions are correlated with the unobserved, user-specific, time-invariant factor, as would be expected in most situations, our estimators will be inconsistent. Since the u_i in equation (1) is correlated with δ_i in equation (2), but uncorrelated with $\eta_{i,t}$ for t = 2, 3,…, T, we specify u_i as follows:[12]

$$u_i = \theta\delta_i + \eta_{i,1}. \tag{3}$$

where θ is a initial condition parameter. We estimate the RE dynamic probit model using Maximum Likelihood Estimation methods based on Stewart (2006, 2007).[13]

4.2 Main Equations: Content Generation and Content Usage Frequencies

We specify a fixed effect (hereafter FE) model in the content generation and usage equations.[14] To account for sample selection bias, we insert a selection correction term into content generation and content usage equations, respectively. To incorporate the temporal interdependence, we included a lagged content download frequency variable in a content upload equation and a lagged content upload frequency variable in a content download equation. We included each one of the four mobility metrics in our main equations in separate estimations to demonstrate robustness to both the mean and the variance of the mobility parameter, at both the local and national levels. In addition, to account for the individual-level social network effect, as before, we included *lagged* social network variables to alleviate the endogeneity bias that can arise with the use of the social network variable.[15] We used normalized number of content uploads and content downloads by network neighbors of each user in the model, respectively. We included the number of voice calls as a control variable in the content generation and usage equations to control for users' inherent propensity to make calls. Further, we included control variables such as time-invariant user-specific random coefficient, time-period dummies, and time-period and location specific fixed effects at the user level to control for endogeneity from using a social network variable as a regressor.

[12] We checked serial autocorrelation in the error term. We found the estimate for serial autocorrelation is not statistically significant (p-value is 0.627).

[13] Given that there are several user-specific, time-invariant variables that may affect one's mobile Internet session initiation, we use a RE dynamic probit model for the selection equation.

[14] Verbeek and Nijman (1996) showed that a fixed effect estimator in the main equations in their two-step approach is more robust to selection biases than a random effects estimator. Moreover, they also show that the conditions for the consistency of a fixed effect estimator are weaker than that for a consistent random effects estimator.

[15] We also specify an alternative model allowing for contemporaneous social network effects by using an instrumental variable approach, similar to Iyengar et al. (2010). The result shows that the contemporaneous social network effect is positive and statistically significant while other estimates qualitatively remain the same. Hence, we find no evidence of misspecification bias from using a lagged social network variable. Details are available upon request.

We took the logarithm on variables to control for their right-skewed nature (i.e., there are some heavy uploaders and heavy downloaders as can be seen from Table 2). We implement a 3SLS estimation on the first-differenced equations of log-transformed content generation and content usage frequencies. The simultaneous estimation method allows for efficiency gain compared to single equation estimation methods by taking the cross-equation error correlation into account.[16] To be specific, the content generation frequency and usage frequency equations are specified as follows, for t = 2, 3,…, T:[17]

$$\begin{aligned}\log\left(\text{Upload}_{i,t}\right) = \beta_0 &+ \beta_1\log\left(\text{Download}_{i,t-1}\right) + \beta_2\log\left(\text{Mobility}_{i,t}\right) + \beta_3\log\left(\text{Social Network Upload}_{i,t-1}\right)\\ &+\beta_4\log\left(\text{Voice}_{i,t}\right) + \beta_5\text{Selection}_{i,t} + \beta_6\log\left(g_{-i,t}\right)\\ &+\beta_7\log\left(\overline{\text{Social Network Upload}_{\imath}}\right) + \kappa_i + \varphi_t + \nu_{i,t}, \end{aligned} \tag{4}$$

$$\begin{aligned}\log\left(\text{Download}_{i,t}\right) = \gamma_0 &+ \gamma_1\log\left(\text{Upload}_{i,t-1}\right) + \gamma_2\log\left(\text{Mobility}_{i,t}\right) + \gamma_3\log\left(\text{Social Network Download}_{i,t-1}\right)\\ &+\gamma_4\log\left(\text{Voice}_{i,t}\right) + \gamma_5\text{Selection}_{i,t} + \gamma_6\log\left(h_{-i,t}\right)\\ &+\gamma_7\log\left(\overline{\text{Social Network Download}_{\imath}}\right) + \psi_i + \tau_t + \varepsilon_{i,t}. \end{aligned} \tag{5}$$

where $\text{Social Network Activity}_{i,t-1} = \sum_{m\in n_{t-1}(i)}\left(w_{i,m,t-1} \times \text{Activity}_{m,t-1}\right)$ and Activity is either Upload or Download. $g_{-i,t}$ and $h_{-i,t}$ are mean uploading and downloading frequencies of all other users in user i's billing zip code are, respectively. In addition, κ_i and ψ_i are user-specific dummies, φ_t and τ_t are time-period dummies, and $\nu_{i,t}$ and $\varepsilon_{i,t}$ are user- and time-specific error terms.

Further, recall that we specified a fixed effect model for equations of content generation and usage frequencies. It is well-known that the estimation of the fixed effect model with a lagged endogenous variable is subject to potential finite-sample bias (Nerlove 1967, Nickell 1981, Godes and Maylin 2004). Our analysis may suffer from this bias because the number of observations per user is 13 for the entire sample and only 5 for the sub-sample. Hence, we take first-differencing transformation on each variable in the model to alleviate the potential bias from the fixed effect model (Wooldridge 2002) as well as to difference out both observed and unobserved user-specific, time-invariant variables (e.g., age, gender, job

[16] When the disturbance covariance matrix is not known, GLS is inefficient compared to full information maximum likelihood (FIML) and three stage-least-squares (Lahiri and Schmidt 1978).

[17] Note that we cannot include a lagged dependent variable as a regressor in equations 4 and 5, because the lagged dependent variable in each equation is correlated with the disturbance term after first-differencing transformation, leading to inconsistency in estimates. In addition, a series of control variables help us capture the impact of any potential omitted variable (i.e., LDV). These control variables include (i) time-period fixed effects, (ii) time-and-location specific mean upload frequency variable at the user-level, and (iii) time mean content uploads by social network neighbors at the user-level can mitigate this bias. That said, our main results are robust even when we include a lagged dependent variable and estimate it using GMM-based dynamic panel data models (see Appendix D).

characteristics, prior internet experience, etc.).[18] Although we have found that there exists no serial correlation in the error term from the selection equation and in the error term from each of the main equations separately, we control for the potential serial correlation in the main simultaneous equations of content generation and usage by using the robust variance matrix (Wooldridge 2002).[19] The robust variance matrix estimator (Arellano 1987) is valid in the presence of serial correlation in error terms in equations 4 and 5 (Wooldridge 2002).

To be specific, the first-differenced content generation frequency and usage frequency equations that we estimate are specified as follows, for t = 2, 3,…, T:

$$\begin{aligned}\Delta\log(\mathrm{Upload}_{i,t}) = {} & \beta_1\Delta\log(\mathrm{Download}_{i,t-1}) + \beta_2\Delta\log(\mathrm{Mobility}_{i,t}) + \beta_3\Delta\log(\mathrm{Social\ Network\ Upload}_{i,t-1}) \\ & +\beta_4\Delta\log(\mathrm{Voice}_{i,t}) + \beta_5\Delta\mathrm{Selection}_{i,t} + \beta_6\Delta\log(g_{-i,t}) + \Delta\varphi_t + \Delta\nu_{i,t}, \qquad (6)\end{aligned}$$

$$\begin{aligned}\Delta\log(\mathrm{Download}_{i,t}) = {} & \gamma_1\Delta\log(\mathrm{Upload}_{i,t-1}) + \gamma_2\Delta\log(\mathrm{Mobility}_{i,t}) + \gamma_3\Delta\log(\mathrm{Social\ Network\ Download}_{i,t-1}) \\ & +\gamma_4\Delta\log(\mathrm{Voice}_{i,t}) + \gamma_5\Delta\mathrm{Selection}_{i,t} + \gamma_6\Delta\log(h_{-i,t}) + \Delta\tau_t + \Delta\varepsilon_{i,t}. \qquad (7)\end{aligned}$$

4.3 Identification

We discuss two issues in the identification of our model: 1) identifying the selection equation and the main equations of content generation and usage frequencies and 2) identifying the social network effect.

4.3.1 Identification of the Selection Equation and the Main Equations

Our identification strategy for the selection equation and the main equations includes normalization of parameters for binary decision variables, exclusion restrictions, and the use of instrument variables throughout our estimation process.

In the selection equation, since a user's mobile Internet session initiation is a binary choice, we need location and scale normalization on the latent dependent variable, $\mathrm{Session}^*_{i,t}$ for identification. For location normalization, we set the user i's utility of not engaging in any mobile Internet sessions in week t, $\mathrm{Session}^*_{i,t} = 0$. For scale normalization, we also set the variance of unobserved, user-specific, time-

[18] This approach is similar to Verbeek's (1990) where he takes the within transformation to eliminate the incidental parameters and maximizes the likelihood of the transformed data. He also shows that the corresponding estimator is consistent, even when only a few time series observations are available.

[19] We included the time-based control variables to control for underlying time trends that help eliminate serial correlation in our data. That said, we have estimated our main equations separately using the Generalized Estimating Equations method (GEE) and found that the estimated AR(1) coefficients in the upload equation and in the download equations are 0.006 and 0.007, respectively. Obviously, with such small AR(1) values, the other estimates qualitatively remained the same as in the result of our main model.

specific effect, σ_{η}^{2} to 1. In addition, sample selection issues require explicitly estimating the selection equation using both time invariant and time varying regressors (such as age, sex, handset age, mobile Internet session initiation by social network, etc.). Let us call these variables Z. Since a selection correction term is a nonlinear function of the variables included in the selection equation, our main equations of content generation and usage frequencies with regressors X is identified due to this nonlinearity even if Z = X. However, this nonlinearity arises from the assumption of normality in the probit model. Thus, we further checked an exclusion restriction by including variables in Z that are not included in X, which makes the identification cleaner (Puhani 2000). The exclusion restriction is satisfied because we included some time-invariant variables (e.g., age, gender, and handset age) as well as a time-varying variable (e.g., mobile Internet session initiation by social network) only in the selection equation but excluded them in main equations.

In the main equations, in the absence of better data, we used time-series based instruments for identification in the main equations. Content upload and download frequency variables in a given week are taken to be endogenous to the system of equations, while all other variables in the system are treated as exogenous to the system or predetermined. For example, in the content upload frequency equation, variables such as lagged download frequency, geographical mobility and lagged social network (see Section 4.3.2) are exogenous or predetermined. This is true for the following reasons. First, the geographical mobility is exogenous because it is very unlikely that one's mobility is determined by one's propensity to generate and use content. Instead, it is far more likely that a user's propensity to generate and use content is driven by the extent of their geographical mobility. Second, the download frequency variable at time t-1 is predetermined because it cannot be determined at time t. This is true because the error term at time t in equation (4) is uncorrelated with current and lagged values of the predetermined variable (i.e., $\mathrm{E}\left[\log\left(\mathrm{Download}_{\mathrm{i,t-1}}\right)\nu_{\mathrm{i,t}}\right] = 0$) but may be correlated with future values (i.e., $\mathrm{E}\left[\log\left(\mathrm{Download}_{\mathrm{i,t}}\right)\nu_{\mathrm{i,t}}\right] \neq 0$. This claim holds true because we have found that there is no serial autocorrelation in the error terms after controlling for time trend control variables (see Section 4.3.2). For the same reason, log upload at t-1 in equation (5) is a predetermined variable. Third, the lagged social network variable is predetermined at time t, because we have controlled for other sources for endogeneity from using a social network variable as a regressor (see Section 4.3.2). Hence, these are valid instruments for the endogenous variable because they are uncorrelated with the unobservable error term. A similar set of arguments applies to the content download frequency equation.

