Distilling network effects from Steam

José F. Tudón*

February 17, 2021

Abstract

This paper develops a method to estimate the demand of network goods, using minimal network data, but leveraging within-consumer variation. I estimate demand for video games as a function of individuals' social networks, prices, and qualities, using data from Steam, the largest video game digital distributor in the world. I separately identify price elasticities on individuals with and without friends with the same game, conditional on individual fixed effects and games' characteristics. I then use the discrepancies between estimated price elasticities to identify the impact of social networks. I compare my method to "traditional-IV" strategies in the literature, which require detailed network data, and find similar results. A 1% increase in friends' demands, increases demand by .3%. These results suggest firms have incentives to employ "influencers," and to merge their gaming networks. I evaluate counterfactual game mergers, and find network effects must be considered to sign consumer welfare changes.

1 Introduction

Even with the recent success of the empirical literature on estimating network effects,[1] two main problems need to be tackled in the field: one practical, and one methodological. First, consider the practical problem of collecting network data. In many settings, access to network data is prohibitively expensive, because of

*ITAM, Río Hondo 1, Ciudad de México 01080, México (e-mail: jtudon@itam.mx). Thanks to Michael Dinerstein, Ali Hortaçsu, Mauricio Romero, Adrián Rubli, and ITAM's applied micro workshop participants. Adrián Martínez provided outstanding research assistance.

[1] See, e.g., Björkegren (2019), De Giorgi, Frederiksen and Pistaferri (2019), Ryan and Tucker (2012).

privacy concerns, ownership, or scale (Breza et al., 2020). Second, without a proper research strategy, the reflection problem prevents the identification of direct network effects (Manski, 1993; Rysman, 2019).[2]

I study demand network effects and propose a simple identification strategy without access to detailed network data. Instead, I leverage within-consumer variation in a three-step identification process. In the first step, the "pure" price elasticity is identified from isolated individuals' behavior in the network. In the second step, the "compounded" price elasticity is identified from the rest of the individuals. And in the third step, the network effect is identified from the discrepancy between the previous two steps' price elasticities within consumers. Therefore, this strategy has simpler data requirements—repeated observations for an individual—than most of the literature, which requires a complete mapping of the network. I use this strategy to study the online video game industry.

The proposed method is simple and intuitive. To fix ideas, consider Becker (1991)'s framework for social interactions, where individual demand, $d_i(p, D)$, is a function of prices, and aggregate demand, $D = \sum_i d_i(p, D) = F(p, D)$. That is, aggregate demand is a function of prices and aggregate demand, where we expect $F_p < 0$, but $F_D > 0$ for positive network effects. Therefore,

$$\frac{\partial D}{\partial p} = F_p + F_D \frac{\partial D}{\partial p} \qquad \Rightarrow \frac{\partial D}{\partial p} = \frac{F_p}{1 - F_D}.$$

This paper's strategy consists of identifying a subset of data where $F_D = 0$, so we can estimate F_p from the observed $\frac{\partial D}{\partial p}$. Given F_p, the network effect, F_D, can then be estimated from the total effect, $F_p/(1 - F_D)$, which is identified in the complementary subset. In the context of online video games, the individual demand is boosted by the proportion of friends who own the game. In this case, F_p can be identified from individuals who do not have friends who own the same game, and $F_p/(1 - F_D)$ can be identified from the individuals with at least one friend with the game. We do not need detailed network data; we only need to know "how many friends own the game." In other words, we only need "Aggregated Relational Data (ARD)" (Breza et al., 2020) to identify network effects and price elasticities.[3]

Moreover, individual demand might be a function of other variables; for

[2]For instance, if a positive network effect is ignored, the price elasticity of demand will be overestimated, because the total effect of a price change is composed of a price effect *and* a network effect.

[3]ARD refers to answers to questions of the form "how many of your links have trait X?"

instance, $d_i(p, r, D)$, where r is quality. Then, $D = F(p, r, D)$,

$$\frac{\partial D}{\partial p} = \frac{F_p}{1 - F_D}, \qquad \text{and} \qquad \frac{\partial D}{\partial r} = \frac{F_r}{1 - F_D}.$$

Thus, as long as we can identify either F_p or F_r, we can identify F_D. If we can identify both, then they must agree in their prediction of F_D. As a result, the estimation of the network effect can be improved by using the corresponding set of moments. This paper shows this is the case when estimating F_p and F_r separately for the online video game industry.

An important concern is that consumers with and without friends might be different. I address this issue by considering consumers' purchasing behavior at the individual level, and observing their decisions for many games, some of which have also been purchased by friends, some of which have not. Moreover, individual fixed effects are also considered to absorb remaining contextual and correlated effects (Hartmann et al., 2008).

The results suggest the presence of strong network externalities: if the proportion of friends with some game j increases in 1%, then demand for game j increases in .3%. If we ignore network effects, price elasticity is significantly overestimated, and we can substantially underestimate consumer welfare. Moreover, with separate price and network elasticities, managers and policy makers can simulate counterfactual pricing and marketing policies. In counterfactuals, I find strong incentives to promote the game through "influencers," and some incentives for games to merge their social networks.

Network effects must be considered to sign consumer welfare changes when evaluating gaming network mergers. In a counterfactual, the merger of two games in the sample increases consumer surplus by 0.87 %, due to network effects, even after upward pricing pressure is taken into account. However, a policymaker who ignores network effects predicts a consumer surplus change of -6.10 %.

I also compare this paper's strategy to "traditional" identification strategies in the context of video game demands. Traditional strategies use detailed network data, such as second-order friends' characteristics, as instruments for first-order friends' characteristics. I arrive to qualitatively similar conclusions with both methodologies.

