Available online at www.sciencedirect.com

ScienceDirect

Journal of Interactive Marketing 42 (2018) 1 – 17

JOURNAL OF INTERACTIVE MARKETING

www.elsevier.com/locate/intmar

Uncovering Patterns of Product Co-consideration: A Case Study of Online Vehicle Price Quote Request Data☆

Sina Damangir [a,*] & Rex Yuxing Du [b] & Ye Hu [b]

[a] College of Business, San Francisco State University, San Francisco, CA 94132, United States
[b] Bauer College of Business, University of Houston, Houston, TX 77204, United States

Abstract

Consumers often consider multiple alternatives from the same product category prior to making a purchase. Uncovering the predominant patterns of such co-considerations can help businesses learn more about the competitive structure of the market in the mind of the consumer. Extant research has shown that various types of online and offline consumer activity data (e.g., shopping baskets, search and browsing histories, social media mentions) can be used to infer product co-considerations. In this paper, we study a case of uncovering co-consideration patterns using a massive dataset of online price quote requests from U.S. auto shoppers. The main challenge we face is that, for privacy protection, no unique individual identifier (anonymous or otherwise) is contained in the data. Such a data deficiency prevents us from using existing methods such as affinity analysis for inferring co-considerations. However, by leveraging spatiotemporal patterns in the data, we manage to probabilistically uncover the predominant patterns of co-considerations in the U.S. auto market. As a validation and illustration of its usefulness, we embed the inferred market structure in a sales response model and show a substantial improvement in predictive performance.

Keywords: Big Data; Market structure analysis; Automotive industry; Sales response models; Privacy

Motivation

Consumers often consider multiple alternatives from the same product category prior to making a purchase. Uncovering the predominant patterns of such co-considerations can help businesses learn more about the competitive structure of the market as perceived by the consumer, which in turn can help managers make decisions such as product design (Bloch 1995), pricing (Choi 1991), and store layout (Brijs et al. 2004). However, it can be cost prohibitive to gather consumer self-reports on co-considerations at a large scale and on a regular basis. Extant research has shown that product co-considerations can be inferred from various types of online and offline consumer activity data, where an individual consumer's activities regarding multiple competing alternatives can be tied together via a unique individual identifier, e.g., frequent shopper card numbers, phone numbers, credit card numbers, mail addresses, email addresses, and Internet Protocol addresses. Examples of such applications include the use of shopping basket/retail panel data (Lattin and McAlister 1985), Internet browsing history data (Park and Fader 2004), credit card transaction records, and mobile device log (Chen, Chiang, and Storey 2012).

Affinity analysis, a class of techniques using the co-occurrences of events to uncover meaningful associations between them, is one of the methods commonly used to take advantage of such consumer activity data. By design, affinity analysis requires unique individual identifiers in order to establish co-occurrences. For example, a retailer can use the fact that a significant number of customers buy shampoo and conditioner in the same shopping basket to infer that these two products are close complements.

☆ We thank Autometrics for sharing with us the data used in this study. We carried out all the analyses independently and did not receive any financial support from Autometrics.
* Corresponding author.
E-mail addresses: damangir@sfsu.edu (S. Damangir), rexdu@bauer.uh.edu (R.Y. Du), yehu@uh.edu (Y. Hu).

https://doi.org/10.1016/j.intmar.2017.11.002
1094-9968© 2017

Similarly, by tracking browsing and searching records of individual visitors, an e-commerce website can figure out which products are frequently considered together by the same individual and infer the intensity of competition between products (Ringel and Skiera 2016).

In this research, we conduct a case study to infer product co-consideration from a massive dataset of online price quote requests (OPQRs) made by U.S. auto shoppers. By providing an unobstructed view of the U.S. auto shoppers at the late stages of the purchase funnel, the OPQR data provides an excellent opportunity to study the consideration sets of U.S. auto shoppers. Despite its richness, due to privacy concerns, the OPQR data has a "deficiency": it does not contain any unique individual identifiers, which precludes the application of affinity analysis. In this paper, we demonstrate that by leveraging spatiotemporal patterns in OPQR data, which is aggregated to five-minute level by vehicle and Zip code, we can overcome this data "deficiency" and uncover the predominant patterns of co-considerations in the U.S. auto market. As a validation and illustration of its usefulness, we embed the inferred market structure in a sales response model and show a substantial improvement in predictive performance.

More specifically, the price quote request data used in our study are gathered through a popular service offered by most of the major automotive shopping websites in the U.S. Fig. 1 shows the user interface of this service at the Kelley Blue Book website. Visitors to the site who are interested in requesting an online price quote from a local dealer can do so by selecting the brand (or "make" in auto industry lingo, e.g., Chevrolet) and model (e.g., Malibu), entering the Zip code, and clicking on the "get your quote" button (see highlighted rectangular area in Fig. 1). After clicking on the "get your quote" button, the visitor would be asked to provide his/her email address in order to receive a price quote from a local dealer. Autometrics (www.autometrics.com), a startup "Big Data" company, has entered agreements with most of the major automotive shopping websites in the U.S. to receive a record every time a visitor to a site selects a brand and model, enters a Zip code and clicks on the "get your quote" button. Each record contains four pieces of information: the time stamp, the brand, the model, and the Zip code associated with each OPQR. Every year Autometrics receives about 200 million such records.

Collecting and sharing individual level data on online activities poses serious security and privacy concerns, especially when it is done at a massive scale (de Montjoye et al. 2015). Due to such concerns, Autometrics has no access to any information identifying individual website visitors (e.g., the email address provided by a site visitor is not shared with Autometrics). In other words, from the records received by Autometrics, one can tell when and where a car shopper is interested in getting a price quote for which vehicle. However, because there is no unique individual identifier in Autometrics' data, one cannot tie multiple OPQRs to any individual car shopper, even though the same shopper may send in multiple OPQRs as she compares alternative vehicles in her consideration set.

The kind of online consumer activity data gathered by Autometrics is obviously less informative than the raw records

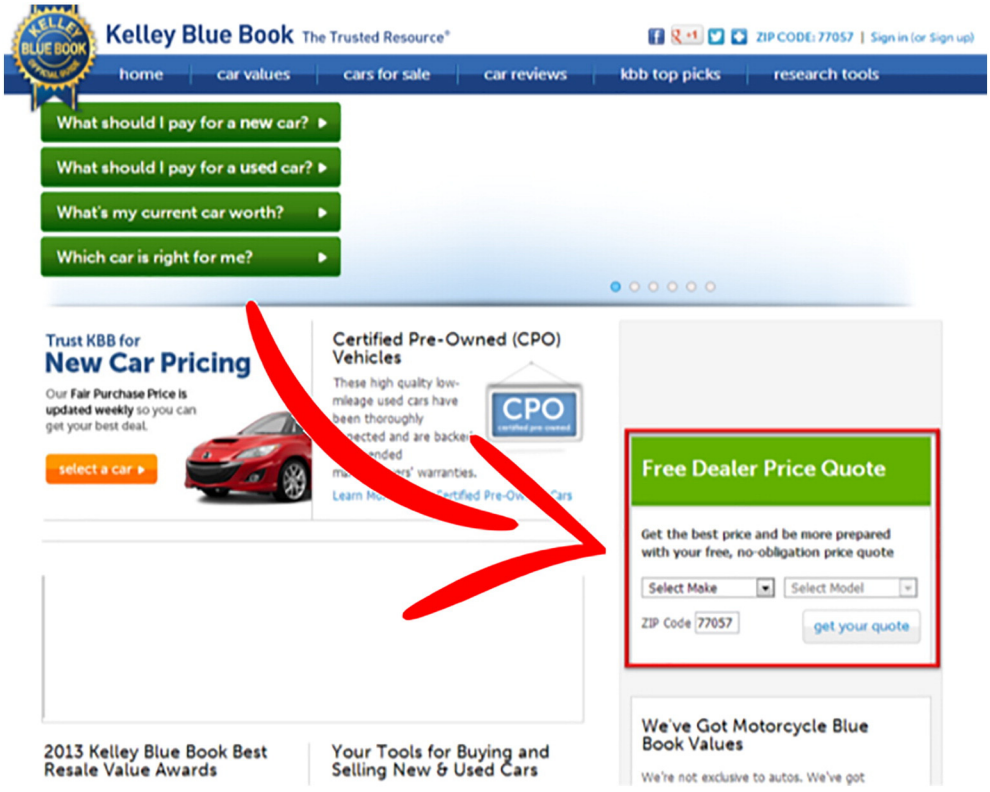


Fig. 1. Screenshot of a typical interface for submitting an online price quote request. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

that have a unique individual identifier. On the other hand, stripping away all uniquely identifiable information provides an extra layer of protection for privacy, a feature that can become increasingly desirable in the era of "Big Data" from the perspective of consumers and regulators (e.g., the Gramm–Leach–Bliley Act in the United States, the European Union legislation on cookies). However, it also poses a major challenge for generating actionable business intelligence, as techniques such as affinity analysis would be rendered infeasible when no two activities can be tied back to the same consumer. It is this seemingly unavoidable tradeoff between protecting consumer privacy and generating actionable insights from consumer activity data that has motivated us to investigate whether it is possible, in our particular empirical context, to infer co-occurrences of activities even though there is no unique identifier that ties any pair of activities to the same individual.