As a robustness check, we also included a non time-series based variable as an instrument. In particular, we included the "handset age variable" as an additional instrument in the content generation equation, excluding it from the content usage equation. The key assumption behind this argument is that the age of the handset is more likely to impede users from uploading multimedia content than downloading content. Users can download multimedia content irrespective of how advanced their handset features are such as the number of pixels in their mobile camera, the technical sophistication of the software and applications installed in the handset – the kinds of features which are required to experience multimedia content downloaded from the web. In contrast, the lack of handset functionality and advanced features is more likely to prevent users from creating and uploading content. This can happen because older phones are not equipped with advanced digital cameras and audio/video/photo editing applications – the kinds of features users need to upload content on the Internet. The qualitative nature of all our results remains the same with the inclusion of this variable as an instrument.

Furthermore, we examined whether both necessary order condition and sufficient rank condition are satisfied for our main equations of content generation and usage frequencies. The order condition is met because each equation has excluded exogenous or predetermined variables (i.e., a mobility variable, a lagged social network variable, a lagged temporal interdependent variable, mean mobile Internet activity of all other users in user i's billing zip code area, etc.) while it has no right-hand-side endogenous variable. Further, we checked the rank condition using Baum (2007)'s Stata code and found that the rank condition for each main equation is satisfied.

4.3.2 Identification of the Social Network Endogeneity

Mobile internet behavior of users and their social network can seem to be correlated regardless of whether it is a causal influence or not. To address the endogeneity issue of the network variable, we adopted the identification strategy and modeling approach of Nair et al. (2010). Consistent with the literature (Hartman et al. 2008), that has pointed out the issues in estimation of social interactions, Nair et al. (2010) distinguished causality from correlation by teasing out causal effects from each of the following three sources of correlation: 1) endogenous group formation, 2) correlated unobservables, and 3) simultaneity. Following their approach, we have now incorporated several additional variables to control for correlated unobservables also. These additional variables include (i) time-period fixed effects, (ii) time-and-location specific mean session initiation variables at the user-level, and (iii) time-and-location specific mean upload and download frequency variables at the user-level. We explain these in detail below.

First, regarding the endogenous group formation, the observed correlation in the behavior of an individual and other individuals in the social network could arise from omitted individual characteristics that are correlated within the group (Nair et al. 2010). Their approach to control for this is to include individual-specific effects in the equation. Hence, we included a user-specific random effect in the selection equation (i.e., equation 3) and a user-specific fixed effect in the main equations (i.e., equations 4 and 5).

Second, regarding the correlated unobservables, some unobservables could drive the behavior of an individual and the other individuals in his social network similarly. For example, a common event (e.g., graduation ceremony) for individuals who belong to a same group (i.e., class of 2010) can drive individuals in that group to upload their graduation photos to a social networking site and download each others' photos from that site. If uncorrected, this effect could be mistaken for a social contagion effect. According to Nair et al. (2010), in addition to adding the user-specific effect, there are three additional controls for the correlated unobservables. The first is to include time-period fixed effects. These can control for common factors or shocks to all individuals at a given time (Van den Bulte and Lilien 2001). The second one is to include location fixed effects (i.e., zipcode dummies). These can control for time-invariant spatially correlated unobservables (Nair et al. 2010). The third one is to include time and location effects to control for unobservables that are correlated at the level of zipcode *and* time. Hence, we included time-period fixed effects and time- and location-specific mean session initiation frequency of all other users in the zip code area of user i, denoted by $\overline{\mathrm{Session}_{-\imath,t}}$ in the selection equation. Likewise, we included time-period fixed effects and time- and location-specific mean content upload and download frequencies of all other users in the zip code area of user i, denoted by $\mathrm{g}_{-\mathrm{i},\mathrm{t}}$ and $\mathrm{h}_{-\mathrm{i},\mathrm{t}}$ in the main equations, respectively. Essentially the time- and location-specific variables proxies for all unobserved time-period and location-specific shocks to behaviors that are common to all users in user i's location (Nair et al. 2010).

Third, to control for simultaneity, in accordance with Hartmann et al. (2008), we use an exclusion restriction that imposes a temporal ordering i.e., the focal individual's behavior in time t is affected by the social network behavior up to time t-1. That is, we use the lagged social network variable in accordance with Manski (1993) and other work that has studied the impact of the social network effect on user behavior (Van den Bulte and Lilien 2001, Manchanda et al. 2008, Iyengar et al. 2010).[20]

[20] Lagging avoids the simultaneity problem, unless (1) people are forward-looking not only about their own behavior but also that of others *and* (2) social ties over which influence flows are symmetric (Hartmann et al. 2008, Iyengar et al. 2010). The first condition is quite unlikely in large networks such as ours, and the second condition does not hold in our data. As Iyengar et al. (2010) point out, if the impact of social networks is contemporaneous in our context, lagging creates a misspecification bias. To address this issue, we also specify our main models allowing for such simultaneous effect between a user and social networks of that user. We followed an instrumental variables

Our econometric specification is able to address all three issues required to identify causality from correlation in the social network effect - endogenous group formation, correlated unobservables, and simultaneity. Hence, our empirical estimates are suggestive of a causal effect of the social network on user behavior in mobile Internet usage and consumption. However, we would like to be cautious in our interpretation. In the absence of controlled variation using natural or field experiments during our sampling period, we interpret the relationship between social network behavior and user behavior in our paper as being a descriptive one that establishes an upper bound on the causal effect of the social network.

5. Results

In this section, we discuss our results on (i) temporal interdependence between content generation and usage, (ii) geographical mobility effect, and (iii) social network effect in the mobile Internet space.

We briefly summarize the key findings from the selection equation. Results show that there exist positive state dependence and positive social network association in initiating mobile Internet sessions. Also, users' mobile Internet initiation behavior greatly varies by age, implying an inverted U-shaped relationship between age and mobile media usage with a peak around approximately 21 years old. Estimates and detailed explanation are provided in Appendix C.

In Section 5.1, we present the 3SLS estimation results of content generation and usage equations using a 13-week sample. These results shed light on temporal interdependence and to some extent on social network effect.[21] Thereafter, in Section 5.2 we present the estimation results using a 5-week sample that has data on the communication strength between social network neighbors and data on the geographical mobility metrics of a user. These analyses shed light on mobility effect.[22] In Section 5.3, we discuss a series of robustness check results that further demonstrate robustness of our main result in

approach, similar to Iyengar et al. (2010). For example, for the social network upload frequency variable at time t ($\text{Social Network Upload}_{it}$), we regress it on an intercept, the social network upload frequency at time t-1, and a lagged cumulative social network upload frequency up to time t-1 ($\sum_{s=1}^{t-1} \text{Social Network Upload}_{is}$). We then compute the predicted value for the dependent variable (R^2= 83% in the upload regression and R^2= 70% in the download regression), and used it as the instrumented value for contemporaneous social network variable in our main model. We find that imposing the contemporaneous social network effect while avoiding the simultaneity bias does not change any of the coefficients in our main results. Details are available upon request.

[21] Recall that we do not observe voice call records during the entire 13-week period, but observe them only over a 5-week period. Thus, in order to measure the amount of content generation and content usage of network neighbors of a given user at a given time period, we defined user i's network neighbors based on 5-week voice call records and treated them fixed throughout the 13-week period. That is, the group of network neighbors of user i is denoted as $\mathrm{n(i)}$ in the 13-week sample, rather than $\mathrm{n_t(i)}$.

[22] We implemented all models here without the social network variable as well, to alleviate the remaining concerns about endogeneity even from the use of a lagged social network variable. Hence, tables in this section show both results based on a lagged network effect variable and based on no network variable. Overall, we find qualitatively the same result between with and without the lagged network variable.

Appendix D. In Section 5.4, we look at cohort analysis results to gain additional insights of results through sub-sample analyses. In Section 5.5, we discuss the economic implications of our findings by converting them into revenue changes.

5.1 Results Using the Total Sample

Recall that we (researchers) can observe the content upload and download frequency of users only if they initiate their mobile Internet sessions. Results show that the estimates for a selection correction term are indeed positive and statistically significant in content generation and usage equations (the coefficient estimates are 0.0216 and 0.7267, respectively). This suggest that some people who initiate their mobile Internet sessions more frequently are more prone to uploading and downloading content as opposed to those who less frequently initiate their mobile Internet sessions. We controlled for this sample selection bias by including a selection correction term in main equations (i.e., content upload and download frequency). Moreover, results from the selection equations (see Appendix C for detail) provide evidence that male users and younger people in early twenties who are prone to upload or download content using their mobile phones, initiate mobile internet sessions more frequently. These results confirm that controlling for the sample selection bias is crucial in our setting.

The main results from our simultaneous equations of content generation and usage using a total sample are given in Table 3. We find that there is a negative and statistically significant temporal interdependence between content generation and usage.[23] This implies that an increase in content usage in a previous period is associated with a decrease in content generation in a current period and vice-versa (the coefficient estimates are -0.0091 and -0.0178, respectively). This finding provides evidence that the resource constraint (e.g., time and money constraint) binds at least for some people. We also find a positive and statistically significant lagged social network association in content generation and usage equations (the coefficient estimates are 0.0115 and 0.0143, respectively). Also, we find statistically significant estimates for our control variables like time-period dummies, mean uploading and downloading frequency of all other users in user i's billing zip code variables, etc.

We discuss the impact of each effect using marginal effects.[24] For example, a 1 standard deviation increase in the number of previous period's content downloading decreases the number of content uploading in a current period by 3.6% when evaluated at the mean. Similarly, a 1 standard

[23] We use a week as a temporal unit in our study. We find qualitatively the same results when we used two weeks as a temporal unit.

[24] Given the log-log specification in equations 4 and 5, the coefficients represent elasticities. In addition to these elasticities, we interpret the coefficients using marginal effects as well as economic implications (see Section 5.5).

deviation increase in the number of previous period's content uploading decreases the number of content downloading in a current period by 5.1%. Thus, the marginal effect of temporal interdependence is asymmetric and stronger in content downloading than in content uploading.

In addition, the marginal effect of lagged social network effect is larger in content usage equation than in content generation equation (11.0% and 1.5%, respectively). Recall that this result of lagged social network effect does not incorporate the communication strength between users in imputing the structure of the social network for a given user. When we incorporate such information, we can obtain a more accurate measure of social network association with user behavior. This is discussed below in Section 5.2.

5.2 Results Using the Communication Strength and Geographical Mobility Sub-Sample

It is possible though that a user's content generation propensity is more strongly associated with that of his family, close friends or colleagues (whom the user calls more frequently or speaks for a longer duration) rather than that of acquaintances (whom the user calls less frequently or speaks for a shorter duration). To fully incorporate the dynamic, weighted social network effect on user behavior, we next present results from 3SLS estimation using a 5-week sample that has time-varying data on the extent of communication strength between users and their network neighbors. In addition, the 5-week sample includes the four different mobility metrics of users described before in Section 3.

We conduct analyses with both frequency-based and duration-based models in which we incorporate call frequencies and call durations, respectively, in determining the magnitude of the communication strength. Our results are robust to the use of either factor (call frequencies and call duration) as a weight for computing the strength of social network effect. Our results are robust to the exclusion of the number of voice calls as a control.

