Customer Overlap and Diversion Ratios

Liran Einav, Stanford and NBER
Mariana Guido, Stanford
Pete Klenow, Stanford and NBER

Conference on Antitrust Economics and Competition Policy
Northwestern University, September 19, 2025

Motivation

- Diversion ratios have become quite central in competition policy over the last decade or so

$$D_{j\to k} = \frac{D'_k - D_k}{D_j - D'_j} = \frac{\textit{Customers lost by firm } j \textit{ who switched to firm } k}{\textit{Customers lost by firm } j}$$

  - More informative than cross-price elasticities about Upper Pricing Pressure (UPP) in horizontal mergers
  - Featured quite prominently in the 2010 DOJ Merger guidelines

- But obtaining estimates of diversion ratios is not always easy, and certainly not scalable
  - Often relies on “full-blown” estimated demand systems or second choice surveys

- This paper: given the advent of various forms of new data sets (credit cards, Comscore, SafeGraph), try to propose a much simpler, data-driven, scalable initial screen / alternative (**at least in some sectors/contexts**)

Basic idea

- Define a measure of **customer overlap**

$$C_{j\to k} = \frac{\sum_{i\in C_j} 1(c_{it}=k)}{\sum_{i\in C_j} 1(c_{it}\neq j)} = \frac{transactions\ at\ k\ by\ j's\ customers}{total\ non{-}j\ transactions\ by\ j's\ customers}$$

- Customer overlap could be used to approximate substitutability, and in particular diversion ratios
  - At least in the right context (with multiple episodes of buyer-seller "relationships")
  - Intuitive approximation for mixed logit model, where $D_{j\to k} \propto \int s_{ik}s_{ij}di$ (Conlon and Mortimer 2021)

- Unlike diversion ratios (or cross-price elasticities or other measures of substitution), customer overlap is directly observed in the data
  - No need for any modeling assumptions
  - Easy and fast to compute, so much more scalable

Log-normalized measures

The main value added is "on top" of the logit diversion ratios, which are captured by HHI.

To capture this explicitly, it will be useful to also work with log-normalized measures:

$$ND_{j\to k} = ln\left(D_{j\to k}\right) - ln\left(\frac{D_k}{1-D_j}\right)$$

$$NC_{j\to k} = ln\left(C_{j\to k}\right) - ln\left(\frac{\sum 1(\,c_{it} = k)}{\sum 1(\,c_{it} \neq j)}\right)$$

- Positive values: closer substitutes (conditional on size)
- Negative values: further substitutes (conditional on size)

Related literature

- Horizontal merger guidelines, Upper pricing pressure, and the importance of diversion ratios (Conlon and Mortimer 2021)

- Ideas that are similar to customer overlap
  - Win/Loss ratio (Qiu, Sawada, and Sheu 2024)
  - “Co-purchasing” rates (Atalay et al. 2025)
  - “Cross-shopping” data
  - Second-choice data

- Loosely connected to demand systems with consumer panel data

Outline

- Motivating theory
  - Why would it make sense for overlap and diversion ratios to be related?
- Validation exercise (coffee …)
  - How well are diversion ratios approximated by overlap?
- Illustrative use cases
  - Overlap as a screen
  - Hotels …

Outline

- **Motivating theory**
  - **Why would it make sense for overlap and diversion ratios to be related?**
- Validation exercise (coffee …)
  - How well are diversion ratios approximated by overlap?
- Illustrative use case
  - Overlap as a screen
  - Hotels …

Setting

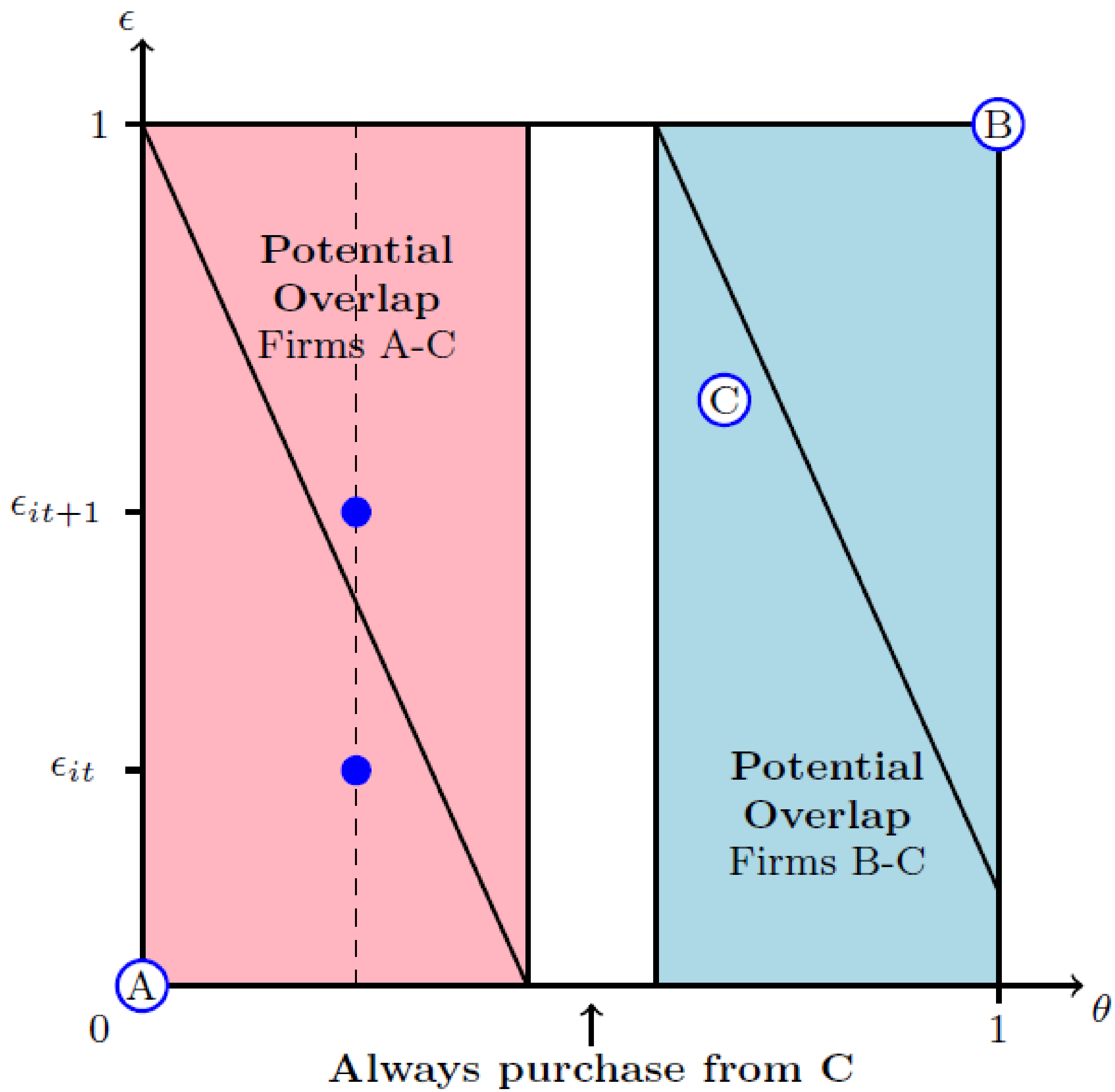
- Hotelling-style setting with two dimensions
  - θ: preferences
  - ϵ: physical location