This paper is related to the empirical literature of the effects of networks. In particular, the idea of achieving identification from isolated individuals goes back at least to Fortin and Boucher (2016) for the linear-in-means model. However, my paper is the first to fully develop this method, describe the outstanding econometric problems, and offer a solution to them, especially in a structural, discrete-choice model.

This paper is also related to the empirical literature that uses minimal network data. Breza et al. (2020) use Aggregated Relational Data (ARD) to recover parameters of a network formation model. They focus on studying the network structure. I also use ARD (e.g., "how many of your friends have game j?"), but focus on quantifying direct network externalities. Indeed, ARD studies have a large practical upside, because mapping whole networks is expensive and time-consuming, as described by Breza and Chandrasekhar (2019), Banerjee et al. (2013) or Chandrasekhar, Kinnan and Larreguy (2018).

Finally, this paper is also related to work emphasizing network effects on video game demand. Dubé, Hitsch and Chintagunta (2010) find strong indirect network effects in the two-sided market of video game consoles, and Lee (2013) studies vertical integration in the presence of indirect network effects in the video game industry. By contrast, this paper studies *direct* network effects.

2 Data and empirical context

Steam is the largest video game digital distribution service in the world, with over 90 million monthly active users shopping from over 30,000 PC video games on offer.[4] On 2017, Steam generated 4.3 billion USD in sales, not counting in-app microtransactions or downloadable content, and in 2020 it recorded 7.25 million users concurrently playing games.[5]

Steam users form a social network where they can "friend" each other in a similar manner to Facebook. On the Steam platform, users can connect with friends and communicate with each other, send and receive games as gifts, or coordinate to play together. Depending on privacy settings, users can see each others' game achievements, video game libraries, and time spent playing video games, as well as other personal information such as name, location, or friends.

The data, published by O'Neill et al. (2016), contain the universe of 108.7 million user accounts, who collectively own 384.3 million games, as of March 2013. For each user, I observe their game library, friends, date of account creation, date of last activity on the platform, and time spent on each video game. And for

[4]See `store.steampowered.com/about`, `store.steampowered.com/stats`, and `pcgamesn.com/steam-player-count`.

[5]See `pcgamesn.com`, `videogameschronicle.com`.

each video game, I observe its price, rating, type of game (full game, DLC, and others), genre, developers, publishers, date of release, among other information.[6]

TABLE 1: SUMMARY STATISTICS

	mean	sd	min	max
Users				
Friends	7	20	0	1085
Owned games	13	40	0	3797
Expenditure	169	498	0	40294
Price	13	9	0	100
Rating	83	7	20	96
Time played	360	813	0	288784
Games				
Price	11	14	0	450
Rating	72	11	20	96
Multiplayer	.24	.43	0	1

Notes: Expenditure and prices in US dollars. Time Played in hours, cumulative. Users considered if active in 2013.

Table 1 shows summary statistics. On average, users have 7 friends, 13 games, spend 169 dollars in total, pay for each game 13 dollars, and play 360 hours total. However, statistics are skewed, showing a long tail of users with much larger consumptions, expenditures, and social networks. For games, the average price on the market is 11, with an average rating of 72 "likes" out of 100, and 24% of games have a multiplayer mode. We can also see that users lean towards pricier and higher-rated games.

Because of the nature of the video game consumption, we expect positive social network effects. Video games can be single-player or multiplayer (or a mix). Both types of games enjoy positive network effects, but the argument can be made for a stronger effect on multiplayer games. Notably, we can construct the complete friends network with the data, but, as this paper shows, we do not need the whole network mapping to estimate network effects.

In this paper, I focus on the top 100 games by popularity, which account for almost 50% of the market share. In this way, I cut out the long tail of games with very small market shares. Also, I only consider active users in 2013 for the analysis; I define "active" as logging on at least once in 2013, and having at least one friend or one game. We have over 53 million active users, though we only

[6]Data and replication files can be found in the author's website, jtudon.com, or by request. The raw data can be found in steam.internet.byu.edu.

require a small random sample for consistency.

3 Linear-in-means model

I follow the peer effects literature by assuming that the demand for a network good depends on the probability of peer consumption. Specifically, I consider the following linear-in-means model:

$$q_{ij} = \mu_i - \alpha \log p_j + \beta n_{ij} + \gamma \log r_j + \boldsymbol{\delta}_0 \boldsymbol{x}_j + \varepsilon_{ij}, \tag{1}$$

where the quantity q_{ij} indicates if i purchased game j, μ_i is an individual fixed effect, p_j is the price of game j, n_{ij} is the proportion of friends of i that own game j, r_j is the quality of the game, measured by a rating index, ε_{ij} is a shock unobservable to the econometrician, and $\boldsymbol{x}_j$ are characteristics of game j that include genre dummies (action, adventure, RPG, etc.), days since game j was released, and indicators for multiplayer and age restrictions. The set of covariates $\boldsymbol{x}_j$ allow me to control for other demand shifters that could correlate with price or quality. Consumers are indexed by $i = 1, \ldots, I$, and games by $j = 1, \ldots, J$.

Equation (1) has the advantage that it can be derived from a utility-maximization problem (Fortin and Boucher, 2016), and can be interpreted as the best-response function for an individual who responds to complementary social interactions. We expect $\beta > 0$ for a positive network effect. Section 4 presents a binary choice model instead of equation (1)'s linear probability model.

3.1 Identification

Equation (1) is not identified, because $\mathbb{E}\left[q_{ij} | \boldsymbol{z}_{ij}\right] = \mathbb{E}\left[n_{ij} | \boldsymbol{z}_{ij}\right]$, for an appropriate vector of instruments, $\boldsymbol{z}_{ij}$. Taking conditional expectations and simplifying,

$$\mathbb{E}\left[q_{ij} | \boldsymbol{z}_{ij}\right] = \frac{\mu_i}{1-\beta} - \frac{\alpha}{1-\beta} \log p_j + \frac{\gamma}{1-\beta} \log r_j + \frac{\boldsymbol{\delta}_0}{1-\beta} \boldsymbol{x}_j, \tag{2}$$

even if $\mathbb{E}\left[\varepsilon_{ij} | \boldsymbol{z}_{ij}\right] = 0$.