The main results are given in Table 4. As before, the estimates for selection correction terms are positive and statistically significant in both equations, reassuring that controlling for the sample selection bias is crucial in our setting. Our results show that there exists a negative and statistically significant interdependence between content usage in a current period and content generation in the previous period (for example, the coefficient estimates are -0.0098 and -0.0239, respectively in a model with the local location mobility variable). This result is consistent irrespective of the use of different mobility variables. As before, this finding lends support to the claim that the resource constraint (e.g., time and money constraint) binds at least for some people. These effects are also asymmetric. The marginal effect of temporal interdependence is stronger in content downloading than in content uploading (-6.8% and -4.3%, respectively, in a model with a local location mobility variable) – the negative impact of previous period's content uploading on current period's content downloading propensity is higher than vice-versa.

We find that our mobility metrics are positively associated with content generation and usage activities of users. For example, local location mobility of a user is positively associated with content generation and usage (the coefficient estimates are 0.0087 and 0.0193, respectively). We find that the marginal effect of local location mobility on content downloading is higher than that on content uploading (1.2% and 0.5%, respectively) when evaluated at the mean. However, we find that the marginal effect of national location mobility on content downloading and content uploading is similar (0.5% and 0.3%, respectively). These results on location mobility metrics suggest that when travelling and visiting different places but not necessarily faraway places, users seem to more frequently engage in content downloading compared to content uploading.

In addition, local mobility dispersion is positively associated with content generation and usage (the coefficient estimates are 0.0057 and 0.0328, respectively). We also find that its marginal effect on content downloading is much higher than on content uploading (3.3% and 0.6%, respectively). Similarly, the marginal effect of national mobility dispersion on content downloading is higher than on content uploading (1.0% and 0.3%, respectively). These results on mobility dispersion suggest that when traveling to a location that is different from their regular travel (i.e., away from home or office), users also seem to more frequently engage in content downloading compared to content uploading.

The relationship between content generation and usage behaviors of lagged network and of users is positive and statistically significant (the coefficient estimates are 0.0162 and 0.0348, respectively, in a model with the local location mobility variable). This result is also consistent irrespective of the use of different mobility variables. We find that the marginal effect of the lagged social network effect on content downloading is much greater than on content uploading (26.8% and 1.4%, respectively). We discuss the implications of these results in Section 6. In addition, we find statistically significant estimates for our control variables like time-period dummies, mean uploading and downloading frequency of all other users in user i's billing zip code variables, etc. We discuss the economic impact of our findings in Section 5.5.

5.3 Robustness Check Results

Furthermore, we implemented a series of robustness checks. Because of the evidence of positive state dependence from the selection equation results, we conducted tests to check the robustness of the results by estimating the main equations separately with a lagged dependent variable to control for the state dependence using *GMM-based dynamic panel data model*. To further capture unobserved heterogeneity amongst users, we also estimated *a mixed effect model* where we include a random coefficient for a constant term (i.e., β_0 in equation 4 and γ_0 in equation 5).[25] We find that the results are qualitatively the same as that in our main results (for details, see Tables D1 and D2 in Appendix D).

Strictly speaking, users' content uploads and downloads variables in our sample take on nonnegative integer values. We have used various types of *linear* models (i.e., 3SLS and GMM dynamic panel data models). However, for count data, linear models have shortcomings. Hence, one could argue that we should examine our questions using *count data models*. However, count data models with fixed effects are known to suffer from the "incidental parameters problem" except the Poisson model. Therefore, we implemented a Poisson fixed effect model in which the incidental parameter problem is not a problem (Lancaster 2000, Greene 2007). We find that the results are qualitatively the same as that in our main results from 3SLS estimation (for details, see Table D3 in Appendix D).

Since the total amount of money and time resources of a user can potentially vary every week, an alternative model would be to model the user's share of content uploads with respect to total amount of content uploads and downloads (*content share model*). Besides the frequency of content uploads and downloads, costs that users will incur also depend on how many bytes a user uploads and downloads. Towards this, we used the amount of bytes uploaded and downloaded for each user instead of the frequency of uploading and downloading (*content size model*). We find that these results remain the same as in the 3SLS estimation result of our main model (for details, see Tables D4 and D5 in Appendix D).

In addition, we implemented our main 3SLS model using network variables based on different kinds of networks – text message-based social network and offline location-based spatial network. For example, people who live in the same geographical vicinity can have a higher probability of communicating with each other offline. We used zip code information from the billing address of a user to identify the spatial network neighbors of the user. We find that the impact of the voice call-based social network on users' mobile Internet usage behavior is higher than the impact of location-based spatial network. Specifically, we find that the marginal effect of the voice-based social network on content downloading is 26.8% vs. 1.4% on content uploading. The marginal effect of the location-based spatial

[25] Recall that in the 3SLS estimation based on first-differenced simultaneous equations, we eliminated these constant terms. Hence, we did not estimate them.

network on content downloading is 10.5% vs. 1.2% on content uploading. Furthermore, we used both voice call data and the zip code data simultaneously to identify the overlapping users between one's social network and spatial network. We find the impact of the overlapping network (that is, combination of the social and spatial networks) is much larger than that in either the social network or spatial network. Specifically, the marginal effect of the combined social and spatial networks on content downloading is 77.9% vs. 14.4% on content uploading (for details, see Tables D6, D7, and D8 in Appendix D).[26]

5.4 Additional Check: Cohort Analysis Results

We implemented four sets of cohort analyses. First, a useful test for the central notion of economic behavior under resource constraints espoused in this paper would be to examine if users who appear closer to a binding constraint on resources show a stronger (negative) temporal interdependence between content generation and usage. Towards this, we divided the sample based on the age of the user into two cohorts: "younger" users (below the age of 22) and "older" ones (above the age of 22). Results were robust with respect to this age cutoff point. The assumption is that younger users are more likely to face monetary constraints compared to the older ones due to a lower amount of discretionary income. Hence, we would expect a greater level of negative temporal interdependence between content generation and usage behavior of such users. 3SLS estimation results show that this holds true. For example, the marginal effect of temporal interdependence in content downloading (content uploading) is -8.7% (-4.1%) in the younger users cohort, whereas it is -2.7% (-2.0%) in the older users cohort. This finding implies that the resource constraint binds more tightly on younger users than on older users.[27]

Second, we divided the sample based on the location of the user into two cohorts: "urban" users (who live in major 6 cities in South Korea) and "sub-urban" users (who live in other areas). The premise is that urban users are more likely to have better 3G broadband coverage as well as traveling-related discretionary time via public transportations (i.e., subways or buses) compared to sub-urban users. Hence, we would expect a higher impact of location mobility on urban users' content generation and usage behavior. 3SLS estimation results show that this holds true. For example, the marginal effect of local location mobility on content downloading (content uploading) is 2.2% (1.8%) in the urban user cohort, whereas it is 1.8% (0.3%) in the sub-urban user cohort.

[26] We also tried alternative specification of social network effects – lagged cumulative effect and lagged binary indicator. Neither of these led to any change in the qualitative nature of the results. Details are available upon request.

[27] Note that the marginal effect of temporal interdependence in content downloading (content uploading) in the *total* sample was -5.1% (-3.6%). The qualitative nature of our results remains the same regardless of different kinds of geographical mobility metrics used in the model.

Third, we implemented a sub-sample analysis by excluding users who either uploaded or downloaded disproportionately, to alleviate the potential bias from including these outliers (i.e., UGC junkies or free-riders). The sub-sample consists of those users whose upload frequency is in a similar range to their download frequency. To be specific, we included a user into our sub-sample if the absolute difference between download frequency and upload frequency of the same user is less than a given cutoff value (e.g., 3, 5 or 10). We find qualitatively the same result as in the main result.

Lastly, we ran analyses on a sub-sample consisting of only those users who have engaged in both uploading and downloading activities in the same week at least once in the sample, to mitigate the potential bias from including users who either uploaded or downloaded but did not engage in both activities. This sub-sample constitutes 15.7% of the total sample. We find support for the negative temporal interdependence between content uploading and downloading behavior. The results for the geographical mobility effect and the social network effect are qualitatively the same as in our main results.

5.5 Economic Implications

It is true that the elasticity estimates are small. However, to understand their economic importance in the context of this industry, it is not that interesting to study the impact of a 1% change in the variables, In particular this is the case because the mean of several variables is small but the standard deviation is proportionately much higher. For example, the mean frequency of content uploading in our sample is 0.27 times a week while its standard deviation is 3.54 times a week. Although the elasticity of content download frequency means that 1% increase in upload frequency at t-1 is associated with about -0.02% decrease in download frequency at t), it is important to note that the 1% increase in upload frequency from its mean corresponds to only 0.0027 times change in content upload. Therefore, it is less meaningful to assess the impact of 1% increase in usage frequency when we interpret their economic effects.

To address this, we evaluate the impact of different percentile changes in independent variables for different sizes of the user base, and convert that impact into monetary values. We look at the economic impact on three different user groups chosen based on their overall content upload and download frequencies – (i) top 1 percentile user group, (ii) top 10 percentile user group, and (iii) top 25 percentile user group. Each user group represent 100,000 users, 1,000,000 users, and 2,500,000 users, respectively, of the company in our data (the company has about 10 million users). And for each of these user groups, we look at changes in their mobile internet content generation or usage that may shift them from the top 90 to the top 10 percentile, from the top 75 to the top 25 percentile, from the top 50 to the top 25 percentile and from the top 25 to the top 10 percentile.[28]

In our data context, users are charged by the amount of traffic transmitted during their content uploading and downloading activities @ $1.5 per 1 mega byte of data transmission. Based on the top 25 percentile user group data (2,500,000 users), a one-time content upload leads to an average of 0.14 mega bytes of data transmission and a user uploads about 0.49 times on an average during a week. As noted above, the elasticity of current content upload frequency with respect to the lagged content download frequency is -0.0098. Suppose there is an increase in the usage frequency of this group such that it takes them from the top 25 percentile (48.8 times/week) to the top 10 percentile (82 times/week).

Then, the annualized economic impact from an increase in the download frequency is computed as follows: elasticity x percentage change in usage x average number of bytes uploaded per instance x price per byte x number of users x average number of times downloaded per week x number of weeks in a year. It is -0.0098 x (82 – 48.8) / 48.8 x 0.14 mega bytes per time x $1.5 per mega byte x 2,500,000 users x 48.8 times per week x 52 weeks = -$8,882,328. That is, the company can incur a loss of approx $8.88 million in upload-traffic revenues from the top 25 percentile user group due to the negative interdependence between content generation and usage. Similarly, we calculate the annualized economic impact for other independent variables. From the same top 25 percentile of its users, the company can incur a gain of more than $4.93 million in revenues by a 15th percentile increase in local location mobility and local mobility dispersion collectively.

As a proportion of the number of consumers in the top-25 percentile user base, an increase in content download frequency from the top 25 to top 10 percentiles, which leads to a loss of $ 8.88 million (as calculated above), translates to more than 3.5% of its gross annual revenues. In addition, the economic impact from other percentile changes in mobile internet content generation and usage frequencies are given in Table 5.

6. Discussion and Implications

Mobile Internet content services constitute one of the fastest-growing applications on the web. However, little is known about how the content generation behavior of users is related to their content usage behavior, and how users' mobile Internet use is related to the extent of their geographical mobility and their mobile social network behavior. To examine such issues, we develop an econometric model for individual-level user behavior in the mobile Internet space and analyze it using an unprecedented user-

[28] These percentile ranges are chosen arbitrarily and are only meant for illustrative purposes. The main implications remain consistent irrespective of the percentiles.

level data. Our data comes from a setting where users are enrolled in usage-based data pricing. Recently media reports have pointed out that the major wireless carriers in the US, AT&T and Verizon Wireless, are getting ready to implement usage-based data pricing scheme for data/internet usage (Rethink Wireless 2010). These mobile carriers have profited from low fixed-fee penetration pricing. However, these carriers had seen enormous growth in data traffic, which often outpaced the capacity of their networks. For example, AT&T had experienced 5,000% growth in data traffic over last three years, but 40% of that traffic was consumed by just 3% of its smart-phone users (Rethink Wireless 2010). A global survey of mobile telecom executives across 50 countries and 6 continents conducted by Economist in 2010 revealed that 60% of respondents believed usage-based data pricing is the way of the future in mobile data and content services. Our results based on usage-based data pricing scheme can provide useful managerial insights for the companies that are contemplating the such pricing schemes.