- Firms $j = 1,2,\dots,J$:
  - Firm j is defined by a fixed (two-dimensional) location (θ_j, ϵ_j)
  - Firms set prices in a Nash Equilibrium

- Continuum of consumers, each denoted by i:
  - Consumer i defined by (θ_i, ϵ_i): fixed preferences and a "baseline" location
  - Consumers make a sequence of choices in periods $t = 1,2,\dots,T$, where locations are stochastic: $\epsilon_{it} \sim G(\epsilon_i)$
  - They make discrete choices that maximize period-t utility

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

Illustration



Numerical parameterization

- 4 firms (A, B, C, and D)

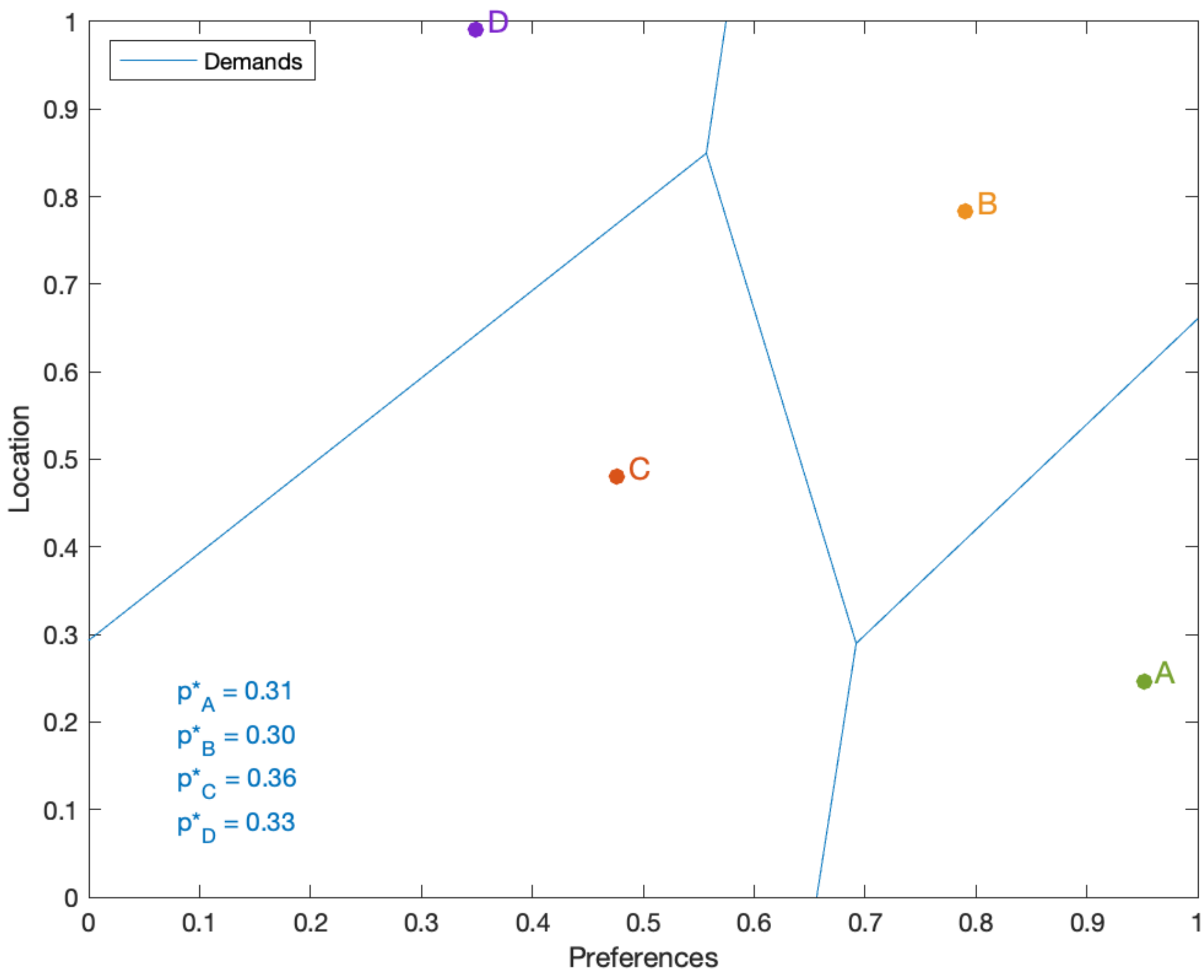
- 100 markets (randomly drawn firm locations)
- 100,000 x 30 choices in each market

- Consumers:

