Customer Overlap and Diversion Ratios*

Liran Einav, Mariana Guido, and Pete Klenow†

January 8, 2026

Abstract

We define the concept of customer overlap of product j with product k as the share of j's customers who buy k. We then argue that, *in appropriate contexts*, customer overlaps are an excellent proxy for diversion ratios, a useful and popular way to summarize competition between sellers of substitute products. Unlike diversion ratios, which are often challenging to estimate, customer overlaps are straightforwardly observed in many data sets. We show theoretically, and then validate empirically, the close connection between customer overlaps and diversion ratios. We then illustrate the potential use of customer overlap in contexts where estimation of diversion ratios could be prohibitive.

*We thank Chris Conlon, Ignacio Cuesta, Ali Yurukoglu, and seminar participants in Alpine IO, the Northwestern conference on Antitrust Economics, and Stanford for useful comments and suggestions. We are very grateful to Nick Scott-Hearn for superb research assistance.

†Einav: Stanford University and NBER, leinav@stanford.edu; Guido: Stanford University, mdguido@stanford.edu; Klenow: Stanford University and NBER, klenow@stanford.edu.

1 Introduction

In oligopolistic markets, the diversion ratio from product j to product k is the share of consumers who switch from product j to product k out of the entire pool of consumers who switch out of product j. More formally, if D_j and D_k are the demands faced by two substitute products j and k, respectively, and $D'_j < D_j$ and $D'_k > D_k$ are the new demands after, say, a price increase of product j or a store closure, the diversion ratio from j to k is

$$D_{j\to k} = (D'_k - D_k)/(D_j - D'_j). \tag{1}$$

Diversion ratios, which were prominently featured in the 2010 Department of Justice merger guidelines, are broadly appreciated as one of the most useful ways to summarize competition between sellers of substitute products. Diversion ratios are central in the theoretical analysis of optimal pricing and are intuitive to understand, so they play a key role in the study of horizontal mergers (Conlon & Mortimer, 2021).

While diversion ratios are theoretically well understood, measuring them empirically is often challenging. The two common ways by which antitrust authorities approach this challenge are either by recovering diversion ratios from estimated demand systems or by using consumer surveys about (hypothetical) second choices.

In this paper, we propose the use of "customer overlap" as an alternative approach to proxy for diversion ratios. We show that, when applicable, the concept of customer overlap proxies for diversion ratios extremely well. At the same time, it can be read directly off the data and does not require any estimation. It is, therefore, easy to measure and very scalable, so it can be used in many contexts where estimating a "full blown" demand system is prohibitive, either due to computational or data constraints.

Our measure of customer overlap leverages repeated customer choices in panel data to compute the share of a firm's customers that also purchase from their competitor. We define the customer overlap of product j with product k as the market share of product k among product j's customers — that is, the universe of consumers who ever purchased from product j — when these customers do not purchase product j. Formally, let c_{it} denote the product chosen by consumer i at time t. Product j's customers are given by $C_j = \{i \,|\, \exists\, t \text{ s.t. } c_{it} = j\}$, and the customer overlap of product j with product k is then given by

$$C_{j\to k} = \sum_{\{it|i\in C_j\}} 1(c_{it} = k) \;/\; \sum_{\{it|i\in C_j\}} 1(c_{it} \neq j). \tag{2}$$

One way to see why customer overlap is a natural approximation for diversion ratios is to consider mixed logit models, in which the diversion ratio is proportional to $\int s_{ik}s_{ij}di$ (Conlon

& Mortimer, 2021), where s_{ik} and s_{ij} are customer i's individual purchase probabilities of products k and j, respectively. In such a case, being j's customer approximates s_{ij}, and the customer overlap essentially integrates over the s_{ik}'s.[1]

We start in Section 2, where we use a stylized model to illustrate the basic idea and the relationship between customer overlaps and diversion ratios. We consider a two-dimensional Hotelling-style framework, in which consumers have fixed (over time) brand preferences and a geographic location, which is idiosyncratic to each consumer but varies stochastically across purchase occasions. Consumer optimal choice could, therefore, change from one purchase occasion to another, thus generating customer overlap. We then simulate multiple consumer panel data sets from 100 parametric configurations of this setting and use them to show how correlated the resultant customer overlaps are to the analytically computed diversion ratios.

Much of the variation in diversion ratios (and customer overlap) is driven by brand market shares, which are captured by simple concentration measures. The key advantage of panel data is that it allows us to capture the "horizontal" component of the diversion ratio (and customer overlap) that is orthogonal to market shares. This motivates the use of log-normalized versions of both diversion ratios and customer overlap, which are given by

$$ND_{j\to k} = \ln(D_{j\to k}) - \ln(s_k/(1-s_j)), \tag{3}$$

where s_j and s_k are the overall market shares of products j and k, respectively, and

$$NC_{j\to k} = \ln\left(C_{j\to k}\right) - \ln\left(\sum_{it} 1(c_{it}=k) / \sum_{it} 1(c_{it}\neq j)\right). \tag{4}$$

The log-normalized versions are intuitive to interpret: they capture the market share of product k among product j's customers relative to product k's overall market share. Positive values of the log-normalized overlap (or diversion ratio) suggest that j and k are stronger substitutes (conditional on j's and k's market shares), while negative values suggest weaker substitutes (conditional on j's and k's market shares). Indeed, using the same theoretical simulations, we show that the log-normalized version of customer overlap tracks the (log-normalized) diversion ratios quite closely.

After laying out the motivating theory in Section 2, in the rest of the paper we use detailed transaction-level data from a large payment card network (described in Section 3) to first empirically validate the close connection between diversion ratios and customer

[1] This mixed-logit intuition may also suggest that an even better approximation would weight the overlap by each customer's propensity to purchase product j (in order to approximate s_{ij} more closely). As it turns out, however, this does not make a meaningful difference; we have replicated all the results presented in the paper with the weighted overlap, and all the results are qualitatively and quantitatively essentially the same.

overlap (Section 4), and then to illustrate its potential use (Section 5).

To validate the theoretical results, we construct a large sample of ready-to-drink coffee transactions across eight local markets of varying sizes. We then use this sample to estimate a rich demand system for coffee in each market, use the demand estimates to compute diversion ratios, and then compare them to the (directly observed) corresponding customer overlaps. As in the theory, the two objects track each other closely in this data.

Finally, in Section 5, we illustrate two potential uses of the overlap measure. First, we use the ready-to-drink coffee results to illustrate quantitatively how the high correlation between customer overlaps and diversion ratios implies that — similar to the use of the Herfindahl-Hirschman Index (HHI) — customer overlap could be used quite effectively as a screen for horizontal mergers. Second, we use the concept of customer overlap to illustrate its potential use in a context where demand estimation could be challenging. We compute customer overlaps across the top 50 hotel brands in the United States and use it to highlight substitution patterns which would have been difficult to isolate otherwise.

We conclude by offering a note of caution. Customer overlap is easy to calculate, but can be misapplied. Its use to proxy for diversion ratios is appropriate *only* after a careful market definition of clearly substitute products. We discuss this point in more detail in our final section (Section 6).

Our paper contributes to the recent literature on diversion ratios, which was highlighted by the emphasis on upward pricing pressure in the 2010 merger guidelines (Farrell & Shapiro, 2010). The most natural approach to estimating diversion ratios is to estimate a rich demand system, which allows flexible substitution, and use it to compute the implied diversion ratios. We are not the first to note that this is difficult to do at scale. Indeed, Conlon and Mortimer (2021) propose using second-choice survey data to estimate diversion ratios, and Qiu et al. (2024) propose relying on consumer churn. A related idea is in Atalay et al. (forthcoming), which uses "co-purchasing" rates across products to generate a data-driven way to group products for a nested-logit demand estimation.

