Local Network Effects in the Adoption of a Digital Platform

Jin-Hyuk Kim[*] Peter Newberry[†] Liad Wagman[‡] Ran Wolff[§]

Abstract

We demonstrate the importance of local network effects by estimating the impact of a county-level installed user base on the adoption of a fantasy sports platform. Using the past success of the nearest professional team as an instrument, we find that the size of the local installed user base significantly impacts the number of people who join the platform, and the local network effect is stronger in high-income counties. We demonstrate that a seeding strategy that places moderate weights on high-income counties grows the overall network size faster than either a strategy that heavily weighs high-income counties or one that places more weight on low-income counties.

Keywords: Network effects; installed base; social platform; fantasy sports

JEL codes: L83, M21, O33

[*] Department of Economics, University of Colorado at Boulder, Economics Building, Boulder, CO 80309. Email: jinhyuk.kim@colorado.edu.

[†] Department of Economics, Pennsylvania State University, 403 Kern Building, University Park, PA 16801. Email: pwnewberry@psu.edu.

[‡] Stuart School of Business, Illinois Institute of Technology, 565 W Adams St, Suite 412, Chicago, IL 60661. Email: lwagman@stuart.iit.edu.

[§] Yahoo! Research, Oath, MATAM, Advanced Technology Park, Tower 3 - Floor 7, Haifa, 31905, Israel. Email: ranw@oath.com

1 Introduction

The concept of network externalities predates the digital economy and has been often used to explain the adoption of innovative products and technologies. For instance, Klemperer (2008) says that "direct network effects arise if each user's payoff from the adoption of a good, and his incentive to adopt it, increase as more others adopt it; that is, if adoption by different users is complementary." These sorts of consumption externalities were intuitively applied to explain the adoption of network-based products such as telephones and fax machines (e.g., Katz and Shapiro, 1985).

Our inquiry begins from the observation that network effects may be limited by social and/or geographic boundaries that determine the value of a network. For example, while it may be that a telephone is useful for connecting people who are even randomly distributed across wide areas, it seems to be plausible that the use of a network-based good is most intense within a user's immediate social and/or geographic circle. In this paper, we detect these 'local' network effects and study the role of income in local network growth using an online fantasy sports platform as a case study.[1]

An online social platform provides an interesting testing ground for network effects. In the digital economy, social platforms purport to provide effective means for maintaining relationships, independent of physical boundaries. This suggests that a platform's global network (or total subscribers) is often considered to be a key factor in facilitating network growth. However, it may be the case that users on the platform value a network closer to them in a geographic or social proximity sense, as there may be benefits of possible offline interactions and lower coordination costs.

We construct a data set that includes the number of individuals who played Yahoo Fantasy Baseball in each US county during the 2017 and the 2018 seasons. Using these data, we estimate the effect of the 2017 county-level user base on the 2018 county-level

[1] Sports channels as well as pro leagues themselves regularly cover fantasy sports. According to the Fantasy Sports Trade Association, fantasy sports are estimated to be worth more than $7 billion a year in the US and Canada, with 59 million participants.

adoption rates while controlling for county characteristics and other factors that may impact adoption. We instrument for the 2017 network size using the closest professional baseball team's performance in the 2016 baseball season and its interaction with the distance from the county's centroid to the team's stadium.

The relevance of the instruments come from the fact that the local team's season performance is positively correlated with fantasy baseball adoption in the next season through an increased interest in the sport.[2] We argue that the instruments are also excluded because the nearest team's performance two seasons ago is not likely to be correlated with the unobserved error term in this season's adoption equation, since any such effect would already be absorbed by the last season's installed base.

We find that the local network size has a statistically and economically significant effect on fantasy baseball adoption rates. We believe that this result can be interpreted as coming from consumption externalities rather than from social learning, contrasting Moretti (2011)'s analysis of box-office sales as a function of the movie's opening weekend performance. This is because online fantasy platforms have been around for almost two decades, meaning that most sports fans are not necessarily learning about fantasy sports from their peers.

Next, we include interaction terms between the user base and county characteristics into our regressions and find that the network effect is heterogeneous with respect to the county's income level, but not with respect to other characteristics. This finding is related to Bonus (1973), who argued that income is an important factor in the diffusion of consumer durables because it determines actual versus potential ownership. Bonus predicts that more diffusion would occur in low-income areas, as the network effect would lower the initially-binding income threshold for purchase. In contrast, we find that the network effect is higher in high income areas. This is likely because the only cost of adoption in fantasy sports is time, and time is more expensive for higher-income households.

[2] Users can only sign up for a fantasy league about a month before the new season starts. Hence, at the time of making their decisions, the performance from the last season is a new piece of information that enters an individual's adoption decision.

In order to demonstrate the importance of the local network effect more clearly, we perform a robustness check that estimates the effect of the network size of surrounding counties. We find that the own-county network size continues to be significant while the network size of nearby counties is not. There may be several reasons for why adopters may value their local network, but we think that the primary mechanism is the benefits from playing the game with a larger local network of individuals, such as friends and acquaintances, who communicate more often. For instance, users in the same area may share similar sports interests and can have offline interactions.[3]

Using the estimates of our model, we demonstrate the optimal strategy for a platform operator who seeks to grow their total network size as quickly as possible, taking into account the heterogeneous network effects. Specifically, we simulate the product adoption curve under different 'seeding' strategies, where a seeding strategy determines the initial distribution of a fixed number of users across counties.