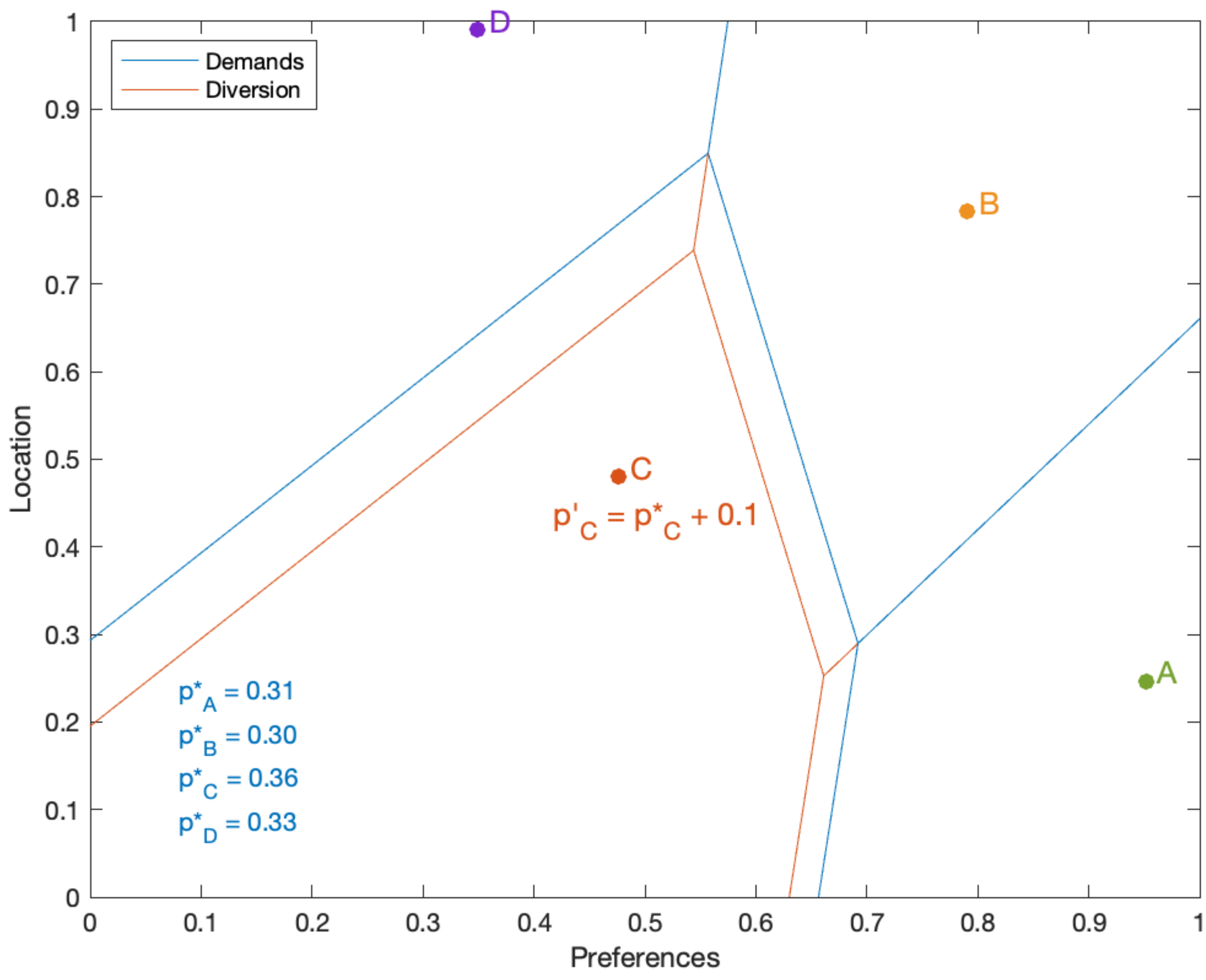
$$u_{ijt} = v - \alpha p_j - \lambda_\theta \left(\theta_j - \theta_i\right)^2 - \lambda_\epsilon \left(\epsilon_j - \epsilon_{it}\right)^2$$

  - $\theta_i \sim \mathbb{U}[0,1]$
  - $\epsilon_i \sim Beta(2,2)$
  - $\epsilon_{it} \sim Beta(1 + \sigma\epsilon_i, 1 + \sigma(1 - \epsilon_i))$ with σ=10
  - $v \gg \lambda_\theta + \lambda_\epsilon$(outside option not relevant)
  - $\lambda_\theta = 4, \lambda_\epsilon = 1$

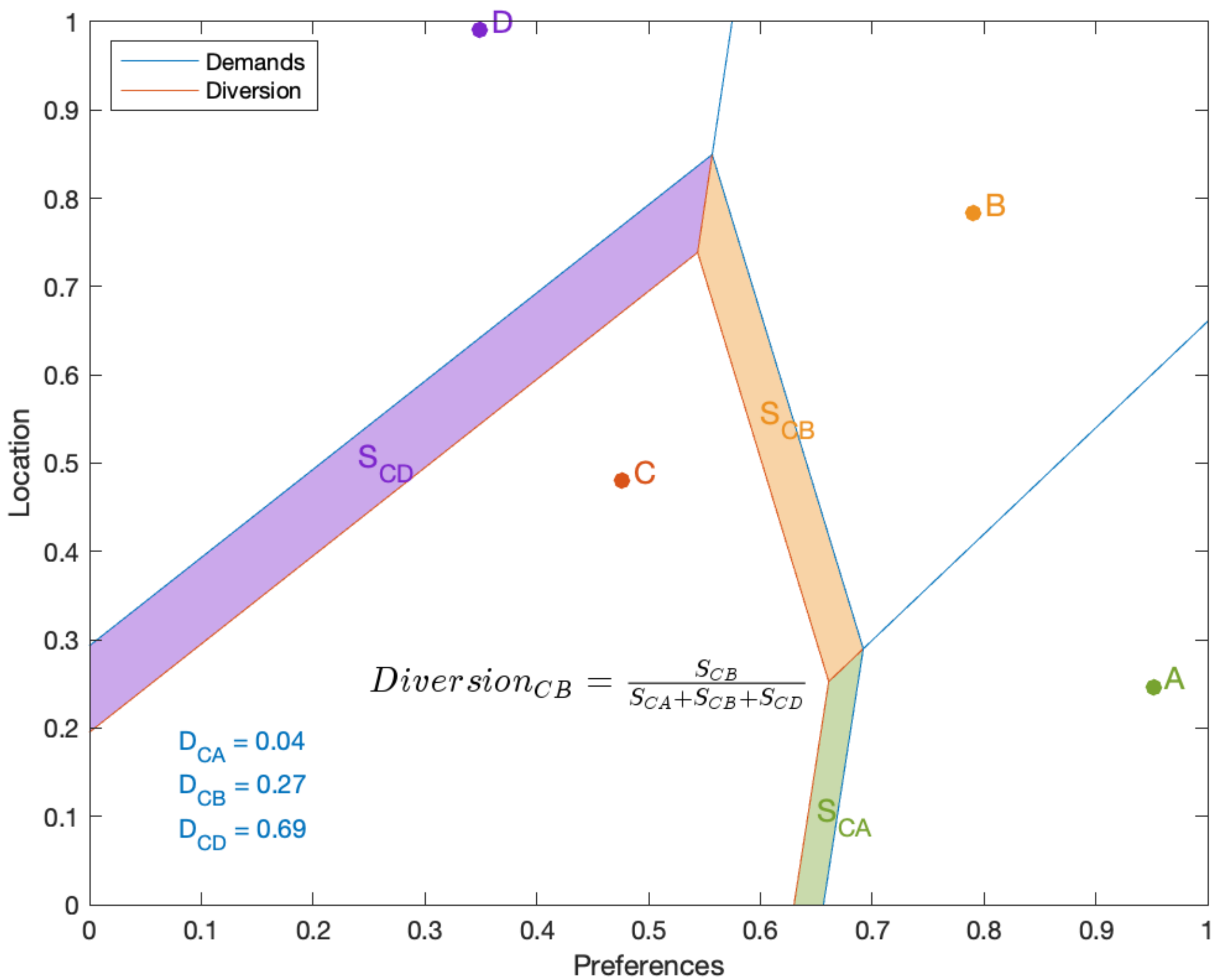
Graphical illustration (one market)