Our paper is related to several other recent papers that propose complementing market-level data with more granular, increasingly available individual-level data to estimate flexible substitution patterns. For example, Raval et al. (2017) estimate substitution separately for each group of consumers who are similar on observables, and aggregate across groups. Our concept of customer overlap is similar, but is used within an individual rather than across similar-on-observables individuals. Conlon et al. (2023) offers a similar approach but, like us, proposes to use a panel structure using second-choice data. Finally, McClure (2025) uses hotel recommendation systems to inform estimates about hotel substitution. To the extent that the underlying recommendations are trained on individual-level browsing or choice data,

the conceptual idea is similar to ours.

We contribute to this literature by pointing out that consumer panel data sets are increasingly available, such as through payment card data (as we use in this paper), cell phone data (e.g., SafeGraph), or browsing data (e.g., Comscore). Our main observation is that these types of data sets allow for a straightforward, scalable approach to generate measures of customer overlap, which (we argue) is a good proxy for diversion ratios (in appropriate contexts, as we discuss in the concluding section). A "full blown" estimation of diversion ratios may still be required in many applications, but this is hard to scale. Thus, we envision the use of customer overlap as akin to how IO economists often use measures of market concentration, namely as an easy-to-produce indicator that could motivate more careful (but less scalable) estimation.

2 Motivating theory

We begin by illustrating the tight connection between diversion ratios and customer overlap in the context of a stylized theoretical framework. This simple framework may help explain why it is plausible that customer overlap between firms j and k could proxy well for the diversion ratios between the two firms. We then use the stylized framework to generate a numerical simulation that provides some general guidance as to how tightly correlated these two objects may be in practice.

Setting. We consider a two-dimensional spatial model of price competition among J firms. It is natural to think of the first dimension, denoted by θ, as capturing brand attributes, while the second dimension, denoted by ϵ, captures physical location. Each firm $j = 1, 2, ..., J$ is associated with a fixed location, denoted by (θ_j, ϵ_j), and firms simultaneously set prices $p_1, p_2, ..., p_J$ in a Nash equilibrium.

Consumers are continuously distributed along the two-dimensional space, with each consumer i defined by their fixed brand preference and baseline location, (θ_i, ϵ_i). Importantly, we consider a panel data structure in which consumers make T distinct choices. While consumer i's brand preference remains fixed (at θ_i) for each of these choices, their physical location gravitates from choice to choice "around" their baseline location ϵ_i (e.g., because of other "local" activities), according to $\epsilon_{it} \sim G(\epsilon_i)$.

Consumers then make T distinct discrete choices. In each period t, consumer i has quadratic costs from deviating from both its ideal brand and current location, and they select product j that maximizes their period-t utility (there is no outside option):

$$u_{ijt} = v - \alpha p_j - \lambda_\theta(\theta_j - \theta_i)^2 - \lambda_\epsilon(\epsilon_j - \epsilon_{it})^2. \quad (5)$$

Appendix Figure A1 provides a simple illustration of the choices implied by the model and the resulting overlaps. It depicts the two-dimensional space defined by θ (horizontal axis) and ϵ (vertical axis). Three firms are located in the space, as indicated by the A, B, and C points in the figure. Consumers are distributed across the space with a fixed brand preference θ_i but a stochastic location ϵ_{it}, which varies across choice occasions.

Equilibrium prices imply the two indifference sets depicted in the figure. One is the set of $(\theta_i, \epsilon_{it})$ pairs that imply indifference between firms A and C, and one is the set of pairs that imply indifference between firms C and B. Given this particular market configuration, consumers may either exhibit positive overlap between firms A and C (red region), always purchase from firm C (white region), or exhibit positive overlap between firms C and B (blue region). The extent of the overlap depends on the distribution of ϵ_i and on $G(\epsilon_i)$, which governs how much consumer locations fluctuate (up and down in the figure, along the ϵ_{it} margin) across choice occasions.

Numerical simulations. We then use this model to generate panel data of choices for 100 randomly generated market configurations. For the purpose of the simulation, we set $\lambda_\theta = 4$, $\lambda_\epsilon = 1$, and assume that each market has four firms. θ_j and ϵ_j are drawn iid (across markets as well as across firms within a market) from a uniform distribution over the unit square. Consumer θ_i values are distributed uniformly over $[0, 1]$, their ϵ_i is assumed to be drawn (iid) from a distribution of $Beta(2, 2)$, independently of θ_i, and ϵ_{it} is assumed to be drawn (independently, period by period) from $Beta(1 + \sigma\epsilon_i, 1 + \sigma(1 - \epsilon_i))$ with $\sigma = 10$.

We solve (numerically) for the equilibrium prices in each market, draw 100,000 consumers in each market, and simulate $T = 30$ discrete choices for each consumer. These choices could be different because each decision is associated with a new draw of ϵ_{it}. We thus end up with a panel of 10 million consumers who are evenly split across 100 markets, each making 30 discrete choices from among four products, which allows us to compute customer overlaps. In each market, we can also compute the 4-by-4 matrix of diversion ratios as a function of equilibrium prices and compare them to the corresponding overlaps. As explained below, we do this for both the absolute levels (of diversion ratios and overlaps) as well as for log normalized values, which may facilitate more straightforward economic interpretation.

Results. We begin by illustrating a single market configuration, which is shown in Appendix Figure A2. Panels (a) and (b) of Appendix Table A1 show the 4-by-4 overlap (left panel) against the 4-by-4 diversion ratios (right panel). The $j \to k$ overlap ($C_{j\to k}$) is defined

as the propensity of firm k to be selected among all the purchase occasions in which firm j's consumers — that is, consumers who have selected firm j at least once — do not select firm j. The $j \to k$ diversion ratio ($D_{j\to k}$) is defined as the market share of firm k out of the consumers lost by firm j due to a marginal price increase. The correlation between the two metrics is almost 1 (0.99).[2]

One may worry that the high correlation merely reflects underlying heterogeneity in market share as both the overlap and diversion ratios are increasing in market share. It is therefore instructive to normalize both the overlap and the diversion ratio by the overall market share of each firm.[3] After taking logs, the normalized measure is centered around zero. A value of zero occurs when consumer diversion or customer overlap follows the unconditional market share, with positive values implying closer substitutes, and negative values implying weaker substitutes. Panels (c) and (d) of Appendix Table A1 reports these log-normalized measures. The correlation remains 0.99.

Figure 1 presents the full results based on all 100 simulated market configurations. It shows bin-scatter plots of overlap on diversion ratio (top panel) and their corresponding (log) normalized versions (bottom panel). The overall correlations are very high (0.90 for the absolute measures and 0.78 for the log-normalized ones).

3 Dataset

Our primary data source, which is common to both (distinct) empirical exercises that we present in the following two sections, is transactions data from a large payment card network in the United States.[4]

An observation in our data is a credit or debit card transaction, and the information on each transaction is similar to the typical information one would find on monthly credit card statements: the name of the merchant, a unique card identifier, a transaction amount, and a date. Importantly, there is no information on the specific goods or services that were purchased nor their prices.

[2]Specifically, the overlap of firm j with firm k is given by $C_{j\to k} = \frac{\sum_{i=1} T_{(i,k)}\,\mathbb{I}_{(i,j)}}{\sum_{k'\in\mathcal{J}\setminus\{j\}}\sum_{i=1} T_{(i,k')}\,\mathbb{I}_{(i,j)}}$, where $\mathbb{I}_{(i,j)}$ is an indicator for individual i having ever selected firm j, $T_{(i,j)}$ is the total number of transactions between individual i and firm j, and $\mathcal{J}$ is the set of all firms in the market. Similarly, the diversion ratio of firm j with firm k is given by $D_{j\to k} = -\frac{s_k^* - s_k'}{s_j^* - s_j'}$, where s_k^* is the equilibrium market share of firm k and s_k' is counterfactual market share of firm k when the price of firm j marginally increases.