The insights from this study can have some managerial implications. *First*, the asymmetric, negative temporal interdependence between content generation and usage provides mobile phone companies with insights on how to stimulate content generation and usage. This implies that users face more difficulties when they upload (e.g., technical difficulties with respect to use content uploading function on their phones, lack of prior experience in uploading any content using their phones, etc.) as opposed to when they download content. There is suggestive evidence of technical difficulties on user content generation in an online setting (Stoeckl et al. 2007). Hence, companies provide easy-to-use content preprocessing tools and less complicated content uploading procedures for the mobile web and look to monetize such user-generated content through advertising.

Further, the provision of monetary incentives might be effective in triggering users' content generation behavior. In many ways, content diffusion in the mobile Internet environment is similar to that in P2P networks because free-riders can create supply-side constraints. Hence, firms engaging in mobile content provision and advertising could think of offering distribution referrals in the form of monetary payments to users who generate and distribute content, similar to that in P2P networks (Hosanagar et al. 2008). Implementation can be done by providing discounts in data transmission charges to such users.

Second, the asymmetric association of the extent of users' geographical mobility with their content usage compared to content generation behavior can provide mobile phone companies with some insights on targeted mobile advertisements. It is now increasingly recognized that location has become a key catalyst to attracting local investments for conversions in mobile ad campaigns. Either through a map-based experience or through a generic location-based search, geography is a critical element of the mobile experience today and will be in the future. Previous research on location-based mobile advertising shows that users find mobile advertisements distressful when they receive those ads in private locations (e.g., home) and in work-related situations (Banerjee and Dholakia 2008). Our results imply that users are more likely to engage in mobile Internet usage when they travel. Moreover, we find that when users travel to different places but not necessarily faraway places (i.e., high location mobility) or when they travel to a location that is different from their regular travel (i.e., high mobility dispersion), they engage more frequently in content usage compared to content generation. Using the geographical mobility information, firms could engage in personalized and dynamically generated mobile advertisements.

Third, our results on the positive association between different metrics of social network behaviors (based on voice, text, and spatial network data) and user behavior can provide insights into how companies could benefit by targeting the right users who can then influence user behavior in the mobile Internet space. Our results show that the social network has a stronger effect on user behavior. Hence, mechanisms designed to instantly update a user about the frequency with which her network neighbors have downloaded or uploaded a certain type of content can affect the incentives of the user to do the same. Furthermore, we find that a combination of spatial co-location and voice call-based social network information generates the highest network neighbor effect compared to using each of these networks in isolation. Hence, firms could personalize offers to users by combining their offline geo-location data with voice or text-based information data. All of these can contribute towards the larger objective of increasing revenues by increasing network traffic as well as by monetizing user-generated content.

Our paper has limitations. These limitations arise mainly from the lack of data. For example, we do not have information about the specific type of content uploaded or downloaded (e.g., photo, audio, text, etc.) and the destination websites (e.g., social networking sites, mobile portal sites, etc.). Future work could examine this information. Another area for future work is to study how content generated via mobile phones diffuses through social networks. We hope that our study will generate further interest in the emerging literature on the economics of user-generated content and more broadly, in mobile commerce.

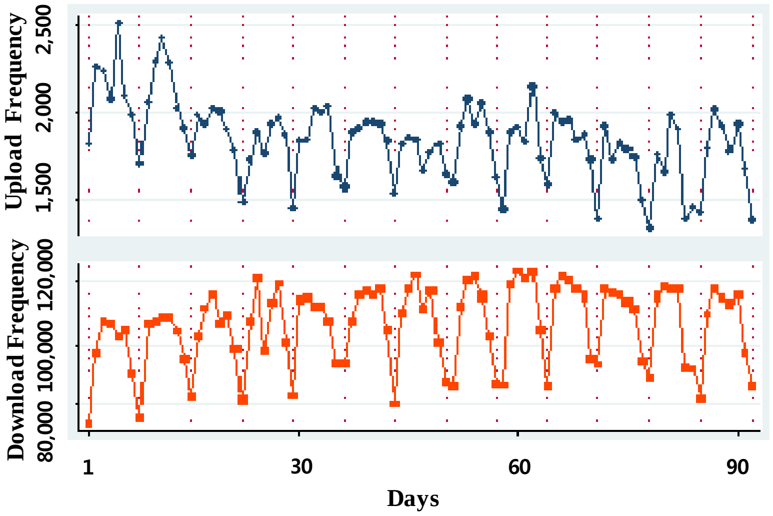
References

- Aral. S., L. Muchnik, A. Sundararajan 2009. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks, *Proceedings of the National Academy of Sciences, 106(51).*
- Arellano, M. 1987. Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics*, **49**(4), 431-434.
- Asvanund, A., K. Clay, R. Krishnan, M. Smith. 2004. An empirical analysis of network externalities in peer-to-peer music sharing networks. *Information Systems Research*, **15**(2), 155-174.
- Banerjee, S., R. R. Dholakia. 2008. Does location based advertising work? *International Journal of Mobile Marketing*, **3**(2), 68-74.
- Baum, C. F. 2007. CHECKREG3: Stata module to check identification status of simultaneous equations system. Statistical software components S456877, Boston College Department of Economics.
- Baye, M., J. Rupert, J. Gatti, P. Kattuman, J. Morgan. 2009. Clicks, discontinuities, and firm demand online. *Journal of Economics and Management Strategy*, **18**(4), 935-975.
- Becker, G. S. 1965. A theory of the allocation of time. *Economic Journal*, **75**(299), 493-517.
- Bell, D., S. Song. 2007. Neighborhood effects and trial on the Internet: evidence from online grocery retailing. *Quantitative Marketing and Economics*. **5**(4), 361-400.
- Bucklin, R. E., C. Sismeiro 2003. A model of Web site browsing behavior estimated on clickstream data. *Journal of Marketing Research*, **40**(3), 249-267.
- Dasgupta, K., R. Singh, B. Viswanathan, D. Chakraborty, S. Mukherjea, A. Nanavati, A. Joshi. 2008. Social ties and their relevance to churn in mobile telecom networks. *Proceedings of the 11th International Conference on Extending Database Technology Conference*, Nantes, France.
- Ellison, G., S. F. Ellison 2005. Lessons about markets from the Internet. *Journal of Economic Perspective*, **19**(2), 139-158.
- Erdem, T., Keane, M., B. Sun. 2008. A Dynamic Model of Brand Choice When Price and Advertising Signal Product Quality, *Marketing Science*, **27**(6), 1111-1125.
- Ghose, A., S. Han. 2009. A Dynamic Structural Model of User Learning in Mobile Media Content, Working paper, SSRN.
- Godes, D., D. Mayzlin. 2004. Using online conversations to study word-of-mouth communication. *Marketing Science*, **23**(4), 45-560.
- Greene, W. 2007. Fixed and random effects models for count data. Working Paper NYU EC-07-16, Stern School of Business. Available at SSRN: http://ssrn.com/abstract= 1281928.
- Hartmann, W. R., P. Manchanda, H. Nair, M. Bothner, P. Dodds, D. Godes, K. Hosanagar, C. Tucker. 2008. Modeling social interactions: Identification, empirical methods and policy implications. *Marketing Lett*ers, **19**(3), 287-304.
- Heckman, J. J. 1979. Sample selection bias a specification error. *Econometrica*, **47**(1), 153-161.
- Hill, S., F. Provost, C. Volinsky. 2006. Network-based marketing: Identifying likely adopters via consumer networks. *Statistical Science*, **21**(2), 256-276.
- Homans, G. C. 1958. Social behavior as exchange. *American Journal of Sociology*, **63**(6), 597-606.
- Hosanagar, K., Y. Tan, P. Han. 2008. Optimal dynamic referrals in peer-to-peer media distribution, Working paper, SSRN.
- Iyengar, R., C. Van den Bulte, T. Valente. 2010. Opinion leadership and social contagion in new product diffusion, forthcoming, *Marketing Science.*
- Jacoby, J., G. J. Szybillo, C. K. Berning. 1976. Time and consumer behavior: An interdisciplinary overview. *Journal of Consumer Research*, **2**(1), 320-339.
- Lahiri, K., P. Schmidt. 1978. On the estimation of triangular structural systems. *Econometrica*, **46**(5), 1217-1221.
- Lancaster, T. 2000. The incidental parameters problem since 1948. *Journal of Econometrics*, **95**(2), 391-414.
- Manchanda, P., Y. Xie, N. Youn. 2008. The role of targeted communication and contagion in product

adoption. *Marketing Science*, **27**(6), 961-976.

- Manski, C. 1993. Identification of endogenous social effects. *Review of Economic Studies*, **60**(3), 531-542.
- McAlister, L. and E. Pessemier. 1982. Variety-Seeking Behavior: An Interdisciplinary Review. *Journal of Consumer Research*, **9**(3), 311-322.
- Mundlak, Y. 1978. On the pooling of time series and cross section data. *Econometrica*, **46**(1), 69-85.
- Nair, H., P. Chintagunta, J.-P. Dubé. 2004. Empirical analysis of indirect network effects in the market for personal digital assistants. *Quantitative Marketing& Economics* 2(1) 23–58.
- Nair, H., P. Manchanda, T. Bhatia, A. 2010. Asymmetric Social Interactions in Physician Prescription Behavior: The Role of Opinion Leaders. *Journal of Marketing Research*, Forthcoming.
- Nerlove, M. 1967. Experimental evidence on the estimation of dynamic economic relations from a time series of cross sections. *Econometrica*, 39(March), 359-387.
- Nickell, S. 1981. Biases in dynamic models with fixed effects. *Econometrica*, 49(6), 1417-1426.
- O'Hara, K., A. Mitchell, A. Vorbau. 2007. Consuming Video on Mobile Devices. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, San Jose, California, USA.
- Osborne, M. 2007. Consumer Learning, Switching Costs, and Heterogeneity: A Structural Examination, No 200710, EAG Discussions Papers, Department of Justice, Antitrust Division, http://econpapers.repec.org/RePEc:doj:eagpap:200710.
- Puhani, P. 2000. The Heckman correction for sample selection and its critique. *Journal of Economic Surveys*, **14**(1), 53-68.
- Rethink Wireless. 2010. AT&T will use new device formats to introduce usage-based pricing. http://www.rethink-wireless.com/article.asp?article_id=2722.
- Shim, J. P., S. Park, J. M. Shim. 2008. Mobile TV phone: current usage, issues, and strategic implications. *Industrial Management and Data Systems*, **108**(9), 1269-1282.
- Stewart, M. B. 2006. Maximum simulated likelihood estimation of random–effects dynamic probit models with autocorrelated errors. *The Stata Journal*, **6**(2), 256-272.
- Stewart, M. B. 2007. The interrelated dynamics of unemployment and low-wage employment. *Journal of Applied Econometrics*, **22**(3), 511-531.
- Susarla, A., J. Oh, Y. Tan. 2010. Social Networks and the Diffusion of User-Generated Content: Evidence from YouTube. *Information Systems Research*, Forthcoming.
- Trusov, M., A. V. Bodapati, R. E. Bucklin. 2006. Your members are also your customers: Marketing for Internet social networks. Working paper, UCLA, CA.
- Tucker, C. 2008. Identifying formal and informal influence in technology adoption with network externalities. *Management Science*, **54**(12), 2024-2038.
- Tucker, C., J. Zhang. 2010. Growing two-sided networks by advertising the user base: A field experiment, forthcoming *Marketing Science*.
- Van den Bulte, C., G. Lilien. 2001. Medical innovation revisited: Social contagion versus marketing effort. *American Journal of Sociology*, **106**(5), 1409-1435.
- Verbeek, M. 1990. On the estimation of a fixed effects model with selectivity bias. *Economics Letters*, **34**(3), 267-270.
- Verbeek, M., T. Nijman. 1996. Incomplete panels and selection bias, in Mátyás, L., P. Sevestre (Eds.). *The Econometrics of Panel Data: A Handbook of the Theory with Applications*. Kluwer Academic Publishers, Dordrecht, 449-490.
- Wooldridge, J. M. 2002. Econometric Analysis of Cross section and Panel Data, MIT Press, MA.
- Xia, M., Y. Huang, W. Duan, A. B. Whinston. 2007. To share or not to share? An empirical analysis on user decisions in online sharing communities. Working paper, SSRN.
- Yang, S., G. Allenby. 2003. Modeling interdependent consumer preferences. *Journal of Marketing Research*, **40**(3), 282-294.
- Zabel, J. E. 1992. Estimating fixed and random effects models with selectivity. *Economics Letters*, **40**(3), 269-272.