However, identification can be achieved if we assume $\mathbb{E}\left[\varepsilon_{ij} | \boldsymbol{z}_{ij}, n_{ij} = 0\right] = 0$, so we get

$$\mathbb{E}\left[q_{ij} | \boldsymbol{z}_{ij}, n_{ij} = 0\right] = \mu_i - \alpha \log p_j + \gamma \log r_j + \boldsymbol{\delta}_0 \boldsymbol{x}_j. \tag{3}$$

The identification argument has three steps.

The first step is to identify price and ratings coefficients from equation (3). Software development represents the main production cost of a game, but consumers are rarely aware of the identity of the developers, as opposed to, say, publishers. If the game is developed "in-house," I expect costs to be lower compared to outsourced development. Then, to identify the price coefficient, α, I instrument prices with the status of integration between developers and publishers of game j as cost shifters. Therefore, conditional on the observable characteristics of game j, the developers of j introduce exogenous variation to the price of j. On the other hand, I assume ratings are exogenous, conditional on individual fixed effects and observable game characteristics.

A threat to identification comes from the fact that people with no friends are different to people with some friends. However, I have three answers for this problem. First, I include individual fixed effects to control for individual heterogeneity, contextual effects, correlated effects, and, to some extent, for network formation (Hartmann et al., 2008). As long as people with no friends are not systematically different in their pricing sensitivities than people with some friends, we have identification. Second, $n_{ij} = 0$ does not imply that i has no friends, but only that i has no friends with game j. That is, people who have $n_{ij} = 0$ are not rare. Indeed, no one game has perfect penetration, nor one individual has every game. We only require that for some game j, i's friends do not have that particular game. In the data, consumer i has game j with .06 probability, and has 7 friends, on average. Therefore, having games with no friends is not uncommon. Third, we can also identify the network effect from other coefficients, such as the ratings coefficients. In this case, even if people with no friends have different rating sensitivities, we would not expect such bias to be exactly the same as the bias in price sensitivities.

The second step is to identify the compounded price and rating coefficients, $\alpha/(1-\beta)$ and $\gamma/(1-\beta)$, from equation (2), but conditional on non-isolated individuals:

$$\mathbb{E}\left[q_{ij} | \boldsymbol{z}_{ij}, n_{ij} > 0\right] = \frac{\mu_i}{1-\beta} - \frac{\alpha}{1-\beta} \log p_j + \frac{\gamma}{1-\beta} \log r_j + \frac{\boldsymbol{\delta}_0}{1-\beta} \boldsymbol{x}_j, \tag{4}$$

where I assume $\mathbb{E}\left[\varepsilon_{ij} | \boldsymbol{z}_{ij}, n_{ij} > 0\right] = 0$. Again, I use developer integration to identify the compounded coefficient $\alpha/(1-\beta)$.

The third step is to identify the network effect, β. With estimates of α and $\alpha/(1-\beta)$, we can back out an estimate of β. Note that we do not need γ and

$\gamma/(1-\beta)$, but if we use these rating coefficients instead, the resulting β should agree with the one derived with the price coefficients.

I also compare my result with the "traditional" peer-effect instrumental-variables strategy. That is, I estimate equation (1) directly,

$$q_{ij} = \mu_i - \alpha \log p_j + \beta n_{ij} + \boldsymbol{\delta}_0 \boldsymbol{x}_j + \varepsilon_{ij}, \tag{1}$$

where n_{ij} is instrumented with $n_{ij}^{(2)}$, the probability that second-degree friends have game j. In other words, let $N_i^{(2)}$ be the total number of i's second-degree friends (friends of my friends, who are not my friends), and let $N_{ij}^{(2)}$ be the number of i's second-degree friends who have purchased game j. Then, $n_{ij}^{(2)} \equiv N_{ij}^{(2)}/N_i^{(2)}$ is an instrument for n_{ij} (e.g., De Giorgi, Frederiksen and Pistaferri (2019)). The disadvantage of this method is that it requires detailed network data to construct $n_{ij}^{(2)}$, as opposed to my method, which only requires n_{ij}.

As a final note, equations (2) and (3) should not be estimated in a single equation model using a within transformation (even with appropriate interaction terms), because even the fixed effects differ by a factor of $(1-\beta)^{-1}$ between these equations. A full set of individual dummies, with all covariates interacted with $\mathbb{1}\{n_{ij} = 0\}$ is theoretically appropriate, but not practical.

3.2 Results

To be clear, the pseudo-panel follows a random sample of .2% of active users who consider buying games $j = 1, \ldots, 100$. Table 2 shows the main results. All models include individual fixed effects and the controls described in section 3. In all models, prices have been instrumented with an indicator for outsourced game development.

The benchmark model in column 1 shows a 2SLS estimation where the network effect has been ignored. As expected, the coefficients are compounded with the network effect. Column 2 shows the "traditional" IV approach, where n_{ij} has been instrumented with the proportion of second-order friends with game j. The network effect corresponds to the coefficient of n_{ij}, which is 0.87, and implies a .1 increase in the proportion of friends with game j increases the probability of buying j in by 0.09.