[3]To normalize both the overlap measure and diversion ratio, we divide these measures by the normalized market share of the competitor firm k, $s_k/(1-s_j)$. If either of these metrics is uninformative relative to a simple logit model of demand, then the normalized metric will equal 1.

[4]The entire database spans the 2019–2024 period and covers approximately 200 billion transactions per year, but we use only a small subset of the data by constructing sub-samples for specific exercises.

Each merchant is classified (by the card network) into the industry in which it operates and — for most physical card transactions — to the location (longitude and latitude) where the transaction occurred. In contrast, the card identifier is depersonalized and does not contain the cardholder's name, address, or any other personally identifiable information. Yet, as described below, we can use the entirety of the card transactions to construct proxies for certain cardholder characteristics.

4 Validation exercise

In this section, we attempt to validate the theoretical predictions in an empirical context. Specifically, we construct a sample of ready-to-drink coffee transactions in local markets, estimate a demand system in each market, and compute the diversion ratios implied by the demand estimates. We then compare these estimated diversion ratios to the corresponding customer overlaps, which we directly observe in the data. We find that our theoretical predictions replicate well in this empirical context, with a correlation of over 0.9 between the estimated diversion ratios and the corresponding customer overlaps.

Sample construction. We begin constructing our sample by defining a list of ready-to-drink coffee merchants in the transaction data. We start by restricting our sample to card transactions associated with merchants that are classified (by the data provider) as belonging to the food and drink market segment. These include ready-to-drink coffee, but also many other restaurants and food vendors. We then use the merchant names to identify coffee-related terms,[5] which we consider as the "universe" of ready-to-drink coffee transactions.

We also select a set of local geographical markets for our analysis. We define a market area as a specific county in 2019[6] and define the choice set to be all ready-to-drink coffee vendors operating in that county at some point in 2019.[7] To focus on a small and manageable set of counties of different size and affluence, we focus on counties with at least three coffee stores (in 2019) and select the 9 counties at the 10th, 20th, ..., and 90th percentiles of the

[5] Via partial string matching, we identify names with keywords (or partial keywords) including "coffee," "cafe," "tea," "bean," "brew," "caffee," "caffe," or "dunkin." Partial string identification on these keywords still yields some non-coffee merchants in our sample since only part of the name is required to match. So, we exclude merchant names with keywords (or partial keywords) such as "brewery," "steak," or "caribbean," which would otherwise be included based on the previous list of keywords but do not primarily serve coffee.

[6] We use 2019, which is the last full year before the Covid pandemic. The data from more recent, post-Covid years is less complete and is often missing exact store location for newer outlets. Results are very similar when we use 2023 data and stores with complete location information (not reported).

[7] Any vendor with a merchant name corresponding to one of the ready-to-drink coffee merchants as previously defined is considered a coffee vendor.

income distribution.[8] We drop the market that corresponds to the 10th percentile due to a limited number of coffee transactions, which leaves us with 8 markets that we analyze. These markets are quite heterogeneous, with 3 to 173 unique coffee stores in each market, representing 1 to 17 distinct chains.

Our final sample thus contains all coffee transactions in these 8 markets during 2019. Each transaction is identified by a market, a card identifier, and a store identifier.[9] In addition, for each card in the data, we construct three cardholder characteristics: (a) monthly average spending on each card (excluding all coffee purchases) that proxies for income, which we then map (county by county) to one of four income quartiles; (b) an indicator for ever transacting at a gas station, which proxies for car ownership; and (c) the average longitude and latitude of all within-county card transactions in 2019, which proxies for the cardholder location.

We assume that all coffee stores are optional choices for each cardholder. We construct the distance between every cardholder location and every coffee store in the market,[10] and we record the store chosen for each transaction.

Summary statistics on the final sample are given in Appendix Table A2. Overall, it contains 8 distinct markets, which are highly heterogeneous in size, affluence, and the number of coffee stores. The average market has 47 coffee stores, 6 coffee chains, almost 400,000 cardholders, and just over 3 coffee transactions per card on average. The sample has all the standard information — markets, choice sets, choices, and cardholder characteristics — that would allow us to estimate demand for ready-to-drink-coffee, where the travel distance between the cardholder and the store plays the role of the price.[11]

Demand estimation. We estimate, market-by-market, a standard random coefficients demand system for coffee stores using the transaction-level data described above. For simplicity, we only rely on observed heterogeneity across cardholders.

Specifically, consider the utility of consumer i from transacting at coffee store j to be

$$u_{ij} = \delta_j^{g(i)} - \alpha^{g(i)} d_{ij} + \epsilon_{ij}, \tag{6}$$

where $g(i)$ indicates the consumer segment/group to which consumer i belongs, and d_{ij} is the distance (in kilometers) between the consumer and the retailer locations. $\delta_j^{g(i)}$ and

[8] For county income, we take a county's total average gross income divided by the number of tax returns in that county to serve as a proxy. These measures are reported by the Internal Revenue Service Statistics of Income (2015).

[9] As mentioned, we also have information on the transaction amount, but this is not being used.

[10] Distance is measured (in kilometers) as the straight-line distance between a card's location and the location of the store using the haversine formula.

[11] Recall that we only observe the total transaction amount but do not have any information on what items have been transacted and their prices.

$\alpha^{g(i)}$ — the group-specific average utility of store j and the group-specific disutility from travel, respectively — are parameters to be estimated, and ϵ_{ij} are iid error terms, which are distributed type 1 extreme value.

To define consumer segments, we classify cardholders into eight groups, defined by the combination of their income quartile and car ownership proxies. Estimating this model is straightforward, as it implies standard logit demand for each pair of market and consumer segments. Yet, using the cardholder heterogeneity, the demand estimates generate reasonably rich substitution patterns.

Results. The key demand parameters — the estimates of $\alpha^{g(i)}$ for each market and consumer segment — are presented in Table 1. We estimate that higher-income cardholders are associated with lower sensitivity to distance and that, as expected, car owners are less sensitive to distance than non-owners (conditional on income).

We then use these data and results to generate an empirical analog to the theoretical exercise shown in Figure 1. Specifically, we generate two measures for each (ordered) pair of stores $j \to k$ in a given market. First, we use the data to directly compute the customer overlap; that is, the share of store j's customers who buy from store k. Second, we use the demand system to estimate the diversion ratio from j to k. To do this, we mimic a price increase by artificially increasing store j's distance from all its customers by one kilometer and use the demand parameters to estimate the resulting market share of each other store in the market.

Figure 2 presents the results in a format analogous to the theoretical exercise. The top panel presents a bin-scatter plot (using the universe of store pairs across all eight local coffee markets) of overlap against the diversion ratio, and the bottom panel repeats the same plot for the normalized measures. Both panels show a remarkably high correlation. This may not be surprising — after all, it is the same underlying overlap that drives the demand estimates — but the very high correlation between the two measures underscores how well the simple and easy-to-compute customer overlap tracks the much less scalable diversion ratio.

5 Illustration of potential use

5.1 Overlap-based screens for horizontal mergers

Overlap is clearly an imperfect measure, but the high correlation between customer overlaps and diversion ratios, and the simplicity and scalability of capturing overlap, could make it a tractable screen. We illustrate this point with a simple exercise, using the ready-to-drink

coffee demand estimates from the last section.

As Conlon and Mortimer (2021) show, under simplifying assumptions (no efficiency gains, similar markups across products, and abstracting from equilibrium response), the impact of a merger between firms j and k on firm j's price is driven by the diversion ratio, $D_{j\to k}$. We can thus approximate the "true" price impact of a (j,k) merger as proportional to $q_j D_{j\to k} + q_k D_{k\to j}$, where q_j and q_k are the number of customers of firms j and k, respectively. For the purpose of this exercise, we assume that if we knew the diversion ratios, we would be able to rank all potential mergers according to their price effect (as defined above) and block those that would lead to the largest increase in prices.