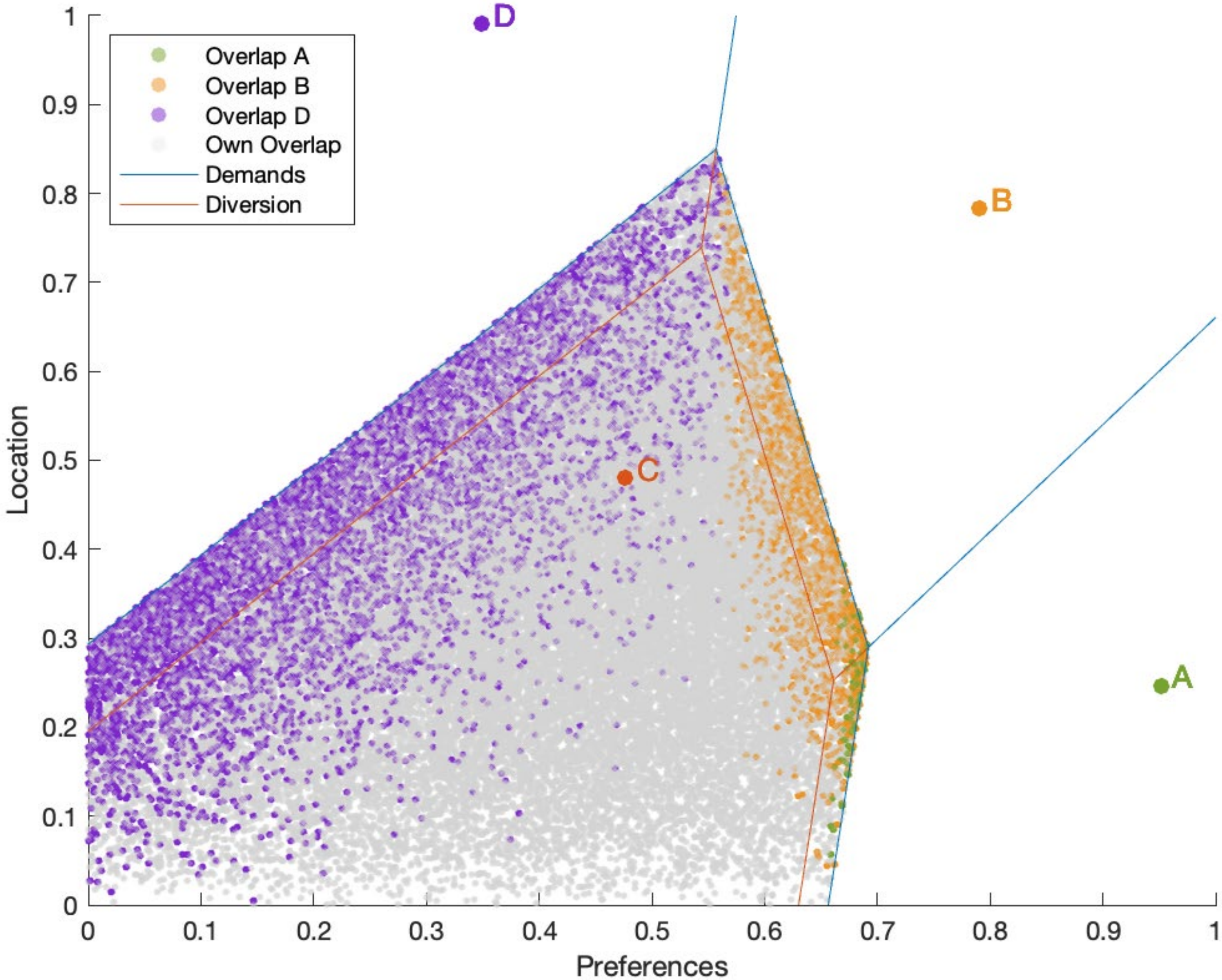
Graphical illustration (one market)



Graphical illustration: diversion ratios



Graphical illustration: overlap



Results from that market

- Solve for equilibrium, compute diversion ratios, simulate overlap

(a) Overlap

	Firm A	Firm B	Firm C	Firm D
Firm A	–	0.973	0.027	0
Firm B	0.709	–	0.289	0.002
Firm C	0.011	0.188	–	0.801
Firm D	0	0.002	0.998	–

(b) Diversion Ratio

	Firm A	Firm B	Firm C	Firm D
Firm A	–	0.901	0.099	0
Firm B	0.591	–	0.387	0.022
Firm C	0.043	0.268	–	0.689
Firm D	0	0.016	0.984	–

(c) Log Normalized Overlap

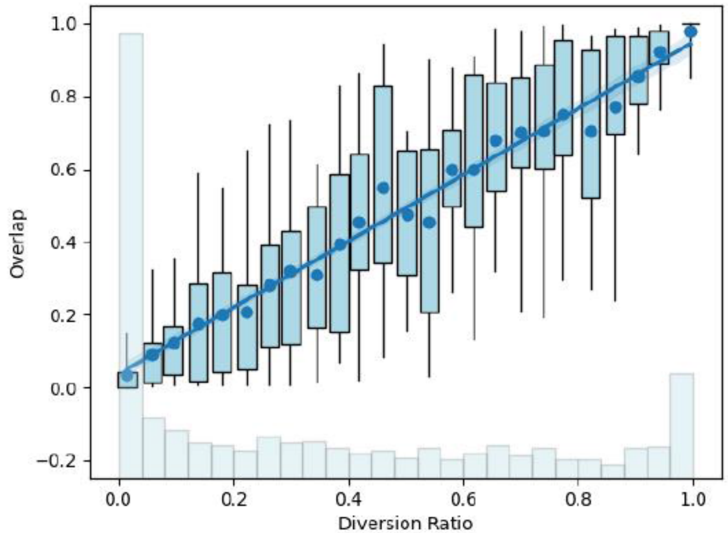
	Firm A	Firm B	Firm C	Firm D
Firm A	–	1.33	-2.88	$-\infty$
Firm B	1.34	–	-0.59	-4.97
Firm C	-3.12	-0.68	–	0.73
Firm D	$-\infty$	-4.78	0.64	–