Columns 3 and 4 show the results of my method. Column 3 restricts the sample to $n_{ij} = 0$, and column 4 to $n_{ij} > 0$. As expected, the benchmark model yields estimates in between those of columns 3 and 4. Column 3's coefficients

TABLE 2: ESTIMATION OF DEMAND AND NETWORK EFFECT

DEP VAR: q_{ij}	BENCHMARK (1)	TRADITIONAL IVS (2)	NO FRIENDS (3)	SOME FRIENDS (4)
log price$_j$	-0.20	-0.06	-0.10	-0.73
	(0.013)	(0.014)	(0.010)	(0.101)
log rating$_j$	0.19	0.04	0.08	0.59
	(0.010)	(0.011)	(0.008)	(0.063)
n_{ij}		0.87		
		(0.019)		
IMPLIED NET EFF		0.87		
USING PRICES			0.87	
USING RATINGS			0.87	
IMPLIED PRICE ELASTICITY	-3.48	-1.11	-1.64	
CONTROLS	YES	YES	YES	YES
FIXED EFFECTS	YES	YES	YES	YES
2SLS	YES	YES	YES	YES
FIRST-STAGE IV'S F-STAT				
log price$_j$	1,819	1,643	1,713	53
n_{ij}	.	124,526	.	.
OBS	428,076	428,076	339,198	88,878

Notes: Standard errors clustered at individual level. In all columns, prices are instrumented with an indicator for outsourced game development. All collumns include as controls: individual fixed effects, a set of genre dummies, days since release, and indicators for multiplayer and age restrictions. Column 1 shows the benchmark model where network effects are ignored. Column 2 includes n_{ij} and instruments it with the proportion of second degree friends with game j. Column 3 is estimated in the subsample where $n_{ij} = 0$, and Column 4 in the subsample where $n_{ij} > 0$. The implied network effect of column 2 is just the coefficient of n_{ij}, while the implied network effect of my method combines the coefficients of columns 3 and 4 as $1 - \alpha_3/\alpha_4$, where α_3 and α_4 are the coefficients from columns 3 and 4, respectively.

correspond to the price and rating effects *without* network effects, and align with those of column 2. Column 4 shows the coefficients compounded with the network effects. The implied network effects are backed out using columns 3 and 4: let α_3 and α_4 be the coefficients from columns 3 and 4, respectively, then, the network effect is $1 - \alpha_3/\alpha_4$.

The estimated network effect from my method is 0.87 using the price coefficients, and 0.87 using the rating coefficients. Not only they agree with each other, but they are also very close to column 2's 0.87. In terms of elasticities, the implied network effects are around 0.9, which suggest a significant direct network externality in the platform.[7]

Moreover, the estimated price elasticity from my method is -1.64, but, if we ignore network effects, the implied price elasticity from the benchmark model is -3.48. As expected, focusing only on isolates yields a more inelastic demand

[7] To boot, most of the unreported coefficients of the controls also yield estimates of the network effect around 0.9. I do not report these results, because I do not have a strong identification argument for these coefficients.

than the compounded elasticity from the whole sample. Therefore, we could significantly overestimate price elasticities, and the same can be said about rating elasticities, for example.[8]

For managers and policy makers, this means that efforts to attract consumers pay off. It also means that willingness-to-pay is significantly higher at any point of the demand curve. Thus, a two-part tariff could be more efficient.

While the linear probability model has advantages, it can overestimate marginal effects, specially when outcomes have probabilities close to the extremes, 0 or 1. In the data, consumers purchase any random game j with a probability of .06, which warrants the exploration of a nonlinear model.

4 Discrete choice model

With individual data we can model the binary choice of purchasing a game or not. Consider the following analogue of equation (1):

$$u_{ij} = \mu_i - \alpha \log p_j + \beta n_{ij} + \gamma \log r_j + \boldsymbol{\delta}_0 \boldsymbol{x}_j + \varepsilon_{ij}, \tag{5}$$

where u_{ij} is the utility of consumer i from purchasing game j, and the remaining variables are as in section 3: fixed effects, prices, proportion of friends with game, ratings, and game characteristics.

Consumer i purchases j iff $u_{ij} > 0$. Let $y_{ij} \equiv \mathbb{1}\{u_{ij} > 0\}$. Assuming iid $\varepsilon \sim F$, the probability of buying j is

$$P\left[y_{ij} = 1\right] = F\left(\mu_i - \alpha \log p_j + \beta n_{ij} + \gamma \log r_j + \boldsymbol{\delta}_0 \boldsymbol{x}_j\right), \tag{6}$$

where F will later be specialized to a logit model. Thus, i's demand for j, is defined as $s_{ij} \equiv P\left[y_{ij} = 1\right]$.

Then, the log-likelihood of the data becomes

$$L = \sum_{i=1}^{N} \sum_{j=1}^{J} y_{ij} \log s_{ij} + (1 - y_{ij}) \log(1 - s_{ij}). \tag{7}$$

4.1 Identification

As with the linear-in-means model, we have a reflection problem. In particular, a likelihood maximization program will not identify the local maximum by using

[8] Nair (2007) reports short-run price elasticities similar to those found in this paper.

first order conditions from L, because the derivatives will not take into account the feedback from network effect.