We highlight two results from the simulations. First, we are able to generate various kinds of S-shaped product adoption curves consistent with Cabral (1990), who theoretically argued that an S-shaped adoption curve can be generated by a dynamic model of adoption with network effects. Our setting has specific advantages for studying this issue, as fantasy sports is free to play and the platform also faces negligible marginal costs. In particular, the falling price of a product and/or its marginal cost over time was one of the complicating factors in prior studies of adoption curves, following the pioneering works of Griliches (1957) and Mansfield (1961).

Second, given our heterogeneous network effects, we find that the platform can grow its total network size faster by moderately focusing its seeding on high-income counties. However, placing too much weight on high-income counties can be worse than seeding more evenly across the counties. The reason is that although user growth is faster in high-income coun-

[3] We cannot distinguish whether it is a pure benefit, peer pressure or conformity. Young (2009) attempts to distinguish among such different mechanisms, which may lead to different aggregate diffusion paths. However, our data do not allow us to speak to possible mechanisms.

ties, over-seeding those counties entails slow growth elsewhere. Hence, there is a U-shaped relationship between the seeding strategy (i.e., the relative weights assigned to different counties) and the length of time it takes to achieve a certain network penetration.

This paper contributes to the literature on direct network effects. Despite the intuitively plausible hypothesis that many products, especially communication and information devices, are thought to have network effects, direct evidence on consumption externalities has not been easy to find. Most of the existing evidence on network effects comes indirectly through firms' adoption decisions: for example, computer platform/software, automated machining tools, automated teller machines, and automated clearinghouse systems (e.g., Greenstein, 1993; Karshenas and Stoneman, 1993; Gandal, 1994; Saloner and Shepard, 1995; Gowrisankaran and Stavins, 2004; Ackerberg and Gowrisankaran, 2006).[4]

More recently, Ryan and Tucker (2012) study the adoption of a video-calling technology at the individual level. The main difference between their work and the current study is that our focus is on the effects of the installed user base across geography while they focus on the adoption pattern across employee groups. Björkegren (2019) examines comprehensive transaction data of a nation's mobile phone subscribers over multiple years and finds that a requirement to serve rural areas may lower operator profits. He leverages a global network of calls made, where the variation in the cellular coverage across locations affects a user's marginal cost, but the local-ness of the call network is not examined.

A local property of network effects is documented by Goolsbee and Klenow (2002). They show that a household's adoption of a home computer in 1997 is an increasing function of the fraction of households in the city that had a computer in the previous year.[5] Like Goolsbee and Klenow, we cannot distinguish between consumption externalities and knowledge spillovers, but as previously mentioned, we think that knowledge spillovers are not likely in

[4] Another type of network effects, named indirect network effect, comes from two-sided markets where the amount and variety of goods supplied on a platform is increasing in the number of participating consumers: for example, yellow pages and console games (e.g., Rysman, 2004; Lee, 2013).

[5] One difference is that the network size enters directly into the user's utility function in our model, to be more in line with the definition of a direct network effect, whereas Goolsbee and Klenow use the percentage of households owning a home computer as an endogenous regressor.

our setting because fantasy sports have been around for a couple of decades.

The localization of technology diffusion has also been discovered elsewhere. For instance, Keller (2002) finds that spillovers from R&D expenditures are to a substantial degree local, not global, contributing to the persistently different levels of economic outputs across regions. Another example is Jaffe et al. (1993), who find that US patent citations are geographically localized. Our paper shows a parallel result between knowledge spillovers and consumption externalities for a platform product in a borderless online environment.

Our findings also complement the nascent literature examining the relationship between local network effects and platform competition and strategy. For example, Fjeldstad et al. (2010) and Xu (2014) show that a single platform is less likely to dominate the market when consumers value a local network more than a global one, and that the market structure as well as allocative efficiency might be different from the standard (global) platform competition model. This paper renders some suitable empirical support to these findings.

The remainder of the paper proceeds as follows. Section 2 presents the model and empirical specification. Section 3 describes the dataset, and Section 4 contains the estimation results. Section 5 examines implications for seeding strategies, and Section 6 concludes.

2 Model

Consider a platform that offers a service to users located in market c. User i's deterministic utility of adopting the service in period t, denoted by $v_i(m_{ct-1}, x_{ic}, z_{ct}, \xi_{ct})$, is a function of the installed user base (i.e., network size) up to period $t-1$, m_{ct-1}, the user's demographic and socioeconomic characteristics, x_{ic}, market-level demand shifters in period t, z_{ct}, and an unobserved market-level demand shock in period t, ξ_{ct}. In addition, user i observes a random utility shock in each period, denoted by ε_{ict}, which we assume is an i.i.d. random variable that follows the type I extreme value distribution. This shock represents any unobserved stochastic variation in the adoption utility at the user level.

At the beginning of period t, each user who has not yet adopted the service decides whether or not to adopt. We assume for simplicity that once a user adopts, he or she will not drop out. Incorporating an exogenous rate of dropout would not change the qualitative nature of our findings because the dropout rate in the data is very small. In addition, we assume that consumers do not internalize their impact on the network, in the sense that each consumer is sufficiently 'small' not to consider the influence of their adoption decision on future network growth. Therefore, the decision to adopt is a one-time choice based on the current market state, which can be thought of as reduced form for a dynamic process where the consumer takes an expectation of the future state based on the current state.

