Identifying Demand with Network Externalities[*]

Michael Rolland Sullivan

University of British Columbia

September 14, 2026

Please find the Online Appendix here.

Abstract

This article characterizes the identifiability of BLP-style demand models with network externalities. Market-level data are generally insufficient for identification because product characteristics cannot be varied independently of market shares to identify substitution patterns separately from network externalities. Micro data permit identification: within-market variation in consumer characteristics identifies substitution patterns holding market shares fixed, and across-market variation in the distribution of these characteristics identifies network externalities under certain conditions. I apply the identification logic in estimating demand for dating websites. In counterfactual analysis, I find that website mergers may benefit consumers when they make websites interoperable.

[*]Email address: michael.sullivan@sauder.ubc.ca. I thank Adam Dearing, Phil Haile, Steven Berry, Katja Seim, Mitsuru Igami, and Judy Chevalier as well as seminar participants at Yale, the Competition Bureau Canada, the International Industrial Organization Conference, and the University of British Columbia for helpful feedback. I additionally thank the Tobin Center for Economic Policy at Yale University and the Yale School of Management for their support in acquiring data used in this project. This article draws on research supported by the Social Sciences and Humanities Research Council.

1 Introduction

Examples abound of markets in which a product's appeal to a consumer depends on its popularity among other consumers. Users adopt messaging applications to communicate with others; individuals wear particular brands because they are fashionable among their peers; and people use face masks in part due to social pressure generated by their widespread uptake. In each case, a product becomes more attractive as others adopt. In other settings, popularity may reduce a product's appeal: fishers may avoid crowded fishing locations and commuters may dislike oversubscribed transit modes. In either case, consumer choice depends directly on the behaviour of other consumers. These pervasive effects are typically called *network externalities* or *network effects*.

Network externalities matter for market structure, regulation, and welfare. Mergers that combining separate networks, for instance, may generate surplus by expanding the set of users with whom each user can interact. Yet, by increasing willingness to pay for the merged network, such mergers may induce large price increases. Negative network effects likewise matter for market structure: the desirability of entry depends on the extent to which entry mitigates congestion. Network externalities also provide a rationale for interoperability mandates requiring competing platforms to let users interact across platforms, as in the European Union's Digital Markets Act.

Furthermore, accounting for network externalities is important for correctly characterizing other market features. Across-market variation in market structure driven by differences in fundamentals such as cost or quality may be amplified by positive network effects, which induce herding on popular alternatives. Failing to account for network externalities may thus lead the researcher to overstate the role of fundamentals in explaining across-market differences.

This article studies the implications of network externalities for the identification of differentiated-product demand models of the type pioneered by Berry et al. (1995), commonly called BLP. I show that with only market-level data, network effects are generally not separately identified from direct responses to product characteristics, but that micro data on consumer choices and characteristics can restore identification under suitable exclusion restrictions. I then apply this logic to the US dating websites industry, wherein I estimate a model of demand with network externalities that I use to analyze market structure in the industry.

Network externalities complicate the identification of demand because, when they are present, changes in product characteristics affect consumer choices through two distinct channels. First, changes in characteristics directly affect choice, reflecting consumers' intrinsic valuation of these characteristics. Additionally, network externalities give rise to indirect feedback effects: consumers attracted to a product by changes in its characteristics raise (or, if network externalities are negative, reduce) the product's appeal to others, thus amplifying (respectively, attenuating) the direct effect of the change in characteristics. With only market-level data on product characteristics and sales, these two channels are generally inseparable. This identification failure reflects that product characteristics cannot be varied independently of market shares.

I show, however, that the reduced-form mapping from product characteristics to market shares conditional on the observed market structure is identified with market-level data under the standard instrumental variable conditions of Berry and Haile (2014). This mapping captures quantity responses to product characteristics owing to both intrinsic tastes for product characteristics and network externalities without distinguishing between these two mechanisms. Although this mapping cannot be used to perform welfare analysis or to simulate counterfactuals involving changes in market structure, it can be used to measure markups and perform pricing-related counterfactuals that hold the observed market structure fixed.

One way to achieve identification with market data is to specify network externalities as a function of quantities (i.e., market size times market shares) as opposed to shares. Identification in this case relies on variation in market size, which controls the magnitude of network externalities, and relies on the strong assumption that market size does not directly enter demand.

The availability of micro data linking individual consumers' choices to their characteristics improves the prospects for identification. Variation in consumer characteristics within and across markets conveys distinct information about (i) the direct effect of consumer characteristics on choice probabilities and (ii) the strength of network externalities. The direct effect is identified from within-market variation in characteristics holding the network fixed. For example, within a given city, consumers with faster internet connections may be more likely to use a dating website than otherwise similar consumers with slower connections. Cross-market variation in the distribution of consumer characteristics, by contrast, affects product usage through two channels. First, it generates the same direct effect identified from within-market variation. Second, it produces an indirect effect operating through network externalities; when a larger share of consumers in a city has fast internet, for instance, overall usage of the dating website increases, which in turn raises the product's appeal to other consumers. Once within-market variation identifies the direct effect, cross-market variation identifies the indirect network effect.

This argument adapts the micro data identification analysis of Berry and Haile (2024) for the case in which market shares rather than prices are the endogenous variables of interest. I show that when market shares directly affect consumer choice, variation across markets in the distribution of consumer characteristics is generally necessary for identification. The identification argument relies on instrumental variables that shift market shares without directly affecting tastes for products, conditional on the observables entering consumer preferences. The first class of instruments consists of functions of market-level distributions of consumer characteristics, such as the share of consumers in a city with high-speed internet access. The validity of these *Waldfogel instruments* requires that market-level consumer characteristics neither enter preferences directly, conditional on the consumer's own characteristics, nor correlate with unobserved product quality. Other available instruments include the characteristics of competing products within the market, i.e., the *BLP instruments* of Berry et al. (1995), and—in the version of the model in which network externalities depend on total usership rather than market shares—market size.

The role of instruments here relates to the finding of Berry and Haile (2024) that micro data eliminate the need for instruments in identifying substitution patterns, which within-market variation identifies directly. Given that they are not required for identifying substitution patterns, the instruments can be used to identify network externalities, which—like substitution patterns—are encoded in the relationship between market shares and utility.

The identification framework accommodates demographic heterogeneity: distinct demographic groups may have distinct preferences for the composition of each product's user base. For example, men and women on a dating platform may place different values on the participation of each group, so that the framework permits both within-group congestion (e.g., aversion to competition from similar users) and across-group complementarities (e.g., one gender's value for the participation of the other). The same structure extends naturally to two-sided platforms in which one side consists of buyers and the other of sellers, with each side's utility depending on the participation of the opposite side. In such two-sided settings, particularly powerful instruments arise from variables that affect one side's utility from the platform directly but affect the other side's utility only indirectly through network externalities.

To illustrate the identification logic and how network externalities shape the analysis of market competition, I empirically study the US dating websites industry in 2007–2008. The industry witnessed considerable consolidation in the decade following this time period, with market leader Match acquiring competitors OKCupid in 2011 and Plenty of Fish in 2015, raising questions about the implications of dating website mergers.

I study demand for dating websites using online browsing micro data that provide consumer locations, characteristics, and internet browsing records. Using these data and Waldfogel instruments, I estimate a demand model with network externalities. The estimates suggest that network externalities are substantial, with an inframarginal user of a site valuing a 10% increase in the site's usership at $5.32 per three-month subscription period, about 9% of the price. Moreover, the estimates provide evidence of age-based homophily in dating website choice.

Using the estimated model, I analyze sources of variation in dating websites' market shares across geography. Although dating websites offered the same user interface across the US, their market shares varied significantly across cities. I find that network externalities account for about half of this variation. They amplify differences in sites' underlying local appeal, reinforcing the competitive positions of the leading sites. Network externalities also raise market concentration: removing them lowers the mean Herfindahl–Hirschman index across markets from about 4200 to 3000.

Last, I assess the welfare effects of consolidation by simulating counterfactual regimes that vary whether the leading paid websites, `eharmony.com` and `match.com`, set prices jointly and whether they are interoperable, i.e., whether users of one site can interact with users of the other. The results suggest that the welfare effects of consolidation depend critically on this distinction. When interoperability is absent, common ownership raises prices and reduces consumer welfare, as in a standard horizontal merger. With interoperability, however, common ownership lowers prices and raises consumer welfare because the merged firm internalizes cross-site network spillovers. These results suggest network externalities alter the typical logic of horizontal mergers and, depending on the structure of a merger, may make mergers beneficial to consumers.

Related literature

The article's primary contribution is to extend the identification analysis of network interactions to the canonical BLP differentiated products demand model. In doing so, I draw upon insights from the literatures on peer effects and social interactions as well as those from the literature on the identification of demand systems (Berry et al. 2013, Berry and Haile 2014, Berry and Haile 2016, Berry and Haile 2024, Dearing 2026). In particular, I adapt the argument of Berry and Haile (2024) in establishing identification with micro data to reflect unique challenges that arise when market shares are the endogenous variables of interest. In related contemporaneous work, Dearing (2026) characterizes the identifying power of various sorts of micro data. Whereas that article focuses on the extent to which micro data relax conditions required for identification, I focus on the extent to which they permit the introduction of a new model element, namely network externalities. Both articles indicate the centrality of variation in consumer characteristics across markets in identification.

One literature analyzing network interactions is that on peer effects (Manski 1993; Graham 2008; Graham 2018; Bramoullé et al. 2009), which studies how peers affect members of a group. Peer effects models typically feature individual covariates whose effects on group outcomes are amplified by peer effects as measured by a *social multiplier* (Becker et al. 2000; Carrell et al. 2008). Glaeser et al. (2003) and Angrist (2014) represent this multiplier as the ratio of two regression coefficients: the numerator from a regression of group-mean outcomes on group-mean covariates, and the denominator from a regression of individual outcomes on individual covariates; I similarly argue that network externalities may be identified by combining across- and within-market variation. One contribution is to extend the peer effects literature's insights to the BLP model.

My work is closely related to that analyzing discrete choice models of social interactions with a finite number of agents (Brock and Durlauf 2001a; Brock and Durlauf 2001b; Brock and Durlauf 2007; Durlauf and Ioannides 2010; Aradillas-López 2021; Blume et al. 2015; Bhattacharya et al. 2024). My article instead adopts a framework closer to differentiated-products demand models in industrial organization; relative to the social interactions literature, I analyze a continuum of consumers; compare what is identified with market-level data to what is identified with micro data; allow for unrestricted BLP-type product-market unobservables; and assess the suitability of various instruments.

The article additionally relates to the empirical literature on consumer choice with network externalities, including Timmins and Murdock (2007), Bayer and Timmins (2007), Bayer et al. (2004), Guiteras et al. (2025), and Allende (2021). Much of the literature considers two-sided markets with indirect network externalities (e.g., Rysman 2004; Ackerberg and Gowrisankaran 2006; Dubé et al. 2010; Farronato et al. 2024; Chatterjee et al. 2026). I focus on direct network externalities in demand in which—as argued by Rysman (2019)—identification is especially challenging, although my framework applies to indirect network externalities in two-sided markets as well. Last, my article relates to the literature on online dating. Whereas this literature has focused on within-platform activity (e.g., Hitsch et al. 2010a; Fong 2024), I characterize how network externalities shape consumer choice among dating websites.

2 Model

This section proposes a semi-nonparametric model of demand with network externalities that extends the model of Berry and Haile (2024). The model features markets t, each with a measure M_t continuum of consumers. Each consumer chooses between J products and an outside option. Let $\mathcal{J}$ denote the set of products *excluding* the outside option. Consumers i choose between alternatives j in a market t based on observable individual characteristics W_{it}, observable market characteristics X_t (including market size M_t), products' unobserved demand shifters Ξ_{jt}, and the size of products' networks. The two measures of network size that I consider are market shares S_{jt} and total quantities $M_t S_{jt}$, the latter of which capture not only the relative popularity of a product but also the overall size of the market. As I will discuss, the choice of the network size measure is likely to depend on details of the setting. I generally denote $\{\zeta_{jt}\}_{j\in\mathcal{J}}$ by ζ_t when I have defined random variables ζ_{jt} for each $j \in \mathcal{J}$. I will later consider a generalization of the model in which consumers belong to distinct demographic groups whose members have differential tastes for the demographic composition of products' networks.

Consumer i's probability of choosing product j is

$$\mathscr{s}_{ijt} = \mathscr{s}_j(C_{it}),$$

where $\mathscr{s}_j$ is the choice probability function and $C_{it} = \{W_{it}, S_t, X_t, \Xi_t\}$ is consumer i's choice environment (i.e., the variables affecting choice). I additionally use $\mathscr{s}$ to describe the $\mathbb{R}^J$-valued function that provides choice probabilities for each product.