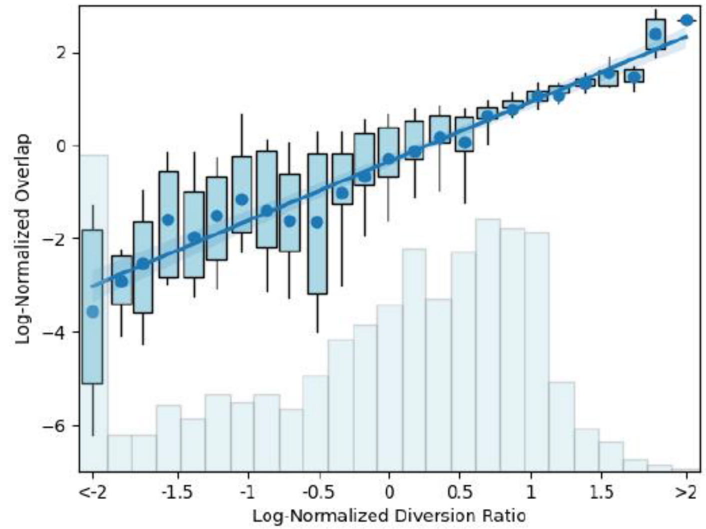
(d) Log Normalized Diversion Ratio

	Firm A	Firm B	Firm C	Firm D
Firm A	–	1.26	-1.57	$-\infty$
Firm B	1.16	–	-0.30	-2.59
Firm C	-1.73	-0.32	–	0.58
Firm D	$-\infty$	-2.85	0.62	–

- Correlations of 0.990 (top) and 0.994 (bottom)
  - but this was (slightly!) cherry picked

Full simulation results





- R^2 of 0.81 (left) and 0.55 (right)

Outline

- Motivating theory
  - Why would it make sense for overlap and diversion ratios to be related?
- **Validation exercise (coffee …)**
  - **How well are diversion ratios approximated by overlap?**
- Illustrative use case
  - Overlap as a screen
  - Hotels …

Overview of the validation exercise

- Create a sample with a set of geographical markets
- Estimate demand in each market (computationally-intensive, requires customer and store locations)
- Use the demand estimates to calculate diversion ratios
- Compare the calculated diversion ratios to the corresponding customer overlap measures

Sample

- Focus on coffee shop market as working example: no concern about multi-stop trips
- Define market as county-year
- Choose counties on 2 conditions:
  - Number of coffee stores in county ≥ 3
  - Represent all consumer income deciles (10^{th}, 20^{th}, etc. pctiles (10^{th} is dropped due to data sparsity))
- Get all coffee transactions in these counties in 2019
  - Observations at credit card – store level
  - Keep only stores for which we have latitude-longitude coordinates
- Get credit card covariates (monthly spending, indicator for gas station spending, and location) from all non-coffee transactions

Sample (cont.)

- Markets (8) vary in size
- Smallest include from 3 stores and 1 brands while largest market include 170 stores and 16 brands
- Covers from 25k to 1.8m cards per market
- Average card has about 3 to 4 coffee-related transactions of an average $8-11 per transaction

Demand estimation

- Standard discrete choice estimation with individual-level data

$$u_{ij} = \delta_j^{g(i)} - \alpha_{g(i)} d_{ij} + \epsilon_{ij}$$

where g(i) is one of eight income-by-car-owner groups

(income quartiles are inferred from monthly non-coffee spending; car ownership is inferred from any gas station spending)

- Estimate the α_gs and the δ_j^gs using maximum likelihood

- Estimates of α_g vary in the -0.8 to -2.3 range, tend to be a bit higher (in abs value) for lower-income cardholders and for non-car owners

Parameter estimates

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)

Compute diversion ratios

- Compute predicted markets shares using observed distances

$$s_j^o = \sum_i \frac{exp\left(\delta_j^{g(i)} - \alpha_{g(i)} d_{ij}\right)}{\sum_k exp\left(\delta_k^{g(i)} - \alpha_{g(i)} d_{ik}\right)}$$

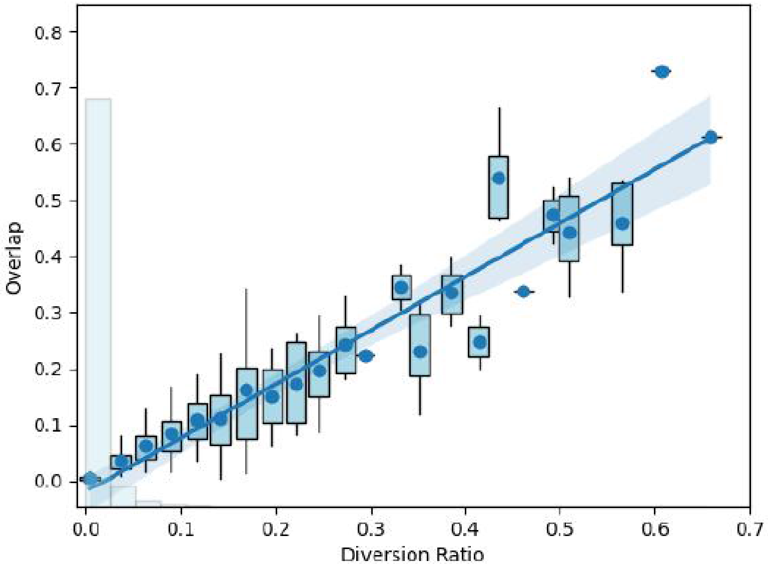
- Compute counterfactual shares using counterfactual distances

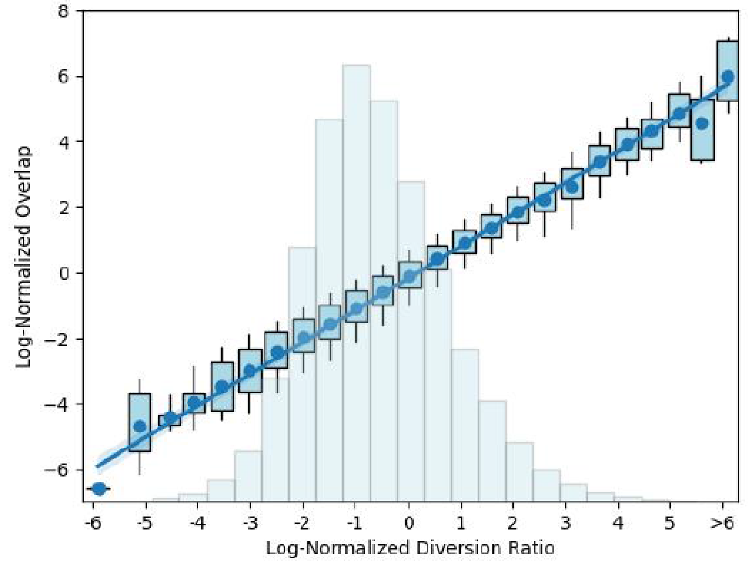
$$s_j^l = \sum_i \frac{exp\left(\delta_j^{g(i)} - \alpha_{g(i)}(d_{ij} + 1(j = l))\right)}{\sum_k exp\left(\delta_k^{g(i)} - \alpha_{g(i)}(d_{ik} + 1(k = l))\right)}$$