We specify the utility of adoption as a separable function of the deterministic utility and the random shock, $u_{i1ct} = v_i(m_{ct-1}, x_{ic}, z_{ct}, \xi_{ct}) + \varepsilon_{ict}$, and we normalize the utility of non-adoption (i.e., the outside option) so that $u_{i0ct} = \varepsilon_{i0ct}$. The probability of user i adopting the service in period t is then

$$P_{i1ct} = \frac{exp(v_i(m_{ct-1}, x_{ic}, z_{ct}, \xi_{ct}))}{1 + exp(v_i(m_{ct-1}, x_{ic}, z_{ct}, \xi_{ct}))}, \tag{1}$$

and the number of adopters in period t is

$$q_{1ct} = M_{ct} \left(\int_i P_{i1ct} dF_c(x_{ic}) \right), \tag{2}$$

where M_{ct} is the number of people in market c who have not yet adopted and $F_c(x_{ic})$ is the distribution of user characteristics in market c. The total number of users as of period t across markets can then be written as $Q_t = \sum_{\tau=0}^{t} \sum_c q_{1c\tau}$.

We assume that the platform's per-period profit is an increasing function of the total number of users in period t, which can be influenced by a seeding strategy implemented in period 0. One can think of the seeding strategy as a promotion that creates an initial distribution of demand across markets according to a distribution of weights $\mathbf{w} = \{w_1, \ldots, w_c, \ldots, w_C\}$. For instance, the platform may implement an advertising campaign with a fixed budget in order to encourage adoption, and the seeding strategy would define how to allocate the advertising budget to each market c. We can thus denote the per-period profit as $\pi(Q_t(\mathbf{w}))$ for some arbitrary function π.

The discounted sum of profits in period 0 is $\sum_{t=0}^{\infty} \beta^t \pi(Q_t(\mathbf{w}))$, where $\beta \in (0,1)$ is a discount factor. The platform chooses $\mathbf{w}$ in period 0 to maximize its profit. Because future profits are discounted, for a given level of eventual market saturation, the optimal seeding strategy is one that increases the aggregate network size the fastest.

In order to estimate the demand model, we make the following assumptions. First, the deterministic utility function is linear in its components: $v_i(m_{ct-1}, x_{ic}, z_{ct}, \xi_{ct}) = \alpha_c m_{ct-1} + x'_{ic}\beta + z'_{ct}\gamma + \xi_{ct}$. Second, user i takes on the average characteristics in the market (i.e., $x'_{ic} = x'_c \quad \forall i \in c$). Third, preferences are the same across markets, with the exception of the network effect parameter, α_c, which can be heterogeneous across markets c. Under these restrictions, the log share of adoption, s_{1ct}, less that of non-adoption, s_{0ct}, takes the well-known form (Berry, 1994):

$$\ln(s_{1ct}) - \ln(s_{0ct}) = \alpha_c m_{ct-1} + x'_c\beta + z'_{ct}\gamma + \xi_{ct}. \tag{3}$$

If ξ_{ct} is i.i.d. across markets and periods, then equation (3) can be estimated using OLS. However, this assumption may fail to hold due to the persistence in ξ_{ct}. For instance, suppose there are markets where the service is popular for some unobservable reason, then α_c cannot be consistently estimated because the network size is a function of the unobservable popularity. To address this issue, we instrument for m_{ct-1} using market demand shifters in period $t-1$, z_{ct-1}. The identifying assumption is that the lagged demand shifters are independent of the market-level demand shock in period t, ξ_{ct}. Under this assumption, we can consistently estimate the model via two-stage least-squares.

3 Data

We estimate our model using proprietary data from Yahoo Fantasy Baseball (YFB). Specifically, we extract information on all YFB users who signed up for a 'traditional' fantasy league during two baseball seasons (2017 and 2018), where a traditional league means a season-long league rather than a 'daily' fantasy league.[6] In a traditional league, a group of users establish a fictional league by each creating a team of professional players. Teams then compete against one another, where the winner is based on the chosen players' real-life performance statistics. The competition takes place over the length of the Major League Baseball season, from early April until late September.

To protect the business interests of Yahoo (now Oath, which is part of Verizon Wireless), we cannot state the exact number of users, but in our data there are about one million users who have participated in YFB in the 2017 season and the number increased by about 5% for the 2018 season. We refer to the 2017 season participants as the installed base and the new participants in the 2018 season as the adopters. To be precise, we define an adopter as a user who signed up by the end of the first week of the 2018 baseball season (March 29 to April 8, 2018), so that the installed base is clearly defined as the 2017 user base. A small fraction of users signed up after the first game week, which we do not include in our analysis.

Obeying a strict anonymization protocol, we are able to match a large (but undisclosed) fraction of the users to a location (ZIP code) by reverse IP address lookup (i.e., involving information associated with cookies) during the first game week of the baseball season. In the event that there are multiple IP addresses associated with a user during the 11-day period, we choose the ZIP code that is most frequently associated with that user.[7] We define the relevant market as a county or a county equivalent in our empirical analysis; hence, we use the US Census Bureau's relationship files to map each user's ZIP code to the county to which

[6] Daily fantasy sports are paid competitions funded by entry fees, and were ruled as illegal gambling, and hence banned, in some US states. Traditional (season-long) fantasy leagues are legal in all US states.

[7] We dropped the IP addresses associated with the city of Sunnyvale, CA, where Yahoo is headquartered. This is not a big problem for our purposes because our market is at the US county level and there are many other IP addresses located in Santa Clara County.

the largest ZIP population belongs.[8]

We create our primary variable of interest, the network size, by determining the installed user base of each county in the 2017 season. We drop counties in Alaska and Hawaii due to their remote locations; and we fold some of the cities in Virginia into counties following the Bureau of Economic Analysis (BEA) reporting convention.[9] To allow for a diminishing network effect, we take the log of the installed user base and denote this as m_{jt-1}. Similarly, in order to determine the number of adopters, we count the number of people in a county who did not register for YFB in 2017 but did so in 2018. The share of adopters is then calculated by dividing the adopters by the potential market size of fantasy baseball players.