Table 1: Summary of notation

Symbol	Description	In market data?	In micro data?
S_t	Market shares	✓	✓
X_t	Observed market characteristics	✓	✓
Ξ_t	Unobserved market characteristics	✗	✗
$\mathscr{s}_{it}$	Consumer choice probabilities	✗	✓
W_{it}	Consumer characteristics	✗	✓

Note that, with market data, the researcher observes only market shares S_t and characteristics X_t. The object of identification in this case is the average choice probability function

$$\bar{\mathscr{s}}_t(s, x, \xi) = \mathbb{E}\left[\mathscr{s}(C_{it}) \mid S_t = s, X_t = x, \Xi_t = \xi, t\right], \qquad (1)$$

which averages over the consumer characteristics W_{it}. The conditioning on t and the t subscript on $\bar{\mathscr{z}}_t$ reflect that the distribution of consumer characteristics may vary across markets. In what follows, I use F_t^W to denote the distribution of W_{it} in market t.

In the micro data setting, the researcher additionally observes W_{it} and $\mathscr{s}_{ijt}$. The goal in this setting is to identify the choice probability function $\mathscr{s}_j$.[1] Table 1 summarizes the notation.

Demand models typically feature endogenous prices as determinants of choice. Although I largely ignore price endogeneity to focus attention on network externalities, I discuss below how identification with micro data extends to endogenous prices when distinct price instruments are available. The two sources of endogeneity (price endogeneity and network externalities) are otherwise separate problems; they interact only through the supply of instruments, as the same instruments may shift both prices and market shares.

Note that I do not consider the setting with network data (i.e., pairwise connections between economic agents) given the typical lack of network data in industrial organization applications.

Parametric example. To build intuition, I describe a parametric specification of the general model in which a consumer i selects the alternative j that maximizes the indirect utility

$$u_j(X_t, S_t, \Xi_t, W_{it}, \varepsilon_{it}) = \begin{cases} X_{jt}'\beta + f_j(S_t) + \Xi_{jt} + W_{it}'\lambda_j + \varepsilon_{ijt}, & j \neq 0 \\ \varepsilon_{i0t}, & j = 0. \end{cases} \tag{2}$$

where ε_{ijt} are unobservables and f_j are functions unknown to the econometrician. One simple specification of f_j is $f_j(s) = \eta s_j$. In this case, $\eta > 0$ implies that a product is more appealing to a consumer when others choose the same product whereas $\eta < 0$ implies the opposite.

The choice probability function in this model is, for each j,

$$\mathscr{s}_j(s, x, \xi, w) = \Pr\left(j = \arg\max_k u_k(X_t, S_t, \Xi_t, W_{it}, \varepsilon_{it}) \mid S_t = s, X_t = x, \Xi_t = \xi, W_{it} = w\right).$$

Although the model features a choice of one alternative, it can accommodate multi-homing, i.e., consumers adopting multiple platforms. To see how, suppose that a consumer chooses whether to join each of two platforms, and that the consumer receives utility proportional to the share of consumers available through these platforms. Then, I set $J = 3$, letting the final alternative $j = 3$ denote the option of joining both platforms. A natural specification of network externalities here is

[1] Dearing (2026) shows that panel data—richer than the cross-sectional micro data studied here—identify preference heterogeneity in the standard BLP model absent across-market demographic variation. Whether such data could similarly reduce the reliance on demographic variation under network externalities is an open question.

$f_1(s) = \eta \times (s_1 + s_3)$, $f_2(s) = \eta \times (s_2 + s_3)$, and $f_3(s) = \eta \times (s_1 + s_2 + s_3)$. This specification reflects that, e.g., a consumer on platform 1 can interact with single-homing consumers on platform 1 (s_1) and multi-homing consumers on both platforms (s_3).

In standard discrete choice demand models, market shares are simply choice probabilities averaged across consumers. With network externalities, they are instead fixed points of

$$S_t = \bar{\mathfrak{s}}_t(S_t, X_t, \Xi_t). \tag{3}$$

Market shares solving (3) exist by Brouwer's theorem when $\bar{\mathfrak{s}}_t$ is continuous in S_t. The solution may not be unique; the identification results below do not require uniqueness.[2]

Shares and quantities models of network externalities. The model described so far is general enough to nest variants in which the appeal of products to consumers depends on market shares S_t or on total quantities $M_t S_t$. To achieve progress on identification, however, it is necessary to impose additional structure on the model and, in particular, to choose between the shares model and the quantities model of network externalities. The appropriate choice depends on the setting. For instance, if consumers' social circles are of similar sizes across markets with different populations, and consumers consider how many of their social contacts use various messaging applications in choosing an application, then the shares model would likely be appropriate for studying choice of messaging application.

In other applications, the quantities model is appropriate. Consider, for example, a model of transit mode choice with two markets being defined as peak travel times and off-peak travel times, the former of which has a larger market size (number of commuters) than the latter. If commuters have tastes over the total number of travellers on each mode—which determines congestion—then the quantities model would better describe preferences.

One disadvantage of the quantities model is that quantities are sensitive to market definition, which is often chosen somewhat arbitrarily in applications. It is unclear, for example, whether network externalities operate within a neighbourhood, municipality, or metro area in the dating websites setting. Market definition directly determines quantities, whereas its effect on shares is limited by the fact that shares reflect relative rather than absolute popularity; this suggests that quantities

[2]With that said, non-uniqueness may complicate counterfactual analysis. When $I - D_s\bar{\mathfrak{s}}$ is nonsingular at the realized equilibrium, however, the implicit function theorem delivers a unique local mapping from (X_t, Ξ_t) to market shares, permitting coherent analysis of marginal changes (Bayer et al., 2004).

models are less robust to the choice of market definition.

3 Identification

Before formally analyzing the general model, I clarify the object of identification and discuss the primary identification problems in the context of a simple parametric model.

Object of identification

Above, I set the goal of identifying the average choice probability function $\bar{\mathfrak{z}}_t$ in the market data setting and the individual choice probability function $\mathfrak{z}$ in the micro data setting. An alternative object of identification is the reduced-form mapping from market characteristics $\{X_t, \Xi_t\}$ to shares S_t within a given market t, a mapping that is implicitly defined by the fixed-point condition (3).[3] As I will show, this reduced-form mapping may be identified even when the underlying choice function $\bar{\mathfrak{z}}_t$ is not. In that case, the overall effect of X_t on equilibrium demand is pinned down, but its decomposition into direct effects of X_t and feedback through network externalities is not. I pursue the separate identification of network externalities because it is required for welfare analysis and counterfactuals that alter market structure. However, the reduced-form mapping is sometimes enough. In the case in which X_t includes prices, for example, the reduced-form mapping is enough for computing price elasticities that can be used in assessing market power. Together with a pricing model, the reduced-form mapping can also be used to recover firm costs and markups and to simulate counterfactuals holding market structure fixed, such as price regulation and taxes.

Illustrative model

Consider a market t in which consumers choose whether to purchase a single product, where $Y_{it} = \mathbb{1}\{U_{it} \geq 0\}$ denotes an indicator for whether consumer i buys the product. Indirect utility U_{it} is specified as

$$U_{it} = X_t + \eta S_t + \Xi_t - \varepsilon_{it}.$$

Here, X_t is an exogenous characteristic of the product, S_t is the product's market share, and Ξ_t is an unobserved demand shifter in market t. Additionally, ε_{it} is consumer i's idiosyncratic taste for the product; assume that ε_{it} is iid according to the strictly increasing and continuous distribution function $F : \mathbb{R} \to [0, 1]$. I fix the coefficient of X_t at one as a scale normalization. The purchase

[3] This mapping is generally a correspondence given that equation (3) may have multiple fixed points.

probability conditional on S_t, X_t, and Ξ_t is

$$\int \mathbb{1}\{X_t + \eta S_t + \Xi_t - \varepsilon_{it} \geq 0\} dF(\varepsilon_{it}) = F(X_t + \eta S_t + \Xi_t).$$

Imposing that the market share equals the consumer's purchase probability yields

$$S_t = F(X_t + \eta S_t + \Xi_t)$$

or, inverting the function F,

$$F^{-1}(S_t) = X_t + \eta S_t + \Xi_t. \tag{4}$$

Consider now the market data setting, in which only S_t and X_t are observed. The true model primitives $\theta = (F, \eta)$ are observationally equivalent to the alternative primitives $\tilde{\theta} = (G_\Delta, \eta + \Delta)$ for any $\Delta \geq 0$ when $G_\Delta^{-1}(s) = F^{-1}(s) + \Delta s$. As Figure 1 illustrates, increasing Δ raises the dispersion of G_Δ.[4] This dampens the effect of X_t on S_t. This observational equivalence result reflects that the data cannot distinguish between two explanations for the relationship between X_t and S_t: (i) sales rise with X_t solely because consumers like X_t (a direct effect of X_t that is relatively large when Δ is low) and (ii) sales also rise in part because consumers are attracted by the increase in users induced by the increase in X_t (an indirect effect owing to network externalities that is relatively large when Δ is high).

Although market data generally cannot identify models in which market shares S_t directly enter preferences, they may be able to identify models in which consumers instead value quantities $M_t S_t$. To illustrate, suppose indirect utility is instead given by

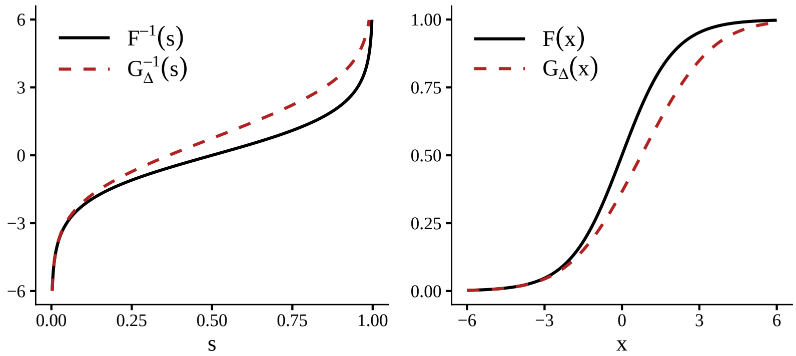
$$U_{it} = X_t + \eta M_t S_t + \Xi_t - \varepsilon_{it}, \tag{5}$$

where M_t denotes the population of market t. Absent network externalities ($\eta = 0$) and assuming $\mathbb{E}[\Xi_t \mid M_t, X_t] = 0$, population size M_t does not directly affect market shares conditional on observables. When $\eta \neq 0$, however, variation in M_t shifts market shares, allowing M_t to be used as an instrument for $M_t S_t$ to identify network externalities. However, this approach relies on the strong restriction that population affects utility only through the network term $M_t S_t$.

Micro data help solve the identification problem that product characteristics and market shares

[4]It is straightforward to verify that G_Δ is a valid distribution function; Online Appendix O.1 records the argument.