We construct the customer overlap analog, $q_j C_{j\to k} + q_k C_{k\to j}$, and (in a similar way to the use of HHI) use this as a screen to decide which mergers to approve, which ones to block, and which ones to further investigate. The higher the correlation between customer overlaps and diversion ratios, the more effective this screen would be.

To quantitatively illustrate the point, we consider the universe of all potential mergers between pairs of coffee retailers that operate within the same market (across all 8 markets we use in the last section). There are approximately 22,000 such potential mergers. For each potential merger, we compute the "true" price effect using our demand estimates and the corresponding diversion ratios, and the customer overlap analog. We then assume that the competition authority must block a fixed share of potential mergers, but does not have the resources to fully investigate each case by estimating the true diversion ratios. Instead, it must rely on the customer-overlap analog in order to determine which potential mergers to approve, block, or investigate, subject to a resource constraint which we assume is given by a fixed share of potential mergers that can be investigated.

Figure 3 illustrates the nature of the exercise, for a case in which 10% of the potential mergers must be blocked and only 5% can be investigated. This leads to upper and lower thresholds for the overlap: potential mergers below the lower threshold are approved, those above the higher threshold are blocked, and those in between get investigated. For the latter group, a fraction of them get blocked (the red dots in Figure 3) based on the "true" price effect (that is, the diversion ratio) that gets discovered.

Panel A of Appendix Table A3 summarizes the results for various combinations of the merger-blocking rate and the investigation-rate. The Table shows the price effect that is prevented relative to the potential price effect that could be prevented if all potential mergers could have been investigated. It suggests that using the overlap as a screen could be quite effective. As a benchmark, Panel B of Appendix Table A3 presents corresponding results when only data on market shares are available, in which case we use the change in concentration — that is, the incremental effect on the Herfindahl-Hirschman Index (HHI) –

as a screen (Nocke & Whinston, 2022). As can be seen, at least in this context, using the ΔHHI screen is useful, but much less effective.

5.2 Hotels

In this subsection, we illustrate a possible use of the overlap measure and potential insights it may help us draw in a context where estimating a full demand system (and the resulting diversion ratios) could be prohibitive. Specifically, we use the same payment card data to generate a rich sample of hotel transactions, compute customer overlap across hotel brands, and derive illustrative insights regarding substitution and competition across hotel chains.

Sample construction. We begin constructing our sample by collecting a list of all hotel chains that we observe in our transactions data. We restrict attention to card transactions that are associated with merchants classified (by the data provider) as belonging to the hospitality market segment. We further restrict to merchants with names containing common hotel-related keywords.[12] Using all observed transactions at these chains in 2023, we then construct national-level market shares, which are based on the count of transactions, and restrict our analysis to the 50 chains with the highest shares.[13]

Analysis. Our sample contains all card transactions at the top 50 chains during 2023. We then sort these 50 chains by an affluence measure that we construct at the chain level. To do so, we calculate the average monthly non-hotel spending (as a proxy for income) for each card that transacts at one of the 50 hotel chains and define the chain's affluence as the average across all the cardholders who transact at the chain in 2023.[14]

We then use the sample to calculate the pairwise overlap measure for all 50-by-50 pairs of hotel chains. As in the last section, the overlap measure is constructed using the count of transactions of each card at each chain (rather than the transaction amount), and we find it more instructive and easier to interpret to use the normalized overlap measure rather than the raw overlap.

[12]The full list of keywords we use includes: inn, hotel, resort, lodge, suite, hyatt, marriott, hilton, westin, aloft, motel, best western, radisson, ritz-carlton, sheraton, la quinta, meridien, iberotel, four seasons, extended stay america, swissotel, howard johnson, wyndham, super 8, and studio 6.

[13]We focus on a demand perspective, so consider each brand as a separate entity even if some brands in the data are co-owned. We also note that some hotel transactions are not for overnight stays. We suspect that, given our normalized measure, the main conclusions from the analysis are unlikely to be much affected.

[14]While we cannot identify chains by name due to data confidentiality, the affluence measure captures conventional wisdom well: luxurious chains at the top, and chains with more basic amenities at the bottom.

Results. The results are summarized in Figure 4, which depicts a heat map of the log-normalized overlap measure for every pair of chains. The figure presents a 50-by-50 matrix for the top 50 chains, where chains are ordered by their chain affluence measure (the most affluent chain is at the top/left and the least affluent is at the bottom/right). Each cell of the matrix presents the (color coded) log-normalized overlap $NC_{j\to k}$, where j is the row chain and k is the column chain. Blue colors represent closer substitutes (positive log-normalized overlap) and red colors represent further substitutes (negative log-normalized overlap).

The figure illustrates several patterns that seem useful and are interesting to point out. First, the most affluent hotels have much higher overlap with each other, and the least affluent hotels have much higher overlap with each other. This is demonstrated by the dark blue cells in the upper left and lower right corners of the figure. Conversely, the mostly red color away from the main diagonal of the figure suggests that high-affluence and low-affluence chains are not close substitutes. This is not surprising, of course, yet provides some validation for our overlap measure. One can imagine many other settings in which the data we use to construct the affluence measure is not available and the overlap measure would be the only measure to rely on.

A second interesting pattern in Figure 4 is that the chains identified by A, B, C, D, and E exhibit especially high overlap with one another (as demonstrated by the very dark blue cells). Upon inspecting the identities of these chains, we find that they all belong to the same large hotel group. Thus, our overlap measure captures some information about the ownership structure in this market, which might be associated with a common loyalty program for example.

A third noticeable pattern is that the chains marked by F and G exhibit unusual overlap patterns. Both chains seem to have log-normalized overlap close to 0 with all other chains (columns are shaded light blue and light pink), suggesting that substitution patterns away from these chains seem to be close to uniform across the other chains. At the same time, other chains seem to have low overlap with these two chains (rows are almost entirely red), suggesting that consumers do not exhibit much substitution away from other hotel chains towards these two, F and G. Again, upon inspecting the identities of these chains, we find that one is a popular ski and entertainment resort and the other is a hotel and casino chain, making them quite differentiated from all other chains in the market. As before, these patterns seem sensible once the nature and identity of the chains are known. But the overlap measure provides a useful metric to capture this, which seems valuable when information about likely substitution is soft, hard-to-quantify, or not available.

6 Conclusions

In this paper, we proposed the use of customer overlap as a proxy for diversion ratios. We illustrate theoretically why the two objects are closely related and use a simple empirical application to validate that the two concepts track each other closely. We then illustrate how this may be used in a context where directly estimating diversion ratios could be prohibitive.

With the increasing availability of digital data sets that offer researchers and policymakers access to large consumer panel data — such as payment card data (as we use in this paper), cell phone data (e.g., SafeGraph), or browsing data (e.g., Comscore) — we think that the concept of customer overlap could be useful quite broadly as an initial proxy for horizontal substitution patterns across products, firms, or market segments.

It is extremely important to caveat that this interpretation of customer overlap as an indicator for substitution patterns is *not* universally applicable. There are many examples in which the use of customer overlap as a proxy for diversion ratios would be wrong and misleading. One case is of complement products; e.g., it is likely common that the same consumer purchases both peanut butter and jelly, but this is obviously because they are complementary products rather than close substitutes. Another case is products that are offered in completely different markets; e.g., there could be a non-trivial customer overlap between the Newark airport Starbucks and the one at O'Hare airport in Chicago, but we do not think that these two coffee places actually compete with each other. A final example concerns markets in which multi-homing is common, when consumers may regularly split their purchases across sellers; for instance, if consumers buy their packaged groceries in Safeway but their produce at Trader Joe's, the customer overlap between the two chains would be high, but they may not be very close substitutes.[15]

Thus, the interpretation of customer overlap should be assessed on a case-by-case basis, and we caution against a "blind use" of this approach. For an appropriate interpretation of customer overlap as a proxy for substitution and diversion ratios, one must first begin with a clearly identified set of choices or products that are natural substitutes to each other, as we tried to do with the coffee and hotel applications in this paper.