The potential market size is based on a 2018 Gallup poll that states that 10% of male and 8% of female adults, aged 18 and older, chose baseball as their favorite sport to watch. Therefore, we define the upper bound or total market size to be 10% of the county male population plus 8% of the county female population, aged 18 to 74 years, and the potential market size to be the total market size net of the installed base in period $t-1$. Accordingly, the adoption share in each county, s_{1ct}, is the number of adopters divided by the potential market size (non-adopters), and given that adoption is a binary decision, the non-adoption share is simply $s_{0ct} = 1 - s_{1ct}$.

In addition to the data from Yahoo, we collect demographic and socioeconomic variables in x_c from the 2010 Decennial Census and American Community Survey (5-year estimate). Specifically, this includes data on population size, the percentage of the population that is nonwhite, the percentage of population that lives in an urban area, the percentage of population that graduated college, median income, and the Gini index for each county.[10]

Included in z_{ct} are the performance of the 'local' professional baseball team at the end of the last season and an interaction between the team's performance and the distance between

[8] We use the Census files to associate each ZIP to a ZIP Code Tabulation Area and to a county.

[9] This is because Virginia is divided into 95 counties and 38 independent cities that are considered county-equivalents for Census purposes. For statistical purposes, the BEA combines some of the independent cities with the county to which they once belonged, which also makes sense for our study.

[10] Using high school graduation and mean income rather than college graduation and median income does not significantly alter the results.

the stadium of the pro team and the centroid of county c. The idea behind including the pro team's performance is that the performance is hard to predict ex ante, so a good performance can spark local interest in the sport, leading more people to join fantasy leagues.

Additionally, we allow the effect of the team's performance to be a function of the county's distance to the stadium, as individuals may be less excited or influenced by the team's performance if the team is located farther away from where they live or work. Playing fantasy sports can be a substitute for going to the stadium to watch a game. This distance also provides a variation in z_{ct} across counties having the same local baseball team.

In order to determine the Major League Baseball (MLB) team that is 'local' to each county and to create the distance measure, we manually collect geo-coordinates of the MLB stadiums from mapdevelopers.com.[11] We then calculate the distance between each county centroid and each stadium using the Haversine formula and match the counties with the team having the closest stadium. The measure of team performance we use is the local teams' winning percentage from the previous season, which are collected from mlb.com. The winning percentage is defined as the number of wins divided by the total number of games played (162 for each team).

To address endogeneity, we instrument for the installed base (m_{ct-1}) using the variables in z_{ct-1}, namely, the local team's performance in period $t-2$ and the interaction between the performance and the distance. Our identifying assumption is that the adoption decision in period $t-1$ is a function of the team's performance in $t-2$, meaning the network size is correlated with z_{ct-1}, but that the team's performance in $t-2$ does not directly impact adoption decisions in period t. Notice that we include the team performance in $t-1$ in the main adoption equation, which may be correlated with the performance in $t-2$, but this does not pose a problem for the IV specification.[12]

Table 1 shows the descriptive statistics of our dataset comprising 3080 counties and

[11] We exclude the Toronto Blue Jays because our data are restricted to the US population.

[12] Nonetheless, pro teams' winning percentages are not significantly auto-correlated. For instance, we estimated an autoregressive dynamic panel model of team performance over several years, and the lagged effect is often insignificant.

county equivalents as described above. While user penetration cannot be disclosed, there is a considerable variation in the geographic distribution of both adopters and installed bases, despite it having been 18 years since Yahoo launched the Fantasy Sports service in 1999. For instance, the 75th percentile county has about six times as many fantasy baseball users by 2017 as the 25th percentile county. Similarly, the 75th percentile county has about ten times as many adopters in 2018 as the 25th percentile county.

4 Estimation Results

Table 2 contains the baseline model estimation of equation (3), in which the network effect is the same across markets (i.e., $\alpha_c = \alpha \;\; \forall c$). Under this specification, the effect of network size does not vary across markets, so the network growth, or adoption, is purely driven by the local network size. The first column shows the OLS estimates, where the network effect coefficient is statistically significant and around 0.72. In terms of an elasticity coefficient, a one-percent increase in the local network size is associated with a 0.72 percent increase in the odds of adoption in that area. This suggests that the local network size is an important determinant of the local network growth.

However, the OLS result can be biased because the installed user base is likely to be endogenous. Hence, in the second and the third columns of Table 2, we present the first and the second stage of the two-stage least-squares estimates, respectively, utilizing the instruments we discussed above. The second column shows that both instruments are statistically significant and the first-stage F statistics is well above 10. Specifically, the local team's 2016 season performance is positively correlated with the installed user base in 2017 and the association becomes weaker as the distance between the county centroid and the local team's stadium gets longer.

The third column tells us that the network effect coefficient is statistically significant and around 0.97. One would think that normally the network effect is biased upwards under the OLS specification, but here we find that it is biased downwards. The reason for this is that the installed user base is not only a function of our instruments, but also the local demographics of the county, as indicated by the first-stage results. Without accounting for this relationship, the network effect could be underestimated in the OLS specification as the demographics absorb some of the true network effect.

The other coefficients suggest that more populous counties tend to have a slower adoption rate, perhaps due to more amenities and/or opportunities for social interaction. Counties that are more remotely located from an MLB team have a lower adoption rate, and this effect is offset by the team's previous season performance. A county's socio-demographic variables, however, do not have a significant standalone effect on adoption.