Consider the more compact notation

$$\boldsymbol{X}_{ij}'\boldsymbol{\theta} \equiv \mu_i - \alpha \log p_j + \gamma \log r_j + \boldsymbol{\delta}_0 \boldsymbol{x}_j,$$

where $\boldsymbol{\theta}$ is the vector of parameters, excluding β, and $\boldsymbol{X}_{ij}$ is the vector of covariates, excluding n_{ij}. We can rewrite the utility function as

$$u_{ij} = \beta n_{ij} + \boldsymbol{X}_{ij}'\boldsymbol{\theta} + \varepsilon_{ij}. \tag{5'}$$

Then, the likelihood's first order conditions with respect to any element θ_k of $\boldsymbol{\theta}$ become

$$0 = \frac{\partial L}{\partial \theta_k} = \sum_{i=1}^{N} \sum_{j=1}^{J} \frac{y_{ij} - s_{ij}}{s_{ij}(1 - s_{ij})} \frac{\partial s_{ij}}{\partial \theta_k}.$$

However,

$$\frac{\partial s_{ij}}{\partial \theta_k} = F'\left(\beta n_{ij} + \boldsymbol{X}_{ij}'\boldsymbol{\theta}\right) \left[X_{ij}^k + \beta \frac{\partial n_{ij}}{\partial \theta_k}\right], \tag{8}$$

where X_{ij}^k is the corresponding covariate to θ_k. In this context, the reflection problem is that $\frac{\partial n_{ij}}{\partial \theta_k}$ would be ignored by a maximization program. But, in equilibrium, $n_{ij} = s_{ij}$, which implies that $\frac{\partial n_{ij}}{\partial \theta_k} = \frac{\partial s_{ij}}{\partial \theta_k}$. Then,

$$\begin{aligned} \frac{\partial s_{ij}}{\partial \theta_k} &= F'\left(\beta n_{ij} + \boldsymbol{X}_{ij}'\boldsymbol{\theta}\right) \left[X_{ij}^k + \beta \frac{\partial s_{ij}}{\partial \theta_k}\right] \\ \Rightarrow \frac{\partial s_{ij}}{\partial \theta_k} &= \frac{F'\left(\beta n_{ij} + \boldsymbol{X}_{ij}'\boldsymbol{\theta}\right)}{1 - \beta F'\left(\beta n_{ij} + \boldsymbol{X}_{ij}'\boldsymbol{\theta}\right)} X_{ij}^k. \end{aligned} \tag{9}$$

A simple solution, as in the linear-in-means model, is to consider first the cases with $n_{ij} = 0$, so that

$$\begin{aligned} u_{ij} &= \boldsymbol{X}_{ij}'\boldsymbol{\theta} + \varepsilon_{ij} \\ \Rightarrow s_{ij} &= F\left(\boldsymbol{X}_{ij}'\boldsymbol{\theta}\right), \\ \Rightarrow \frac{\partial s_{ij}}{\partial \theta_k} &= F'\left(\boldsymbol{X}_{ij}'\boldsymbol{\theta}\right) X_{ij}^k \end{aligned} \tag{10}$$

[9] Note that from the FOC on β, $\frac{\partial s_{ij}}{\partial \beta} = F'\left(\mu_i - \alpha \log p_j + \beta n_{ij} + \gamma \log r_j + \boldsymbol{\delta}_0 \boldsymbol{x}_j\right) \beta \frac{\partial s_{ij}}{\partial \beta}$, which does not identify β.

for the consumer i without friends with game j.

Intuitively, we can identify $\boldsymbol{\theta}$ using equation (10), and then use that information to identify β using (9).[9]

To identify the price and ratings coefficients, I use the same set of instruments from the linear-in-means model. However, because of the nature of the binary choice model, I opt for a control function approach (Petrin and Train, 2010). To create the control function, in a first stage, I regress endogenous variables on instruments and exogenous variables, and I estimate residuals. In a second stage, I include those estimated residuals as extra covariates. Standard errors can be calculated by bootstrapping.

Finally, I use Chamberlain (1980)'s conditional logit model to avoid the incidental parameters problem caused by the individual fixed effects (Lancaster, 2000). The rest of threats to identification that arise here are very similar to those from the linear-in-means model which I have already discussed.

4.2 Estimation

If $\boldsymbol{\theta}$ is of length K, then the first order conditions define the following set of $2K$ moments for $K+1$ parameters $\{\boldsymbol{\theta}, \beta\}$. For $k = 1, \ldots, K$,

$$0 = \frac{\partial L}{\partial \theta_k} = \sum_{i=1}^{N} \sum_{j=1}^{J} \frac{y_{ij} - s_{ij}}{s_{ij}(1-s_{ij})} F'\left(\boldsymbol{X}_{ij}'\boldsymbol{\theta}\right) X_{ij}^k \qquad \text{if } n_{ij} = 0, \tag{11}$$

$$0 = \frac{\partial L}{\partial \theta_k} = \sum_{i=1}^{N} \sum_{j=1}^{J} \frac{y_{ij} - s_{ij}}{s_{ij}(1-s_{ij})} \frac{F'\left(\beta n_{ij} + \boldsymbol{X}_{ij}'\boldsymbol{\theta}\right)}{1 - \beta F'\left(\beta n_{ij} + \boldsymbol{X}_{ij}'\boldsymbol{\theta}\right)} X_{ij}^k \qquad \text{if } n_{ij} > 0. \tag{12}$$

Thus, estimation is done by GMM.

In practice, we can separately estimate $\boldsymbol{\theta}$ from moments (11) at the cost of efficiency. Using $\widehat{\boldsymbol{\theta}}$, we can then use moments (12) to estimate β. For consistency with section 3, I only use the price and rating moments to identify the network effect.

4.3 Results

Table 3 shows the results for a binary-choice logit model. Every column addresses price endogeneity as described. As a benchmark, column 1 shows a logit model which ignores the network effect. Column 2 identifies the network effect using the traditional instrument from the literature, the proportion of second-degree friends with game j. Finally, column 3 presents this paper's method, where the network effect is identified from the moments defined by the first order conditions of a logit model with network effects.

Price elasticities align with those from the linear-in-means model. I find an elastic demand curve, but the elasticity could be overestimated if we ignore network effects, as expected. Demand is also inelastic with respect to the quality of the product, measured by its rating.

Finally, the network effect is estimated to be 0.29, which is lower than the one estimated with a linear probability model. The network externality translates to an increase in demand of .03% when the proportion of friends with the game increases by .1%. For perspective, the average demand of a game in the sample is 6%. The effect is strong, as the experiments from the following section show.