Figure 1: Aggregate-Level Mobile Internet Activity Frequency Series Plot



***Notes*:** Vertical dotted lines represent Sundays. Mobile Internet activity during weekdays is generally higher than that during weekends except the national holidays (e.g., Day 50 – Children's day, Day 57 – Buddha's birthday, Day 83 – Memorial Day).

Figure 2: Individual-Level Mobile Internet Activity Frequency Series Plots

Interrelationship between User Content Generation and Usage

User Mobile Internet Session Initiation and Content Generation

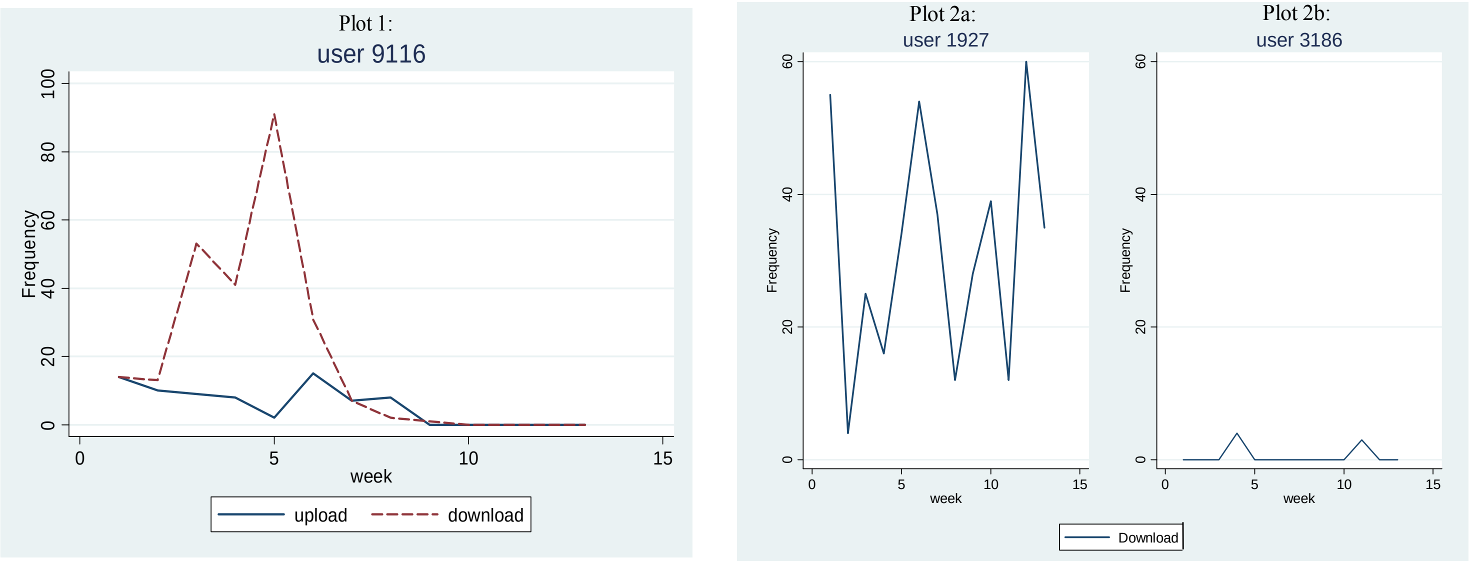


Table 2: Summary Statistics

Variable	Observations	Mean	Std. dev.
Weekly, User-Specific Content Activity Data			
Num. of Mobile Internet session Activation	2,340,000	4.00	41.38
Num. of Uploading	610,809	0.27	3.54
Num of Downloading	610,809	22.57	86.80
Weekly, User-Specific Call Data			
Num. of Calls Made	900,000	11.88	16.32
Call Duration (Hours)	900,000	2.61	5.68
Weekly, User-Specific Geographical Data			
Local Location Mobility	900,000	14.04	13.18
National Location Mobility	900,000	5.91	2.88
Local Mobility Dispersion	900,000	2.85	3.83
National Mobility Dispersion	900,000	0.70	0.80
User Characteristics			
Age	180,000	30.13	5.91
Sex (1 = Male, 0 = Female)	180,000	0.53	0.50
Handset Age (Months)	180,000	9.63	3.97

Notes: We observe content generation and usage data only when a user starts mobile Internet sessions, thus the number of uploading and the number of downloading are lower than the number of sessions.

Table 3: 3SLS Estimation Results on Content Frequency (Total Sample)

Dependent Variable	Explanatory Variable	Coefficient	
		Lagged Network Variable	No Network Variable
Log Upload Frequency (t)	Log Download Frequency (t-1)	-0.0091 (0.0005)***	-0.0091 (0.0005)***
	Log Upload Frequency by NN (t-1)	0.0115 (0.0042)***	
	Selection (t)	0.0216 (0.0007)***	0.0216 (0.0007)***
Log Download Frequency (t)	Log Upload Frequency (t-1)	-0.0178 (0.0073)***	-0.0178 (0.0073)***
	Log Download Frequency by NN (t-1)	0.0143 (0.0047)***	
	Selection (t)	0.7267 (0.0026)***	0.7267 (0.0039)***

Notes: NN refers to network neighbors. Estimates for *time-period* fixed effects and *mean uploading and downloading frequency of all other users in user i's billing zip code* effects are not reported due to brevity. *** denotes significant at 0.01.

Table 4: 3SLS Estimation Results on Content Frequency

Dependent Variable	Explanatory Variables	Coefficient							
		Lagged Social Network	No Social Network	Lagged Social Network	No Social Network	Lagged Social Network	No Social Network	Lagged Social Network	No Social Network
Log Upload Frequency (t)	Log Download Freq. (t-1)	-0.0098 (0.0005)***	-0.0098 (0.0005)***	-0.0098 (0.0005)***	-0.0098 (0.0005)***	-0.0097 (0.0005)***	-0.0097 (0.0005)***	-0.0097 (0.0005)***	-0.0097 (0.0005)***
	Log Local Location Mobility (t)	0.0087 (0.0014)***	0.0087 (0.0014)***						
	Log National Location Mobility (t)			0.0163 (0.0025)***	0.0163 (0.0025)***				
	Log Local Mobility Dispersion (t)					0.0057 (0.0016)***	0.0057 (0.0016)***		
	Log National Mobility Dispersion (t)							0.0057 (0.0012)***	0.0057 (0.0012)***
	Log Number of Voice Calls (t)	0.0010 (0.0007)	0.0011 (0.0007)	0.0013 (0.0007)*	0.0013 (0.0007)*	0.0018 (0.0007)***	0.0019 (0.0007)***	0.0010 (0.0007)	0.0010 (0.0007)
	Log Upload Freq. by NN (t-1)	0.0162 (0.0054)***		0.0163 (0.0054)***		0.0163 (0.0054)***		0.0163 (0.0054)***	
	Selection (t)	0.0118 (0.0004)***	0.0118 (0.0004)***	0.0116 (0.0004)***	0.0116 (0.0004)***	0.0126 (0.0004)***	0.0126 (0.0004)***	0.0130 (0.0004)***	0.0130 (0.0004)***
Log Download Frequency (t)	Log Upload Freq. (t-1)	-0.0239 (0.0092)***	-0.0239 (0.0092)***	-0.0238 (0.0092)***	-0.0238 (0.0092)***	-0.0234 (0.0091)***	-0.0234 (0.0091)***	-0.0234 (0.0091)***	-0.0234 (0.0091)***
	Log Local Location Mobility (t)	0.0193 (0.0057)***	0.0194 (0.0057)***						
	Log National Location Mobility (t)			0.0218 (0.0098)**	0.0220 (0.0098)**				
	Log Local Mobility Dispersion (t)					0.0328 (0.0065)***	0.0328 (0.0065)***		
	Log National Mobility Dispersion (t)							0.0218 (0.0047)***	0.0219 (0.0047)***
	Log Number of Voice Calls (t)	0.0385 (0.0027)***	0.0386 (0.0027)***	0.0401 (0.0026)***	0.0402 (0.0026)***	0.0377 (0.0026)***	0.0378 (0.0026)***	0.0356 (0.0028)***	0.0357 (0.0028)***
	Log Download Freq. by NN (t-1)	0.0348 (0.0057)***		0.0348 (0.0057)***		0.0348 (0.0057)***		0.0348 (0.0057)***	
	Selection (t)	0.3633 (0.0014)***	0.3633 (0.0014)***	0.3635 (0.0014)***	0.3635 (0.0014)***	0.3658 (0.0014)***	0.3658 (0.0014)***	0.3666 (0.0014)***	0.3666 (0.0014)***

Notes: We included each one of the two *location mobility* metric and the two *mobility dispersion* metrics in our main equations at both the local and national levels. NN refers to network neighbors. Estimates for *time-period* fixed effects and *mean uploading and downloading frequency of all other users in user i's billing zip code* effects are not reported due to brevity. *** denotes significant at 0.01, ** denotes significant at 0.05, * denotes significant 0.1.

Table 5: Annualized Economic Impact of Different Percentile Changes in Independent Variable from Different Sizes of User Base

Independent Variable	Percentile Change in Independent Variable (Worse → Better)											
	90→10	75→25	50→25	25→10	90→10	75→25	50→25	25→10	90→10	75→25	50→25	25→10
	Top 1 Percentile User Base (100,000 users)				Top 10 Percentile User Base (1,000,000 users)				Top 25 Percentile User Base (2,500,000 users)			
Lagged Download Frequency	-$9,081,072	-$3,837,288	-$2,644,824	-$5,075,616	-$13,621,608	-$5,755,932	-$3,967,236	-$7,613,424	-$15,891,876	-$6,715,254	-$4,628,442	-$8,882,328
Lagged Upload Frequency	-$2,374,991	-$1,043,952	-$782,964	-$1,278,841	-$5,241,055	-$1,864,200	-$1,398,150	-$2,283,645	-$5,937,477	-$2,609,880	-$1,957,410	-$3,197,103
Local Location Mobility	$5,435,640	$2,415,840	$905,940	$1,207,920	$7,140,744	$3,173,664	$1,190,124	$1,586,832	$9,911,538	$4,405,128	$1,651,923	$2,202,564
Local Mobility Dispersion	$4,227,611	$2,536,566	$845,522	$1,268,283	$6,529,380	$3,917,628	$1,305,876	$1,958,814	$9,110,010	$5,466,006	$1,822,002	$2,733,003

Note: We selected the top 1, 10 and 25 percentile user bases in terms of the total number of user content upload and download frequency. With respect to mobility variables, we summed the economic impacts for both upload and download together.