Turning to Table 3, we allow the network effect to be heterogeneous across counties, meaning the network growth is driven by both the local network size and the local network effect. There exist some theoretical reasons for why the adoption may be particularly related to income (Bonus, 1973). In our case, fantasy sports are free to play, but high-income earners tend to have a higher time (opportunity) cost, so network effects may entice high-income earners on the margin to adopt. Further, the notion of homophily suggests that similarly situated individuals tend to share common activities; hence, network effects may also interact with a county's demographic characteristics.

The first column of Table 3 contains the two-stage least-squares estimates of an interactive model, where the network effect depends on the log median income of the county: $\alpha_c = \alpha_0 + \alpha_1 \times income_c$. It shows that the income-driven network effect, α_1, is statistically significant and positive while the base network effect, α_0, is not statistically different from zero. Thus, in our context, the local network effect is stronger in high-income counties than in low-income counties. We also estimated the same interactive model with the other demographic characteristics (e.g., percent urban, nonwhite, and college graduates), but none of those interactions were significant.

Hence, network growth in a local market is driven by both the network size and the income of the local population. To check the robustness of this finding, we add another interactive term in each of the remaining columns of Table 3. Specifically, there are now two interactive network effects, one with income and another with a chosen county characteristics, and six instruments (i.e., the two original instruments, and each of them interacted with income and another variable).[13] We find that the income interaction effect, α_1, is still statistically significant, though with varying magnitudes, while none of the other interactive effects are statistically significant.

In order to confirm the presence of the local network effect, we perform a robustness check in which we include network size variables in the areas surrounding a given county. To do so, we aggregate the installed user bases across counties whose centroids fall within a certain radius (100, 200, 300, and 400 miles) from that of a given county. This can be thought of as creating spatially lagged variables using a symmetric weight matrix W, whose diagonal elements are zero and off-diagonal elements are one if county i and county j are located within a certain radius and zero otherwise.[14]

Table 4 presents the two-stage least-squares estimates including the installed base in the surrounding areas of gradually varying radii. In all columns of Table 4, the installed base in the nearby counties does not affect the adoption rate in a statistically significant manner, although the magnitude of the own-county network effect tends to decrease somewhat as we broaden the boundaries of the surrounding areas. We also estimated the same model by using a weight matrix that introduces a discount factor, rather than a binary variable, that is an inverse function of the distance between a given pair of counties, and found qualitatively similar insignificant results.

[13] If we include three or more interactive effects at the same time (meaning eight or more instruments), then the first-stage F statistics falls below 10.

[14] Specifically, the installed base in the surrounding areas can be created by Wm_{t-1}, where m_{t-1} is a vector of the network sizes in each county, so this measure does not include the network size of the focal county. Instruments for Wm_{t-1} are created by pre-multiplying W with the two original instruments.

5 Seeding Strategy

We now turn to the seeding problem that the platform faces, where a seeding strategy is represented by a distribution of a fixed number of initial adopters across counties. This can be thought of as an allocation of a fixed marketing budget across counties at the launch of the platform's service. The above results imply that different distributions will affect the overall growth of the network differently, so that the platform can affect network growth through its choice of a seeding strategy. We shed light on optimal seeding strategies by using our model estimates to simulate adoption curves over time, holding constant all other factors of the model. However, note that since the local pro team's winning percentages do vary over time and are history dependent, we set all teams' winning percentages at .5 in all periods for our simulation, so the adoption curves are drawn from an ex-ante standpoint.

In order to perform this dynamic simulation using the the estimates of our model, we make a number of assumptions. First, adoption is a one-time decision that is an absorbing state. This is justified by the data as the dropout rate is very small. Second, the preference parameters for adoption do not change over time. Third, the consumer's adoption decision depends only on the state of the market today. The second and third assumptions are made because we do not have panel data in order to estimate parameters of a dynamic model. To the extent that these assumptions are violated in our setting, we believe that any resulting error will not systematically impact the qualitative comparison of adoption dynamics across different seeding strategies. That is, we do not have reason to believe that violations of these assumptions will impact one seeding strategy differently from the others.

We simulate the adoption curve using the specification in the first column of Table 3, the heterogeneous network effect model, in which a county's market penetration will depend both on the effect size and on the network size. To be more precise, the seeding strategy we analyze is one in which the platform places the initial number of adopters in a county based on the county's median income. The specific class of strategies comprises single-parameter weighing functions that determine a county's share of a given initial set of adopters as follows:

$$w_c(\theta) = \frac{\exp(\theta \times \overline{income}_c)}{\sum_c \exp(\theta \times \overline{income}_c)} \tag{4}$$

where $\overline{income}_c$ is a county's de-meaned log median income (i.e., centered at zero).

Suppose the platform allocates M users across counties based on this weight. Then each county's initial allocation of adopters is $n_c(\theta) = w_c(\theta)M$. Hence, the choice of θ means different allocations of users across counties with different median incomes. For instance, if $\theta = 0$, then all counties have an equal weight and receive the same number of initial adopters. If $\theta > 0$, then high-income counties have more weight than low-income counties, while the reverse holds true if $\theta < 0$. We set M equal to the number of counties as a normalization, so that an equal allocation means one adopter per county. The higher the value of θ, the more skewed the initial distribution is toward counties with higher income.[15]

With an initial number of adopters M and a seeding strategy θ, we can predict the adoption share as well as the number of adopters for each county in period 1 using the estimates of the model. Specifically, we predict the adoption share by substituting the value of $n_c(\theta)$ into our model and then deriving the number of adopters as the adoption share times the potential market size.[16] We then update the size of installed user base and move to period 2, where we predict the adoption share and the number of adopters. We iterate this process accumulating the number of adopters and updating the size of the potential market of adopters. Assuming that the platform maximizes the growth of the total network size Q_t, we aggregate the simulated installed base across counties at each iteration and present the results in Figure 1(a) for a range of θ values.