[15] In many of these examples, additional data (if available) could be deployed to resurrect the usefulness of overlap. For example, for the airport Starbucks example one could condition on location to avoid mixing up distinct geographic markets, and in the grocery store example one could condition on the type of items that are being purchased (if known).

References

Atalay, E., Frost, E., Sorensen, A., Sullivan, C., & Zhu, W. (forthcoming). Scalable demand and markups. *Journal of Political Economy*, forthcoming.

Conlon, C., & Mortimer, J. (2021). Empirical properties of diversion ratios. *RAND Journal of Economics*, *52*(3), 558–579.

Conlon, C., Mortimer, J. H., & Sarkis, P. (2023). *Estimating preferences and substitution patterns from second-choice data alone* (tech. rep.). New York University.

Farrell, J., & Shapiro, C. (2010). Antitrust evaluation of horizontal mergers: An economic alternative to market definition. *The B.E. Journal of Theoretical Economics*, *10*(1), 1–41.

Internal Revenue Service Statistics of Income. (2015). SOI Tax Stats - County Data 2015 [Accessed: 2025-01-28].

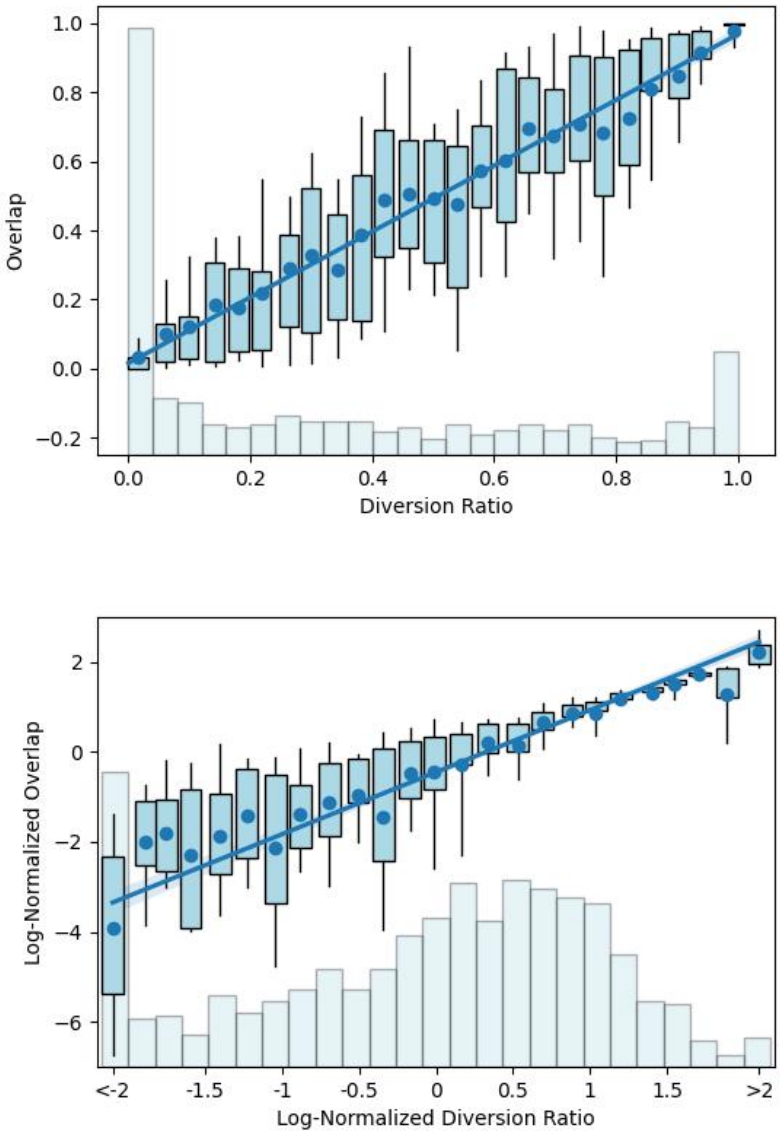
McClure, J. (2025). *Using default recommendations in demand estimation* (tech. rep.). Purdue University.

Nocke, V., & Whinston, M. D. (2022). Concentration thresholds for horizontal mergers. *American Economic Review*, *112*(6), 1915–1948.

Qiu, Y. J., Sawada, M., & Sheu, G. (2024). Win/loss data and consumer switching costs: Measuring diversion ratios and the impact of mergers. *Journal of Industrial Economics*, *72*(1), 327–355.

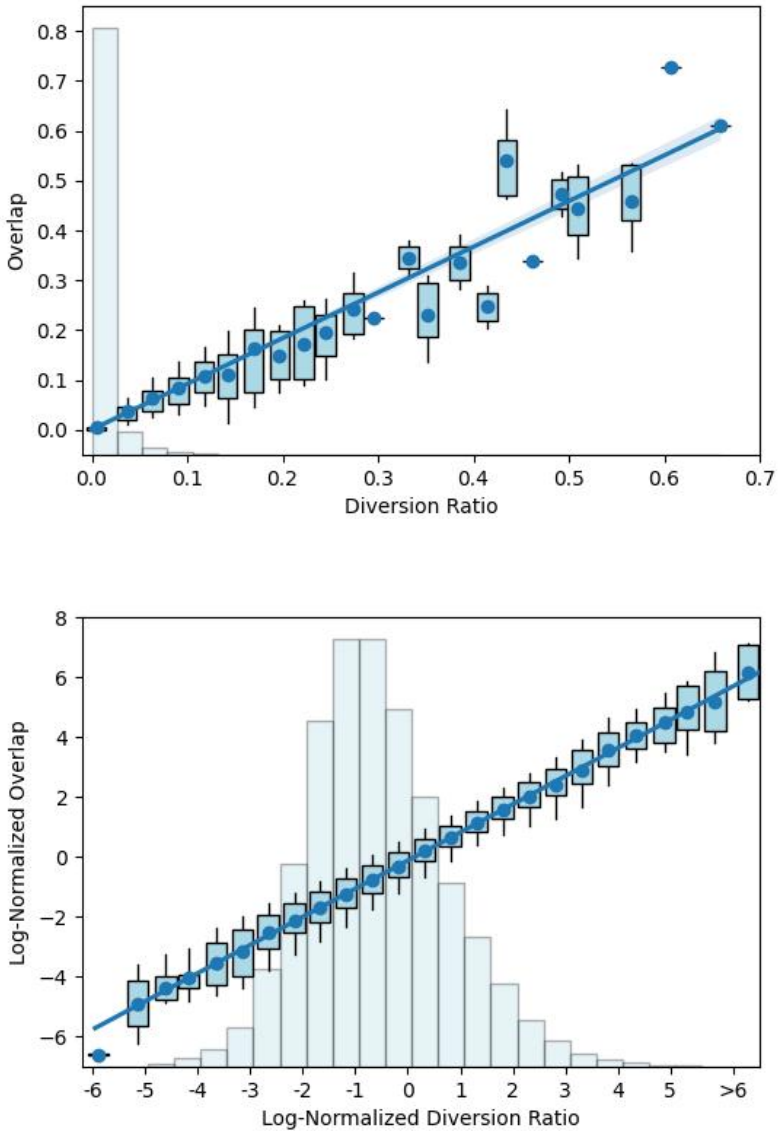
Raval, D., Rosenbaum, T., & Tenn, S. A. (2017). A semiparametric discrete choice: An application to hospital mergers. *Economic Inquiry*, *55*(4), 1919–1944.