The remainder of the paper proceeds as follows. In the next section, we discuss the related literature and our intended contribution, and present more information about our data, highlighting the challenges that we must address. Then we present our method of using spatiotemporal information in the data for inferring product co-considerations probabilistically. To validate our method, we infer the market structure from the predominant patterns of co-considerations and embed the resulting market structure in a sales response model. We use the model to predict market shares in the U.S. automotive market and show considerably improved performances over the benchmark models. We conclude by discussing what our findings mean in terms of striking a balance between privacy protection and insight generation when working with large-scale online consumer activity data. Finally, we discuss the limitations of our proposed method and suggest promising directions for future research.

Research Background and Empirical Context

Market Structure Analysis Based on Consumer Online Activity Data

Knowing which products are closer substitutes to one another constitutes a central part of market structure analysis, which helps shape many strategic decisions, e.g., identification of primary and secondary competitors, product design, positioning, pricing, and supply chain management (Day, Shocker, and Srivastava 1979; Urban, Johnson, and Hauser 1984; see Shugan 2014 for a review of the literature on market structure analysis). In order to uncover the competitive relationships among products, managers have often relied on examining the predominant patterns of product co-considerations (e.g., Shocker et al. 1991; Urban, Johnson, and Hauser 1984). Historically, many approaches have been developed to capture product co-considerations, e.g., tracking brand switching patterns (e.g., Cooper and Inoue 1996), estimating cross-elasticities (e.g., Russell and Bolton 1988), or most commonly, examining consumer self-reports of consideration sets (Urban, Johnson, and Hauser 1984).

In the era of "Big Data," market structure analysis has increasingly relied on uncovering patterns of co-considerations hidden in consumers' digital footprints (Kim, Albuquerque, and Bronnenberg 2011; Lee and Bradlow 2011; Netzer et al. 2012; Park and Fader 2004; Ringel and Skiera 2016). One key advantage of online consumer activity data is that they reflect observed behavior, which can be more reliable than self-report. In addition, online consumer activity data are often gathered as a byproduct/exhaust of the transaction process, which means they can be cost effective for large-scale tracking studies.

Most recent developments in market structure analysis have tapped into two types of digital footprints: user-generated contents (UGCs) and Internet browsing data. For example, Lee and Bradlow (2011) apply text-mining techniques to transform online reviews into an association network of products and attributes. Then, they use those associations to infer the underlying competitive market structure. Similarly, Netzer et al. (2012) apply text mining to posts scraped from online discussion/review forums to infer the network of associations among a large set of competing products. They do so by examining the predominant patterns of co-mentions in individual consumers' posts. Since online discussion/review forum data is publicly available, inferring competitive structure using UGCs is scalable to cover a large number of websites and the resulting market structure reflects the voice of a diverse set of consumers.

Richness in contents aside, UGCs have some limitations. First, only a small percentage of self-selected consumers engage in online conversations, limiting the representativeness of UGCs (Gao et al. 2015; Li and Hitt 2008). As a result, analysis of such data becomes less reliable for small markets or niche products. Second, there is often a time discrepancy between making a purchase and conducting an online discussion or posting a review. Finally, there can be substantial biases and noises in consumer recall and verbalization when they are engaged in online discussions or reviews (Gardial et al. 1994; Hunt, Sparkman, and Wilcox 1982).

Besides UGCs, Internet browsing data, gathered unobtrusively through cookies and server logs, have become an important source of consumer intelligence (Chen, Chiang, and Storey 2012). By capturing online activities of consumers who are actively seeking product information, Internet browsing data can be quite useful for inferring co-considerations among a large number of products (Ringel and Skiera 2016). Kim, Albuquerque, and Bronnenberg (2011) propose a method to use browsing history to understand consumer search behavior and thus infer competitive structure in the market and the composition of consumer consideration sets. Ringel and Skiera (2016) use Internet browsing data from users researching about products to infer associations between the products and develop a method for visualizing the competitive landscape for a market comprised of thousands of products. Compared to UGCs, the large volume of Internet browsing data makes it possible to generate reliable insights about small markets and niche products. For inferring co-considerations from online activity data, the current methods (e.g., Kim, Albuquerque, and Bronnenberg 2011; Ringel and Skiera 2016) rely on calculating the probability of viewing one product conditional on viewing another product, which requires knowing the entire click stream of an individual website visitor. In this research, our empirical context is analogous to situations

where one observes only the number of times a product webpage is browsed in a given time window by visitors from a given geographic area. Our core research question is: given such limited data, whether it is still possible to establish associations between the products and infer the market structure when co-occurrence-based affinity analysis is infeasible.

In summary, through our case study we intend to make the following contributions to the growing literature on inferring market structure from consumer online activity data.

1. Introduce a new type of data that can be used for inferring market structure.
2. Propose a method for calculating associations between pairs of products, tailored for the case where the data do not have a unique individual identifier. Our method for inferring co-consideration is applicable only when the lack of individual identification information can be compensated by granular spatiotemporal information.[1]
3. We extend the previous literature by embedding our inferred information about competitive market structure in a sales response model. We reveal a meaningful link between the inferred market structure and market share covariations. In the absence of ground truth of the competitive structure, we argue that this link is an indirect piece of evidence to establish the validity of the co-consideration measure inferred from online activity data. Furthermore, we show that practitioners can gain more insight about the market by augmenting sales response models with information about market competitive structure yielded from large-scale consumer online activity data.

Empirical Context: Online Price Quote Request Data Without a Unique Individual Identifier

Consumers use the Internet to gather product information as they conduct comparison-shopping (Ratchford, Lee, and Talukdar 2003). The resulting browsing history data on competing products resides in two main sources: 1) the sites of individual product maker (e.g., Ford.com for Ford Motor Company, GM.com for General Motors) and 2) the sites of third-party firms (e.g., kbb.com for Kelley Blue Books, Edmunds.com). Scattered across different sites, consumers' Internet browsing data, despite its richness, can be fragmented from any individual site's standpoint. Analyses based on data from any single site cannot reveal the whole picture of consumer online activities. As a result, there are quite a few third-party data aggregators filling this gap by merging the data across multiple websites. For example, Facebook and Google track users across third-party websites through their digital advertising network. Doubleclick.com, now part of Google, through the extensive use of cookies at multiple websites, can turn site centric data into user centric data.

Autometrics, the company behind the data used in this case study, is one of the third-party aggregators providing a solution to this data fragmentation issue in the U.S. automotive market by assembling OPQRs from all the major automotive shopping websites. The data provide a comprehensive coverage of the population of the U.S. consumers who use OPQR services through many automotive shopping websites. When a site visitor enters an auto brand and model and his/her Zip code, the site records this information along with time of the request and sends it to Autometrics.[2] Despite the fact that Autometrics does not have access to any uniquely identifiable information, such cross-site OPQR data aggregated by Autometrics offer several advantages.

First, the data are voluminous: on average 200 million OPQRs per year. The comprehensive coverage of U.S. auto shoppers makes patterns extracted from the data representative of the target population. As we shall demonstrate later, the scale and comprehensiveness of the data enable us to examine competitive relationships among niche products (e.g., Mazda 6 and Mitsubishi Lancer) in small markets (e.g., Boise, Idaho), where data generated by alternative sources can be sparse and unreliable. This is particularly important when many domains of the economy have shifted towards "long tail" competition (Anderson 2006).

Second, lower purchase funnel activities such as OPQRs are particularly appealing for examining the predominant patterns of co-considerations in the auto market. OPQRs tend to take place when vehicle shoppers' preferences have become less fuzzy than earlier stages of the purchase funnel, where shoppers are more likely to focus on vehicle features as opposed to dealership pricing. Consequently, OPQRs can reveal the substitutability among products close to the final purchase stage.