Figure 1(a) shows the adoption curves for a select number of θ values. They are all

[15] To give a sense of the skewness of the distribution, when $\theta = 10$, the highest 90th percentile county receives 0.439 seed users while the 10th percentile county receives 0.001 seed users. When $\theta = -10$, the 90th percentile county receives 0.003 seed users while the 10th percentile county receives 1.08 seed users.

[16] Specifically, we can predict the adoption share as $\exp(\hat{y} + mse/2)/(1 + \exp(\hat{y} + mse/2))$, where $\hat{y}$ is the predicted value of the left-hand side of equation (3) and mse is the mean squared error. The goodness-of-fit of our model is quite good: The predicted and actual numbers of adopters have a correlation of 0.986.

variants of an S-shaped curve because the network effect kicks in as the size of the installed user base grows and then it tapers off in later iterations as the pool of potential adopters becomes small. It is worth noting that the total-user maximizing seeding parameter θ is not unique across periods. Nonetheless, a parameter value around $\theta = 3$ dominates other seeding strategies for most of the periods. The figure also shows that the growth of the aggregate installed base can dramatically differ based on the geographic distribution of seed users across counties. For instance, the total user base can be as much as 100% larger by following the optimal seeding strategy ($\theta = 3$) relative to, say, $\theta = 15$ or $\theta = -9$, when the network size starts to grow exponentially.

Figure 1(a) also shows that when θ is negative, the overall network grows relatively fast at the beginning and then growth begins to slow down; in contrast, when θ is positive, the network grows relatively slowly early on and then growth speeds up at a higher rate. The reason is that putting more weight on low-income counties compensates for their weaker network effects; however, doing so comes at the cost of dampening some of the network effects in high-income counties. At the same time, putting too much weight on high-income counties can limit the overall network growth because there are too few users elsewhere, even though the installed base in the high-income counties grows relatively quickly.

Figure 1(b) presents the number of iterations it takes to reach a given level of market penetration (e.g., the 75% line refers to the total network size surpassing 75% of the total market size). Indeed, choosing a seeding strategy around $\theta = 3$ would allow the platform to reach a given market penetration in the shortest number of iterations. Further, the curves suggest that the growth of the installed base is relatively sensitive to the seeding strategy, or the geographic distribution of users. Thus, the platform can try to steer the adoption curve in the optimal direction by fine-tuning its advertising strategy across areas based on such characteristics as income.

6 Conclusion

Using unique geo-coded user adoption data, we documented evidence of a local network effect in a digital environment. The estimated model generates a typical S-shaped adoption curve, in which growth or diffusion depends both on the local network effect and on the local network size. Our analysis sheds light on platform strategies, particularly with respect to local income levels. That is, the simulation exercise presented highlights the importance of considering optimal seeding strategies for overall network growth when network effects are local and heterogeneous.

References

[1] Ackerberg, D., and G. Gowrisankaran (2006). "Quantifying Equilibrium Network Externalities in the ACH Banking Industry." *RAND Journal of Economics* 37: 738–761.

[2] Berry, S. (1994). "Estimating Discrete-Choice Models of Product Differentiation." *RAND Journal of Economics* 25: 242–262.

[3] Björkegren, D. (2019). "The Adoption of Network Goods: Evidence from the Spread of Mobile Phones in Rwanda." *Review of Economic Studies* 86: 1033–1060.

[4] Bonus, H. (1973). "Quasi-Engel Curves, Diffusion, and the Ownership of Major Consumer Durables." *Journal of Political Economy* 81: 655–677.

[5] Cabral, L. (1990). "On the Adoption of Innovations with 'Network' Externalities." *Mathematical Social Sciences* 19: 299–308.

[6] Fjeldstad, Ø., E. Moen, and C. Riis (2010). "Competition with Local Network Externalities." Mimeo.

[7] Gandal, N. (1994). "Hedonic Price Indexes for Spreadsheets and an Empirical Test for Network Externalities." *RAND Journal of Economics* 25: 160–170.

[8] Goolsbee, A., and P. Klenow (2002). “Evidence on Learning and Network Externalities in the Diffusion of Home Computers.” *Journal of Law and Economics* 45: 317–343.

[9] Gowrisankaran, G., and J. Stavins (2004). “Network Externalities and Technology Adoption: Lessons from Electronic Payments.” *RAND Journal of Economics* 35: 260–276.

[10] Greenstein, S. (1993). “Did Installed Base Give an Incumbent any (Measureable) Advantages in Federal Computer Procurement?” *RAND Journal of Economics* 24: 19–39.

[11] Griliches, Z. (1957). “Hybrid Corn: An Exploration in the Economics of Technological Change.” *Econometrica* 25: 501–522.

[12] Jaffe, A., M. Trajtenberg, and R. Henderson (1993). “Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations.” *Quarterly Journal of Economics* 108: 577–598.

[13] Karshenas, M., and P. Stoneman (1993). “Rank, Stock, Order, and Epidemic Effects in the Diffusion of New Process Technologies: An Empirical Model.” *RAND Journal of Economics* 24: 503–528.

[14] Katz, M., and C. Shapiro (1985). “Network Externalities, Competition, and Compatibility.” *American Economic Review* 75: 424–440.

[15] Keller, W. (2002). “Geographic Localization of International Technology Diffusion.” *American Economic Review* 92: 120–142.

[16] Klemperer, P. (2008). “Network Goods (Theory).” In: S. Durlauf and L. Blume (eds) *The New Palgrave Dictionary of Economics.* Palgrave Macmillan: London.