- Compute diversion ratios

$$D_{l \to k} = \frac{s_k^l - s_k^o}{s_l^o - s_l^l}$$

Correlation of overlap and diversion ratio





R^2 of 0.74 (left) and 0.75 (right)

Outline

- Motivating theory
  - Why would it make sense for overlap and diversion ratios to be related?
- Validation exercise (coffee …)
  - How well are diversion ratios approximated by overlap?
- **Illustrative use case**
  - **Overlap as a screen**
  - **Hotels …**

1. Overlap as a screen: overview of the exercise

- Define the (approximate) price effect of a merger between firms *j* and *k* to be

$$q_j D_{j\to k} + q_k D_{k\to j}$$

- Use the customer-overlap analog

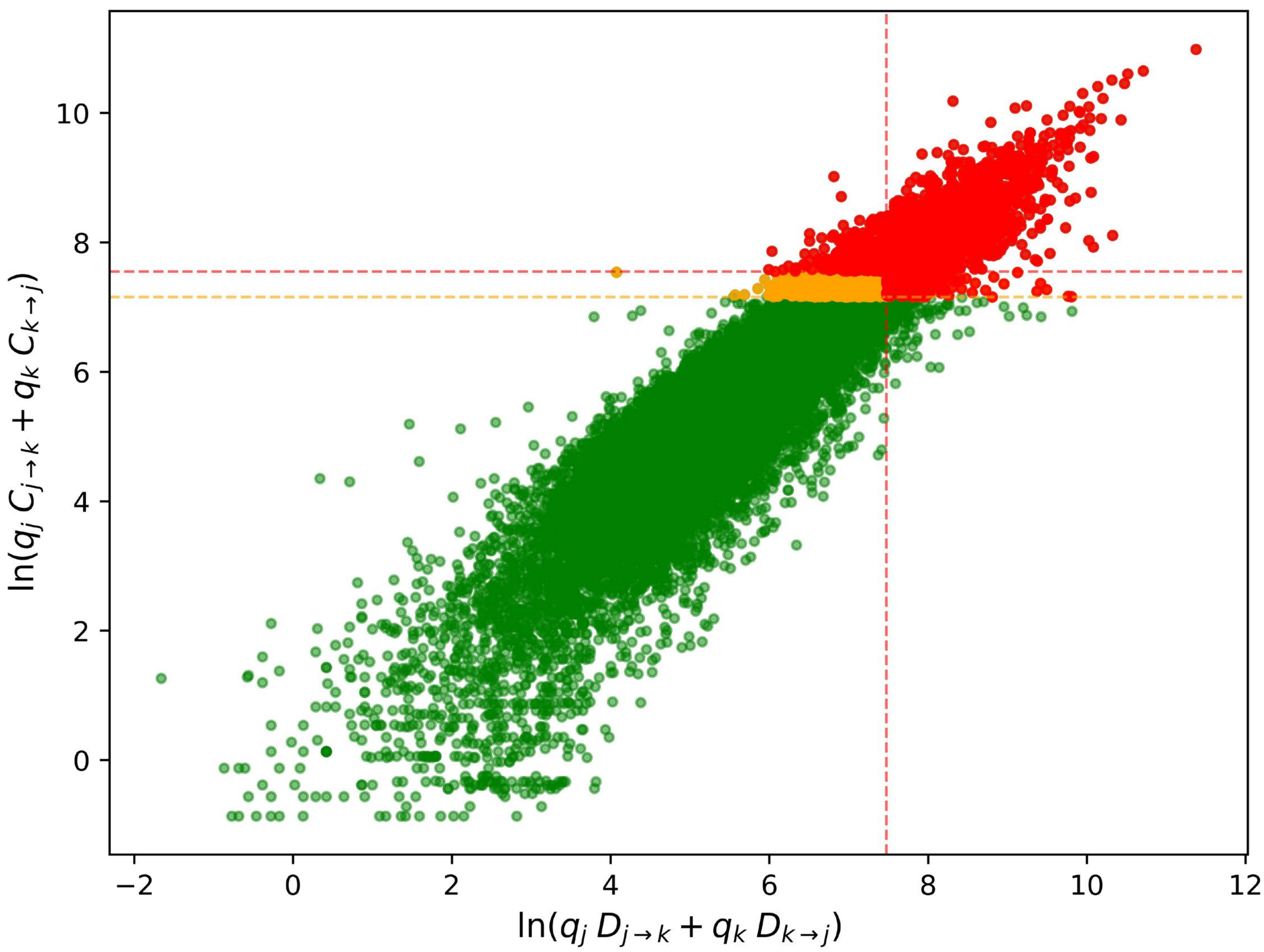
$$q_j C_{j\to k} + q_k C_{k\to j}$$

as a screen (in a similar way to the way we use HHI)

- Assume a fixed share of potential mergers that can get reviewed/investigated due to a resource constraint, and a fixed share of potential mergers that "should" get blocked

- Report the overall "avoided" price effect as a share of the "avoidable" price effect (when all mergers could get reviewed)