Finally, because the user interface and data fields are similar across all the websites, it is relatively easy to maintain consistency over time and across different sources. Through its contracts with all the major auto shopping websites, Autometrics is in a position to aggregate all the OPQRs in the U.S. and share the merged data with automakers and dealerships. Such a capability to integrate and share data across sources is crucial for developing industry-wide business intelligence systems (Zheng, Fader, and Padmanabhan 2012).

Companies sharing consumer activity data with third-party aggregators often need to mitigate privacy concerns with a solution that goes beyond simple anonymization. Otherwise, they risk exposing the identity of consumers through re-identification of anonymized data (de Montjoye et al. 2015; Li and Sarkar 2011, 2014; Menon, Sarkar, and Mukherjee 2005). To avoid such privacy and identity concerns, Autometrics data contain no uniquely identifiable information, retaining only the time stamp and Zip code of each OPQR. Admittedly, the lack of any unique individual identifier is a double-edged sword, putting a severe constraint on how much insight can be extracted from the data.

[1] Our method is not applicable when there is a unique individual identifier.

[2] After recording this data, the website usually proceeds to ask for further contact information that allows nearby auto dealers to contact the site visitor and provide a price quote. The contact information and subsequent price quote are not shared with Autometrics for privacy and data security reasons.

Challenges and Solution

In our attempt to uncover predominant patterns of co-considerations from Autometrics data, the first and foremost challenge lies in that, because our data has no unique individual identifier, we cannot tie any pair of OPQRs to the same individual. Without knowing which two OPQRs come from the same shopper, as we shall demonstrate later, we have to probabilistically infer co-considerations by leveraging the spatiotemporal patterns of the data. We also have to ensure the validity of the inferred co-consideration pattern when co-considerations are never directly observed in the data. This is the main data challenge that differentiates our study from the extant literature.

Furthermore, our data are quite noisy. Although there is little doubt that OPQRs are in general tied with purchase-oriented considerations, other consumer activities can generate OPQRs as well. For example, window shoppers who request a price quote out of curiosity may have no real intention of buying a vehicle. In other words, window shoppers' OPQRs are mixed with those of shoppers who are genuinely in the market for a car. Such noises are inevitable when online activities are used as proxies of intents and no other contextual information is available about the individual behind those activities (Lazer et al. 2014). The challenge lies in how we can ensure the inferred market structure actually captures the pattern of co-considerations among genuine auto shoppers.

Given the deficiencies in our data, to validate the inferred co-consideration patterns, we build on a long tradition in market structure analysis; we shall turn the inferred patterns of co-considerations into a product-positioning map that reflects the competitive landscape geometrically (Shugan 2014). To establish the face validity of the co-consideration measure, we examine whether the resulting product-positioning map conforms to the common beliefs about competition in the U.S. auto market. We also directly compare our product-positioning map with a comparable map Netzer et al. (2012) generated using UGCs about automobiles. Furthermore, we embed the product-positioning map in a sales response model to predict market shares. If the inferred patterns of co-considerations are valid, the resulting product-positioning map should improve the predictive performance of the sales response model over benchmark models that do not take into account information captured in the map.

Inferring Co-considerations Probabilistically

Operationalization

In this section, we develop a method for probabilistically inferring co-considerations from OPQR data. Consider a shopper choosing between two cars: all else being equal, conditional on the shopper having requested price quote for one car, the probability of him requesting price quote for the other car should be higher than the unconditional probability. Furthermore, the second price quote request has an above random chance of happening within a short time period after the first price quote request — based on a conceivable process of comparison shopping or co-consideration. As this happens to millions of individual shoppers who consider multiple vehicles, we expect that, when the number of consumers co-considering two vehicles increases, there should also be an increase in the number of incidences where OPQRs for these two vehicles "co-occur" in short time intervals from the same geographic area.

Empirically, we operationalize this concept by partitioning each day into 288 non-overlapping five-minute intervals (=24 hours times 12 five-minute intervals per hour). We treat OPQRs for two distinct vehicles from the same Zip code within the same five-minute interval as a "co-occurrence" of the OPQRs for the two vehicles. Fig. 2 clarifies the way we operationalize co-occurrences. The horizontal axis in Fig. 2 represents the time during a day. Suppose that in a Zip code we have three OPQRs for three vehicles A, B, and C during minutes 2, 4, and 6 respectively. In our operationalization, there is a co-occurrence of A and B, but C does not co-occur with either A or B.

We use the number of co-occurrences as a proxy for co-considerations as we expect co-occurrences and co-considerations to be positively correlated. Admittedly, using co-occurrences as proxies for co-considerations is far from perfect. For example, some shoppers may consider two vehicles but do not request price quotes for both of them. Some shoppers may request price quotes for both vehicles but do so with a large time interval in-between. In other words, the number of co-occurrences captured in our data is likely very different from the number of actual co-considerations. Nevertheless, it is important to note that as long as the number of co-occurrences is positively correlated with the number of co-considerations and such correlation is consistent across vehicles, using the pattern of co-occurrences as a proxy for the pattern of co-considerations would yield the same insights about the competitive landscape because all that matters is to capture the relative degree of competition among vehicles.

In our application, we do not use the raw count of co-occurrences, because co-occurrences can happen by random chance, which favors vehicles with a large number of OPQRs. For example, consider two shoppers from the same Zip code, one of whom is interested in Toyota Camry and the other interested in Honda Accord, both very popular midsize sedans. These two shoppers can each enter an OPQR for the car they are considering, and by chance, their independent OPQRs can happen within the same five-minute interval. This would result in a co-occurrence of Camry and Accord that happens not due to co-considerations, but due to a random coincidence. To mitigate this issue, following a long tradition in affinity analysis and association rule learning, we adopt the concept of lift (Kotsiantis and Kanellopoulos 2006). More specifically, we operationalize lift as the ratio of observed co-occurrences to expected co-occurrences. The latter represents the number of co-occurrences we would have expected to observe if none of the OPQRs were due to co-considerations. In other words, by using lift as opposed to raw co-occurrences, we identify the associations beyond random chance.

More formally, denote two focal products as A and B. For a given Zip code Z, in a given day D, denote the number of OPQRs for product A as n_{AZD}, the number of OPQRs for product B as

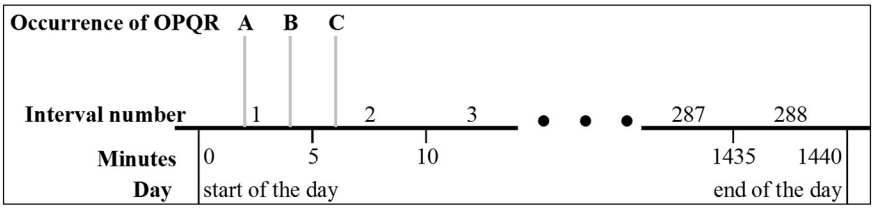


Fig. 2. Example of occurrence and co-occurrences of OPQRs during a day within a Zip code. Note: In our operationalization, A and B are co-occurring, but C is not co-occurring with A or B.

n_{BZD}, and the number of observed co-occurrences as n_{ABZD}. Conditional on n_{AZD} and n_{BZD}, there is a non-zero probability of observing co-occurrences even if none of the OPQRs were due to co-considerations. Denote the expected number of such coincidental co-occurrences as $E[n_{ABZD}|n_{AZD}, n_{BZD}]$, which we shall use as the baseline to determine whether the observed co-occurrences n_{ABZD} are above and beyond random chance and thus signal a meaningful positive association between A and B.

Conditional on n_{AZD} and n_{BZD}, if OPQRs were randomly distributed throughout the day (i.e., no systematic within-day fluctuation), one can use a contingency table to calculate $E[n_{ABZD}|n_{AZD}, n_{BZD}]$. However, as Fig. 3 shows, the traffic of OPQRs fluctuates within a day, peaking during early evening and bottoming after midnight. A higher density of OPQRs in a time interval means a higher probability to observe co-occurrences by chance. Given this, no closed form analytical solution exists for $E[n_{ABZD}|n_{AZD}, n_{BZD}]$. Instead, for each combination of n_{AZD} and n_{BZD}, we use the following Monte Carlo simulation to calculate the numerical value of $E[n_{ABZD}|n_{AZD}, n_{BZD}]$.

1. Calculate the empirical probability of OPQR occurrences for each five-minute interval. The probabilities are based on the percentage of OPQRs occurring in each five-minute interval aggregated over all days and across all Zip codes. They are, in other words, not Zip code or day specific.