Figure 1: Illustration of G_Δ



cannot be independently varied. To see why, suppose that the observed individual characteristic W_{it} is a component of the idiosyncratic tastes: $\varepsilon_{it} = W_{it} + \tilde{\varepsilon}_{it}$. In the dating website setting, W_{it} could be an inverse measure of internet speed. Suppose additionally that the remaining unobservable aspect of idiosyncratic tastes $\tilde{\varepsilon}_{it}$ is distributed according to $\tilde{F}$ independently of W_{it} so that

$$U_{it} = \underbrace{\beta X_t + \eta S_t + \Xi_t}_{=:\delta_t} - W_{it} - \tilde{\varepsilon}_{it}.$$

Here, δ_t is a market-level utility index common to all consumers in market t. Recalling that Y_{it} is an indicator for consumer i's usage of the dating website,

$$\Pr_t\left(Y_{it} = 1 \mid W_{it} = w\right) = \Pr_t\left(U_{it} \geq 0 \mid W_{it} = w\right) = \Pr_t\left(\tilde{\varepsilon}_{it} \leq \delta_t - w\right) = \tilde{F}(\delta_t - w), \tag{6}$$

where $\Pr_t$ indicates conditioning on market t's characteristics. The left-hand side is observable and the final expression on the right-hand side is the distribution function of $\delta_t - \varepsilon_{it}$. When W_{it} has sufficiently large support, variation in W_{it} traces out the entire function $z \mapsto \tilde{F}(z)$ up to a location shift. With a location normalization (e.g. $\mathbb{E}_t[\tilde{\varepsilon}_{it}] = 0$), knowledge of this distribution separately identifies δ_t and $\tilde{F}$.[5]

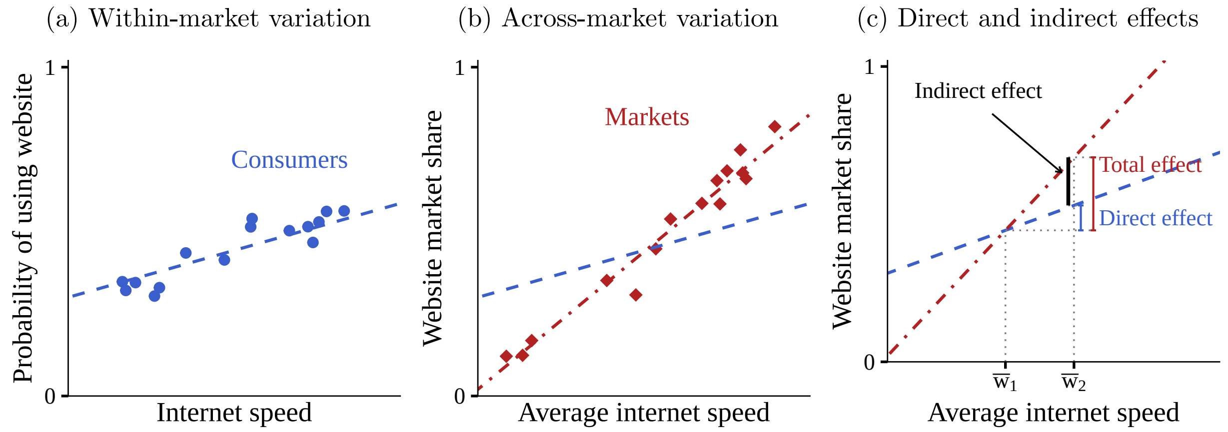
The remaining task of identification is the separate identification of the components of the utility index δ_t. The primary challenge in completing this task is the fact that the regressor of interest S_t and the unobservable Ξ_t are mechanically dependent. Instrumental variables provide a solution to this problem. The available instruments Z_t here are based on market-specific distributions F_t^W of W_{it}, distributions that shift market shares given the effect of W_{it} on individual choice probabilities.[6] Instruments based on market-specific demographic distributions are often called

[5] Although the assumption of a large support simplifies the argument, it is not essential; I do not make a large support assumption in the semi-nonparametric identification analysis.

[6] This logic is suggested in Section 7.1.2 of Jullien et al. (2021).

Waldfogel instruments, following the insights of Waldfogel (2003) and Waldfogel (2008) that local demographic composition shapes product availability and choice sets (Berry and Haile 2016). An example of a Waldfogel instrument in the dating websites example would be average internet speed. I discuss these instruments in detail below.

Figure 2: Illustration of identification with micro data



I now provide a heuristic example to build intuition. Suppose that, within a market t, consumers with faster internet are more likely to use a dating website. Figure 2a, in which consumers are represented by blue circles, expresses this relationship. Based on the observed within-market relationship between internet speed and website usage, one can predict how dating website usage would increase in t if internet speeds rose for all consumers in the market. Under positive network externalities (i.e., $\eta > 0$), this would be an underprediction: it would capture the direct effect of internet speed as identified using within-market variation, but not network externalities. Figure 2b represents this scenario; in the figure, each red diamond is a market. Last, as illustrated by Figure 2c, the gap between the predicted and actual change in adoption reflects the strength of network externalities.

To connect this example to the identification argument, within-market variation identifies the relationship between W_{it} (internet speed) and choice probabilities holding market shares fixed as shown by (6). Raising average internet speeds shifts the distribution of W_{it} across markets. This cross-market variation identifies the extent of network externalities.

Identification with market data

I now formalize the identification arguments developed above, first in the context of the shares model. Consider identification under an index structure closely related to that of Berry and Haile

(2014), a structure under which the average choice probability function takes the form

$$
\begin{aligned}
\bar{\mathscr{s}}(S_t, X_t, \Xi_t) &= \bar{\sigma}(\delta(S_t, X_t, \Xi_t)) \\
\delta_j(S_t, X_t, \Xi_t) &= X_{jt} + h_j(S_t) + \Xi_{jt},
\end{aligned}
\tag{7}
$$

where $X_t = [X_{1t}, \ldots, X_{Jt}]'$ is a vector of observable product characteristics and S_t denotes the vector of market shares. Relative to Berry and Haile (2014), the index restriction (7) additionally imposes that market shares enter demand through an additive index. I make this restriction to establish that the model is generally not identified even in a favourable case with greater structure.[7] Note that the imposition of a coefficient of one on X_{jt} is a scale normalization. As a location normalization, I also impose $\mathbb{E}[\Xi_{jt}] = 0$ for each j. I slightly abuse notation in (7) by writing $\bar{\sigma}$ both as a function of $\{S_t, X_t, \Xi_t\}$ and as a function of the index alone. In practice, I seek to identify $\bar{\sigma}$ as a function of the index and to identify h_j.

To make my identification analysis comparable to that of Berry and Haile (2014), I make an adapted version of their demand invertibility assumption.

Assumption INVERT-MARKET (Invertible demand — market data)**.** The function $\bar{\sigma}$ is injective on the support of $\delta(S_t, X_t, \Xi_t)$.

This assumption implies that $\bar{\sigma}$ admits a well-defined inverse $\bar{\sigma}^{-1}$. Invertibility generally requires a minimal degree of substitutability across products (Berry et al. 2013).

The following proposition establishes the non-identification of $(\bar{\sigma}, h)$ with market data.

Proposition 1. *Fix some model primitives $\theta \coloneqq (\bar{\sigma},\ h)$. For any function $\tilde{\sigma}$ satisfying Assumption INVERT-MARKET, there is a function $\tilde{h}$ such that $\tilde{\theta} = (\tilde{\sigma},\ \tilde{h})$ is observationally equivalent to θ.*

Proof. The primitives $(\bar{\sigma},\ h)$ are consistent with the observable data if and only if

$$
\bar{\sigma}^{-1}(S_t) = X_t + h(S_t) + \Xi_t. \tag{8}
$$

Define $\tilde{h}(S_t) = h(S_t) + \tilde{\sigma}^{-1}(S_t) - \bar{\sigma}^{-1}(S_t)$. Then,

$$
\tilde{\sigma}^{-1}(S_t) = X_t + \tilde{h}(S_t) + \Xi_t.
$$

[7] This approach allows for product characteristics other than the X_{jt} that are required to satisfy the index restriction of (7); as Berry and Haile (2014) suggest, the researcher can condition on exogenous characteristics $x_t^{(2)}$ and suppress them in the notation.

Thus, $\tilde{\theta} = (\tilde{\sigma}, \tilde{h})$ is observationally equivalent to θ. $\square$

Proposition 1 shows that substitution patterns (encoded in $\bar{\sigma}$) and network externalities (encoded in h) cannot be disentangled using market data alone. This is because market shares enter the inverse demand equation both through $\bar{\sigma}^{-1}$ and through h. This non-identification persists even if S_t were exogenous, which is generally implausible. Similarly, the identification problem does not relate to the unavailability of instruments.

The reduced-form mapping from (X_t, Ξ_t) to S_t may be identified even when the primitives $\bar{\sigma}$ and h are not. Identification requires instruments Z_{jt} that satisfy the nonparametric instrumental variables conditions of Newey and Powell (2003), as provided below in Assumption IV-MARKET. These assumptions require that the instruments do not directly affect choice via Ξ_{jt} (exclusion) and shift the endogenous variables (completeness). The only instruments Z_t for market shares S_t consistent with the model are the BLP instruments X_t, which are discussed extensively in Berry and Haile (2014): absent a variable that shifts which of multiple equilibria is selected, shares are determined by (X_t, Ξ_t) alone. Thus, an excluded variable can shift S_t only through X_t.

Assumption IV-MARKET (Instruments for market data setting)**.** Let V_t denote regressors of the model variant under consideration: $V_t = S_t$ in the shares model and $V_t = (S_t, M_t)$ in the quantities model. There is an observable random vector Z_t satisfying

(i) Exclusion restriction: $\mathbb{E}[\Xi_{jt}|Z_t] = 0$ (almost surely), for all j.

(ii) Completeness condition: for all real-valued functions Ψ such that $\mathbb{E}|\Psi(V_t)| < \infty$, $\mathbb{E}[\Psi(V_t)|Z_t] = 0$ (almost surely) implies $\Psi(V_t) = 0$ (almost surely).

Proposition 2 below establishes identification of the reduced form given the availability of valid instruments. Note that the proposition uses the term *observable market shares* S_t. I introduce this term as the fixed point (3) may have multiple solutions for a particular (X_t, Ξ_t), but, due to unmodelled and possibly stochastic equilibrium selection mechanisms, it is possible that not all of these solutions are observable. By *observable market shares* S_t, I mean the fixed points of (3) that may be selected by the unmodelled equilibrium selection mechanism.

Proposition 2. *Suppose that Assumptions INVERT-MARKET and IV-MARKET hold. Then, the correspondence from* (X_t, Ξ_t) *to the set of observable market shares* S_t *satisfying equation (3) is identified.*

Online Appendix O.2 provides the proof.

Identification may be achieved by assuming that $\bar{\sigma}$ is known, a situation that arises when the researcher specifies a distribution of the unobservable ε_{ijt} in (2) without any unknown parameters (e.g., when the researcher specifies a multinomial logit demand system without random coefficients or nesting parameters). Suppose that, in addition to $\bar{\sigma}$ being known, the network externalities applying to product j depend only on product j's market share: in an abuse of notation, $h_j(S_t) = h_j(S_{jt})$. In this case, the inverse demand equation (8) becomes

$$\bar{\sigma}_j^{-1}(S_t) = \beta_j X_{jt} + h_j(S_{jt}) + \Xi_{jt}, \quad j \in \mathcal{J} \tag{9}$$

I introduce the coefficients β_j because setting the coefficient of each x_{jt} to one is no longer a scale normalization when $\bar{\sigma}$ is known. Equation (9) contains S_{jt} as an endogenous regressor, which may be instrumented for by the characteristics X_{-jt} of products other than j, as long as $J > 1$ (thus ruling out binary choice) and these characteristics are assumed to be exogenous. The possibility of identification when $\bar{\sigma}$ is known establishes that an arbitrary functional form assumption rather than the data is pivotal for identification. This suggests the unreliability of estimates obtained using market data with inflexible distributional assumptions on ε_{ijt}.

Quantities model. Now consider identification of the model variant in which network externalities operate through quantities $M_t S_t$ rather than shares S_t. Although the quantities model suffers from similar identification problems as the market shares model, cross-market variation in market size M_t facilitates identification when this variation is assumed to be exogenous.[8]

I study identification of this model under the following index restrictions:

$$\begin{aligned} \bar{s}_{jt}(M_t S_t, X_t, \Xi_t) &= \bar{\sigma}_j(\delta(M_t S_t, X_t, \Xi_t)) \\ \delta_j(M_t S_t, X_t, \Xi_t) &= X_{jt} + h_j(M_t S_t) + \Xi_{jt}. \end{aligned} \tag{10}$$

It is generally not possible to identify the model under a slightly relaxed index restriction wherein $M_t S_t$ need not affect consumer choice through the additively separable δ index, unless $J = 1$; I explore this case in Online Appendix O.3. The restriction that quantities must enter through the linear index δ, thus requiring that they impact choice in the same manner as the vertical characteristic X_{jt}, is substantial: it generally rules out heterogeneity in network externalities

[8] This result relates to that of Lee (2007), who studies a linear model of social interactions that requires variation in group size for identification.

across consumers. As in the analysis of the shares model, I assume that the function $\bar{\sigma}$ mapping the index δ into choice probabilities is invertible.

Assumption INVERT-QUANTITIES (Invertible demand — quantities model)**.** The function $\bar{\sigma}$ is injective on the support of $\delta(M_t S_t, X_t, \Xi_t)$.

I additionally impose the location normalization that there is a known vector q^* such that $h_j(q^*) = 0$ for all j. When considering Assumption INVERT-MARKET in the context of the quantities model, I use $\bar{\sigma}$ to denote the function of the index δ that appears on the right-hand side of the first equation in (10).

Identification also requires several technical conditions that permit the use of calculus in the analysis and ensure continuous variation in total quantities.

Assumption CALC-TOT (Technical conditions for total quantities model)**.** (i) h is differentiable; (ii) $\bar{\sigma}$ is differentiable; (iii) the support of $M_t S_t$ is convex.