1. Overlap as a screen (cont.)



1. Overlap as a screen (cont.)

	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

2. Hotels: overview of the exercise

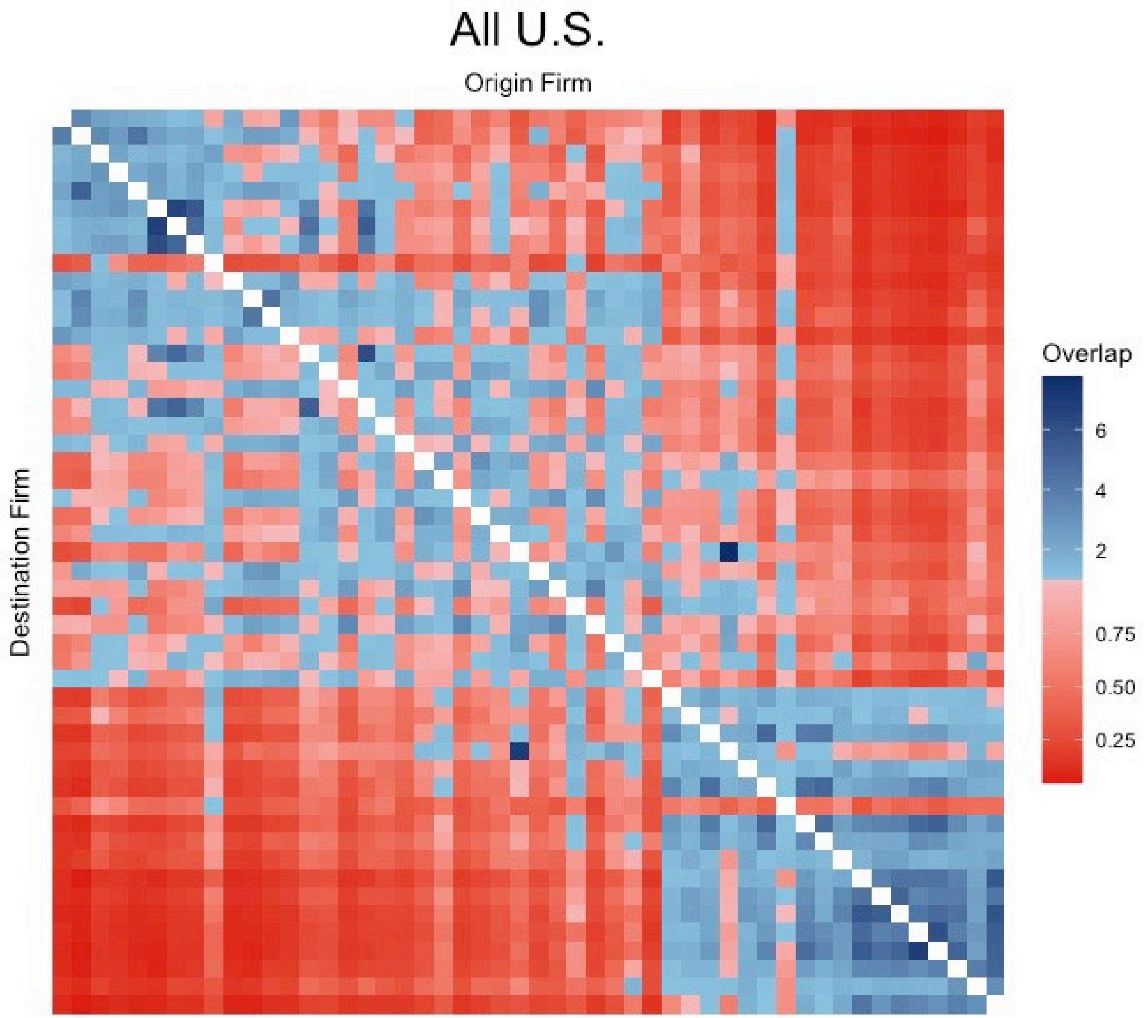
- Create a national sample of all hotel-related transactions across the US in 2023

- Analysis at the *hotel brand* level (*not* specific hotels)

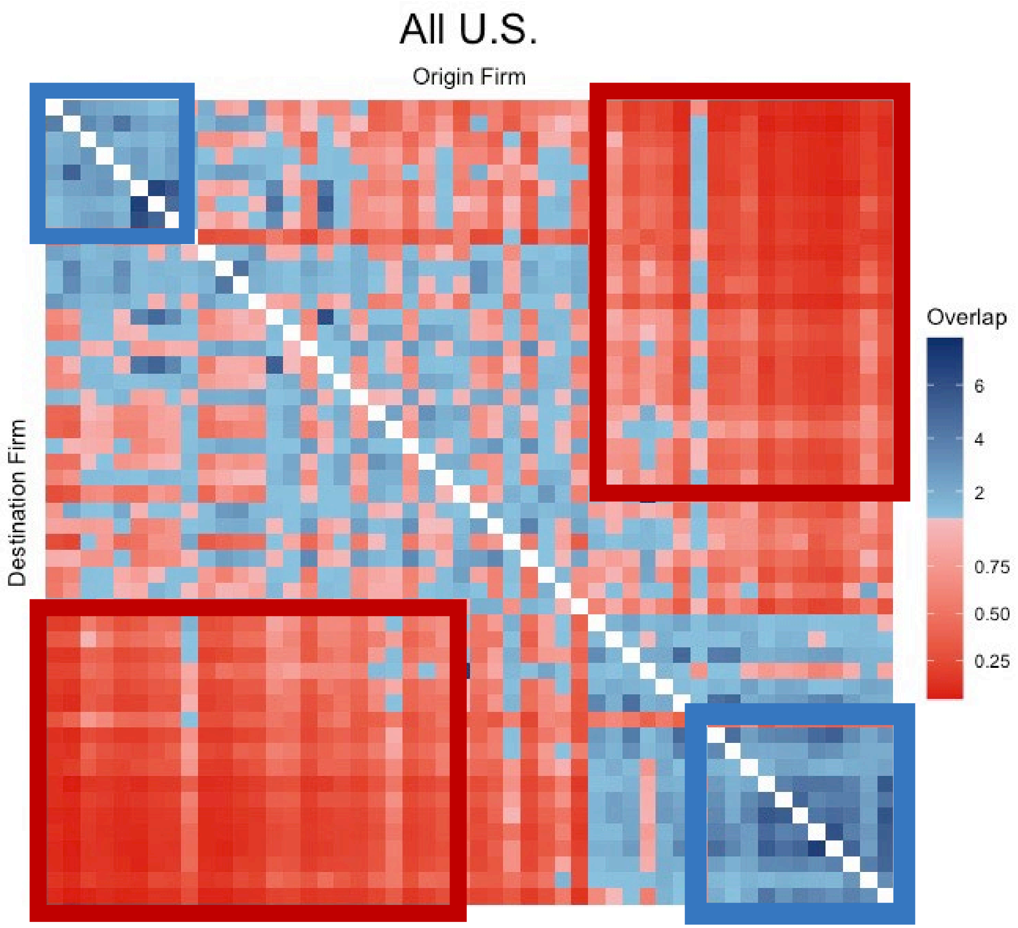
- Restrict attention to top 50 hotel brands
  - Relationship measured by number of transactions (stays)
  - Very similar if we use total dollars or binary relationship indicators

- Order brands by “affluence,” defined as the average monthly non-hotel card spending across all transacting guests