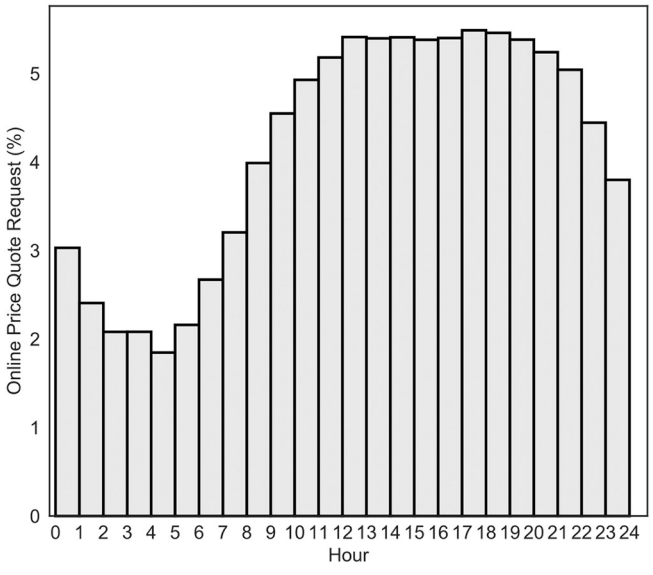


Fig. 3. Average hourly fluctuation of OPQRs during a day.

2. Randomly and independently, generate n_{AZD} and n_{BZD} OPQRs for A and B during the simulated day. The empirical probability calculated in step 1 determines the probability of simulated OPQRs occurring in each time interval.
3. Count the number of co-occurrences of A and B in the simulated day.
4. Repeat steps 2 and 3 until the mean of simulated co-occurrence converges, which we shall use as an estimate of $E[n_{ABZD}|n_{AZD}, n_{BZD}]$.

We define the ratio of n_{ABZD} to $E[n_{ABZD}|n_{AZD}, n_{BZD}]$ as the lift. Since we shall later combine lift with sales data, we obtain an aggregated lift by summing n_{ABZD} and $E[n_{ABZD}|n_{AZD}, n_{BZD}]$ over the period between January 1, 2009 and December 31, 2011, and across all Zip codes in the U.S.

$$Lift_{AB} = \frac{\sum_D \sum_Z n_{ABZD}}{\sum_D \sum_Z E[n_{ABZD}|n_{AZD}, n_{BZD}]} \quad (1)$$

Intuitively, the higher the lift, the more likely OPQR co-occurrences are caused by co-considerations as opposed to random coincidence. In order to establish validity of our approach, we examine the lift metric using several methods:

1. We compare our results on U.S. auto market competitive structure with those reported by Netzer et al. (2012).
2. We investigate the face validity by examining whether the patterns in the lift measures reveal pair-wise competitive relationships that are qualitatively/directionally consistent with conventional wisdom of the U.S. auto market.
3. We examine whether the lift measures lead to a product-positioning map consistent with the competitive structure in the U.S. auto market.
4. We check the robustness of the lift measures to changes in granularity of the data. Specifically, we investigate how lift measures vary as we systematically change the length of time intervals and the size of geographic units.
5. Finally, in the next section, we empirically validate the lift measures by examining the correspondence between the actual market shares of cars in the U.S. auto market and the competitive relationship revealed by our proposed measure of co-considerations.

Comparison with the Results in the Literature

To validate our proposed method for measuring co-considerations, we first compare our results with those from

the existing literature. In our context of the U.S. auto industry, Netzer et al. (2012) provides a reasonable benchmark for comparison. For automobile brands available in the United States, Netzer and colleagues reported a positioning map reproduced in Fig. 4 panel A. To replicate their results, we focus on 29 top automobile brands.[3] These 29 brands form 406 unique pairs (combination of selecting 2 out of 29, $\binom{29}{2} = 406$). For all these pairs, we use Eq. (1) to calculate the lifts. Taking lifts as measures of pair-wise similarity/proximity, we create a co-consideration map using Kruskal's (1964) non-metric Multidimensional Scaling (MDS). Fig. 4 panel B depicts the resulting configuration (stress: 0.180) where each square represents an automobile brand. Brands closer to each other have a higher lift. Fig. 4 panels A and B show great resemblance. For example, we observe that both maps identify three separate clusters of vehicles: domestic non-luxury, international non-luxury, and luxury. This resemblance is obtained despite the fact that the two maps are generated using different methods and sources of data: Netzer et al. (2012) produced the map by applying text-mining methods to data on online UGCs, whereas we get the map by applying our proposed method to OPQR data. This comparison suggests that even without individual identifiers, we can uncover brand-level competitive relationships from Autometrics data, as validated by the consistency with previous research.

Comparison with Alternative Co-occurrence Metrics

A desirable feature of the lift measure described in Eq. (1) is that in the case of our data where individual identifiers are not available, it can account for co-occurrences beyond randomness. Here, we examine how our proposed method compares with alternative methods to create co-occurrence metrics, controlling for the possibility of random co-occurrences. Following Netzer et al. (2012), we use two of such co-occurrence metrics, Jaccard index and cosine similarity, as benchmarks for comparison.

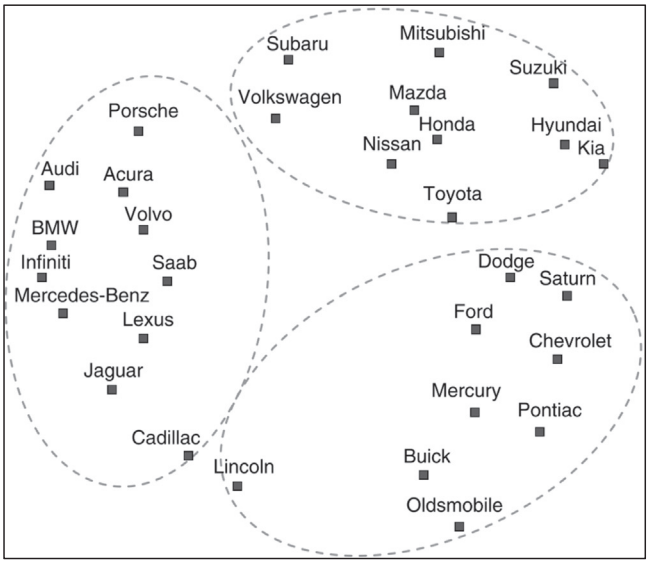
$$\text{Jaccard index}_{AB} = \frac{\sum_D \sum_Z n_{ABZD}}{\sum_D \sum_Z (n_{AZD} + n_{BZD} + n_{ABZD})} \tag{2}$$

$$\text{cosine}_{AB} = \frac{\sum_D \sum_Z n_{ABZD}}{\sqrt{\sum_D \sum_Z n_{AZD} \sum_D \sum_Z n_{BZD}}} \tag{3}$$

Table 1 shows the correlation of the benchmark metrics with the lift measures we used to generate Fig. 4 panel B. The results from Jaccard index and cosine are similar (correlation = .99). However, they are weakly correlated with the proposed lift (.61 and .66). The weak correlations, together with the striking resemblance between our map and that of Netzer et al. (2012), suggest that neither Jaccard index nor cosine similarity could yield a map of comparable quality. Moreover, the normalizing coefficients in Jaccard index and cosine ($\sum_D \sum_Z (n_{AZD} + n_{BZD} + n_{ABZD})$ and $\sqrt{\sum_D \sum_Z n_{AZD} \sum_D \sum_Z n_{BZD}}$ respectively) do not correctly account for expected random co-occurrences when we do not have individual identifiers.

[3] Netzer et al. (2012) reported results from 30 brands. As Oldsmobile has since been shut down, we run the analysis using the 29 remaining brands.

A: generated from online UGCs data (Netzer et al. 2012)



B: generated from OPQR data

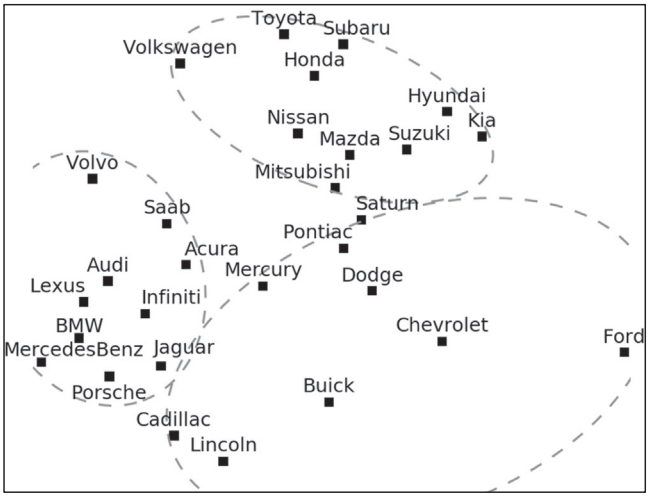


Fig. 4. Brand level product positioning maps.