The following proposition establishes that the assumptions above in addition to a further restriction, either that consumer tastes for a product depend only on the size of the product's own user base or that the support of M_t includes zero, yield identification.

Proposition 3. *Suppose that Assumptions INVERT-QUANTITIES, IV-MARKET, and CALC-TOT hold. Additionally suppose that at least one of the two following assumptions holds:*

(a) Own shares: for each $j \in \{1, \ldots, J\}$, $h_j(ms)$ depends only on ms_j.

(b) Large support: for all s in the support of S_t, the support of M_t conditional on $S_t = s$ includes the interval $(0, \bar{m}(s)]$ for some $\bar{m}(s) > 0$.

Then, h is identified on the support of $M_t S_t$, $\bar{\sigma}$ is identified on the support of $\delta(M_t S_t, X_t, \Xi_t)$, and Ξ_t is identified.

The proof, which appears in Online Appendix O.3, has two parts: identifying the function $\kappa_j(S_t, M_t) = \sigma_j^{-1}(S_t) - h_j(M_t S_t)$ using a nonparametric instrumental variables argument and then disentangling substitution patterns—as captured by $\sigma_j^{-1}(S_t)$—from network externalities, as captured by $h_j(M_t S_t)$. Note that κ_j has $J+1$ arguments (the J market shares and the one-dimensional market size M_t) and thus generally requires $J+1$ instruments. The available instruments are the J exogenous market characteristics X_t and market size M_t. The exogenous market characteristics are the standard BLP instruments; market size M_t becomes necessary to use as an instrument upon introducing network externalities. A major threat to identification is that consumer tastes systematically depend on market size. The required exclusion restriction is violated if, e.g., the value of the outside option varies with local population, a violation that is likely to occur in the commuting mode choice and dating websites examples. It would also fail if larger markets systematically receive greater unobserved firm investment that is reflected in Ξ_t.

Once κ_j has been identified, Proposition 3 provides two alternative ways to disentangle substitution patterns and network externalities: first, to assume that a product's user base affects the appeal of that product alone or, second, to assume that market shares close to zero are observed for each vector of observed market shares. The first assumption is reasonable and typically used in models of network externalities. It would be violated if network externalities operate across products due, e.g., to multi-homing, interoperability, or congestion spillovers (e.g., two fishing sites on the same lake may impose congestion externalities on each other).

The second assumption is helpful because sending $M_t \to 0$ permits the identification of the model excluding network externalities. Then, varying M_t at positive levels permits identification of network externalities and hence the full model. Although helpful, the assumption requires extrapolating substitution patterns in small markets to larger markets and is not tenable in empirical settings without small markets free of network externalities.

Thus, although specifying network externalities as a function of total quantities aids identification with market data, the approach requires strong assumptions. Fortunately, micro data offer a more credible way to identify network externalities.

Identification with micro data

Identification with micro data closely follows Berry and Haile (2024), with one key modification: market shares enter the individual choice probability function through network externalities. In the standard setting without network externalities, a central contribution of Berry and Haile (2024) is that micro data eliminate the need for instruments in identifying substitution patterns, which within-market variation identifies directly. The central problem in identifying network externalities with market data is in disentangling substitution patterns and network externalities. Crucially, when within-market variation identifies substitution patterns, this concern no longer applies: instruments for market shares serve solely to address the endogeneity of shares arising from network externalities, not to pin down substitution patterns.

As will become apparent, the distinction between the shares model and quantities model matters little in the micro data setting, other than that market size becomes available as an instrument under the latter. For this reason, I focus attention on the shares model.

For each market t, let $\varsigma_t(w)$ denote the vector of choice probabilities among consumers with characteristics $W_{it} = w$ in market t. Micro data reveal ς_t on the support of W_{it} in each market, along with (S_t, X_t) and the instruments Z_t introduced below. Throughout this subsection, I take W_{it} to be $\mathbb{R}^J$-valued and assume that its support conditional on $X_t = x$ is the same in every market; I denote this common support by $\mathcal{W}(x)$. The dimension requirement parallels that of Berry and Haile (2024): consumer characteristics must generate variation in the indices of all J products, and when over J consumer characteristics are available, the analysis applies after conditioning on all but J of them. The common support requirement permits comparisons of consumers with identical characteristics who face different market-level conditions. Importantly, it does not restrict the *distribution* F_t^W of W_{it} on that support, which may vary freely across markets sharing the same X_t; this variation plays a central role below.

I study identification under an index assumption analogous to that of Berry and Haile (2024), with market shares replacing prices as the endogenous variables of interest.

Assumption IND-MICRO (Index assumption for micro data case)**.** Consumer i's probability of choosing alternative j conditional on $C_{it} = \{W_{it}, S_t, X_t, \Xi_t\}$ is given by

$$\mathscr{s}_j(C_{it}) = \sigma_j(\gamma(W_{it}, X_t, \Xi_t), S_t, X_t), \quad \forall j,$$

where $\gamma(W_{it}, X_t, \Xi_t) = (\gamma_1(W_{it}, X_t, \Xi_t), \ldots, \gamma_J(W_{it}, X_t, \Xi_t))$ is $\mathbb{R}^J$-valued with[9]

$$\gamma_j(W_{it}, X_t, \Xi_t) = \Gamma_j(W_{it}, X_t) + \Xi_{jt}, \quad \forall j. \tag{11}$$

I also adapt Assumptions 2–6 of Berry and Haile (2024), which I present below.

Assumption INV-DEMAND (Injectivity of demand)**.** The function $\sigma(\cdot, s, x)$ is injective on the support of $\gamma(W_{it}, x, \Xi_t)$ for all $(s, x) \in \operatorname{supp}\{S_t, X_t\}$.

[9] Here, I use γ rather than δ to denote the index (i) to reflect that γ, unlike the δ index from the discussion of identification with market-level data, depends on consumer-level characteristics, and (ii) to follow the notation of Berry and Haile (2024).

Assumption INV-INDEX (Injectivity of index)**.** The function $\gamma(\cdot, x, \xi)$ is injective on $\operatorname{supp} W_{it} \mid X_t = x$ for all $(x, \xi) \in \operatorname{supp}\{X_t, \Xi_t\}$.

Assumption SUPP (Support)**.** For all $x \in \operatorname{supp} X_t$, the support of W_{it} conditional on $X_t = x$ is the same in every market; this common support, denoted $\mathcal{W}(x)$, is open and connected.

Assumption NOND (Nondegeneracy)**.** (a) For each $x \in \operatorname{supp} X_t$, there exists $s \in \operatorname{supp} S_t \mid X_t = x$ such that $\operatorname{supp} \Xi_t \mid S_t = s, X_t = x$ is open and connected. (b) For almost every $(s, x) \in \operatorname{supp}\{S_t, X_t\}$, the set $\operatorname{supp} \Xi_t \mid S_t = s, X_t = x$ is connected.

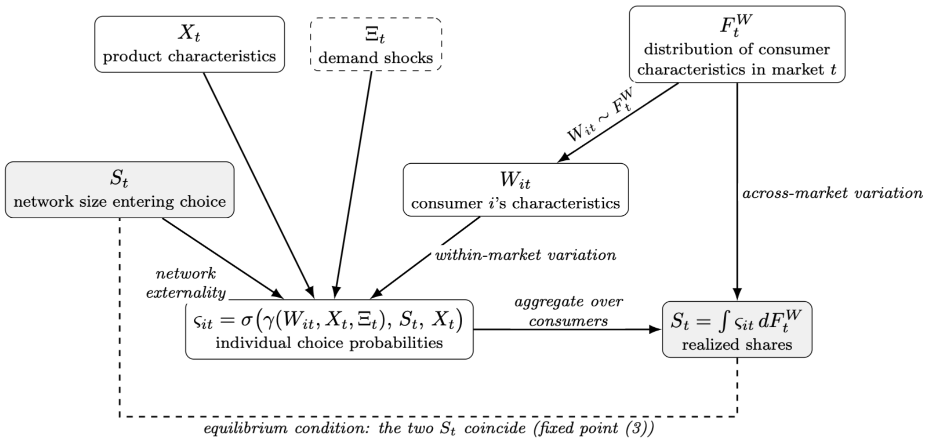
Assumption TECH (Technical conditions)**.** For all $(s, x) \in \operatorname{supp}\{S_t, X_t\}$, (i) $\Gamma(\cdot, x)$ is uniformly continuous and continuously differentiable; (ii) $\sigma(\cdot, s, x)$ is continuously differentiable; (iii) $\partial\Gamma(w, x)/\partial w$ is nonsingular almost surely on $\operatorname{supp} W_{it} \mid X_t = x$; and (iv) $\partial\sigma(\gamma, s, x)/\partial\gamma$ is nonsingular almost surely on $\operatorname{supp} \gamma(W_{it}, X_t, \Xi_t) \mid S_t = s, X_t = x$.

Assumption INV-DEMAND, a standard invertibility assumption, gives rise to an equation that equates a transformation of market shares to an index that is additively separable in a structural unobservable, thus permitting the application of instruments. Next, Assumption INV-INDEX ensures that consumer characteristics W_{it} generate sufficient variation in the index. Assumptions SUPP and NOND then provide support conditions: Assumption SUPP ensures continuous within-market variation in W_{it}, and part (a) of Assumption NOND ensures that, for a given value of X_t, there remains nontrivial cross-market variation in the latent demand shock Ξ_t after conditioning on market shares. Part (b) of Assumption NOND additionally requires that this conditional variation not split into disjoint pieces; I discuss the role of this condition, which is specific to the network externalities setting, below. Assumption TECH enables the use of calculus in the analysis. Under Assumption IND-MICRO, the observed conditional choice probability functions satisfy $\varsigma_t(w) = \sigma(\gamma(w, x_t, \xi_t), s_t, x_t)$ for all $w \in \mathcal{W}(x_t)$, where (s_t, x_t, ξ_t) denote market t's realizations of (S_t, X_t, Ξ_t); the identification question is what these observed functions, combined with instruments, reveal about σ and the demand shocks.

I adopt the location and scale normalizations of Berry and Haile (2024). Because any invertible transformation of the index can be undone by a corresponding transformation of σ, these place no restriction on choice probabilities. Online Appendix O.4 states the normalizations.

The last required assumption is that instruments are available that shift market shares without directly entering utility. Assumption IV-MICRO below parallels Assumption IV-MARKET, with

Figure 3: Causal diagram



the exclusion restriction and completeness condition now stated conditional on X_t.

Assumption IV-MICRO (Instruments for micro data case)**.** There exist instrumental variables Z_t satisfying the following assumptions: (i) $\mathbb{E}[\Xi_{jt} \mid X_t, Z_t] = 0$ almost surely for all j (exclusion restriction); and (ii) for all functions $\Psi(X_t, S_t)$ with finite expectation, $\mathbb{E}[\Psi(X_t, S_t) \mid X_t, Z_t] = 0$ almost surely implies $\Psi(X_t, S_t) = 0$ almost surely (completeness).

These assumptions are relevant exclusion and completeness conditions. The causal diagram in Figure 3 reflects these assumptions and illustrates the model structure. I discuss the available instruments following the statement of the main identification result below.

Proposition 4. *Under Assumptions IND-MICRO, INV-DEMAND, INV-INDEX, SUPP, NOND, TECH, and IV-MICRO, the realization* ξ_t *is identified for all* t*, and the choice probability function* $\mathcal{s}$ *is identified on* $\operatorname{supp} C_{it}$.

Proof. See Online Appendix O.4. □

The proof adapts the arguments of Berry and Haile (2024) to the case in which market shares, rather than prices, are the endogenous market-level variables entering choice probabilities. Here, I sketch the proof to highlight how including market shares as determinants of choice alters the argument. An important initial observation is that micro data reveal, within each market t, the mapping ς_t from consumer characteristics W_{it} to choice probabilities. This mapping does not by itself reveal Γ, which governs how consumer characteristics enter utility in the units of the demand shock Ξ_{jt}; see (11). What a single market reveals is the composition of Γ with the unknown function σ. Within-market variation cannot disentangle the two.

However, Γ and σ can be disentangled by comparing markets. The comparison involves two markets t and t' sharing $(S_t, X_t) = (S_{t'}, X_{t'}) = (s, x)$ in which different consumer characteristics $w_t(\varsigma)$ and $w_{t'}(\varsigma)$ achieve a common vector of choice probabilities ς. For different characteristics to achieve the same choice probabilities at the same (s, x), the two markets must differ in their demand shocks Ξ_t and $\Xi_{t'}$ by exactly the amount that offsets the difference in Γ across the two consumers. Part (a) of Assumption NOND, which requires continuous variation in Ξ_t conditional on (s, x), ensures that such markets exist for any two nearby consumer types.