TABLE 3: DISCRETE-CHOICE LOGIT MODEL ESTIMATION

DEP VAR: y_{ij}	BENCHMARK LOGIT (1)	TRADITIONAL IV LOGIT (2)	LOGIT + GMM (3)
log price$_j$	-2.21	-1.22	-1.90
	(0.345)	(0.201)	(0.300)
log rating$_j$	2.26	1.00	0.75
	(0.283)	(0.145)	(0.107)
n_{ij}		4.79	8.00
		(0.152)	(1.551)
IMPLIED ELASTICITIES			
log price$_j$	-2.15	-1.19	-1.85
log rating$_j$	2.21	0.98	0.73
n_{ij}		0.18	0.29
CONTROLS	YES	YES	YES
FIXED EFFECTS	YES	YES	YES
CONTROL FUNCTIONS	YES	YES	YES
FIRST-STAGE IV'S F-STAT			
log price$_j$	1,539	1,468	3,151
n_{ij}	.	33,770	.
OBS	213,474	213,474	213,474

Notes: Bootstrapped standard errors clustered at individual level with 100 repetitions. In all columns, prices are instrumented with an indicator for outsourced game development. All columns include as controls: individual fixed effects, a set of genre dummies, days since release, and indicators for multiplayer and age restrictions. Instrumented variables are regressed on instruments in a first stage, where residuals are obtained and used as control functions (Petrin and Train, 2010). Column 1 shows a logit model where the network effect is ignored. Column 2 includes n_{ij} and instruments it with the proportion of i's second-degree friends who own game j. Column 3 shows this paper's logit model where the network effect is estimated using GMM and the moments defined by the FOC of a logit with network effects.

5 Counterfactuals

Here, I explore two main counterfactuals. First, I consider the merger of two games, which implies the merger of two networks. If network effects are ignored, the policymaker estimates a drop in consumer surplus from increased prices, but consumer surplus might increase after network effects are taken into account. Second, I consider a giveaway to the top 1% most popular consumers in the network, and find significant increases in revenue due to network effects.

5.1 Antitrust, regulation, and upward pricing pressure

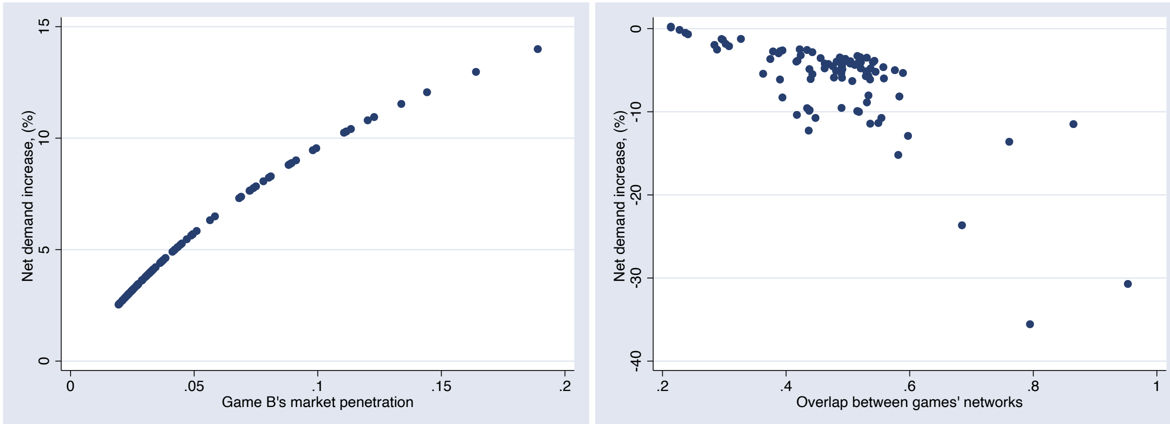
In network industries, policymakers consider the trade-off between competition and monopoly when evaluating mergers. If network effects are strong enough, welfare might be improved by coordinating consumption under a monopolist. In the case of video games, and because marginal costs are essentially zero and further marginal cost reductions are unlikely, the merged video game will almost surely face upward pricing pressure. Then, the resulting change in consumer surplus from the merger becomes an empirical question.

In this experiment, I take the most popular game in the sample and merge it with a smaller game, which increases the size of the network. Assuming games are compatible, and that the merger is technically possible, the merged game will feature the merged network, which is the union of the pre-merger networks. Because of network effects, consumers translate a larger network as if the game were of a higher quality. Indeed, on the intensive margin, consumers who already own the game value it even more. On the extensive margin, marginal consumers now purchase the game. Therefore, demand increases further, which increases demand yet again.

Let game A be "Counter-Strike: Source," the largest game in the sample in 2013, which holds a 20% market penetration; that is, 20% of consumers in the sample own game A.[10] Game A might consider merging networks with game B, a hypothetical smaller game. Owners of game B will be "upgraded" to game A and will join the network of A, forming the merged network $A \cup B$, the union of networks A and B.

Profits might be larger for the merger than for the separate games if the demand of the merger increases by enough. Importantly, the interception $A \cap B$ must be small enough for the merger to be profitable. In the extreme, if $A \cap B = B$,

[10] See store.steampowered.com/app/240.



(A) Assuming no overlap of games' networks (B) Taking overlap into account

FIGURE 1: Counterfactual merger

then $A \cup B = A$, and the merger gains nothing, but loses the revenue of a separate game B. Assuming that games A and B are identical, except for their networks, Figure 1 shows the net increase in revenues from a merger between game A and each of the other games in the sample. That is, the figure shows

$$\frac{D(A \cup B) - D(A) - D(B)}{D(A) + D(B)},$$

for every other game B in the sample, where $D(S)$ is the demand for a game with a network S, assuming the characteristics of the game are equal to those of game A.