Face Validity

To further examine the validity of our method, we focus on lifts at the vehicle-model level. As an illustration, we examine the top-selling sedans between 2009 and 2011, with sizes

Table 1
Correlation between lift and the alternative metrics.

	lift	Jaccard	cosine
lift		.615	.662
Jaccard	.615		.986
cosine	.662	.986	

Table 2
Selected sedans and comparison of their OPQRs and sales from 2009 to 2011.

Car	Body size	Sales	OPQR	OPQR to sales ratio
Chevrolet Impala	Midsize	172,385	4,255,584	24.7
Chevrolet Malibu	Midsize	370,081	4,419,556	11.9
Chrysler 300	Full-size	66,518	2,819,476	42.4
Ford Focus	Compact	355,435	6,059,254	17.0
Honda Accord	Midsize	783,355	11,951,240	15.3
Honda Civic	Compact	712,665	8,853,180	12.4
Hyundai Sonata	Midsize	441,210	5,826,399	13.2
Kia Optima	Midsize	114,282	2,328,163	20.4
Mazda 3	Compact	264,558	2,515,869	9.5
Mazda 6	Midsize	78,393	1,249,185	15.9
Mitsubishi Lancer	Compact	59,816	989,066	16.5
Nissan Altima	Midsize	541,999	6,391,978	11.8
Nissan Maxima	Midsize	153,952	2,824,595	18.3
Nissan Sentra	Compact	230,856	1,610,594	7.0
Subaru Forester	Full-size	220,074	1,972,208	9.0
Subaru Impreza	Compact	119,608	1,375,984	11.5
Subaru Legacy	Midsize	98,148	772,006	7.9
Toyota Avalon	Full-size	77,120	1,469,159	19.1
Toyota Camry	Midsize	814,781	9,427,479	11.6
Toyota Corolla	Compact	612,346	4,792,886	7.8
Toyota Prius	Midsize	380,136	3,604,511	9.5
Volkswagen Jetta	Compact	365,390	3,269,403	8.9
Volkswagen Passat	Midsize	42,025	656,777	15.6
All		7,075,133	89,434,552	12.6

varying from compact to full-size. During the time window of our study, the national level OPQRs vary greatly across these 23 top sedans (Table 2). As Fig. 5 shows, the total OPQRs and sales for each car are highly correlated (correlation = .92).

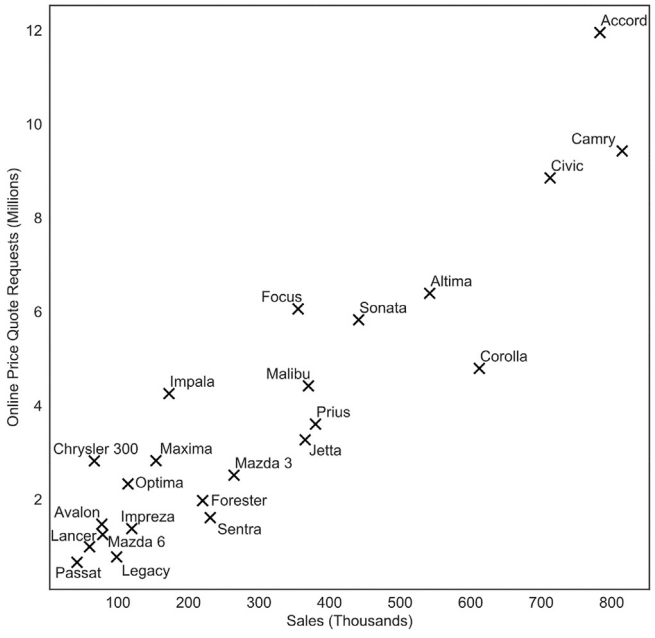


Fig. 5. Scatter plot of total sales vs. total OPQRs from 2009 to 2011 in the United States.

These 23 sedans form 253 unique pairs (combination of selecting 2 out of 23, $\binom{23}{2} = 253$). For all these pairs, we use Eq. (1) to calculate the lifts and report them in Table 3.

The competitive relationship that emerges from the results in Table 3 allows us to investigate the face validity of the lift measure. Since we expect consumers to co-consider pairs with high substitutability, if our proposed lift is positively correlated with the degree of co-considerations, the pairs with the highest lifts should be close substitutes.

Table 4 shows the 20 pairs of sedans with the highest lifts. Two common patterns emerge. First, not surprisingly, 13 out of the 20 highest-lift pairs share the same body size, suggesting consumers are more likely to consider vehicles of the same body size as substitutes as they fulfill the same need for passenger and cargo capacity. Second, seven out of the 20 highest-lift pairs share the same brand. For example, Passat and Jetta of Volkswagen (VW), Mazda 6 and Mazda 3 of Mazda, and Impreza, Legacy, and Forester of Subaru are all pairs with high lifts, indicating consumers are more likely to consider substitutes within the same brand for these vehicles. Both of these patterns – high lifts for sedans of either the same size or brand – are consistent with the conventional wisdom, indicating face validity of our lift measure.

Next, we examine whether the overall lift patterns lead to a plausible portrait of the market structure. Treating the lifts reported in Table 3 as measures of pair-wise similarity/proximity, we apply non-metric MDS to produce a two-dimensional product-positioning map using the Kruskal (1964) method. Fig. 6 depicts the resulting configuration (stress: 0.201) where each "x" represents a car. We will discuss segment sizes and ideal positions (circles) in the Empirical Validation section.

In the map, full-size sedans cluster near the bottom left quadrant, midsize sedans occupy the center, and compact sedans are located in the top right quadrant. Furthermore, the bottom right quadrant represents a niche market of sedans with all-wheel drive and outdoor capabilities. To see the "big picture" revealed in Fig. 6, one can interpret the lower left-upper right diagonal direction as a combination of "vehicle size" and "fuel economy," with the upper right direction representing smaller cars with better fuel economy. The upper left–lower right diagonal direction can be labeled as "performance in complicated terrains," with the lower right direction representing cars with higher performance off-road and in snow.

Robustness

The analysis above provides some evidence that our method can infer co-considerations even though there is no unique individual identifier in our data. The method works because in our empirical context the lack of individual identifiers is compensated by the spatiotemporal information – we used the most granular data available to us as the basis of our analysis – OPQRs at five-minute intervals at Zip code level.

Table 3
Lift at vehicle-model level for all pairs of top 23 sedans in the United States from 2009 to 2011.

Impala	4.74	3.82	1.29	2.85	1.74	2.27	2.24	1.39	2.80	1.70	2.87	3.20	1.77	.70	.81	1.55	2.56	2.65	1.65	1.04	1.08	1.61
	Malibu	3.00	1.76	3.19	2.31	3.29	3.28	2.07	4.18	2.31	3.70	2.82	2.82	.97	1.18	2.17	1.93	3.37	2.43	1.55	1.60	1.93
		Chrysler 300	.93	2.68	1.46	1.99	2.07	1.35	2.83	2.62	2.43	4.07	1.15	.83	1.28	2.49	3.77	2.24	1.22	.90	1.25	2.13
			Focus	1.67	2.48	1.59	3.87	3.59	3.94	3.85	1.95	1.10	3.65	.71	2.00	3.73	.68	1.37	2.75	1.21	1.80	1.46
				Accord	3.48	3.36	3.39	2.33	5.71	2.54	4.61	3.99	2.58	1.07	1.27	4.24	3.22	4.81	2.98	1.50	1.83	3.05
					Civic	1.99	1.87	4.37	3.38	4.25	2.86	2.12	4.63	1.08	2.33	1.74	1.42	2.67	5.44	2.93	2.42	1.47
						Sonata	4.78	1.90	5.45	1.90	3.48	2.05	2.10	.97	1.12	4.49	1.99	3.39	2.14	1.67	1.51	2.22
							Optima	2.16	8.73	2.64	3.95	2.47	2.30	.99	1.44	9.32	1.65	2.99	2.06	1.42	1.66	2.78
								Mazda 3	14.98	7.26	2.67	2.03	5.46	1.35	4.63	2.50	1.24	1.91	5.23	2.10	3.19	1.65
									Mazda 6	6.10	6.68	4.75	3.60	1.27	2.30	13.35	2.83	5.32	3.28	1.54	3.28	6.12
										Lancer	2.95	2.57	6.84	1.39	9.26	3.28	1.68	2.14	5.07	1.38	3.69	1.86
											Altima	5.12	5.00	1.08	1.56	5.24	2.55	4.54	2.81	2.13	2.04	2.87
												Maxima	3.53	1.08	1.49	3.79	5.77	3.61	2.00	1.05	1.62	4.12
													Sentra	1.14	2.72	2.02	2.11	2.63	6.62	1.66	2.67	1.47
														Forester	5.99	5.80	.81	1.05	1.09	1.04	1.46	1.75
															Impreza	8.05	.94	1.20	1.95	1.43	2.45	1.85
																Legacy	1.53	4.53	1.65	1.11	2.76	4.12
																	Avalon	4.80	2.30	1.94	1.30	2.53
																		Camry	4.09	2.63	1.89	2.52
																			Corolla	2.49	2.62	1.70
																				Prius	1.94	1.62
																					Jetta	20.46
																						Passat

Table 4
Top 20 sedan pairs with the highest lifts.