Such pairs of markets exist only if the distribution of consumer characteristics F_t^W varies across markets. Shares are determined by X_t, Ξ_t, and F_t^W alone through the fixed point equation (3). This means that, if F_t^W were common to all markets, conditioning on $(S_t, X_t) = (s, x)$ would generally pin down Ξ_t, leaving no two markets that share (s, x) yet differ in their demand shocks.[10] This would invalidate the identification argument. A similar concern might seem to arise when (as in Berry and Haile 2024) prices rather than shares are the endogenous variables. After all, given that prices are determined in part by X_t and Ξ_t, it could be that conditioning on prices and X_t would pin down Ξ_t. However, cost-side variables excluded from demand allow Ξ_t to vary among markets with common prices and X_t. By contrast, shares have no excluded shifter, leaving F_t^W as the only source of continuous variation in Ξ_t conditional on (s, x). Figure 3 illustrates the role of F_t^W in shifting shares S_t without directly affecting consumer choice.

I now explain how the cross-market comparison aids identification. Within each market t featuring $(S_t, X_t) = (s, x)$, inverting the mapping from utility indices to choice probabilities yields

$$\Gamma(w_t(\varsigma), x) + \Xi_t = \sigma^{-1}(\varsigma; s, x). \tag{12}$$

Now consider the two markets t, t' described above with common (s, x) and within which distinct characteristics $w_t(\varsigma), w_{t'}(\varsigma)$ generate common choice probabilities ς. Given that $\sigma^{-1}(\cdot\,; s, x)$ is common across these markets, it can be eliminated after applying the inversion in (12):

$$\Gamma(w_t(\varsigma), x) + \Xi_t = \Gamma(w_{t'}(\varsigma), x) + \Xi_{t'}. \tag{13}$$

[10] Where there are multiple solutions of the fixed point (3), it may instead be the case that Ξ_t is constrained to a discrete set. However, continuous variation in Ξ_t is required, meaning that identification would still break down absent variation in F_t^W.

Differentiating with respect to the choice probability ς and rearranging terms yields

$$\left(\frac{\partial\Gamma(w_{t'}(\varsigma),x)}{\partial w}\right)^{-1}\frac{\partial\Gamma(w_t(\varsigma),x)}{\partial w}=\frac{\partial w_{t'}(\varsigma)}{\partial\varsigma}\left(\frac{\partial w_t(\varsigma)}{\partial\varsigma}\right)^{-1}. \tag{14}$$

Note that differentiating within each market removes the demand shocks Ξ_t, which are constant within a market. On the right-hand side of equation (14) appear derivatives of the form $\partial w_t(\varsigma)/\partial\varsigma$, which measure the changes in consumer characteristics W_{it} required to achieve a given change in choice probabilities; these derivatives are observable from within-market variation in W_{it}. Thus, the ratio of Γ derivatives on the left-hand side is identified. The identification of Γ involves piecing these derivatives together across consumer types and markets.

With the utility index Γ identified, it is straightforward to identify the derivative $\partial\sigma^{-1}(\varsigma;s,x)/\partial\varsigma$, which captures the relationship between choice probabilities ς (observed) and the utility index (identified). Recovering σ^{-1} itself requires integrating this derivative between distinct choice probabilities ς^0 and ς^1, holding fixed shares s and characteristics x. This requires the set of choice probabilities generated by markets sharing (s,x) to be connected, which is the role of part (b) of Assumption NOND. Network externalities make possible multiple equilibria, which threaten this condition: the same shares can be a high equilibrium for one range of demand shocks and a low equilibrium for another, with no shocks in between generating them, so that the choice probabilities generated by markets sharing (s,x) split into disconnected pieces.

Integration identifies σ^{-1} only up to a constant of integration that depends on (s,x). This constant embeds how utility depends on shares s, i.e., the network externalities. The final step of the argument uses across-market variation in the instruments to identify this constant as a function of (s,x), and with it the network externalities and the demand shocks.

As in Berry and Haile (2024), the argument can be adapted to the case of endogenous X_t. In this case, what is identified is the mapping from consumer characteristics and market shares to choice probabilities holding (X_t,Ξ_t) fixed. Also, it is straightforward to add prices or other endogenous variables as elements of C_{it} that are included as distinct arguments of σ_j (i.e., that are not restricted to enter through the γ index) as long as distinct instruments for these variables are available.

The identification argument above applies equally to the quantities model in which M_tS_t enters σ_j. It equally applies to a mixed model in which both market shares and quantities affect choice, provided sufficient instruments are available. In this case, the researcher can let the data speak to the question of whether the shares model or the quantities model is more appropriate.

Group heterogeneity and multi-sided platforms

In some settings, consumers can be divided into observably distinct groups with different preferences for the number of consumers of each other group that use a product. In the dating websites example, for instance, younger users may value the presence of other young users on a website more than that of older users. More fundamentally, a dating website's appeal is likely to depend on that user's own gender and sexual orientation as well as the number of other users on the site disaggregated by gender and sexual orientation. Another clear example of a setting in which distinct groups exert heterogeneous network externalities on each other is that of multi-sided platforms in areas including e-commerce, food delivery, and ride-hailing. In the food delivery industry, for example, restaurants may prefer platforms hosting a greater number of consumers but a lower number of rival restaurants.

To extend the analysis to such settings, I develop a model variant in which consumers belong to one of N_D demographic groups, each with its own demand system and unobservable Ξ_t. This variant allows consumers in different groups to have different tastes for members of other groups on a network. In the case of dating websites, the demographic groups could be determined based on age, gender, sexual orientation, or any other variable so long as it is observable. In the case of a multi-sided platform, the demographic groups d would be distinct sides of the market, such as buyers and sellers on an e-commerce platform. Online Appendix O.5 generalizes Proposition 4 to this model variant; the generalized result establishes identification of the multi-group model under conditions similar to those required for Proposition 4.

Instrumental variables

The model admits three potential classes of instruments. The first are the Waldfogel instruments, which are functions of the market-specific distribution F_t^W of consumer characteristics W_{it}. One example is $Z_t = \mathbb{E}_t[W_{it}]$, the mean of W_{it} in market t. In the semiparametric specification described in Section 2, another candidate instrument is the $\mathbb{R}^J$-valued vector Z_t with components

$$Z_{jt} = \Pr_t(W_{it}'\lambda_j + \varepsilon_{ijt} = \max_k\{W_{it}'\lambda_k + \varepsilon_{ikt}\}),$$

which equals the predicted share of product j in market t when all market-level shifters (X_{jt}, S_t, and Ξ_t) are removed from the utility equation.

The validity of the Waldfogel instruments requires F_t^W to be mean-independent of the product-market unobservables Ξ_{jt}, an assumption that may fail in several ways. First, consumers may sort across markets based on both their characteristics W_{it} and products' unobserved demand shifters Ξ_{jt}, thereby inducing a spurious cross-market correlation between F_t^W and Ξ_t. In a model of transit-mode choice, for example, highly educated consumers may choose to live in cities with superior public transit, as captured by Ξ_{jt}. The share of highly educated consumers and the unobserved appeal of public transit would then be positively correlated. Second, firms may respond to cross-market variation in F_t^W by providing superior unobserved service quality or by advertising more intensively in markets whose demographics make consumers especially predisposed to use their products, thereby making Ξ_t a function of F_t^W. Third, consumers may value not only the size of a product's network but also its composition—that is, the distribution of W_{it} among consumers who belong to the network. If the distribution of characteristics within a product's network covaries with the distribution in the market as a whole, then F_t^W may be correlated with unmodelled tastes for network composition, which would be captured by Ξ_t. This final concern implies that the researcher must carefully select the characteristics W_{it} used to construct such instruments, restricting attention to characteristics that are unrelated to tastes over network composition. For example, if age is an important dimension of tastes over network composition in online dating, it would be problematic to base Waldfogel instruments on market-specific distributions of age.

Experimental variation can also generate Waldfogel instruments. This approach is taken up by Guiteras et al. (2025), who study demand for latrines in Bangladesh using data from a randomized controlled trial that offered subsidies to a random subset of households with eligibility rates varying across neighbourhoods and villages. To instrument for the share of households in a neighbourhood purchasing latrines, they use a Waldfogel instrument: the share eligible for subsidies.

Two other classes of instruments may also be available. When there are multiple products ($J > 1$), the BLP instruments—that is, exogenous characteristics of competing products—may be available. BLP instruments are used, for example, by Timmins and Murdock (2007) in a study of recreational fishers' choice of fishing sites. BLP instruments require, however, that the observed characteristics X_t vary across markets, a condition that fails in settings wherein platforms with the same exogenous characteristics X_t operate across distinct geographical markets. In addition, when network externalities operate through quantities rather than shares, market size M_t itself becomes available as an instrument, provided that it satisfies Assumption IV-MICRO. As discussed above, however, the exclusion restriction required for this instrument is often questionable.

Multi-sided platforms. As explained above, the logic of identification with micro data applies to multi-sided platforms. In fact, multi-sided settings may offer more available instruments than single-sided ones: any variable affecting one side's utility without affecting that of another can be used as an instrument in identifying network externalities running from the affected to the unaffected side. Variables directly impacting one side's utility but not the other's may be individual-level and thus provide a basis for Waldfogel instruments, or product-level, in which case they provide BLP instruments. As an example of the latter, local governments in the US have enacted regulations capping the commissions that food delivery platforms charge to restaurants (Sullivan 2026). These commission regulations could thus be used as an instrument for the number of restaurants on a platform when estimating network externalities in consumer choice.

4 Empirical application

Setting and data

I now illustrate the application of the article's identification insights to the estimation of a model of dating website choice.[11] The 2007–2008 online dating industry lends itself to the study of network externalities in demand. The industry consisted of a small number of national websites that charged uniform national subscription prices, thus permitting a focus on network externalities rather than price endogeneity. Also, because broadband and e-commerce were still diffusing unevenly across the country, the setting also offers pronounced cross-market variation in internet access and online behaviour, variation that helps identify the model.

I estimate the model using data from the Comscore Web Behavior Database, which includes the browsing and online transactions records for a panel of US households (91689 panelists in 2007 and 57817 panelists in 2008). Each panelist's activity is recorded for an entire year. Most individuals in the Comscore data appear in only one year. When an individual appears in multiple years, I treat that individual's records for each year as a separate panelist. De Los Santos et al. (2012) find that individuals in Web Behavior Database are representative of online US buyers. For each Comscore panelist, I observe each visited domain, the time and duration of the visit, and the number of pages viewed. I also observe panelist characteristics including income group, educational attainment, race, age group, and ZIP code. I supplement the Comscore data with geographical and population data from the US Census Bureau.

[11] Network externalities in demand for dating websites can be microfounded by a search-and-matching model; Online Appendix O.6 provides such a microfoundation based on the model of Smith (2006).

The analysis focuses on four largest dating sites in the 2007–2008 period: `match.com`, `eharmony.com`, `pof.com` ("Plenty of Fish"), and `okcupid.com`. The shares of Comscore panelists spending at least 5 minutes on each of these sites were 8.61%, 4.63%, 1.93%, and 0.82%, respectively.

Whereas `eharmony.com` and `match.com` required payment for use, `pof.com` and `okcupid.com` were free to use. These free sites relied on advertising and paid premium features for revenue. The paid sites sell subscriptions in bundles of one, three, six, and twelve months, with the price per month of service declining in bundle length; one-month subscriptions in 2007 were priced at \$59.95 on `eharmony.com` and \$34.99 on `match.com`. I construct the price P_j of each paid site from the subscription purchases appearing in the Comscore transaction records. For each site, I compute the mean price paid for each bundle length, convert these bundle prices to prices per month of service, and average the per-month prices across bundle lengths, weighting each length by the number of months of service that it provides. I then multiply the weighted-average per-month price by three, expressing P_j as the price of a three-month subscription period. I adopt the three-month horizon because the median subscription length is three months at both paid sites. This procedure yields $P_j = \$63.00$ for `eharmony.com` and $P_j = \$57.66$ for `match.com`. Table O.1 of Online Appendix O.7 reports the mean price and the number of observed purchases for each site and bundle length.

The data do not contain information on gender or sexual orientation. Thus, I specify a model in which the decomposition of website usership along these dimensions does not impact consumer choice. This limitation is a function of the data, not the identification framework; as explained in Section 3, identification with micro data holds in the presence of observably distinct groups under assumptions analogous to those used in identifying the model without distinct groups.