Figure 1: Simulation Results



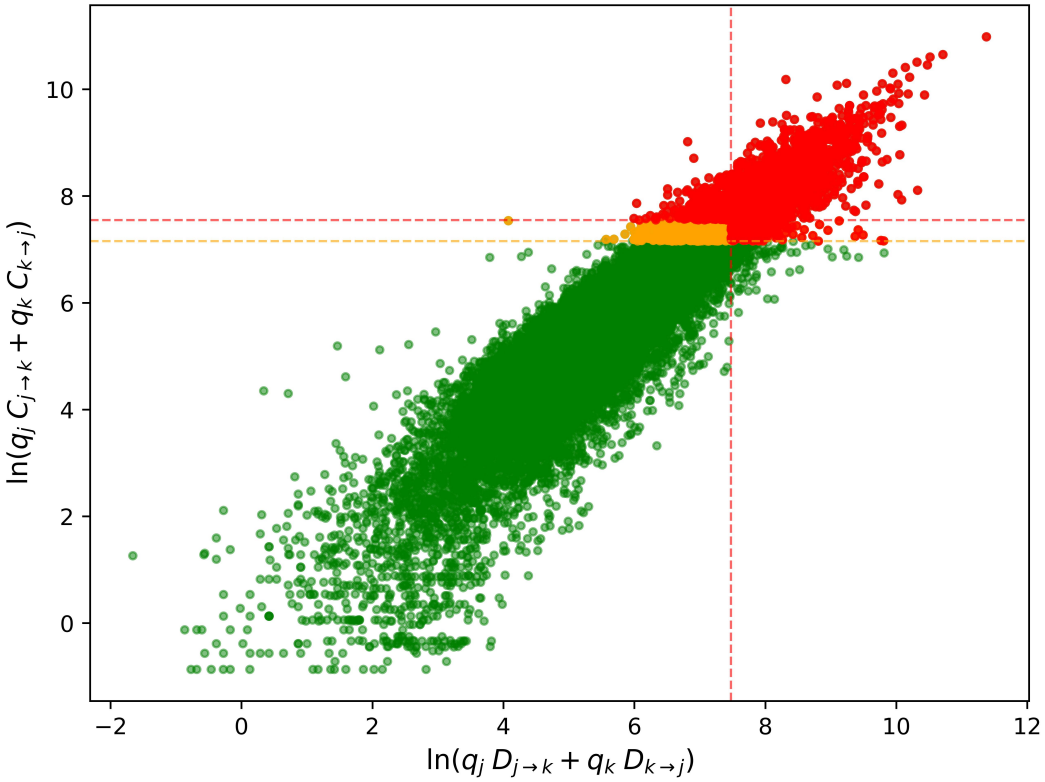
Note: These figures depict the correlation between overlap and diversion ratios computed from 100 market simulations of 4 firms. In each market, we simulate 100,000 consumers with fixed θ_i and $\epsilon_{it} \sim Beta(10\epsilon_i + 1, 10(1-\epsilon_i) + 1)$ across 30 periods. Overlap is calculated by using consumers' choices across all periods to determine the customer base of each firm. Diversion ratios are calculated by changing the equilibrium price of each firm in the first period and simulating consumers' new optimal choices. In panel (a), we regress overlap on diversion ratio and estimate a coefficient of 0.95 with a corresponding R^2 of 0.81. In panel (b), we regress the log-normalized overlap on log-normalized diversion ratio (excluding values where either measure is $-\infty$) and also estimate a coefficient of 0.99 with a corresponding R^2 of 0.61. We split each sample into 25 equidistant bins on the x-axis and plot the mean, 10th, 25th, 75th, and 90th percentiles of (log-normalized) overlap in each box-and-whisker plot. The light bars underneath each bin represent the distribution of (log-normalized) diversion ratios between firms across all simulated markets. In total, we have 1,200 observations in the raw sample and 869 observations after normalizing and removing observations with values of -∞ for either measure.

Figure 2: Coffee Results



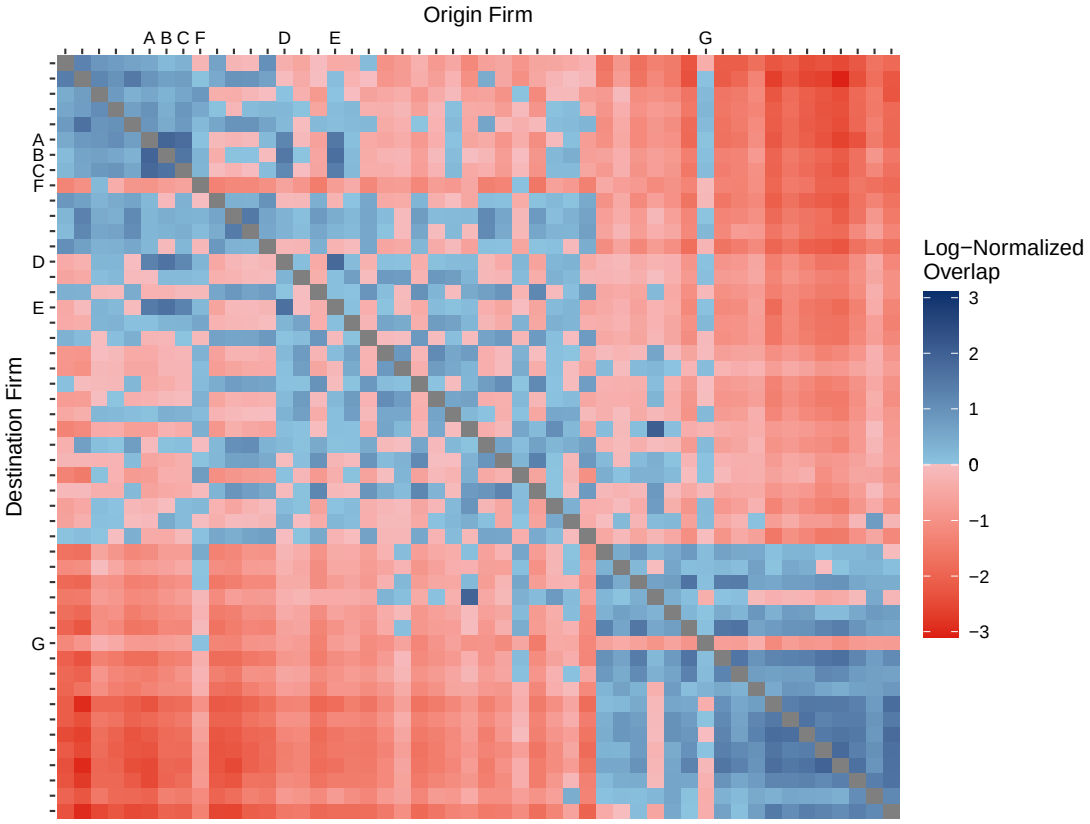
Note: These figures depict the correlation between overlap and diversion ratios computed from the 8 markets for ready-to-drink coffee (as described above). In each market, we estimate a logit model of demand (with merchant fixed effects) using all transactions of coffee consumers, accounting for the distance between vendors and consumers and heterogeneity across consumers by income quartiles and car-ownership status. Overlap is calculated directly from consumers' transaction choices among coffee vendors in 2019. Diversion ratios are calculated by mechanically increasing the distance between all consumers and each firm by 1 kilometer and then simulating consumers' new optimal choices. In panel (a), we regress overlap on diversion ratio and estimate a coefficient of 0.92 with a corresponding R^2 of 0.74. In panel (b), we regress the log-normalized overlap on log-normalized diversion ratio (excluding values where either measure is $-\infty$) and also estimate a coefficient of 0.94 with a corresponding R^2 of 0.75. We split each sample into 25 equidistant bins on the x-axis and plot the mean, 10th, 25th, 75th, and 90th percentiles of (log-normalized) overlap in each box-and-whisker plot. The light bars underneath each bin represent the distribution of (log-normalized) diversion ratios between firms across all simulated markets. In total, we have 43,952 observations in the raw sample and 42,738 observations after normalizing and removing observations with values of -∞ for either measure.