Car 1	Size	Car 2	Size	Match		Lift
				Brand	Size	
VW Passat	Midsize	VW Jetta	Compact	√		20.46
Mazda6	Midsize	Mazda3	Compact	√		14.98
Subaru Legacy	Midsize	Mazda6	Midsize		√	13.35
Subaru Legacy	Midsize	Kia Optima	Midsize		√	9.32
Mitsubishi Lancer	Compact	Subaru Impreza	Compact		√	9.26
Mazda6	Midsize	Kia Optima	Midsize		√	8.73
Subaru Legacy	Midsize	Subaru Impreza	Compact	√		8.05
Mitsubishi Lancer	Compact	Mazda3	Compact		√	7.26
Mitsubishi Lancer	Compact	Nissan Sentra	Compact		√	6.84
Mazda6	Midsize	Nissan Altima	Midsize	√		6.68
Nissan Sentra	Compact	Toyota Corolla	Compact		√	6.62
VW Passat	Midsize	Mazda6	Midsize		√	6.12
Mitsubishi Lancer	Compact	Mazda6	Midsize			6.10
Subaru Impreza	Compact	Subaru Forester	Full-size	√		5.99
Subaru Legacy	Midsize	Subaru Forester	Full-size	√		5.80
Toyota Avalon	Full-size	Nissan Maxima	Full-size		√	5.77
Mazda6	Midsize	Honda Accord	Midsize		√	5.71
Mazda6	Midsize	Hyundai Sonata	Midsize		√	5.45
Nissan Sentra	Compact	Mazda3	Compact		√	5.46
Toyota Corolla	Compact	Honda Civic	Compact		√	5.44

To examine the robustness of our method and provide evidence that granularity of the spatiotemporal information is essential in inferring co-considerations, we set out to investigate the boundary conditions by expanding the spatio-temporal horizon to suppress granularity. Specifically, we conducted the following robustness checks.

1. We set the geographic granularity to Zip code, Designated Market Area (DMA, as defined by Nielsen Corporation) or state level.
2. We set the time interval for co-occurrence to be 5, 10, 20, 30, or 60 minutes.

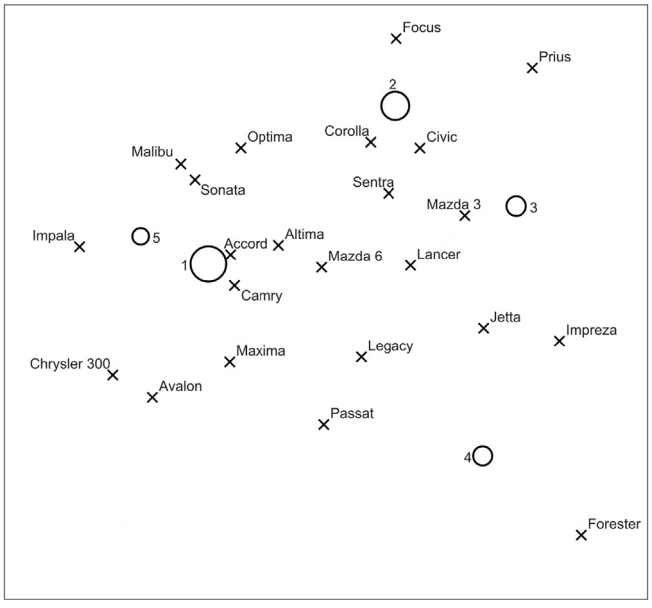


Fig. 6. Vehicle model level product positioning based on co-consideration measures and the relative position of segment ideal points.

This results in 15 sets of lifts, the Pearson correlation matrix of which is reported in Table 5. The results suggest that lift is very sensitive to changes in the level of geographic aggregation, from Zip code to DMA or state. It is, however, not sensitive to the selection of time intervals up to 60 minutes. This is likely because within each Zip code, OPQR data are relatively sparse over time. Therefore, changing the length of time interval has a relatively small effect on co-occurrence counts.

Piecing all the evidence together – the resemblance of our brand-level positioning map to that of Netzer et al. (2012), the consistency between industry conventional wisdom and the patterns of the vehicle pairs with the highest lifts, the interpretability of brand-level and vehicle-model-level positioning maps, and the robustness of the results – provides multi-facet validations for the use of lifts as proxies of co-consideration intensity. In the next section, we provide further evidence that lift measures in fact capture the underlying competitive structure of the U.S. auto market.

Empirical Validation

Sales Response Model

We embed the product-positioning configuration of Fig. 6 in a sales response model. Following the literature on random utility theory and ideal-point models (MacKay, Easley, and Zinnes 1995), we formulate the utility a consumer derives from purchasing a vehicle as a function of marketing activities (e.g., advertising, incentives), vehicle characteristics (e.g., price and features), and the proximity between the consumer's ideal point and the perceived location of the vehicle in a two-dimensional preference space (Elrod 1991; Grover and Dillon 1985). We assume the consumer incurs a disutility from preference mismatch, which is the square of the Euclidean distance between the coordinates of consumer k's ideal point, $\boldsymbol{s_k}$ ($k = 1\ldots K$), and the coordinates of vehicle j's location, $\boldsymbol{r_j}$ ($j = 1\ldots J$) (Cooper and Inoue 1996). The utility consumer k derives from purchasing vehicle j during time period t is given by

$$u_{kjt} = \rho_k\left(\boldsymbol{x_{kjt}}\boldsymbol{\beta}\right) - \delta_k\left(\left|\boldsymbol{r}_j - \boldsymbol{s_k}\right|\right)^2 + \varepsilon_{kjt} \tag{4}$$

where $\boldsymbol{x_{kjt}}$ is a vector of the marketing activities and $\boldsymbol{\beta}$ is a vector of their effects. ρ_k and δ_k are positive scaling parameters, the ratio of which accounts for the relative sensitivity of consumer k to marketing activities versus preference mismatch. ε_{kjt} is an independent and identically distributed extreme value I random disturbance. We derive vehicle positions $\boldsymbol{r_j}$ based on OPQR co-occurrences and estimate the ideal points $\boldsymbol{s_k}$.

Instead of estimating Eq. (4) at the individual level, we aggregate it to the regional-level, resulting in a heterogeneous aggregate logit model similar to Besanko, Dube, and Gupta (2003). More specifically, we assume there are I segments to which a consumer may belong. All consumers belonging to the same segment i ($i = 1\ldots I$) share parameters ρ_i, δ_i and ideal point $\boldsymbol{s_i}$. In addition, within the same region d, all consumers

Table 5
Correlations between lift metrics calculated at various geographic spans and time intervals.