Appendix A:

Table 1. Notations and Variable Descriptions

$\text{Session}_{i,t}$	Whether user i started mobile Internet sessions in week t (1 = Yes, 0 = No)
$\text{Social Network Session}_{i,t-1}$	Weighted average number of mobile Internet session initiations by network neighbors of user i at time t-1. That is, $\sum_{m\in n_{t-1}(i)}\left(w_{i,m,t-1} \times \text{Session}_{m,t-1}\right)$
$n_{t-1}(i)$	User i's social network neighbors based on voice call records (i.e., users called by user i) in week t-1
$\text{Session}_{m,t-1}$	Whether user i's social network neighbor m started his/her mobile Internet sessions in week t-1(1 = Yes, 0 = No)
$w_{i,m,t-1}$	Normalized number of calls user i made to user m in week t-1, that is, $w_{i,m,t}$ is a fraction of voice calls from user i to user m in week t-1 with respect to the total voice calls originated from user i in week t-1
Age_i	User i's age
Sex_i	User i's sex (1 = Male, 0 = Female)
Handset Age_i	Months elapsed since user i's handset was launched in the market
$\overline{\text{Social Network Session}_i}$	Time mean mobile Internet session initiation by social network neighbors of user i
δ_i	Unobservable, user-specific, time-invariant effect ($\delta_i \sim \text{IIN}\left(0, \sigma_\delta^2\right)$) where IIN is independent identical normal
λ_t	Time-period fixed effects
$z_{-i,t}$	Mean mobile Internet session initiation of all other users in user i's billing zip code
$\eta_{i,t}$	Unobservable, individual-specific, time-specific effect ($\eta_{i,t} \sim \text{IIN}(0, \sigma_\eta^2)$, σ_η^2 is set to 1 for normalization and δ_i are independent of $\eta_{i,t}$)
θ	Initial conditions parameter
$\text{Upload}_{i,t}$	Number of times user i uploaded content in week t
$\text{Upload}_{m,t-1}$	Number of times user i's network neighbor m uploaded content in week t-1
$\text{Download}_{i,t}$	Number of times user i downloaded content in week t
$\text{Download}_{m,t-1}$	Number of times user i's network neighbor m downloaded content in week t-1
$\text{Location Mobility}_{i,t}$	Any one of following two location mobility metrics - 1) Number of unique zip code-level locations from where user i placed calls in week t and 2) Number of unique province-/state-level locations from where user i placed calls in week t
$\text{Mobility Dispersion}_{i,t}$	Any one of following two mobility dispersion metrics - 1) Fraction of geographical deviation from one's commonly visited places at zip code-level for a user in week t and 2) Fraction of geographical deviation from one's commonly visited places at province-/state-level for a user in week t
$\text{Selection}_{i,t}$	Selection correction term for user i at time t
$g_{-i,t}$, $h_{-i,t}$	Mean uploading and downloading frequencies of all other users in user i's billing zip code area, respectively
β_0, γ_0	Intercepts
κ_i, ψ_i	User-specific dummies
φ_t, τ_t	Time-period dummies
$\eta_{i,t}$, $\nu_{i,t}$, $\varepsilon_{i,t}$	Unobservable, user-specific, time-specific effect, $\eta_{i,t} \sim \text{IIN}(0, \sigma_\eta^2)$, $\nu_{i,t} \sim \text{IIN}(0, \sigma_\upsilon^2)$ and $\varepsilon_{i,t} \sim \text{IIN}(0, \sigma_\varepsilon^2)$

Appendix B: Computing a Selection Term for Content Generation and Usage Equations

The selection correction term is common across both content generation and content usage equations. It is corresponding to the conditional expectations of $\nu_{i,t}$ and $\varepsilon_{i,t}$, respectively given a user mobile Internet session initiation decision, as follows:

$$\mathrm{E}\left(\nu_{i,t}\middle|\mathrm{Session}_{i,t}\right) = \sigma_{\eta\nu} \times \mathrm{Selection}_{i,t} \quad \text{and} \quad \mathrm{E}\left(\varepsilon_{i,t}\middle|\mathrm{Session}_{i,t}\right) = \sigma_{\eta\varepsilon} \times \mathrm{Selection}_{i,t}. \tag{B1}$$

According to Verbeek and Nijman (1996), the selection correction term is computed as follows:

$$\mathrm{Selection}_{i,t} = \mathrm{E}\left\{\delta_i + \eta_{i,t}\middle|\mathrm{Session}_{i,t}\right\} - \frac{\sigma_\delta^2}{1 + T\sigma_\delta^2}\sum_{s=1}^{T} \mathrm{E}\left\{\delta_i + \eta_{i,s}\middle|\mathrm{Session}_{i,s}\right\} \tag{B2}$$

where $\sigma_\eta^2 = 1$ for normalization and T = total number of observable time periods. Further, the conditional expectation $\mathrm{E}\{\delta_i + \eta_{i,t}|\mathrm{Session}_{i,t}\}$ is as follows:

$$\mathrm{E}\left\{\delta_i + \eta_{i,t}\middle|\mathrm{Session}_{i,t}\right\} = \int_{-\infty}^{\infty}\left[\delta_i + \mathrm{E}\left\{\eta_{i,t}\middle|\mathrm{Session}_{i,t}, \delta_i\right\}\right] f\left(\delta_i|\mathrm{Session}_{i,t}\right)d\delta_i \tag{B3}$$

where

$$\mathrm{E}\left\{\eta_{i,t}|\mathrm{Session}_{i,t}, \delta_i\right\} = (2 \times \mathrm{Session}_{i,t} - 1)\frac{\Phi\left(\frac{z'_{i,t}\times\gamma+\delta_i}{\sigma_\delta}\right)}{\Phi\left((2\times\mathrm{Session}_{i,t}-1)\frac{z'_{i,t}\times\gamma+\delta_i}{\sigma_\delta}\right)},$$

$$z'_{i,t} \times \gamma = \alpha_0 + \alpha_1\mathrm{Session}_{i,t-1} + \alpha_2\mathrm{Social\ Network\ Session}_{i,t-1} + \alpha_3\mathrm{Age}_i + \alpha_4{\mathrm{Age}_i}^2 + \alpha_5\mathrm{Sex}_i$$
$$+\alpha_6\overline{\mathrm{Social\ Network\ Session}_{i,t-1}} + \lambda_t + \alpha_7 z_{-i,t},$$

$$\mathrm{Social\ Network\ Session}_{i,t-1} = \textstyle\sum_{m\in n_{t-1}(i)}\left(w_{i,m,t-1} \times \mathrm{Session}_{m,t-1}\right),$$

$$f\left(\delta_i|\mathrm{Session}_{i,t}\right) = \frac{\prod_{s=1}^{T}\Phi\left((2\times\mathrm{Session}_{i,s}-1)\frac{z'_{i,s}\times\gamma+\delta_i}{\sigma_\delta}\right)\frac{1}{\sigma_\delta}\phi\left(\frac{\delta_i}{\sigma_\delta}\right)}{\int_{-\infty}^{\infty}\prod_{s=1}^{T}\Phi\left((2\times\mathrm{Session}_{i,s}-1)\frac{z'_{i,s}\times\gamma+\delta_i}{\sigma_\delta}\right)\frac{1}{\sigma_\delta}\phi\left(\frac{\delta}{\sigma_\delta}\right)d\delta}.$$

We use numerical integration to compute (B3). Specifically, we use the Geweke-Hajivassiliou-Keane (GHK) algorithm (Hajivassiliou et al. 1996).

Appendix C: Selection Equation Results

This following table shows the results from the RE dynamic probit model. In the second week and thereafter (i.e., t ≥ 2), we find the estimate for Session(t-1), is positive (0.4602) and statistically significant, implying a positive state dependence in initiating mobile Internet sessions. The estimate for Session by NN(t-1), is positive (0.0020) and statistically significant, suggesting a positive impact of lagged social network effect. An interesting aspect is that user behavior greatly varies by age, given that the coefficient of Age is positive (0.0960) and statistically significant and the coefficient of Age Square is negative (-0.0023) and statistically significant, implying an inverted U-shaped relationship between age and mobile media usage with a peak around approximately 21 years old. Results also show that male users are more likely to engage in mobile Internet content activities than female users. For the first week of observation window (i.e., t = 1), we observe similar results regarding age and gender as above. In addition, the estimate for Handset Age is negative (-0.0007) and statistically significant. This implies that the oldness of 3G mobile handset is negatively associated with users' propensity to engage in mobile content activities. Further, note that the significant estimate for θ suggests that the exogeneity of initial conditions is strongly rejected (refer to equation 3 for θ). Finally, based on these selection equation estimates, we compute a selection correction term as demonstrated in Appendix B, which is later inserted into content generation and usage equations, respectively.

Equation	Explanatory Variable	Coefficient
Session ($t \geq 2$)	Session (t-1) (1 = Yes, 0 = No)	0.4602 (0.0232)***
	Social Network Session (t-1) (1 = Yes, 0 = No)	0.0020 (0.0008)**
	Age	0.0960 (0.0217)***
	Age Square	-0.0023 (0.0004)***
	Sex (1 = Male, 0 = Female)	0.1442 (0.0379)***
	$\overline{\text{Social Network Session}}$	0.0001 (0.0002)
	Constant	-1.6300 (0.2898)***
Session (t = 1)	Age	0.0548 (0.0219)**
	Age Square	-0.0014 (0.0006)**
	Sex (1 = Male, 0 = Female)	0.1754 (0.0642)***
	Handset Age (Months)	-0.0007 (0.0003)**
	Constant	-0.9163 (0.3206)***
	θ	0.8176 (0.0339)***
	σ_{δ}^{2}	0.2499 (0.0098)***

***Notes*:** $\overline{\text{Social Network Session}}$ is time mean mobile Internet initiation by social network neighbors. Estimates for *time-period* fixed effects and *mean mobile Internet session initiation of all other users in user i's billing zip code* effects are not reported due to brevity. *** denotes significant at 0.01, ** denotes significant at 0.05.