The markets in the analysis are regions based on combined statistical areas (CSAs), which consist of counties that are economically connected. I construct CSA/state pairs by dividing each CSA up into its parts that belong to different states. I then construct the geographical units underlying the markets in my analysis by adding each county that is within 50 miles of a CSA/state and that does not belong to a CSA/state to this county's closest CSA/state.[12]

I assign each panelist in my estimation sample to either a primary site or the outside option. In particular, I consider a panelist to be a user of a dating website if they visit at least five pages within the site, visits the site during at least two distinct sessions, and spends at least five minutes

[12] Hitsch et al. (2010b) similarly treat the metropolitan area as the relevant market for online dating, restricting their analysis to users in Boston and San Diego.

Table 2: Summary statistics

Panel A: Sample and markets

	N	Mean	SD	P10	Median	P90
Comscore panelists	149506					
Estimation sample	23697					
Markets	35					
Consumers per market		677	453	295	483	1098
Market population (1000s)		4708	3443	1996	3673	8095

Panel B: Dating websites

	Users	Share of sample (%)	Share of users (%)
`match.com`	4715	19.90	57.64
`eharmony.com`	2221	9.37	27.15
`pof.com`	963	4.06	11.77
`okcupid.com`	281	1.19	3.44
Outside option	15517	65.48	

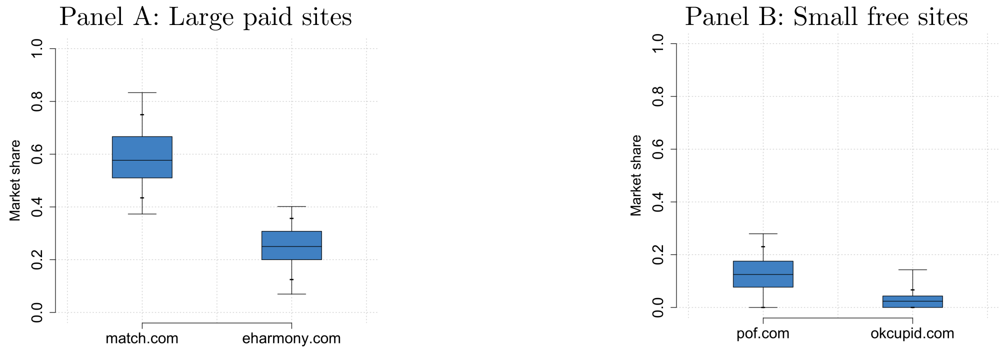
Notes: this table describes the estimation sample as described in the text. In Panel A, the "Consumers per market" and "Market population" rows report distributions across the markets in the estimation sample. Market populations are computed from US Census Bureau county population estimates for 2010. In Panel B, "Share of sample" provides the percentage of the estimation sample using the indicated site and "Share of users" provides the percentage of dating website users using the indicated site.

on the site qualifies as a user of that site. When a panelist qualifies as a user of multiple sites, I determine their primary site as that on which the panelist spends the most time. To capture substitution into dating websites by users who did not previously use these websites, I specify an outside option consisting of panelists who visit a dating website but do not qualify as a user under the criteria above. I drop all other panelists. Furthermore, I drop markets with under 100 observed users, leaving 35 markets in the estimation sample.

Table 2 reports summary statistics describing the estimation sample whereas Figure 4 describes the cross-market distribution of sites' market shares. Online Appendix O.8 reports summary statistics describing the characteristics of the consumers in this sample. Consumers who visit dating websites are demographically similar to the Comscore panel as a whole, but they browse the web roughly twice as intensively as the typical panelist.

Consumers are able to multi-home (i.e., use multiple websites), although the majority do not. Among panelists who use at least one site, 81% use only one website. Online Appendix O.9 provides additional information on the extent of multi-homing in the data. I do not incorporate multi-homing in the main specification, although I estimate and analyze a version of the model with multi-homing in Online Appendix O.10.

Figure 4: Cross-market variation in sites' market shares in 2007–2008



Notes: These plots display the 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, and 0.95 quantiles of sites' market shares across markets in the 2007–2008 time period. The plotted market shares for each site equal that site's user base divided by the total number of dating website users in the market (i.e., the shares exclude the outside option).

I use web browsing records in the data to generate variables describing consumers' web usage that I include in W_{it}. These variables include the log of the total number of webpages that the consumer viewed in the course of the year, the log of the browsing duration (in minutes) aggregated across all of the consumer's web sessions, and the number of pages viewed within various categories, including finance, gaming, information, media, retail, social media, video, weather, etc.; Online Appendix O.11 provides a complete list. These variables are intended to measure consumers' interests and the extent of their internet access. The consumer characteristics W_{it} also include the log of the consumer's local population. I define a consumer's local population as the total population of the ZIP codes in the consumer's state whose centroids lie within five miles of the centroid of the consumer's home ZIP code; I compute this variable using ZIP-code-level population counts from the 2010 US Census. Local population measures the density of a consumer's immediate area, which may shift the appeal of offline alternatives to dating websites.

Estimation

Estimation of the model proceeds in two steps. In the first *micro step*, I estimate site/market-specific utility indices and preference loadings on consumer characteristics. In the second *market step*, I estimate the contribution of site characteristics and market shares to mean tastes for sites. Note that I estimate the model under two sorts of demographic group specifications: one in which there are not distinct demographic groups, and the other in which consumers below and above the age of 35 constitute two distinct demographic groups. I estimate both shares models in which consumers value market shares and quantities models in which consumers value total user bases defined by market shares times market size.

Micro step. The estimating equation of the micro step is

$$U_{ijt} = \delta_{jt}^{d(i)} + W_{it}'\lambda_j + \varepsilon_{ijt}, \tag{15}$$

where δ_{jt}^d are mean tastes for site j among members of demographic group d in market t, $d(i)$ is the demographic group of consumer i, and W_{it} are the observed consumer characteristics described above, which include demographic variables, web usage variables, and the log of the consumer's local population. Last, ε_{ijt} is assumed to follow a type 1 extreme value distribution. I estimate the parameters δ_{jt}^d and λ of (15) via maximum likelihood. Note that the micro step is the same for each of the share and quantities models.

Market step. I estimate the market step via a linear instrumental variables regression. The estimating equation of the market step of the shares model is

$$\delta_{jt}^d = \psi_j + f_j^d \left(S_t^{\varnothing}, S_t^1, \ldots, S_t^D; \eta\right) + \Xi_{jt}^d. \tag{16}$$

Here, the f_j^d network externality function is known up to the finite-dimensional parameter vector η. The arguments include overall market shares $S_t^{\varnothing}$ and demographic-group-specific market shares S_t^d. Given that the utility indices δ_{jt}^d are not directly observed, I substitute estimates of these indices from the micro step for the true quantity in the estimation routine. The ψ_j terms are firm-time period indicators. Last, I estimate (16) using several different parametric forms of the network externality function $f_j^d(\cdot;\eta)$ as specified below.

I use Waldfogel instruments in estimating (16), specifying a distinct instrument for each market share appearing in the equation. Each instrument is the predicted value of its associated market share based on the micro step estimates. To purge these predictions of endogenous across-market variation in demand shocks, I replace each utility index δ_{jt}^d by the average of the estimated indices for site j and group d across markets in the same time period when computing predictions.[13] I include the characteristics W_{it} to endow the predictions with cross-market variation driven by heterogeneity in the distribution F_t^W of consumer characteristics across markets. However, given that potential endogeneity of distributions of certain demographic variables (as detailed below), I set the effects of all consumer characteristics other than the internet access and web usage measures set to zero. I apply the same transformation to the predicted market shares as to the market shares

[13]This approach follows the logic of approximating the optimal instrument of Reynaert and Verboven 2014.

entering (16). To illustrate, when $D = 1$ and the estimating equation is

$$\delta_{jt} = \psi_j + \eta \log(S_{jt}) + \Xi_{jt},$$

the Waldfogel instrument for S_{jt} is

$$Z_{jt} = \log\left(\frac{1}{M_t}\sum_{i=1}^{M_t}\frac{e^{\bar{\delta}_j + \tilde{W}_{it}'\tilde{\lambda}_j}}{1 + \sum_k e^{\bar{\delta}_k + \tilde{W}_{it}'\tilde{\lambda}_k}}\right),$$

where M_t is the sample size of consumers in market t, $\bar{\delta}_j$ is the average of the estimated δ_{jt} across markets, and $\tilde{W}_{it}$ collects the internet access and web usage measures: an indicator for broadband access, browsing intensity and category usage, browsing on paid subscription services, and the number of online transactions. Because market shares are estimated from a finite number of panelists per market, I estimate (16) by weighted two-stage least squares; the weight on each observation, $n_t^d S_{jt}^d/(1 - S_{jt}^d)$, is the inverse of the delta-method sampling variance of its log market share, where n_t^d is the number of sampled consumers in group d of market t.

The instruments deliberately use no demographic variation. As noted in my discussion of identification, Waldfogel instruments based on characteristics whose market-specific distributions directly affect consumer tastes violate the exclusion restriction required for these instruments to identify the model. Markets' distributions of characteristics including age, race, and education may directly affect tastes for dating websites because people value these characteristics in potential mates. The distributions of private indicators of internet access and online interests, by contrast, are unlikely to shift the appeal of dating itself. Instead, they proxy for web access and web usage behaviour, factors that plausibly shift dating website usage without shifting mate desirability conditional on included covariates such as income and education. The cost of discarding the demographic variation is power: the first-stage F-statistic of the preferred specification is 18.1, which is adequate by conventional rules of thumb but below the level required for undistorted t-based inference. I therefore report Anderson–Rubin confidence sets that remain valid under weak instruments (Anderson and Rubin, 1949; Andrews et al., 2019) alongside conventional standard errors.

Online Appendix O.12 reports a robustness specification in which the instruments do use demographic variation, interacted with the internet measures through the estimated choice model, with the market-level mean of every consumer characteristic included as a control so that the exclusion restriction is imposed conditional on mean market demographics that directly affect tastes.

That specification has a stronger first stage and delivers a statistically indistinguishable estimate. Furthermore, the two instrument bases jointly pass a Sargan overidentification test. These results suggest that variation in demographics potentially underlying mate desirability does not drive the estimates.

The estimation of the market step of a quantities model proceeds similarly. For the quantities model, quantities replace market shares in (16) and predicted quantities are used in constructing instruments instead of predicted market shares.

Price sensitivity

To this point, I have ignored price competition. Consumer price sensitivity is needed for computing pricing equilibria in counterfactuals and for expressing welfare figures in dollar terms. But the fact that the dating websites in the sample charge uniform prices across geography means that I observe minimal price variation, which precludes the estimation of price sensitivity alongside the other preference parameters. Instead, following Gentzkow (2007), I infer price sensitivity from the supply side of the market. Under this approach, the paid sites' pricing first-order conditions and an assumption of zero marginal costs pin down the price sensitivity parameter. As previously mentioned, I use site/time indicators as the x_j in (16) and let ψ_j denote the fixed effect for site j. I then make the decomposition $\psi_j = \bar{\psi}_j - \alpha P_j$ and assume that the observed prices $\{P_j^*\}$ constitute a Nash equilibrium in pricing with marginal costs of zero in that

$$P_j^* = \arg\max_{P_j} \sum_t M_t \sigma_{jt}(P_j, P_{-j}^*; \alpha) P_j \qquad \forall j \text{ s.t. } P_j^* > 0. \tag{17}$$

Here, M_t is the number of consumers in my estimation sample from market t, so that $M_t\sigma_{jt}$ counts site j's users in that market. Because $\hat{\alpha}$ depends on the M_t only through their relative magnitudes, the scale on which market size is measured does not affect the estimate. The profit maximization problem in (17) yields the first-order conditions (FOCs)

$$\sum_t M_t \left[\frac{\partial \sigma_{jt}}{\partial P_j}(P_j^*, P_{-j}^*; \alpha) P_j^* + \sigma_{jt}(P_j^*, P_{-j}^*; \alpha) \right] = 0. \tag{18}$$

To recover $\hat{\alpha}$, I substitute empirical analogues/estimates for population objects/parameters in each paid site's FOC (18) and then solve for α. Each paid site's FOC yields a separate value of α; I take $\hat{\alpha}$ to be the average of these site-specific values. I compute a standard error for $\hat{\alpha}$ using a parametric bootstrap that samples from the estimated asymptotic distribution of the parameters estimated in the market step of estimation and repeats the procedure. I include additional details on the recovery of α in Online Appendix O.13.

Some sites are free to use. I do not include these sites' FOCs in the estimation of α and I assume that free sites remain free in my counterfactuals.