Geo.	Time interval (minutes)	Zip code					DMA					State				
		5	10	20	30	60	5	10	20	30	60	5	10	20	30	60
Zip code	5		1.00	.98	.97	.91	.06	.03	.01	.00	−.01	.02	.01	.01	.01	.01
	10	1.00		.99	.98	.94	.06	.03	.01	.00	−.01	.01	.01	.00	.00	.00
	20	.98	.99		1.00	.97	.07	.04	.02	.01	.00	.01	.01	.00	.00	.00
	30	.97	.98	1.00		.98	.09	.06	.04	.03	.02	.02	.02	.01	.01	.01
	60	.91	.94	.97	.98		.15	.12	.10	.09	.07	.07	.06	.06	.06	.06
DMA	5	.06	.06	.07	.09	.15		1.00	1.00	1.00	.99	.93	.93	.93	.93	.93
	10	.03	.03	.04	.06	.12	1.00		1.00	1.00	1.00	.93	.93	.93	.93	.93
	20	.01	.01	.02	.04	.10	1.00	1.00		1.00	1.00	.93	.93	.93	.93	.93
	30	.00	.00	.01	.03	.09	1.00	1.00	1.00		1.00	.93	.93	.93	.93	.93
	60	−.01	−.01	.00	.02	.07	.99	1.00	1.00	1.00		.93	.93	.93	.93	.93
State	5	.02	.01	.01	.02	.07	.93	.93	.93	.93	.93		1.00	1.00	1.00	1.00
	10	.01	.01	.01	.02	.06	.93	.93	.93	.93	.93	1.00		1.00	1.00	1.00
	20	.01	.00	.00	.01	.06	.93	.93	.93	.93	.93	1.00	1.00		1.00	1.00
	30	.01	.00	.00	.01	.06	.93	.93	.93	.93	.93	1.00	1.00	1.00		1.00
	60	.01	.00	.00	.01	.06	.93	.93	.93	.93	.93	1.00	1.00	1.00	1.00	

receive the same treatment of marketing activities $\boldsymbol{x_{djt}}$. As a result, Eq. (4) becomes[4]

$$u_{idjt} = \rho_i\left(\boldsymbol{x_{djt}}\boldsymbol{\beta}\right) - \delta_i\left(\left|\boldsymbol{r_j} - \boldsymbol{s_i}\right|\right)^2 + \varepsilon_{kjt} \tag{5}$$

Thus, the probability that consumers of segment i buy product j is

$$p_{idjt} = \frac{\exp\left(\rho_i\left(\boldsymbol{x_{djt}}\boldsymbol{\beta}\right) - \delta_i\left(\left|\boldsymbol{r_j} - \boldsymbol{s_i}\right|\right)^2\right)}{\sum_{j=1}^{J} \exp\left(\rho_i\left(\boldsymbol{x_{djt}}\boldsymbol{\beta}\right) - \delta_i\left(\left|\boldsymbol{r_j} - \boldsymbol{s_i}\right|\right)^2\right)} \tag{6}$$

Following the extensive literatures on latent variable models (e.g., Dillon and Mulani 1989; Kamakura and Russell 1989), we use λ_{id} to represent the probability of finding a consumer belonging to segment i in region d, and λ_{id} must fulfill the condition $\sum_{i=1}^{I}\lambda_{id} = 1\ \forall\ d$. The unconditional probability of product j being chosen by a consumer from region d during time period t is therefore

$$p_{djt} = \sum_{i=1}^{I} \lambda_{id} p_{idjt} \tag{7}$$

We observe the market share of vehicle j in region d during time period t, y_{djt}. Given the choice probability defined in Eq. (6), the log-likelihood function is defined as

$$LL(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{r}; \boldsymbol{\beta}, \boldsymbol{s}, \lambda, \rho, \delta) = \sum_{t}\sum_{j}\sum_{d} y_{djt}\ log\left(p_{djt}\right) \tag{8}$$

We use the EM algorithm to maximize the log-likelihood defined in Eq. (8), yielding estimates for the model parameters: the pooled effect of marketing activities $\boldsymbol{\beta}$, segment specific ideal-point locations $\boldsymbol{s_i}$, the relative weights of market activities ρ_i vs. proximity to ideal point δ_i, and segment sizes in each region λ_{id}.

We can identify the model because the data have temporal and geographic variations. The parameters $\boldsymbol{\beta}$ and ρ_i are identified through the spatiotemporal co-variations in market share (y_{djt}) and marketing activities ($\boldsymbol{x_{djt}}$). The residual variations in market share not captured by marketing activities are captured through $-\delta_i(|\boldsymbol{r_j} - \boldsymbol{s_i}|)^2 + \varepsilon_{kjt}$, which identifies the segment-specific ideal points.

Data

In the empirical analysis, we focus on the 23 sedans in Table 2, which account for 80% of the total sedan sales in the U.S. For each of the 23 cars in each DMA, we compiled historical monthly sales data. The data spans 41 periods from January 2009 to May 2012 and covers the top 50 DMAs, which account for 62% of sales and 75% of advertising spending in the U.S. auto industry. To control for marketing-related effects, we use five variables:

- Advertisement spending, available at monthly DMA level for each vehicle.
- Total incentive expenditure per vehicle, including all promotional expenditures such as cash back to the consumers, costs of offers for financing the cars at promotional APRs, and trade promotions paid to the dealers, available monthly at the region level, where the U.S. is divided into five regions: East, Southeast, Great Lakes, Central, and West.
- Three features of the vehicles, manufacturer suggested retail price (MSRP), fuel economy (miles per gallon or MPG), and power (horse power), which have been shown to be prominent in driving auto sales in the U.S. (Du, Hu, and

[4] An alternative to Eq. (5) is to relax the assumption that the effects of marketing activities are pooled and allow $\boldsymbol{\beta}$ to vary by segment resulting in the following utility function

$$u_{idjt} = \boldsymbol{x_{djt}}\boldsymbol{\beta_i} - \delta\left(\left|\boldsymbol{r_j} - \boldsymbol{s_i}\right|\right)^2 + \varepsilon_{idjt}$$

We have estimated both models and the substantive findings remain robust. In the rest of the paper, we present the results of the model with pooled $\boldsymbol{\beta}$ (Eq. (5)), which performs better in goodness-of-fit.

Damangir 2015). The features vary by year and are constant across all geographic areas.

All the data at lower frequency and granularity are mapped to monthly DMA level. As shown in Table 6, across DMAs, time, and different vehicle models, there are large variations in sales and marketing activities. Product features are comparably stable, as they vary by vehicle, but change slowly over time.

Benchmarks

A key innovation of our sales response model lies in that we augment the sales and marketing data with a product positioning map (Fig. 6) generated from the patterns of co-considerations inferred from the co-occurrences of hundreds of millions of OPQRs without the benefit of any unique individual identifiers. To evaluate how well the proposed product-positioning map captures the underlying market structure, we compare the performance of our model to two benchmark models using alternative product positioning maps. Comparing the goodness-of-fit performance of these models should provide further evidence regarding the validity of using OPQR co-occurrences as proxies for product co-considerations.

The first benchmark model uses vehicle features to create a product-positioning map ("feature-based map" hereafter). Following previous research (e.g., Du, Hu, and Damangir 2015), we focus on four prominent features: brand, fuel economy, price, and power, using data collected from Edmunds.com. In order to create a similarity measure from the combination of categorical and continuous vehicle features, we first categorize the continuous features into quartiles. For example, if the price of a car falls into the first, second, third, and fourth quartiles of the distribution, the price value is transformed to a categorical variable with four possible nominal values. The transformed features along with brand give us four categorical dimensions. Then, we measure similarities between the cars using simple matching coefficient: the total number of dimensions with matching categorical attributes divided by the number of attributes (we have four attributes here). Finally, we apply non-metric MDS to these measures to generate the feature-based map.

The second benchmark model uses a product-positioning map that is based on the similarity in market share variations ("share-based map" hereafter). We measure similarity using the correlation between the market shares of pairs of cars over time and across DMAs. The rationale for this approach is that higher correlation in demand is driven by similarity of preferences across DMAs and over time. For example, consumers in San Francisco Bay Area tend to purchase a higher share of environment-friendly cars. Consumers in suburban environments may prefer more powerful and larger cars. As a result, a high correlation in market shares across regions conveys similarity between a pair of vehicles. Moreover, the seasonal purchase patterns of consumers with similar preferences are more likely to coincide. For example, college graduates usually purchase new cars in the summer before the start of a new job. Such behavior gives rise to similar seasonality for similar products. Again, we use non-metric MDS to transform market share correlations into a two-dimensional product-positioning map. We expect this benchmark model to be difficult to outperform because it enjoys an "unfair" advantage by using patterns of correlations in market share as an input to a model for predicting market.

Results

We create three sets of data by augmenting the sales and marketing data with the alternative product-positioning maps generated from co-consideration patterns, feature similarity, and market-share variation similarity. To examine the robustness of the results we perform the analysis using 25 and 50 top DMAs. At the same time, we systematically vary the number of latent segments from two to five. Table 7 reports the performance of the resulting 24 models (3 product positioning maps × 2 DMA counts × 4 latent segment counts).