[17] Lee, R. (2013). “Vertical Integration and Exclusivity in Platform and Two-Sided Markets.” *American Economic Review* 103: 2960–3000.

[18] Mansfield, E. (1961). "Technical Change and the Rate of Imitation." *Econometrica* 29: 741–766.

[19] Moretti, E. (2011). "Social Learning and Peer Effects in Consumption: Evidence from Movie Sales." *Review of Economic Studies* 78: 356–393.

[20] Ryan, S., and C. Tucker (2012). "Heterogeneity and the Dynamics of Technology Adoption." *Quantitative Marketing and Economics* 10: 63–109.

[21] Rysman, M. (2004). "Competition between Networks: A Study of the Market for Yellow Pages." *Review of Economic Studies* 71: 483–512.

[22] Saloner, G., and A. Shepard (1995). "Adoption of Technologies with Network Effects: An Empirical Examination of the Adoption of Automated Teller Machines." *RAND Journal of Economics* 26: 479–501.

[23] Xu, L. (2014). "Platform Competition with Local Network Effects." Mimeo.

[24] Young, H. (2009). "Innovation Diffusion in Heterogeneous Populations: Contagion, Social Influence, and Social Learning." *American Economic Review* 99: 1899–1924.

Table 1: Descriptive Summary Statistics

Variable	Mean	Std. dev.	Min	Max	N
No. of adopters	(redacted)	256.6	(redacted)	(redacted)	3080
Installed base	(redacted)	733.5	(redacted)	(redacted)	3080
% Urban	.4107	.3111	0	1	3080
% Nonwhite	.1665	.1628	.008	.971	3080
% College	18.93	8.565	3.7	70.1	3080
Median income	44062	11282	19351	115574	3080
Gini index	.4316	.0363	.207	.645	3080
County pop.	99570	315522	82	9818605	3080
Dist. to stadium	278.5	168.2	2.101	1006	3080
2017 team % wins	.5010	.0600	.395	.642	3080
2016 team % wins	.4835	.0685	.364	.640	3080

The unit of observation is the county or county-equivalent entity spanning the contiguous United States (i.e., the 48 adjoining states plus Washington DC), where some of the cities in Virginia are merged with counties following the Bureau of Economic Analysis. Adopters are the number of users who participated in Yahoo Fantasy Baseball in the 2018 season and not in the previous seasons; and installed base is the number of users who participated in the 2017 YFB season. County characteristics are from the 2010 Decennial Census and American Community Survey (5 year estimate). Distance between county centroid and the nearest Major League Baseball stadium is calculated by using the Haversine formula. Note that data in the first two rows are redacted to protect the business interest of Yahoo.

Table 2: Homogeneous Network Effects

Variable	OLS	First stage	Second stage
Log(installed base)	.7161		.9693
	(.0189)***		(.1113)***
% Urban	.1826	.5088	.0710
	(.0714)**	(.0915)***	(.0947)
% Nonwhite	-.1321	-1.580	.2649
	(.0979)	(.1387)***	(.2012)
% College	.0047	.0324	-.0036
	(.0023)**	(.0031)***	(.0041)
Log(median income)	-.0424	.2531	-.1303
	(.0979)	(.1313)*	(.1127)
Gini index	.0455	-1.007	.2577
	(.5352)	(.6869)	(.5605)
Log(county pop.)	-.7603	1.117	-1.044
	(.0269)***	(.0216)***	(.1245)***
Dist. to stadium	-.0015	.0042	-.0015
	(.0006)**	(.0010)***	(.0007)**
2017 team % wins	.0714	-.3267	-.0264
	(.2946)	(.4478)	(.3134)
Dist. × 2017 wins	.0027	.0007	.0031
	(.0012)**	(.0018)	(.0014)**
2016 team % wins		2.781	
		(.4192)***	
Dist. × 2016 wins		-.0113	
		(.0014)***	
Constant	.4567	-12.53	3.651
	(1.159)	(1.537)***	(1.933)*
Adjusted R^2	.6297		
First-stage F		33.05	
N	2285	2285	2285

The dependent variable is log of adopters divided by potential market minus log of non-adopters divided by potential market, where potential market is defined as 10% of male and 8% of female county population aged 18 to 74 years old minus the installed base. The instruments for the log of installed base in 2017 are the nearest pro team's 2016 season winning percentage and its interaction with the distance from the county centroid. First-stage F is the Kleibergen-Paap Wald F statistic for the test of weak identification. In all three columns, robust standard errors are shown in parentheses. * denotes statistical significance at the 10% level, ** at the 5% level, and *** at the 1% level.