In markets with positive network externalities, platforms may price below the static optimum to enlarge their networks (Dubé et al., 2010; Cabral, 2011), in which case the FOC (18) would rationalize low observed prices by overstating price sensitivity. Various features of the dating websites industry limit this concern. First, dating platforms' user bases are largely transient as users who successfully find monogamous partners exit the market. Also, the sample period falls well after the launches of `match.com` (1995) and `eharmony.com` (2000), so that observed pricing plausibly reflects a mature phase of competition rather than the introductory phase in which penetration pricing is most effective.

Parameter estimates

This section reports and discusses the parameter estimates. The micro step involves a large number of parameters whose presentation I relegate to Online Appendix O.11. These estimates indicate significant taste differences across individuals with different observable characteristics.

Panel A of Table 3 displays the parameter estimates of a share-type model with the "Overall" demographic specification and the network externality function specification $f_j(S_t; \eta) = \eta \log(S_{jt})$. Instrumenting for market shares with the Waldfogel instruments decreases the estimated network externality coefficient relative to ordinary least squares (OLS). This reflects the mechanical positive correlation between the unobservables Ξ_{jt} and market shares S_{jt}. Panel A also reports the estimated value of α for this specification. The rows with names of websites (e.g. "eharmony") provide estimated site intercepts.

As a robustness check, Online Appendix O.10 estimates a specification in which consumers may use up to two sites; the network externality estimate rises to 0.73, with an Anderson–Rubin confidence set that contains the single-homing estimate. Robustness likewise extends to the specification of the network externality function f_j. Under a quantities model wherein preferences depend on log quantities rather than log shares yields an estimate $\hat{\eta} = 0.54$ (standard error 0.15). Online Appendix O.14 reports the full quantities-model estimates. I also estimate a specification with a linear externality function, $f_j(S_t) = \eta S_{jt}$, under which the welfare ranking of the counterfactual regimes studied below is preserved. See Online Appendix O.15 for results from the linear specification.

Panel B reports the first stage of the instrumental variables regression whose results are displayed in Panel A. The first stage here is a regression of the log market share on the instrument, with both residualized on the site indicators so that the F-statistic reflects the strength of the excluded instrument alone. The first-stage F-statistic takes on the moderate value of 18.1.

The results presented above allow the calculation of the value of an increase in a site's usership to its inframarginal users. The estimates for the baseline model of Table 3 imply that a 10% increase in the usership of a site is worth \$5.32 per three-month subscription period to an inframarginal user of that site, about 9% of the price.

Table 4 reports estimates from the "Age" demographic group specification under various specifications of the network externality function. In particular, column (1) reports estimates from a specification in which consumers care only about the market share of a site within their own demographic group; column (2) reports estimates from a specification in which a consumer's tastes depend on the market share of a site within their own demographic group and on the other demographic group in a distinct fashion; column (3) reports estimates from a specification in which a consumer's tastes depend on the market share of a site within their own demographic group only, but members of different demographic groups have different preferences for their own-group market shares; and column (4) reports estimates from a specification in which a consumer's tastes depend on the market share of a site both within their own demographic group and within the entire population. These tables suggest considerable homophily within groupings defined by age, as consumers more highly value market shares within their own age group than shares within the other group. Column (3) is consistent with similar network externality strength across groups, although the own-group taste for the younger group is imprecisely estimated. The estimates of column (4) are imprecise due to collinearity.

Online Appendix O.14 provides estimates of a model in which overall quantities $M_t S_{jt}$ rather than market shares S_{jt} enter consumer utility. The results are similar to those for the baseline model.

Table 3: Market step parameter estimates – "Overall" demographic group specification

Panel A: Parameter estimates

	OLS	IV
$\log(s_{jt})$	0.99	0.55
	(0.03)	(0.14)
eharmony	0.56	-0.48
	(0.07)	(0.33)
match	-0.67	-1.38
	(0.04)	(0.22)
okcupid	-2.92	-4.86
	(0.12)	(0.61)
pof	0.37	-1.01
	(0.09)	(0.43)
p_j $(\hat{\alpha})$		0.0099
		(0.0028)

Panel B: First stage of IV regression

	$\widetilde{\log(s_{jt})}$
$\tilde{z}_{jt}$	2.52
	(0.59)
F	18.14

Notes: the first-stage regression of Panel B is residualized on site indicators. Standard errors appear in parentheses. Weak-instrument-robust Anderson–Rubin confidence sets are reported in the text.

Table 4: Market step parameter estimates – "Age" demographic group specification

	(1)	(2)	(3)	(4)
Own-group $\log\left(s_{jt}^{d}\right)$	0.322	0.374	-	-2.400
	(0.160)	(0.333)	-	(6.464)
Other-group $\log\left(s_{jt}^{d'}\right)$	-	-0.038	-	-
	-	(0.283)	-	-
Own-group $\log\left(s_{jt}^{\text{younger}}\right)$	-	-	0.041	-
	-	-	(0.396)	-
Own-group $\log\left(s_{jt}^{\text{older}}\right)$	-	-	0.390	-
	-	-	(0.178)	-
$\log\left(s_{jt}^{\text{overall}}\right)$	-	-	-	2.624
	-	-	-	(6.138)
p_{jt} $(\hat{\alpha})$	0.0140	0.0138	0.0135	0.0158
	(0.0028)	(0.0028)	(0.0033)	(0.0095)

Decomposition of variance in market shares

As illustrated by Figure 4, dating websites exhibit substantial variation in popularity across cities. This raises the question of how websites establish market-specific competitive advantages. Cross-market variation in shares could reflect differences in unobserved mean tastes for particular platforms, captured by Ξ_{jt}, or differences in demographic composition across markets. A third possibility is that network externalities amplify otherwise modest differences in platforms' underlying appeal across markets, generating herding on sites that enjoy comparatively small initial advantages. This mechanism parallels the insight of Glaeser et al. (1996) that social interactions can magnify small underlying differences across locations. In their context, higher crime rates induce additional criminal activity, causing observed cross-city variation in crime to exceed what fundamentals alone would predict. Here, even modest differences in sites' fundamentals may be magnified by network effects.

To assess the importance of these forces, I recompute equilibrium market shares under a set of counterfactual environments. For each site, I first compute the standard deviation across markets of the site's market share among the four sites considered (i.e., excluding the outside option) in the baseline regime. I then recompute the same statistic under counterfactuals that remove, one at a time relative to the baseline, network externalities, unobserved mean tastes Ξ_{jt}, and cross-market differences in observed demographics. In doing so, I use the estimated shares model without multiple demographic groups. Table 5 reports the resulting counterfactual standard deviations normalized by the baseline standard deviation for each site, along with the baseline standard deviation itself. This procedure characterizes how much geographic dispersion in market shares each force generates on its own.

Each counterfactual modifies the baseline in a single dimension. In the "No network externalities" column, I remove the network externality term from consumers' indirect utilities and recompute equilibrium market shares in each market. Removing network externalities also changes sites' average utilities and, mechanically, the market share of the outside option. To hold fixed the overall attractiveness of online dating relative to the outside option, I therefore add a common constant $f^{\dagger}$ to each site's mean utility in every market, where $f^{\dagger}$ is chosen so that the implied outside share matches its baseline value. In the "No Ξ_{jt}" column, I instead set all market-level unobserved taste shocks Ξ_{jt} to zero, retaining network externalities and demographic differences. In the "Equalize demographics" column, I recompute market shares using the pooled distribution of demographic characteristics across all markets in the sample, rather than each market's own demographic composition, again retaining the other two forces.[14]

Table 5 indicates that network externalities are the leading source of cross-market variation in market shares. Eliminating network externalities reduces the standard deviation of market shares by about half on average, with particularly large effects for `match.com` and `eharmony.com`; even for `pof.com`, the site for which the effect of network externalities is smallest, the remaining vari-

[14] Online Appendix O.16 reports the corresponding sequential decomposition, in which the three forces are removed cumulatively, together with an order-free (Shapley) attribution that accounts for interactions between forces. The Shapley attribution confirms the ranking found here, indicating that network externalities and unobserved tastes play leading roles in explaining across-market variation in site popularity.

ation is about two-thirds of its baseline level. The main exception is `okcupid.com`, for which the standard deviation rises when network externalities are removed. This pattern suggests that network effects tend to reinforce the competitive positions of the more established platforms and to compress the relative geographic variation of a smaller platform whose baseline shares are low. Unobserved market-specific taste shifters Ξ_{jt} play an important secondary role. Indeed, setting $\Xi_{jt} = 0$ on its own reduces the cross-market standard deviation to about two-thirds of its baseline level on average. Equalizing demographics across markets, by contrast, leaves dispersion nearly unchanged, at about 0.94 of baseline on average, so observed demographic composition accounts for little geographic dispersion by itself. Taken together, network externalities and unobserved tastes are the principal drivers of cross-market variation in market shares, whereas observed demographic differences contribute little on their own. Because the three forces interact, their individual contributions need not sum to the total. The order-free attribution in Online Appendix O.16 apportions this interaction and reaches the same ranking. The "Average" row in Table 5 is computed using sites' national baseline market shares as weights, so it places greater weight on the dispersion of the larger platforms.

These forces also have implications for market concentration. Table 6 reports the mean and standard deviation across markets of the Herfindahl–Hirschman index (HHI) computed from inside shares under the same counterfactual environments. In the baseline, the mean HHI is 4198, with a cross-market standard deviation of 443. Eliminating network externalities lowers the mean HHI sharply, to about 2990, and also greatly reduces its cross-market dispersion, to 124. This result reflects that, when consumers value being on the same platform as others, even small market-specific advantages become self-reinforcing, allowing a site that is slightly more attractive in a given market to capture a disproportionately large share there. Network externalities therefore both raise concentration on average and magnify differences in concentration across markets. Removing unobserved tastes Ξ_{jt} or equalizing demographics, by contrast, leaves the mean HHI almost unchanged (4168 and 4213, respectively) and reduces its cross-market dispersion only modestly (to 290 and 415). Concentration and its geographic variation are thus driven overwhelmingly by network externalities: absent the self-reinforcing feedback they create, neither unobserved tastes nor demographic differences generate highly concentrated or cross-sectionally dispersed market structures on their own.

Table 5: Decomposition of geographic variation in market shares

Site	Standard deviation in market shares Baseline	Share of baseline standard deviation in market shares Baseline	No network externalities	No Ξ_{jt}	Equalize demographics
match.com	0.056	1.00	0.48	0.65	0.94
eharmony.com	0.042	1.00	0.47	0.66	0.95
pof.com	0.041	1.00	0.69	0.71	0.89
okcupid.com	0.014	1.00	1.32	0.65	1.00
Average	0.049	1.00	0.53	0.66	0.94

Notes: the table is based on market shares among the inside goods. For each site, the second column reports the standard deviation across markets of its baseline inside share. Each of the remaining columns reports the standard deviation under a counterfactual that removes the indicated force alone, relative to the baseline, divided by that site's baseline standard deviation. The "Average" row reports the weighted average of the site-level entries, where the weights are sites' national baseline market shares.

Table 6: Decomposition of geographic variation in market concentration

	Herfindahl–Hirschman index (/10,000) Baseline	No network externalities	No Ξ_{jt}	Equalize demographics
Mean	4198	2989	4168	4213
SD	443	124	290	415

Notes: the table reports the mean and standard deviation across markets of the Herfindahl–Hirschman index (HHI) computed from market shares among the inside goods. Each column corresponds to the same counterfactual environments as in Table 5. The baseline column reports the HHI in the estimated model, and the remaining columns report the HHI after removing, one at a time relative to the baseline, network externalities, unobserved mean tastes Ξ_{jt}, and cross-market demographic differences.

Mergers and interoperability

Next, I use the estimated shares model without multiple demographic groups to evaluate the effects of a merger between the leading sites, `eharmony.com` and `match.com`. There are several possible effects of such a merger. First, as in a standard horizontal merger, the merged firm would gain market power and therefore have an incentive to raise prices. In addition, if the merged firm combined its sites' networks, it would increase consumers' willingness to pay for either site, thereby creating a further incentive to raise prices. At the same time, consumers would enjoy access to a larger network on the merged platforms, which could increase their welfare.

This discussion suggests two distinct elements of a platform merger with interacting effects. First, a merger could involve joint pricing, whereby the two websites set prices to maximize the profits of the merged entity. Second, a merger could involve interoperability, whereby the firm allows users of each website to access users on the other. Huang et al. (2026), who study the effects of interoperability without joint pricing, note that rival firms benefit from the expansion of a given firm's network under interoperability, thereby reducing the focal firm's incentive to expand its network by lowering its price. Rather than lowering prices to build a network, each platform free-rides on rivals. The upshot is that interoperability tends to raise prices.