Figure 1a assumes no overlap in their networks, $A \cap B = \emptyset$, and shows the theoretical gains from merging two networks into a single one as a function of B's size. Demand for the merger is predicted to increase between 2.5% and 14%, depending on the size of the smaller game. However, in reality, networks often intersect. Figure 1b presents net gains as a function of the intersection of the networks, and shows that actually the merger only makes sense for the two games with the least degree of intersection in the data.[11]

Therefore, the merger stands to (net) gain at most .23% of market penetration when game A merges with game B^*, "Total War: Empire," a game with a market penetration of 4%.[12] In terms of consumer surplus, I estimate that consumers

[11] To be sure, most of the mergers in the graph are not feasible, as games would be too different from one another for the merger to make technical sense. Nevertheless, the graphed mergers can be interpreted as the case where game A buys out the network of game B.

[12] See store.steampowered.com/app/10500.

enjoy a combined benefit of 12.3 million USD from both games A and B^* before the merger. After the merger, consumer surplus increases by 6.40 % to 13.1 million USD, due to the larger network.[13]

The merged firm however faces upward pricing pressure, not because competition has eased, but because aggregate demand has expanded. The problem of firm A can be written as $\max_{p_A}(p_A - c_A)s_A(p_A)$, where we expect marginal costs close to zero. The first order condition is the usual Lerner index equation:

$$\frac{p_A - c_A}{p_A} = -\left(\frac{\partial s_A}{\partial p_A}\frac{p_A}{s_A}\right)^{-1}.$$

However, the pertinent price elasticity of demand in the Lerner equation is not $\alpha(1 - s_A)$, which is the elasticity of the willingness-to-pay curve. The firm takes into account the complete effect from a price change. Figure 2 distinguishes between willingness-to-pay and demand. At p_0, willingness-to-pay is D_0, but at p_1, quantity increases, which increases the willingness-to-pay to D_1. The direct price effect is a movement along D_0, but the network effect is the horizontal movement of the willingness-to-pay. The observed demand, D, traces the changes in consumed quantity due to the price decrease, and shows a higher elasticity than the underlying willingness-to-pay.

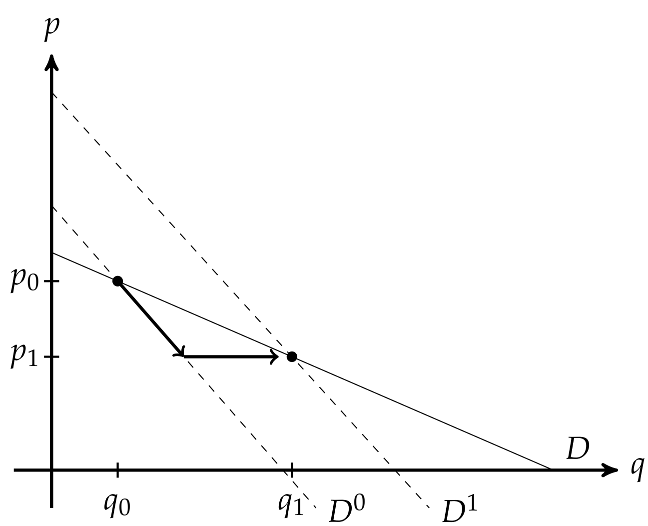


FIGURE 2: Network effects: Demand (D) vs. willingness-to-pay (D_0, D_1)

Therefore, the relevant curve is the aggregate demand curve, D. The bench-

[13] To measure consumer surplus, we consider the ex ante expected utility for consumer i, $\mathbb{E}\left[\max\{u_{ij}(n_{ij}, \boldsymbol{X}_{ij}), 0\}\right] = \log(1 + \exp\{\beta n_{ij} + \boldsymbol{X}_{ij}'\boldsymbol{\theta}\})$.

mark model in table 3 gives us the appropriate price elasticity for the Lerner equation, which is -2.21. Due to the merger, s_A has increased by about 20%. Therefore, the Lerner equation implies that prices should increase by 9%.[14]

After price changes are taken into account, consumer surplus increases to 12.4 million USD after the merger. That is, consumer surplus increases by 0.87%, which implies the merger should be allowed.

Importantly, a policymaker who ignores network effects would block the merger; the merged demand would have been predicted as the static union of demands, instead of allowing the network effect to increase demand by a further .23%. After accounting for the upward pricing pressure, but without network externalities, the policymaker predicts a consumer surplus change of -6.10%.

5.2 Promotional giveaway

Consider giving a game for free to the top 1% most popular consumers as a way to promote the game. Because of network effects, promoting the consumption among some "influencers" might have net gains. The top 1% accounts for about 20% of friendship connections; that is, the probability for a random member of the network to be linked with the top 1% is 20%. Because popular consumers are also more likely to buy any game, this counterfactual consists of ensuring that the top 1% has the game.

Specifically, the probability that a random member of the top 1% buys any one game is around 14%. Then, in this counterfactual, the average consumer sees its number of friends with the game increasing by 1 with a probability of $.2(1 - .14)$. Figure 3 shows the net gains from the giveaway as a function of market penetration (gains net of top 1% demands). For small games, a giveaway to the top 1% represents a big effort relative to their market share. However, for the largest game in the data, with 20% market penetration, the giveaway represents 5% of its sales but yields a 21% increase in sales. To put it in perspective, a game would have to drop its price by 9% to achieve a 20% increase in demand.

6 Concluding remarks

The main contribution of this paper is to show that, with minimal network data, network effects can be credibly estimated. My method is robust in two

[14] $20/|-2.21| = 9$.