Figure 3: Overlap Threshold Example



Note: This figure depicts the log-scaled, post-merger price effects (x axis) against the customer-overlap analog (y axis) for the 21,976 potential mergers in our sample of ready-to-drink coffee markets. As described in the main text, the (log-scaled) "true" price effects is defined by $\ln(q_j D_{j\to k} + q_k D_{k\to j})$ and the (log-scaled) customer-overlap analog overlap is given by $\ln(q_j C_{j\to k} + q_k C_{k\to j})$. The color of each point corresponds to whether the merger would be rejected (red), investigated and ultimately accepted (orange), or accepted without review (green) for a particular investigation rate (5% of all potential mergers) and rejection rate (10% of all potential mergers). The dashed horizontal lines represent the (optimally chosen) customer-overlap thresholds: above the higher threshold a merger is automatically rejected, below the lower threshold is accepted, and in between it is investigated and gets rejected (red) or accepted (orange) based on the (discovered through investigation) diversion ratios.

Figure 4: Hotel Overlap



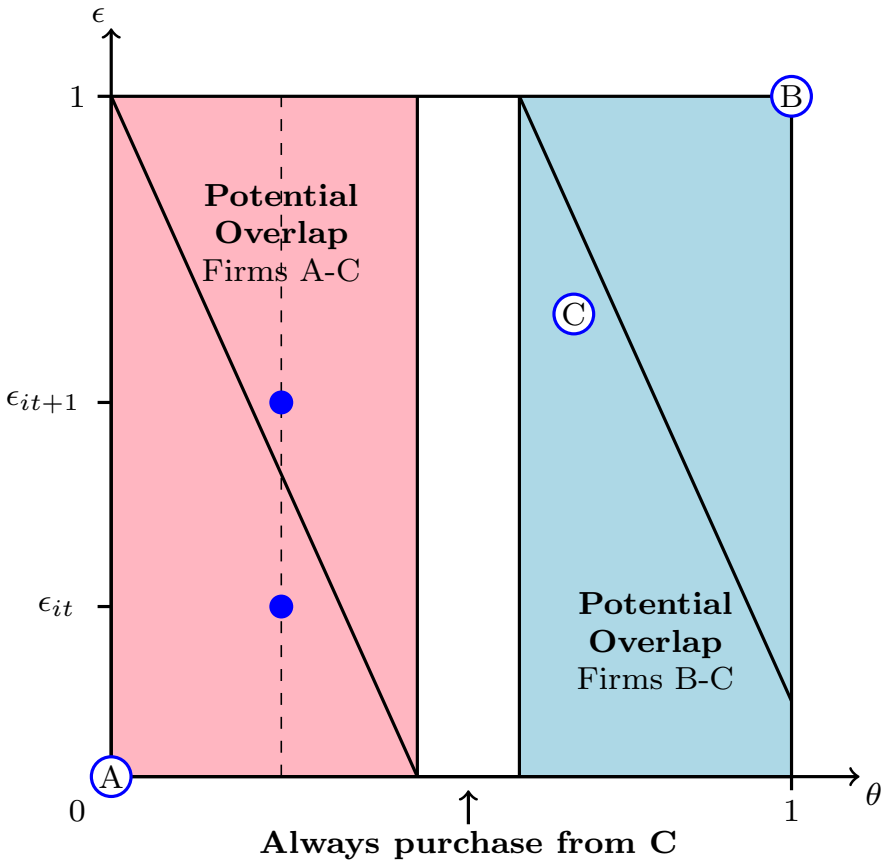
Note: This picture visualizes all pairwise log-normalized overlap measures for the top 50 hotel chains in the US. Hotel chains are sorted from highest non-hotel spending rank to lowest non-hotel spending rank across both axes, going from left to right and from top to bottom (i.e. highest ranked chains are in the top left and lowest ranked chains are in the bottom right). A single cell of this matrix can be interpreted as the log-normalized overlap of the column chain with the row chain (i.e. $NC_{col \to row}$). Thus, any column can be interpreted as how much that chain's customers relatively overlap with other chains. Similarly, any row can be interpreted as how much all other chains' customers overlap with the row chain. Values above zero are colored blue, with darker colors indicating higher log-normalized overlap. Similarly, values below zero are colored red, with darker shades indicating lower log-normalized overlap.

Table 1: Elasticity of Demand wrt Distance by Market Segment

Market	1st Income Quartile		2nd Income Quartile		3rd Income Quartile		4th Income Quartile	
	Car	No Car	Car	No Car	Car	No Car	Car	No Car
2	-1.40	-1.53	-1.37	-1.70	-1.18	-0.86	-1.12	-1.47
	(0.03)	(0.06)	(0.02)	(0.06)	(0.02)	(0.04)	(0.02)	(0.07)
3	-2.13	-2.10	-1.94	-2.26	-1.97	-1.79	-1.77	-1.99
	(0.03)	(0.04)	(0.02)	(0.05)	(0.02)	(0.04)	(0.01)	(0.05)
4	-1.95	-2.10	-2.01	-1.92	-1.79	-1.93	-1.62	-1.76
	(0.02)	(0.03)	(0.02)	(0.04)	(0.01)	(0.04)	(0.01)	(0.03)
5	-2.13	-2.29	-2.11	-2.14	-2.10	-2.07	-2.00	-2.10
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
6	-1.88	-2.04	-1.86	-1.91	-1.86	-1.81	-1.83	-1.79
	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)
7	-2.17	-2.33	-2.02	-2.29	-2.00	-2.20	-1.69	-2.12
	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.02)	(0.00)	(0.02)
8	-1.78	-1.84	-1.68	-1.67	-1.69	-1.66	-1.68	-1.64
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
9	-1.79	-2.10	-1.84	-1.93	-1.78	-1.78	-1.64	-1.72
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.01)

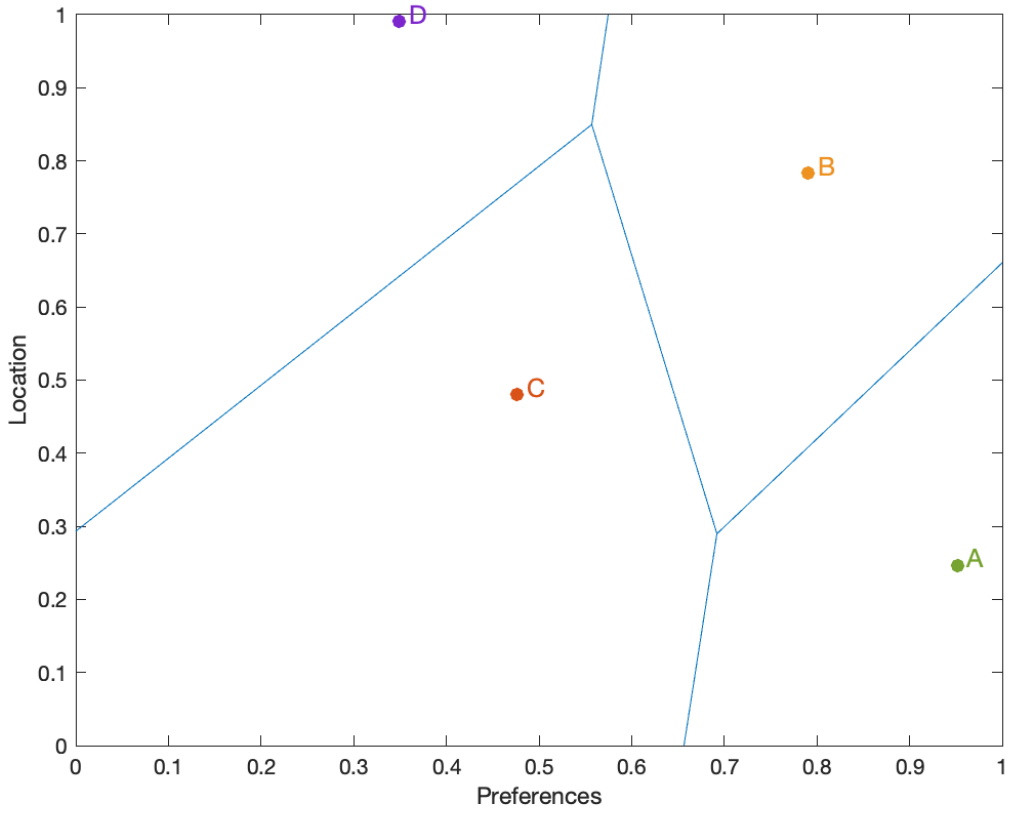
Note: In this table, we report the estimated coefficient on distance, $\alpha^{g(i)}$, in our model of consumer demand for ready-to-drink coffee. We report an estimate for each market segment, defined at the market-by-income-quartile-by-car-ownership-status level. Thus, we report eight estimates for each of our eight markets. The first income quartile corresponds to the lowest income quartile and the fourth income quartile corresponds to the highest. We report the standard errors for each estimate in parentheses below. These standard errors are calculated analytically using the gradient of our loss function evaluated at the estimate. More specifically, we report the square root of the inverse of the Hessian of our negative log-likelihood loss evaluated at the coefficient estimate as our analytic standard error. Conditional on estimating our coefficients via this loss function (which we do), this yields consistent estimates of the standard errors.