Appendix D: Robustness Check Results

Table D1: GMM-based Dynamic Panel Data Model Results on Content Frequency

Dependent Variable	Explanatory Variables	Coefficient							
		Lagged Social Network	No Social Network	Lagged Social Network	No Social Network	Lagged Social Network	No Social Network	Lagged Social Network	No Social Network
Log Upload Frequency (t)	Log Upload Freq. (t-1)	0.0744 (0.0054)***	0.0743 (0.0054)***	0.0748 (0.0057)***	0.0747 (0.0057)***	0.0735 (0.0051)***	0.0734 (0.0052)***	0.0735 (0.0051)***	0.0735 (0.0051)***
	Log Download Freq. (t-1)	-0.0234 (0.0011)***	-0.0234 (0.0011)***	-0.0234 (0.0011)***	-0.0234 (0.0011)***	-0.0234 (0.0011)***	-0.0234 (0.0011)***	-0.0234 (0.0011)***	-0.0234 (0.0011)***
	Log Local Location Mobility (t)	0.0127 (0.0038)***	0.0128 (0.0038)***						
	Log National Location Mobility (t)			0.0190 (0.0076)***	0.0190 (0.0076)***				
	Log Local Mobility Dispersion (t)					0.0068 (0.0040)*	0.0069 (0.0040)*		
	Log National Mobility Dispersion (t)							0.0037 (0.0019)*	0.0037 (0.0019)*
	Log Number of Voice Calls (t)	0.0040 (0.0019)**	0.0040 (0.0019)**	0.0030 (0.0015)**	0.0030 (0.0015)**	0.0055 (0.0026)**	0.0056 (0.0026)**	0.0045 (0.0016)***	0.0045 (0.0016)***
	Log Upload Freq. by NN (t-1)	0.0300 (0.0098)***		0.0298 (0.0099)***		0.0301 (0.0098)***		0.0302 (0.0098)***	
Log Download Frequency (t)	Log Download Freq. (t-1)	0.1093 (0.0054)***	0.1092 (0.0054)***	0.1109 (0.0055)***	0.1108 (0.0055)***	0.1051 (0.0054)***	0.1051 (0.0054)***	0.1054 (0.0057)***	0.1239 (0.0052)***
	Log Upload Freq. (t-1)	-0.1840 (0.0157)***	-0.1841 (0.0157)***	-0.1842 (0.0156)***	-0.1843 (0.0156)***	-0.1820 (0.0156)***	-0.1821 (0.0156)***	-0.1823 (0.0156)***	-0.1866 (0.0158)***
	Log Local Location Mobility (t)	0.0149 (0.0018)***	0.0149 (0.0018)***						
	Log National Location Mobility (t)			0.0314 (0.0034)***	0.0314 (0.0034)***				
	Log Local Mobility Dispersion (t)					0.0231 (0.0039)***	0.0232 (0.0039)***		
	Log National Mobility Dispersion (t)							0.0112 (0.0030)***	0.0112 (0.0030)***
	Log Number of Voice Calls (t)	0.0175 (0.0063)***	0.0175 (0.0063)***	0.0209 (0.0061)***	0.0209 (0.0061)***	0.0357 (0.0077)***	0.0357 (0.0077)***	0.0271 (0.0085)***	0.0271 (0.0085)***
	Log Download Freq. by NN (t-1)	0.0432 (0.0105)***		0.0438 (0.0107)***		0.0428 (0.0106)***		0.0432 (0.0110)***	

Notes: We included each one of the two *location mobility* metric and the two *mobility dispersion* metrics in our main equations at both the local and national levels. NN refers to network neighbors. Estimates for *time-period* fixed effects and *mean uploading and downloading frequency of all other users in user i's billing zip code* effects are not reported due to brevity. *** denotes significant at 0.01, ** denotes significant at 0.05.

Table D2: Mixed Effects Model Results (with Random Coefficients on Constant Terms) on Content Frequency

Dependent Variable	Explanatory Variables	Coefficient							
		Lagged Social Network	No Social Network	Lagged Social Network	No Social Network	Lagged Social Network	No Social Network	Lagged Social Network	No Social Network
Log Upload Frequency (t)	Log Upload Freq. (t-1)	0.6512 (0.0022)***	0.6513 (0.0022)***	0.6517 (0.0022)***	0.6524 (0.0022)***	0.6527 (0.0022)***	0.6524 (0.0022)***	0.6517 (0.0022)***	0.6524 (0.0022)***
	Log Download Freq. (t-1)	-0.0069 (0.0007)***	-0.0068 (0.0007)***	-0.0068 (0.0007)***	-0.0068 (0.0007)***	-0.0068 (0.0007)***	-0.0067 (0.0007)***	-0.0068 (0.0007)***	-0.0067 (0.0007)***
	Log Local Location Mobility (t)	0.0064 (0.0017)***	0.0066 (0.0017)***						
	Log National Location Mobility (t)			0.0083 (0.0038)**	0.0085 (0.0038)**				
	Log Local Mobility Dispersion (t)					0.0002 (0.0018)	0.0002 (0.0018)		
	Log National Mobility Dispersion (t)							0.0012 (0.0012)	0.0014 (0.0012)
	Log Number of Voice Calls (t)	-0.00011 (0.00015)	-0.00010 (0.00015)	-0.00006 (0.00005)	-0.00005 (0.00005)	-0.00003 (0.00004)	-0.00001 (0.00004)	-0.00006 (0.00005)	-0.00005 (0.00005)
	Log Upload Freq. by NN (t-1)	0.0405 (0.0061)***		0.0408 (0.0061)***		0.0409 (0.0061)***		0.0408 (0.0061)***	
Log Download Frequency (t)	Log Download Freq. (t-1)	0.6671 (0.0021)***	0.6670 (0.0021)***	0.6673 (0.0021)***	0.6672 (0.0021)***	0.6672 (0.0021)***	0.6672 (0.0021)***	0.6673 (0.0021)***	0.6672 (0.0021)***
	Log Upload Freq. (t-1)	-0.0281 (0.0108)***	-0.0283 (0.0108)***	-0.0279 (0.0108)***	-0.0282 (0.0108)***	-0.0277 (0.0108)**	-0.0279 (0.0108)***	-0.0276 (0.0108)***	-0.0278 (0.0108)***
	Log Local Location Mobility (t)	0.0326 (0.0067)***	0.0357 (0.0067)***						
	Log National Location Mobility (t)			0.0439 (0.0144)***	0.0465 (0.0144)***				
	Log Local Mobility Dispersion (t)					0.0092 (0.0069)	0.0098 (0.0069)		
	Log National Mobility Dispersion (t)							0.0104 (0.0045)***	0.0134 (0.0045)***
	Log Number of Voice Calls (t)	0.0009 (0.0002)***	0.0010 (0.0002)***	0.0011 (0.0002)***	0.0013 (0.0002)***	0.0013 (0.0002)***	0.0015 (0.0002)***	0.0010 (0.0002)***	0.0011 (0.0002)***
	Log Download Freq. by NN (t-1)	0.0275 (0.0037)***		0.0283 (0.0037)***		0.0285 (0.0037)***		0.0278 (0.0037)***	

Notes: We included each one of the two *location mobility* metric and the two *mobility dispersion* metrics in our main equations at both the local and national levels. NN refers to network neighbors. Estimates for *time-period* fixed effects and *mean uploading and downloading frequency of all other users in user i's billing zip code* effects are not reported due to brevity. *** denotes significant at 0.01, ** denotes significant at 0.05.

Table D3: Poisson Fixed Effect Model Results

Dependent Variable	Explanatory Variables	Coefficient							
		Lagged Social Network	No Social Network	Lagged Social Network	No Social Network	Lagged Social Network	No Social Network	Lagged Social Network	No Social Network
Log Upload Frequency (t)	Log Upload Freq. (t-1)	0.0011 (0.0001)***	0.0011 (0.0001)***	0.0011 (0.0001)***	0.0011 (0.0001)***	0.0011 (0.0001)***	0.0011 (0.0001)***	0.0010 (0.0001)***	0.0010 (0.0001)***
	Log Download Freq. (t-1)	-3.6e-4 (6.0e-5)***	-3.6e-4 (6.0e-5)***	-3.7e-4 (6.0e-5)***	-3.6e-4 (6.0e-5)***	-3.5e-4 (6.0e-5)***	-3.5e-4 (6.0e-5)***	-3.6e-4 (6.0e-5)***	-3.6e-4 (6.0e-5)***
	Log Local Location Mobility (t)	0.0313 (0.0044)***	0.0313 (0.0044)***						
	Log National Location Mobility (t)			0.2621 (0.0210)***	0.2622 (0.0210)***				
	Log Local Mobility Dispersion (t)					0.0124 (0.0028)***	0.0124 (0.0028)***		
	Log National Mobility Dispersion (t)							0.0678 (0.0091)***	0.0678 (0.0091)***
	Log Number of Voice Calls (t)	0.0073 (0.0006)***	0.0073 (0.0006)***	0.0073 (0.0006)***	0.0073 (0.0006)***	0.0066 (0.0008)***	0.0066 (0.0008)***	0.0075 (0.0006)***	0.0075 (0.0006)***
	Log Upload Freq. by NN (t-1)	0.0052 (0.0015)***		0.0052 (0.0015)***		0.0046 (0.0015)***		0.0047 (0.0015)***	
Log Download Frequency (t)	Log Download Freq. (t-1)	5.4e-6 (7.7e-7)***	5.5e-6 (7.7e-7)***	5.7e-6 (7.7e-7)***	5.8e-6 (7.7e-7)***	5.3e-6 (7.7e-7)***	5.4e-6 (7.7e-7)***	5.2e-6 (7.7e-7)***	5.2e-6 (7.7e-7)***
	Log Upload Freq. (t-1)	-0.0011 (0.0003)***	-0.0011 (0.0003)***	-0.0011 (0.0003)***	-0.0011 (0.0003)***	-0.0012 (0.0003)***	-0.0012 (0.0003)***	-0.0012 (0.0003)***	-0.0012 (0.0003)***
	Log Local Location Mobility (t)	0.0057 (0.0006)***	0.0057 (0.0006)***						
	Log National Location Mobility (t)			0.0432 (0.0020)***	0.0432 (0.0020)***				
	Log Local Mobility Dispersion (t)					0.0007 (0.0002)***	0.0007 (0.0002)***		
	Log National Mobility Dispersion (t)							0.0120 (0.0009)***	0.0120 (0.0009)***
	Log Number of Voice Calls (t)	0.0010 (0.0001)***	0.0010 (0.0001)***	0.0010 (0.0001)***	0.0010 (0.0001)***	0.0011 (0.0001)***	0.0011 (0.0001)***	0.0011 (0.0001)***	0.0011 (0.0001)***
	Log Download Freq. by NN (t-1)	7.0e-5 (2.7e-5)***		7.0e-5 (2.7e-5)***		7.0e-5 (2.7e-5)***		7.0e-5 (2.7e-5)***	

Notes: We included each one of the two *location mobility* metric and the two *mobility dispersion* metrics in our main equations at both the local and national levels. NN refers to network neighbors. Estimates for *time-period* fixed effects and *mean uploading and downloading frequency of all other users in user i's billing zip code* effects are not reported due to brevity. *** denotes significant at 0.01.

Table D4: GMM-based Dynamic Panel Data Model Results on Content Share

Dependent Variable	Explanatory Variables	Coefficient							
		Lagged Social Network	No Social Network	Lagged Social Network	No Social Network	Lagged Social Network	No Social Network	Lagged Social Network	No Social Network
Upload Share (t)	Download Share (t-1)	-0.1190 (0.0120)***	-0.1191 (0.0120)***	-0.1199 (0.0120)***	-0.1197 (0.0120)***	-0.1199 (0.0120)***	-0.1199 (0.0120)***	-0.1191 (0.0120)***	-0.1191 (0.0120)***
	Local Location Mobility (t)	0.0006 (0.0002)***	0.0006 (0.0002)***						
	National Location Mobility (t)			0.0033 (0.0009)***	0.0033 (0.0009)***				
	Local Mobility Dispersion (t)					0.0002 (0.0001)**	0.0002 (0.0001)**		
	National Mobility Dispersion (t)							0.0004 (0.0002)**	0.0004 (0.0002)**
	Number of Voice Calls (t)	8.0e-5 (2.0e-5)***	8.0e-5 (2.0e-5)***	6.0e-5 (2.0e-5)***	6.0e-5 (2.0e-5)***	7.0e-5 (2.0e-5)***	7.0e-5 (2.0e-5)***	4.0e-5 (1.0e-5)***	4.0e-5 (1.0e-5)***
	Upload Share by NN (t-1)	0.0797 (0.0047)***		0.0798 (0.0047)***		0.0801 (0.0047)***		0.0800 (0.0047)***	

Notes: We included each one of the two *location mobility* metric and the two *mobility dispersion* metrics in our main equations at both the local and national levels. NN refers to network neighbors. Estimates for *time-period* fixed effects are not reported due to brevity. *** denotes significant at 0.01.