Table 3: Heterogeneous Network Effects

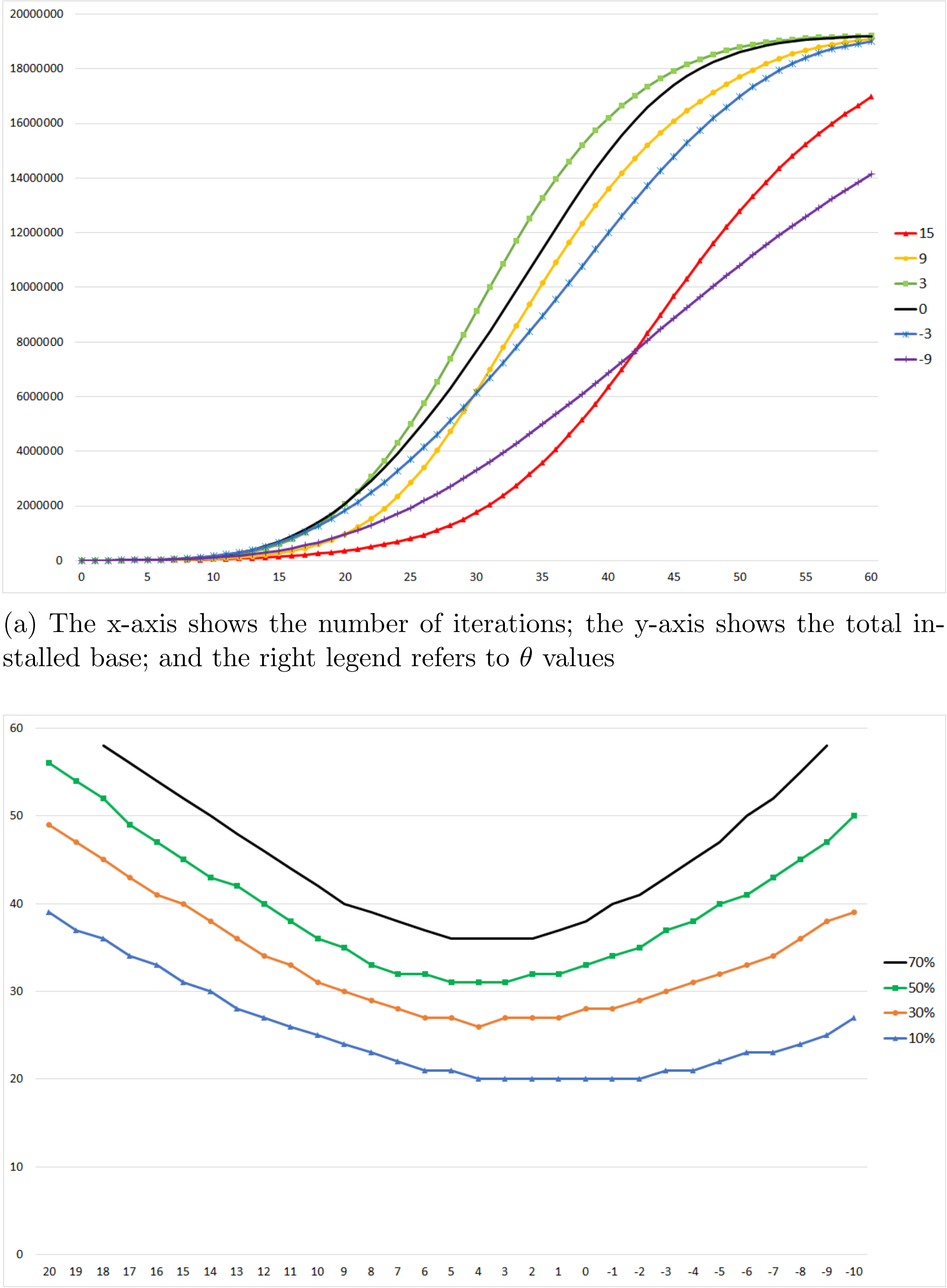
Variable	(1)	(2)	(3)	(4)
Log(installed base)	-.4271	-.0710	-.5689	-.8177
	(.5460)	(.6818)	(.5451)	(.7759)
Log(installed base)	.1317	.1114	.1436	.1705
× Log(income)	(.0489)***	(.0534)**	(.0488)***	(.0745)**
Log(installed base)		-.3005		
× Gini index		(.3702)		
Log(installed base)			.0946	
× % Nonwhite			(.0705)	
Log(installed base)				-.0018
× % College				(.0022)
Controls	Yes	Yes	Yes	Yes
First-stage F	18.13	12.31	12.48	12.19
N	2285	2285	2285	2285

The dependent variable is the same as that in Table 2. Each column only shows the second-stage least-square coefficients of the log of installed base in the county and its interactions with the county income and other characteristics; and the control variables are not shown here for brevity. The instruments for the endogenous variables are the nearest pro team's 2016 season winning percentage, its interaction with the distance from the county centroid, and the same variables interacted with the respective county characteristics. First-stage F is the Kleibergen-Paap Wald F statistic for test of weak identification. In all four columns, robust standard errors are shown in parentheses. * denotes statistical significance at the 10% level, ** at the 5% level, and *** at the 1% level.

Table 4: Robustness to Nearby Counties

Variable	(1)	(2)	(3)	(4)
Log(installed base)	1.003	.9217	.8828	.8790
	(.1017)***	(.1039)***	(.0932)***	(.0850)***
Log(installed base	.0060			
in 100 miles)	(.0335)			
Log(installed base		.0133		
in 200 miles)		(.0319)		
Log(installed base			-.0047	
in 300 miles)			(.0251)	
Log(installed base				-.0148
in 400 miles)				(.0234)
Controls	Yes	Yes	Yes	Yes
First-stage F	18.64	16.40	19.88	23.82
N	2285	2285	2285	2285

The dependent variable is the same as that in Table 2. Each column only shows the second-stage least-square coefficients of the log of installed base in own county and the log of installed base in the neighboring counties within a certain radius; and the control variables are not shown here for brevity. The instruments for the endogenous variables are the nearest pro team's 2016 season winning percentage, its interaction with the distance from the county centroid, and the same variables pre-multiplied by the respective spatial matrix. First-stage F is the Kleibergen-Paap Wald F statistic for test of weak identification. In all four columns, robust standard errors are shown in parentheses. * denotes statistical significance at the 10% level, ** at the 5% level, and *** at the 1% level.



(a) The x-axis shows the number of iterations; the y-axis shows the total installed base; and the right legend refers to θ values

(b) The x-axis shows the θ value; the y-axis shows the number of iterations; and the right legend refers to total market penetration

Figure 1: Simulated Adoption Outcomes