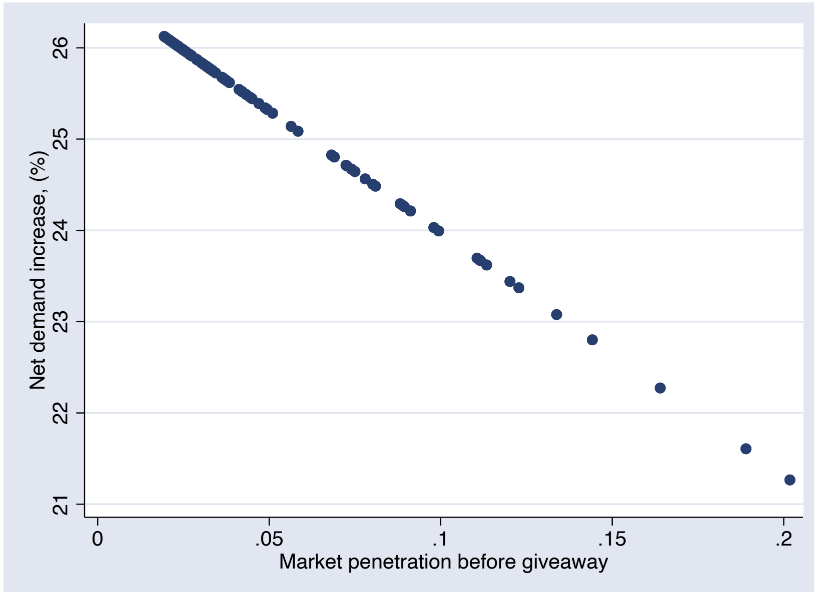


FIGURE 3: Counterfactual giveaway

important dimensions. First, it agrees internally by achieving similar estimates with different variables. Second, it aligns with other data-intensive strategies from the literature. Moreover, using a discrete choice model and several moments from the model, we can estimate more precisely the network effect.

In the case of Steam, price elasticity is about -1.85, while the network effect is about 0.29. Disentangling the price elasticity from the network effect allows managers and policy makers to separately evaluate counterfactual pricing and marketing policies. In particular, firms have strong incentives to promote the game through influencers. At the same time, firms have some incentives to merge their networks when the degree of interception between them is low. Policymakers must take network effects into account to put a sign on changes in consumer surplus.

References

Banerjee, Abhijit, Arun G. Chandrasekhar, Esther Duflo, and Matthew O. Jackson. 2013. "The diffusion of microfinance." *Science*, 341(6144).

Becker, Gary S. 1991. "A Note on Restaurant Pricing and Other Examples of Social Influences on Price." *Journal of Political Economy*, 99(5): 1109–1116.

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

Breza, Emily, and Arun G. Chandrasekhar. 2019. "Social Networks, Reputation, and Commitment: Evidence From a Savings Monitors Experiment." *Econometrica*, 87(1): 175–216.

Breza, Emily, Arun G. Chandrasekhar, Tyler H. McCormick, and Mengjie Pan. 2020. "Using Aggregated Relational Data to feasibly identify network structure without network data." *American Economic Review*, 1–28.

Chamberlain, Gary. 1980. "Analysis of Covariance with Qualitative Data." *The Review of Economic Studies*, 47(146): 225.

Chandrasekhar, Arun G., Cynthia Kinnan, and Horacio Larreguy. 2018. "Social networks as contract enforcement: Evidence from a lab experiment in the field." *American Economic Journal: Applied Economics*, 10(4): 43–78.

De Giorgi, Giacomo, Anders Frederiksen, and Luigi Pistaferri. 2019. "Consumption Network Effects." *The Review of Economic Studies*, , (May 2019): 130–163.

Dubé, Jean-Pierre H., G. J. Hitsch, and Pradeep K. Chintagunta. 2010. "Tipping and Concentration in Markets with Indirect Network Effects." *Marketing Science*, 29(2): 216–249.

Fortin, Bernard, and Vincent Boucher. 2016. "Some Challenges in the Empirics of the Effects of Networks." In *The Oxford Handbook of the Economics of Networks*. , ed. Yann Bramoullé, Andrea Galeotti and Brian Rogers, 276–302. Oxford University Press.

Hartmann, Wesley R., Puneet Manchanda, Harikesh Nair, Matthew Bothner, Peter Dodds, David Godes, Kartik Hosanagar, and Catherine Tucker. 2008. "Modeling social interactions: Identification, empirical methods and policy implications." *Marketing Letters*, 19(3-4): 287–304.

Lancaster, Tony. 2000. "The incidental parameter problem since 1948." *Journal of Econometrics*, 95(2): 391–413.

Lee, Robin S. 2013. "Vertical Integration and Exclusivity in Two-Sided Markets." *The American Economic Review*, 103(7): 2960–3000.

Manski, Charles F. 1993. "Identification of Endogenous Social Effects: The Reflection Problem." *The Review of Economic Studies*, 60(3): 531.

Nair, Harikesh. 2007. "Intertemporal price discrimination with forward-looking consumers: Application to the US market for console video-games." *Quantitative Marketing and Economics*, 5(3): 239–292.

O'Neill, Mark, Elham Vaziripour, Justin Wu, and Daniel Zappala. 2016. "Condensing Steam: Distilling the Diversity of Gamer Behavior." 81–95. New York, New York, USA:ACM Press.

Petrin, Amil, and Kenneth Train. 2010. "A Control Function Approach to Endogeneity in Consumer Choice Models." *Journal of Marketing Research*, XLVI(February): 3–13.

Ryan, Stephen P., and Catherine Tucker. 2012. *Heterogeneity and the dynamics of technology adoption.* Vol. 10.

Rysman, Marc. 2019. "The reflection problem in network effect estimation." *Journal of Economics and Management Strategy*, 28(1): 153–158.