Appendix Figure A1: Overlap Illustration



Note: This figure demonstrates consumer choices implied by the model described in Section 2. Given that consumer preferences are fixed at θ_i but their location ϵ_{it}) fluctuate from one choice occasion to another, consumers could either be changing their choices between firms A and C (red region), always choose firm C (white region), or changing their choices between firms C and B (blue region), as illustrated above.

Appendix Figure A2: Graphical depiction of a single simulated market



Note: This is a graphical illustration of one of the 100 simulated markets described in Section 2, which corresponds to the overlap and diversion ratios reported in Appendix Table A1. Each blue line represents an indifference curve between two "adjacent" firms. The equilibrium prices in this market that generate these indifference curves are $p_A^* = 0.31$, $p_B^* = 0.30$, $p_C^* = 0.36$, $p_D^* = 0.33$.

Appendix Table A1: Numerical Simulation Example

(a) Overlap

	Firm A	Firm B	Firm C	Firm D
Firm A	–	0.972	0.028	0
Firm B	0.712	–	0.286	0.003
Firm C	0.012	0.185	–	0.804
Firm D	0	0.003	0.997	–

(b) Diversion Ratio

	Firm A	Firm B	Firm C	Firm D
Firm A	–	0.899	0.101	0
Firm B	0.589	–	0.389	0.022
Firm C	0.051	0.262	–	0.687
Firm D	0	0.019	0.981	–

(c) Log Normalized Overlap

	Firm A	Firm B	Firm C	Firm D
Firm A	–	1.63	-2.18	$-\infty$
Firm B	1.02	–	-0.24	-4.69
Firm C	-3.70	-1.01	–	0.44
Firm D	$-\infty$	-4.55	0.94	–

(d) Log Normalized Diversion Ratio

	Firm A	Firm B	Firm C	Firm D
Firm A	–	1.55	-0.90	$-\infty$
Firm B	0.83	–	0.07	-2.55
Firm C	-2.21	-0.66	–	0.28
Firm D	$-\infty$	-2.75	0.92	–

Note: Each cell in Panel (a) represents the overlap between a pair of firms. That is, the element (j, k) reports the share of firm k out of non-j purchases of consumers who ever bought from firm j. Panel (b) reports the corresponding diversion ratios from firm j to form k. The correlation between the cells in these two panels is 0.989. Panels (c) and (d) normalize the overlap and diversion ratios above by dividing each cell by $s_k/(1-s_j)$ (where s_k and s_j are the overall market shares of firm k and j) and then taking the log of each normalized measure. The correlation between cells in these two panels (excluding the values equal to $-\infty$) is 0.992.

Appendix Table A2: Summary Statistics of Coffee Vendor Markets

Market	Unique Cards (000s)	Share with a Car	Stores	Brands	Txns per Card	Amount per Txn ($US)
2	20.7	0.78	3	1	2.80	8.13
3	33.5	0.74	3	1	3.49	9.10
4	42.0	0.74	6	1	2.92	9.21
5	1603.2	0.77	173	16	4.24	9.01
6	212.5	0.82	26	9	3.29	10.93
7	166.7	0.76	14	6	3.33	8.62
8	854.8	0.76	110	11	3.73	8.71
9	167.1	0.65	37	7	2.96	8.96
Avg	387.6	0.75	46.5	6.5	3.35	9.08

Market	Store HHI	Brand HHI	Avg Monthly Spend ($US)	Travel Distance (km)		
				10^{th} pctile store	Median store	90^{th} pctile store
2	0.380	1	1719	6.67	9.96	14.42
3	0.335	1	2168	3.38	4.92	6.62
4	0.218	1	2338	8.04	13.57	20.23
5	0.008	0.879	3335	12.67	28.58	47.52
6	0.052	0.408	2559	3.70	9.31	16.21
7	0.100	0.567	2310	5.68	10.51	17.10
8	0.011	0.599	3227	8.98	24.82	42.49
9	0.033	0.838	2591	10.50	17.22	33.71
Avg	0.142	0.786	2532	14.86	7.45	24.79

Note: The definitions of a card, car owner, store and brand are described in detail in Section 3. *Txns per Card* is defined as the total number of transactions made by cardholders in the market divided by the number of unique cards. Similarly, Amount per Txn is the defined as the sum of transaction amounts (in $s) divided by the number of transactions in the market. HHIs are calculated from the share of transactions in a market (as opposed to revenues), both at the store-level and brand-level. Monthly spending is defined as the average monthly spending that we observe for each card. Therefore, *Average Monthly Spending* is defined as the average across all cards within a market of each card's average monthly spending. At the card level, travel distance to the median store is defined as the median distance between a cardholder and all available coffee vendors in their market, irrespective of whether a transaction occurred. Thus, this measure should not be interpreted as the median distance traveled, but the median distance among potential vendors. We report the average of these median distances across all cardholders within each market. Similarly, we report the average 10^{th} percentile of distance and the average 90^{th} percentile of distance in each market.

Appendix Table A3: Proportion of Avoidable Price Effects

Panel A: Customer Overlap

	Share blocked		
Share investigated	1%	4%	8%
0%	0.812	0.887	0.928
0.2%	0.829	0.895	0.931
0.6%	0.869	0.906	0.935
1%	0.893	0.914	0.938
2%	0.926	0.932	0.945
5%	0.968	0.966	0.964
50%	> 0.999	> 0.999	> 0.999

Panel B: ΔHHI

	Share blocked		
Share investigated	1%	4%	8%
0%	0.564	0.533	0.550
0.2%	0.573	0.538	0.551
0.6%	0.597	0.547	0.556
1%	0.614	0.563	0.560
2%	0.686	0.585	0.573
5%	0.803	0.652	0.610
50%	0.989	0.973	0.954

Note: This table reports the proportion of avoided post-merger price effects as a share of of the maximum avoidable price effect for the given merger rejection rate given by the column header; that is, relative to price effect that would be avoided if all potential mergers were to be investigated. We report this proportion for a set of different investigation rates (in each row) and total rejection rates (in each column). Panel A reports these statistics for the screening exercise using our proposed customer overlap analog. Panel B reports these statistics for the screening exercise using ΔHHI scaled by the total number of transactions within the relevant market.