Table D5: 3SLS Estimation Results on Content Size

Dependent Variable	Explanatory Variable	Coefficient	
		Lagged Social Network	No Social Network
Log Upload Size (t)	Log Download Size (t-1)	-0.0166 (0.0010)***	-0.0166 (0.0010)***
	Log Upload Size by NN (t-1)	0.0027 (0.0039)	
	Selection (t)	0.0159 (0.0023)***	0.0159 (0.0023)***
Log Download Size (t)	Log Upload Size (t-1)	-0.0477 (0.0033)***	-0.0477 (0.0033)***
	Log Download Size by NN (t-1)	0.0034 (0.0072)	
	Selection (t)	0.0534 (0.0041)**	0.0534 (0.0041)**

Notes: We used 13 week sample. NN refers to network neighbors. Estimates for *time-period* fixed effects are not reported only due to brevity. *** denotes significant at 0.01, ** denotes significant at 0.05.

Table D6: 3SLS Estimation Results based on Text Message-based Social Network

Dependent Variable	Explanatory Variables	Coefficient							
		Lagged Social Network	No Social Network	Lagged Social Network	No Social Network	Lagged Social Network	No Social Network	Lagged Social Network	No Social Network
Log Upload Frequency (t)	Log Download Freq. (t-1)	-0.0081 (0.0004)***	-0.0082 (0.0004)***	-0.0082 (0.0004)***	-0.0082 (0.0004)***	-0.0078 (0.0004)***	-0.0078 (0.0004)***	-0.0077 (0.0004)***	-0.0078 (0.0004)***
	Log Local Location Mobility (t)	0.0097 (0.0014)***	0.0098 (0.0014)***						
	Log National Location Mobility (t)			0.0179 (0.0024)***	0.0179 (0.0024)***				
	Log Local Mobility Dispersion (t)					0.0058 (0.0012)***	0.0058 (0.0016)***		
	Log National Mobility Dispersion (t)							0.0058 (0.0016)***	0.0059 (0.0012)***
	Log Number of Voice Calls (t)	0.0009 (0.0007)	0.0010 (0.0007)	0.0012 (0.0007)*	0.0013 (0.0007)*	0.0010 (0.0007)	0.0020 (0.0007)***	0.0009 (0.0007)	0.0011 (0.0007)
	Log Upload Freq. by NN (t-1)	0.1919 (0.0082)***		0.1919 (0.0082)***		0.1919 (0.0082)***		0.1920 (0.0082)***	
	Selection (t)	0.0118 (0.0003)***	0.0118 (0.0003)***	0.0116 (0.0003)	0.0115 (0.0003)***	0.0131 (0.0003)***	0.0128 (0.0003)***	0.0128 (0.0003)***	0.0131 (0.0003)***
Log Download Frequency (t)	Log Upload Freq. (t-1)	-0.0241 (0.0092)***	-0.0241 (0.0092)***	-0.0235 (0.0092)***	-0.0235 (0.0092)***	-0.0221 (0.0092)**	-0.0220 (0.0092)**	-0.0220 (0.0092)**	-0.0220 (0.0092)**
	Log Local Location Mobility (t)	0.0111 (0.0055)**	0.0118 (0.0055)**						
	Log National Location Mobility (t)			0.0062 (0.0095)	0.0073 (0.0095)				
	Log Local Mobility Dispersion (t)					0.0213 (0.0047)***	0.0329 (0.0065)***		
	Log National Mobility Dispersion (t)							0.0328 (0.0065)***	0.0218 (0.0047)***
	Log Number of Voice Calls (t)	0.0386 (0.0027)***	0.0389 (0.0027)***	0.0400 (0.0026)***	0.0404 (0.0026)***	0.0345 (0.0028)***	0.0368 (0.0026)***	0.0364 (0.0026)***	0.0348 (0.0028)***
	Log Download Freq. by NN (t-1)	0.0891 (0.0061)***		0.0891 (0.0061)***		0.0890 (0.0061)***		0.0891 (0.0061)***	
	Selection (t)	0.3578 (0.0013)***	0.3576 (0.0013)***	0.3583 (0.0013)	0.3580 (0.0013)***	0.3607 (0.0013)***	0.3598 (0.0013)***	0.3600 (0.0013)***	0.3606 (0.0013)***

Notes: We included each one of the two *location mobility* metric and the two *mobility dispersion* metrics in our main equations at both the local and national levels. NN refers to network neighbors. Estimates for *time-period* fixed effects and *mean uploading and downloading frequency of all other users in user i's billing zip code* effects are not reported due to brevity. *** denotes significant at 0.01, ** denotes significant at 0.05, * denotes significant 0.1.

Table D7: 3SLS Estimation Results based on Offline Location-based Spatial Network

Dependent Variable	Explanatory Variables	Coefficient							
		Lagged Social Network	No Social Network	Lagged Social Network	No Social Network	Lagged Social Network	No Social Network	Lagged Social Network	No Social Network
Log Upload Frequency (t)	Log Download Freq. (t-1)	-0.0085 (0.0004)***	-0.0085 (0.0004)***	-0.0085 (0.0004)***	-0.0085 (0.0004)***	-0.0080 (0.0004)***	-0.0080 (0.0004)***	-0.0080 (0.0004)***	-0.0080 (0.0004)***
	Log Local Location Mobility (t)	0.0098 (0.0014)***	0.0098 (0.0014)***						
	Log National Location Mobility (t)			0.0182 (0.0024)***	0.0179 (0.0024)***				
	Log Local Mobility Dispersion (t)					0.0059 (0.0012)***	0.0058 (0.0016)***		
	Log National Mobility Dispersion (t)							0.0058 (0.0016)***	0.0059 (0.0012)***
	Log Number of Voice Calls (t)	0.0010 (0.0007)	0.0010 (0.0007)	0.0013 (0.0007)*	0.0013 (0.0007)*	0.0011 (0.0007)	0.0020 (0.0007)***	0.0011 (0.0007)	0.0011 (0.0007)
	Log Upload Freq. by NN (t-1)	0.0082 (0.0014)***		0.0082 (0.0014)***		0.0081 (0.0014)***		0.0081 (0.0014)***	
	Selection (t)	0.0118 (0.0003)***	0.0118 (0.0003)***	0.0115 (0.0003)***	0.0115 (0.0003)***	0.0131 (0.0003)***	0.0128 (0.0003)***	0.0128 (0.0003)***	0.0131 (0.0003)***
Log Download Frequency (t)	Log Upload Freq. (t-1)	-0.0247 (0.0092)***	-0.0246 (0.0092)***	-0.0242 (0.0092)***	-0.0241 (0.0092)***	-0.0223 (0.0092)**	-0.0222 (0.0092)**	-0.0222 (0.0092)**	-0.0223 (0.0092)**
	Log Local Location Mobility (t)	0.0120 (0.0055)**	0.0118 (0.0055)**						
	Log National Location Mobility (t)			0.0078 (0.0095)	0.0073 (0.0095)				
	Log Local Mobility Dispersion (t)					0.0213 (0.0047)***	0.0329 (0.0065)***		
	Log National Mobility Dispersion (t)							0.0328 (0.0065)***	0.0218 (0.0047)***
	Log Number of Voice Calls (t)	0.0390 (0.0027)***	0.0389 (0.0027)***	0.0405 (0.0026)***	0.0404 (0.0026)***	0.0345 (0.0028)***	0.0368 (0.0026)***	0.0364 (0.0026)***	0.0348 (0.0028)***
	Log Download Freq. by NN (t-1)	0.0225 (0.0049)***		0.0225 (0.0049)***		0.0890 (0.0061)***		0.0891 (0.0061)***	
	Selection (t)	0.3575 (0.0013)***	0.3576 (0.0013)***	0.3579 (0.0013)***	0.3580 (0.0013)***	0.3607 (0.0013)***	0.3598 (0.0013)***	0.3600 (0.0013)***	0.3606 (0.0013)***

Notes: We included each one of the two *location mobility* metric and the two *mobility dispersion* metrics in our main equations at both the local and national levels. NN refers to network neighbors. Estimates for *time-period* fixed effects are not reported due to brevity. *** denotes significant at 0.01, ** denotes significant at 0.05, * denotes significant 0.1.

Table D8: 3SLS Estimation Results based on both Voice Call and Offline Location-based Network

Dependent Variable	Explanatory Variables	Coefficient							
		Lagged Social Network	No Social Network	Lagged Social Network	No Social Network	Lagged Social Network	No Social Network	Lagged Social Network	No Social Network
Log Upload Frequency (t)	Log Download Freq. (t-1)	-0.0081 (0.0004)***	-0.0082 (0.0004)***	-0.0081 (0.0004)***	-0.0082 (0.0004)***	-0.0078 (0.0004)***	-0.0080 (0.0004)***	-0.0078 (0.0004)***	-0.0080 (0.0004)***
	Log Local Location Mobility (t)	0.0099 (0.0014)***	0.0098 (0.0014)***						
	Log National Location Mobility (t)			0.0178 (0.0024)***	0.0179 (0.0024)***				
	Log Local Mobility Dispersion (t)					0.0057 (0.0012)***	0.0058 (0.0016)***		
	Log National Mobility Dispersion (t)							0.0053 (0.0016)***	0.0059 (0.0012)***
	Log Number of Voice Calls (t)	0.0006 (0.0007)	0.0010 (0.0007)	0.0009 (0.0007)	0.0013 (0.0007)*	0.0008 (0.0007)	0.0020 (0.0007)***	0.0011 (0.0007)	0.0011 (0.0007)
	Log Upload Freq. by NN (t-1)	0.4640 (0.0106)***		0.4637 (0.0106)***		0.4637 (0.0106)***		0.4637 (0.0106)***	
	Selection (t)	0.0118 (0.0003)***	0.0118 (0.0003)***	0.0116 (0.0003)***	0.0115 (0.0003)***	0.0131 (0.0003)***	0.0128 (0.0003)***	0.0128 (0.0003)***	0.0131 (0.0003)***
Log Download Frequency (t)	Log Upload Freq. (t-1)	-0.0231 (0.0092)***	-0.0238 (0.0092)***	-0.0222 (0.0092)**	-0.0235 (0.0092)***	-0.0209 (0.0092)**	-0.0222 (0.0092)**	-0.0208 (0.0092)**	-0.0223 (0.0092)**
	Log Local Location Mobility (t)	0.0121 (0.0055)**	0.0118 (0.0055)**						
	Log National Location Mobility (t)			0.0069 (0.0095)	0.0073 (0.0095)				
	Log Local Mobility Dispersion (t)					0.0200 (0.0047)***	0.0329 (0.0065)***		
	Log National Mobility Dispersion (t)							0.0303 (0.0065)***	0.0218 (0.0047)***
	Log Number of Voice Calls (t)	0.0363 (0.0027)***	0.0389 (0.0027)***	0.0379 (0.0026)***	0.0404 (0.0026)***	0.0328 (0.0028)***	0.0368 (0.0026)***	0.0346 (0.0026)***	0.0348 (0.0028)***
	Log Download Freq. by NN (t-1)	0.1642 (0.0071)***		0.1642 (0.0071)***		0.1637 (0.0071)***		0.1636 (0.0071)***	
	Selection (t)	0.3578 (0.0013)***	0.3576 (0.0013)***	0.3583 (0.0013)***	0.3580 (0.0013)***	0.3606 (0.0013)***	0.3598 (0.0013)***	0.3599 (0.0013)***	0.3606 (0.0013)***

Notes: We included each one of the two *location mobility* metric and the two *mobility dispersion* metrics in our main equations at both the local and national levels. NN refers to network neighbors. Estimates for *time-period* fixed effects and *mean uploading and downloading frequency of all other users in user i's billing zip code* effects are not reported due to brevity. *** denotes significant at 0.01, ** denotes significant at 0.05, * denotes significant 0.1.