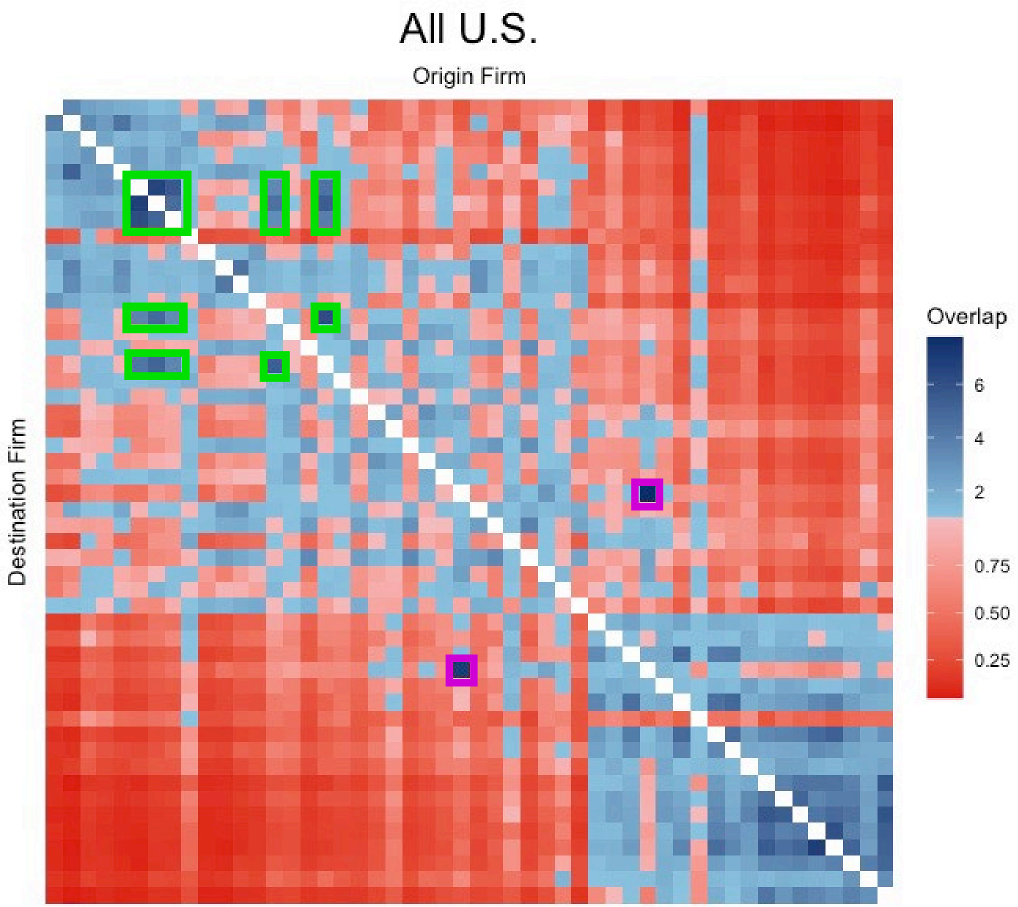
2. Hotel: normalized overlap in the hotels market



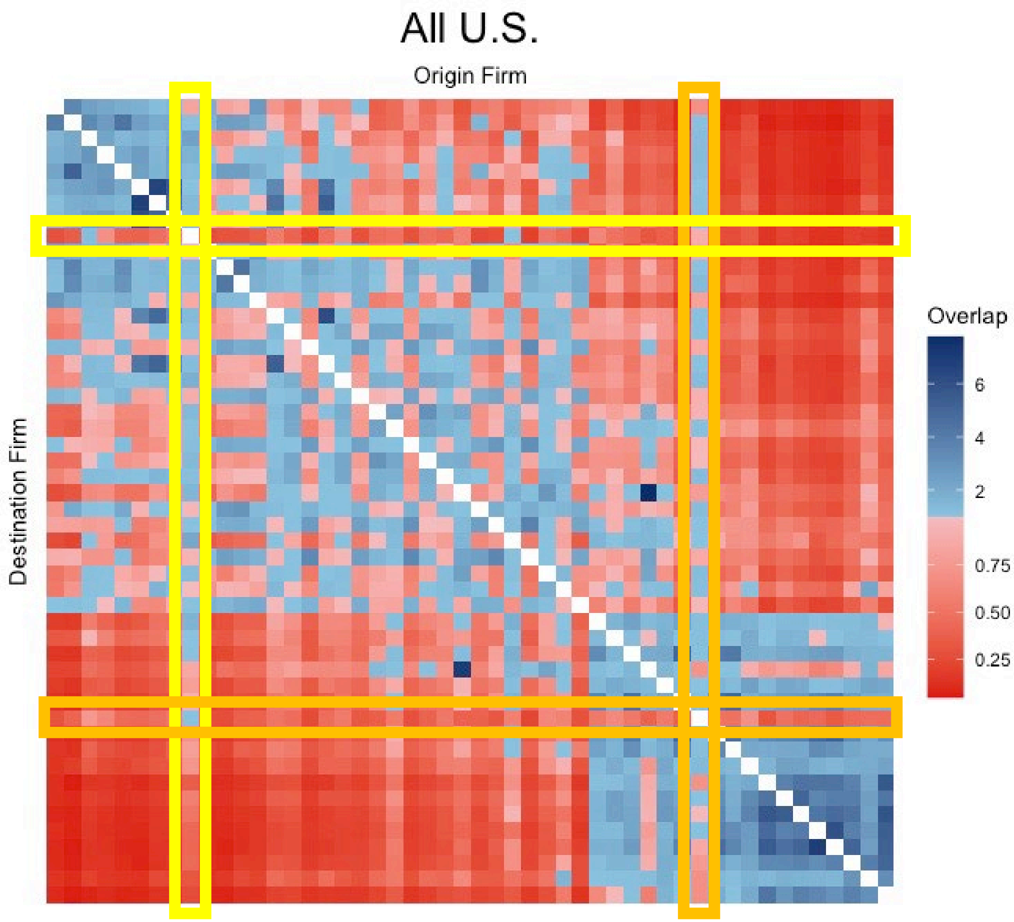
2. Hotels: overlaps pick up reasonable patterns



2. Hotels: overlaps pick up more granular patterns



2. Hotels: overlaps may help in defining markets



Conclusions

- Proposed the idea of customer overlap as a (very good!) proxy for diversion ratios
- With the **appropriate data** and **in the right context**, this is straightforward and scalable to compute, and can be very handy to gauge substitution across products, firms, or market segments
- *Appropriate data* is increasingly available: credit cards, IP addresses (e.g., Comscore), or cell phones (e.g., SafeGraph)
- Need to be careful to use it *in the right context*
  - Subjectively … coffee shops and hotels seem fine
  - Grocery stores not so much

Backup: Summary statistics for coffee 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