Interoperability combined with a merger, however, may reduce prices. This is because the merged firm internalizes the positive effect of reducing one site's price on traffic at the other site, an effect that arises because the price reduction expands the first site's usership and hence the network available to potential users of the second site. This positive effect of a price reduction on the merged firm's profits may lead it to charge lower prices than would competing firms in a setting with interoperability.

To assess the distinct effects of joint pricing and interoperability, I simulate counterfactual regimes corresponding to each possible combination of the two. Throughout, I assume that the paid firms set prices at a national level and that these prices constitute a Nash equilibrium. Additionally, I allow interoperability only between the two leading paid firms whose potential merger I analyze. Under interoperability, the network externality terms of merging sites $j = 1, 2$ are replaced with the externality evaluated at the combined user base: for $j \in \{1, 2\}$, $f_j(S_t) = \eta \log(S_{1t} + S_{2t})$. The terms of the non-merging sites are unchanged.[15]

Computing these counterfactuals requires solving for equilibrium market shares and prices, which raises the issue of equilibrium multiplicity. I compute demand fixed points (i.e., solutions of (3)) by iterating the demand mapping to a fixed point and prices by iterating firms' best responses, in each case starting from the observed market structure. In practice, this selection is innocuous, as the stable demand equilibrium is unique in every market and regime.[16]

Table 7 reports the results. First consider the move from competitive pricing to joint pricing without interoperability, represented by the difference between the first and second rows of the first column of each panel. Joint pricing predictably raises prices and reduces consumer welfare. The other element of a website merger, interoperability, instead has a positive effect on consumer welfare. Under competitive pricing, introducing interoperability raises consumer welfare by \$5.88

[15] This counterfactual extrapolates the logarithmic externality outside the estimation sample. Online Appendix O.15 repeats the analysis under a linear specification, $f_j(S_t) = \eta S_{jt}$. The welfare ranking of the four regimes is unchanged.

[16] I verify uniqueness by searching for all fixed points of the demand mapping from roughly sixty dispersed starting configurations in each of the 175 market–regime cells. The search returns a single stable equilibrium, the reported one. The other fixed points are unstable configurations in which a site collapses.

Table 7: Joint pricing and interoperability

(a) Prices

	Interoperability?	
Joint pricing?	No	Yes
No	\$60.99	\$95.08
Yes	\$76.82	\$91.14

Notes: the table provides the mean price of a three-month subscription period across `eharmony.com` and `match.com` as predicted by the model under each regime.

(b) Consumer welfare (\$/baseline site user)

	Interoperability?	
Joint pricing?	No	Yes
No	0	\$5.88
Yes	-\$17.94	\$11.10

Notes: the table provides the mean change in consumer welfare per dating website user in the baseline equilibrium with neither joint pricing nor interoperability, expressed in dollars per three-month subscription period.

per user in the baseline equilibrium despite raising the average subscription price by about \$34, a result explained by consumers' high valuation of dating-site network size. Last, consider the effect of introducing joint pricing when the two leading websites are interoperable. This change is intended to correspond to a full merger between the sites. In this case, prices fall and welfare rises, in stark contrast to the effects of joint pricing without interoperability. Joint pricing reduces prices under interoperability because, as noted above, the merged firm internalizes the positive spillovers from price reductions that expand one site's user base. Overall, the analysis suggests that a full merger involving both interoperability and joint pricing is the consumer-welfare-maximizing regime and would increase consumer welfare by \$11.10 per user relative to the baseline equilibrium. Online Appendix O.10 repeats this analysis under the multi-homing specification with the same qualitative conclusions.

5 Conclusion

This article studies the identification of BLP-style demand models with network externalities. The central identification result is that, with market-level data alone, network externalities are generally not separately identified from consumers' direct responses to product characteristics. Micro data linking consumers' choices to their characteristics permit identification by allowing the researcher to use within-market variation to identify substitution patterns separately from network externalities. The analysis also clarifies the classes of instrumental variables that can provide the required exogenous variation, in particular Waldfogel instruments and, in some settings, BLP-style instruments. To illustrate the application of the identification analysis, I estimate a model of dating website choice. The results imply that network externalities are substantial, account for most of the geographic variation in sites' market shares, and materially affect the welfare analysis of changes in market structure.

Bibliography

Ackerberg, Daniel A., and Gautam Gowrisankaran. 2006. "Quantifying equilibrium network externalities in the ACH banking industry." *The RAND Journal of Economics* 37 (3): 738–761. https://doi.org/10.1111/j.1756-2171.2006.tb00040.x.

Allende, Claudia. 2021. "Competition under social interactions and the design of education policies."

Anderson, T. W., and Herman Rubin. 1949. "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations." *Annals of Mathematical Statistics* 20 (1): 46–63.

Andrews, Isaiah, James H. Stock, and Liyang Sun. 2019. "Weak Instruments in Instrumental Variables Regression: Theory and Practice." *Annual Review of Economics* 11 727–753.

Angrist, Joshua. 2014. "The Perils of Peer Effects." *Labour Economics* 30 98–108.

Aradillas-López, Andrés. 2021. "Computing semiparametric efficiency bounds in discrete choice models with strategic-interactions and rational expectations." *Journal of Econometrics* 221 (1): 25–42.

Bayer, Patrick, Robert McMillan, and Kim Rueben. 2004. "An equilibrium model of sorting in an urban housing market." (10865): .

Bayer, Patrick, and Christopher Timmins. 2007. "Estimating equilibrium models of sorting." *The Economic Journal* 117 (518): 353–374.

Becker, Gary S., Kevin M. Murphy, and Harvard University Press eBook Package Backlist 2000-2013. 2000. *Social economics: market behavior in a social environment.* Cambridge, Mass: Belknap Press of Harvard University Press.

Berry, Steven, Amit Gandhi, and Philip Haile. 2013. "Connected Substitutes and Invertibility of Demand." *Econometrica* 81 (5): 2087–2111.

Berry, Steven, and Philip Haile. 2014. "Identification in Differentiated Products Markets Using Market Level Data." *Econometrica* 82 (5): 1749–1798.

Berry, Steven, and Philip Haile. 2016. "Identification in Differentiated Product Markets." *Annual Review of Economics* 8 27–52.

Berry, Steven, James Levinsohn, and Ariel Pakes. 1995. "Automobile prices in market equilibrium." *Econometrica* 63 (4): 841–890.

Berry, Steven T, and Philip A Haile. 2024. "Nonparametric identification of differentiated products demand using micro data." *Econometrica* 92 (4): 1135–1162. 10.3982/ECTA20731.

Bhattacharya, Debopam, Pascaline Dupas, and Shin Kanaya. 2024. "Demand and Welfare Analysis in Discrete Choice Models with Social Interactions." *The Review of Economic Studies* 91 (2): 748–784.

Blume, Lawrence E, William A Brock, Steven N Durlauf, and Rajshri Jayaraman. 2015. "Linear social interactions models." *Journal of Political Economy* 123 (2): 444–496.

Bramoullé, Yann, Habiba Djebbari, and Bernard Fortin. 2009. "Identification of peer effects through social networks." *Journal of Econometrics* 150 (1): 41–55.

Brock, William A., and Steven N. Durlauf. 2001a. "Discrete choice with social interactions." *Review of Economic Studies* 68 (2): 235–260.

Brock, William A., and Steven N. Durlauf. 2001b. "Interaction-based models." In *Handbook of Econometrics, Volume 5*, edited by Heckman, James, and Edward Leamer Volume 5. of Handbook of Econometrics 3297–3380, North-Holland.

Brock, William A., and Steven N. Durlauf. 2007. "Identification of binary choice models with social interactions." *Journal of Econometrics* 140 52–75.

Cabral, Luís. 2011. "Dynamic Price Competition with Network Effects." *The Review of Economic Studies* 78 (1): 83–111.

Carrell, Scott E., Frederick V. Malmstrom, and James E. West. 2008. "Peer effects in academic cheating." *Journal of Human Resources* 43 (1): 173 – 207. 10.3368/jhr.43.1.173, Cited by: 128; All Open Access, Bronze Open Access.

Chatterjee, Chirantan, Ying Fan, and Debi Prasad Mohapatra. 2026. "Spillover Effects in Complementary Markets: A Study of the Indian Cell Phone and Wireless Service Markets." Unpublished working paper.

De Los Santos, Babur, Ali Hortacsu, and Matthijs R. Wildenbeest. 2012. "Testing models of consumer search using data on web browsing and purchasing behavior." *American Economic Review* 102 (6): 2955–2980.

Dearing, Adam. 2026. "Identifying Preference Heterogeneity in the BLP Model with Micro Data." Unpublished working paper.

Dubé, Jean-Pierre H, Günter J Hitsch, and Pradeep K Chintagunta. 2010. "Tipping and concentration in markets with indirect network effects." *Marketing Science* 29 (2): 216–249.

Durlauf, Steven, and Yannis Ioannides. 2010. "Social Interactions." *Annual Review of Economics* 2 451–478.

Farronato, Chiaria, Jessica Fong, and Andrey Fradkin. 2024. "Dog Eat Dog: Balancing Network Effects and Differentiation in a Digital Platform Merger." *Management Science* 70 (1): 464–483.

Fong, Jessica. 2024. "Effects of Market Size and Competition in Two-Sided Markets: Evidence from Online Dating." *Marketing Science* 43 (5): 971–985.

Gentzkow, Matthew. 2007. "Valuing New Goods in a Model with Complementarity: Online Newspapers." *American Economic Review* 97 (3): 713–744.

Glaeser, Edward L, Bruce I Sacerdote, and Jose A Scheinkman. 2003. "The social multiplier." *Journal of the European Economic Association* 1 (2-3): 345–353.

Glaeser, Edward L, Bruce Sacerdote, and Jose A Scheinkman. 1996. "Crime and social interactions." *The Quarterly journal of economics* 111 (2): 507–548.

Graham, Bryan S. 2008. "Identifying social interactions through conditional variance restrictions." *Econometrica* 76 643–660.

Graham, Bryan S. 2018. "Identifying and estimating neighborhood effects." *Journal of Economic Literature* 56 450–500.

Guiteras, Raymond, James Levinsohn, and Ahmed Mushfiq Mobarak. 2025. "Demand estimation with strategic complementarities: sanitation in Bangladesh." Unpublished working paper.

Hitsch, Günter J., Ali Hortacsu, and Dan Ariely. 2010a. "Matching and sorting in online dating." *American Economic Review* 100 (1): 130–163.

Hitsch, Günter J., Ali Hortacsu, and Dan Ariely. 2010b. "Matching and Sorting in Online Dating." *American Economic Review* 100 (1): 130–163.

Huang, Jinglei, Guofu Tan, Tat-How Teh, and Junjie Zhou. 2026. "A network approach to interoperability."

Jullien, Bruno, Alessandro Pavan, and Marc Rysman. 2021. "Two-sided markets, pricing, and network effects." In *Handbook of Industrial Organization, Volume 4*, edited by Ho, Kate, Ali Hortacsu, and Alessandro Lizzeri Volume 4. of Handbook of Industrial Organization 485–592, Elsevier.

Lee, Lung-Fei. 2007. "Identification and estimation of econometric models with group interactions, contextual factors and fixed effects." *Journal of econometrics* 140 (2): 333–374.

Manski, Charles. 1993. "Identification of endogenous social effects: the reflection problem." *Review of Economic Studies* 60 531–542.

Newey, Whitney, and James Powell. 2003. "Instrumental Variable Estimation of Nonparametric Models." *Econometrica* 71 (5): 1565–1578.

Reynaert, Mathias, and Frank Verboven. 2014. "Improving the Performance of Random Coefficients Demand Models: The Role of Optimal Instruments." *Journal of Econometrics* 179 (1): 83–98.

Rysman, Marc. 2004. "Competition between networks: a study of the market for Yellow Pages." *Review of Economic Studies* 71 (2): 483–512.

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

Smith, Lones. 2006. "The Marriage Model with Search Frictions." *Journal of Political Economy* 114 (6): 1124–1144.

Sullivan, Michael. 2026. "Fee Optimality in a Two-Sided Market." Unpublished working paper.

Timmins, Christopher, and Jennifer Murdock. 2007. "A Revealed Preference Approach to the Measurement of Congestion in Travel Cost Models." *Journal of Environmental Economics and Management* 53 (2): 230–249.

Waldfogel, Joel. 2003. "Preference Externalities: An Empirical Study of Who Benefits Whom in Differentiated-Product Markets." *The RAND Journal of Economics* 34 (3): 557–568.

Waldfogel, Joel. 2008. "The median voter and the median consumer: Local private goods and population composition." *Journal of Urban Economics* 63 (2): 567–582.