The main takeaway from Table 7 is that the proposed model, using the OPQR-based product-positioning map, outperforms the benchmark models using either feature-based or share-based product-positioning map. As we increase the number of latent segments, the performance improves. Regardless of the number of DMAs included, the best performing model is the one with five latent segments. For all the cases with more than two segments, we find that the proposed model performs better than the alternatives, suggesting our proposed map best captures the underlying competitive relationships among the 23 sedans included in our analysis. This finding provides further evidence for the validity of our probabilistic approach to inferring co-considerations from OPQR co-occurrences. We attribute the superior performance of the proposed product-positioning map to its consumer centric nature, which allows it to better capture nuances in the underlying market structure elusive to the product-centric alternatives. For example, Chevrolet Malibu and Subaru Legacy have high similarity in features (Malibu is slightly

Table 6
Descriptive statistics of the data.

Variable (unit)	Data frequency	Geographic granularity	Min	Max	Mean	Standard deviation
Sales (units)	Monthly	DMA	0	5,525	127.8	236.9
Ad spend ($)	Monthly	DMA	0	2,374,329	57,970	125,923
Incentives ($ per car)	Monthly	Region	0	10,990	2,983	1,898
MSRP ($)	Yearly	National	15,043	31,220	20,500	4726
Fuel economy (MPG)	Yearly	National	20.5	49.5	27.1	5.4
Power (horse power)	Yearly	National	132.0	292.0	182.5	45.0

Table 7
Performance of alternative models with different number of markets and segments.

DMAs	Segments	Product positioning map from		
		Co-considerations	Feature	Share
25	2	138,979 (BIC) .0149 (MAE) 61% (MAPE)	138,853 .0151 61%	138,639 .0136 65%
	3	133,543 .0052 16%	138,547 .0145 59%	134,539 .0096 34%
	4	133,408 .0046 14%	137,968 .0138 55%	134,435 .0088 31%
	5	**132,795** **.0025** **12%**	137,929 .0129 54%	133,477 .0071 27%
50	2	279,199 (BIC) .0155 (MAE) 62% (MAPE)	278,606 .0154 61%	279,006 .0140 67%
	3	267,905 .0055 17%	279,324 .0162 61%	270,463 .0104 36%
	4	267,562 .0045 14%	276,425 .0132 52%	268,891 .0072 24%
	5	**266,462** **.0026** **12%**	276,590 .0129 52%	267,812 .0070 26%

Note: In each cell, the reported figures are BIC, mean absolute error, and mean absolute percentage error respectively. Bold fonts indicate the best performance in each block.

Table 8
Parameters estimates.

Parameters	Segments				
	1	2	3	4	5
s_1 (ideal-point dimension 1)	−.88 ±.03 (.05)	.57 ±.18 (.36)	1.51 ±.19 (.37)	1.25 ±.08 (.16)	−1.41 ±.18 (.34)
s_2 (ideal-point dimension 2)	.03 ±.03 (.05)	1.26 ±.02 (.03)	.48 ±.14 (.28)	−1.47 ±.09 (.17)	.24 ±.20 (.38)
ρ (marketing)	.72 ±.31 (.60)	.62 ±.23 (.44)	1.00 .00 (.00)	.80 ±.24 (.48)	1.90 ±.21 (.41)
δ (proximity to ideal point)	2.43 ±.18 (.35)	3.61 ±.21 (.41)	2.65 ±.23 (.45)	2.04 ±.25 (.48)	1.71 ±.46 (.90)
Parameters common across all segments					
$\boldsymbol{\beta}_{\text{ad}}$	.07 ±.04 (.08)				
$\boldsymbol{\beta}_{\text{incentive}}$	.10 ±.19 (.37)				
$\boldsymbol{\beta}_{\text{MSRP}}$	−.02 ±.15 (.28)				
$\boldsymbol{\beta}_{\text{MPG}}$	.73 ±.14 (.28)				
$\boldsymbol{\beta}_{\text{power}}$	.23 ±.06 (.13)				

Notes: For each parameter, the reported estimates are the mean, margin of error, and standard error, respectively.

more fuel-efficient), and are at the 31st percentile on market share correlation. As a result, Malibu and Legacy are close to each other on both the feature- and share-based maps. However, they clearly occupy distinct perceptive positions and target distinct segments of consumers, a fact that has not eluded our proposed map.

We report the results of the best performing model with 50 top DMAs in Table 8, which summarizes the estimated parameters in Eq. (5): the ideal point for each segment $\boldsymbol{s_i}$, the relative importance of marketing activities to proximity to ideal point for each segment ρ_i and δ_i, and $\boldsymbol{\beta}$, the average overall responsiveness to advertisements, incentive expenditures, and features. Except for the element corresponding to MSRP and incentives, the estimated elements of $\boldsymbol{\beta}$ are positive and significant. The first three sets of parameters ($\boldsymbol{s_i}$, ρ_i, and δ_i) are better interpreted along with the segment results visually in Fig. 6, which provides a complete picture of the overall market structure. We overlay the estimated segment ideal points (circles) on top of the product-positioning map inferred from the patterns of OPQR co-occurrences. The radius of each circle is proportional to the size of each segment aggregated to the national level. In addition to ad spending, incentives, MSRP, fuel economy, and power, the proximity between the segment ideal point and a product's position determines the probability of choice.

In addition, the estimated segment sizes λ_{id} in Table 9 demonstrate sizable heterogeneity across DMAs. To better represent the segment size estimates, we visualize the segment sizes for a few large DMAs on a map (Fig. 7).

To calculate the probability of each segment buying a particular car, we aggregate the probabilities from Eq. (6) across DMAs and over time. The resulting choice probabilities (Table 10) help us quantitatively interpret the graphical representation of the market structure illustrated in Fig. 6.

Segment 1, the largest segment and perhaps not surprisingly, is close to the market leaders in this class, Honda Accord and Toyota Camry. The probability for a consumer in the segment to buy these automobiles (p hereafter) is 26%. To a lesser extent, this segment prefers other mainstream midsize and full-size sedans such as Nissan Altima (p = 14%), Hyundai

Table 9
Estimated share of segment in each DMA.

DMA	Segment					DMA	Segment				
	1	2	3	4	5		1	2	3	4	5
New York	.55	.22	.09	.13	.00	Sacramento	.44	.28	.15	.12	.00
Los Angeles	.44	.27	.20	.08	.00	San Antonio	.47	.27	.18	.05	.03
Chicago	.30	.26	.11	.11	.21	Charlotte	.51	.26	.08	.07	.09
Philadelphia	.43	.24	.12	.14	.08	Raleigh/Durham	.49	.26	.13	.07	.04
Detroit	.00	.34	.04	.07	.55	Indianapolis	.19	.23	.11	.08	.39
Dallas	.50	.26	.11	.08	.04	Portland	.30	.20	.23	.24	.03
Boston	.44	.27	.11	.18	.00	Cincinnati	.29	.28	.12	.09	.23
Houston	.51	.28	.09	.08	.04	Buffalo	.11	.25	.06	.13	.45
Washington DC	.42	.27	.14	.12	.04	Kansas City	.30	.27	.12	.08	.23
Miami	.51	.31	.08	.10	.00	Milwaukee	.26	.28	.09	.15	.21
SF Bay Area	.35	.27	.27	.12	.00	Columbus	.35	.29	.08	.10	.18
Atlanta	.53	.27	.09	.06	.04	Austin	.43	.21	.21	.09	.06
Tampa	.40	.26	.11	.06	.15	Nashville	.52	.26	.11	.05	.07
Cleveland	.29	.28	.08	.11	.23	Salt Lake City	.31	.19	.16	.21	.12
Phoenix	.36	.25	.18	.08	.13	Albany	.35	.25	.10	.20	.10
Minneapolis	.27	.25	.12	.13	.23	Harrisburg	.29	.31	.12	.15	.14
Orlando	.44	.27	.13	.06	.09	Las Vegas	.50	.28	.11	.09	.02
Seattle	.24	.20	.23	.27	.06	Oklahoma City	.42	.25	.10	.06	.17
Pittsburgh	.19	.22	.07	.22	.29	New Orleans	.65	.25	.07	.03	.00
Denver	.24	.18	.15	.33	.09	Norfolk	.43	.28	.12	.09	.08
Baltimore	.41	.25	.13	.11	.10	Greenville/Asheville	.51	.26	.10	.10	.03
St. Louis	.26	.24	.12	.08	.30	Jacksonville	.53	.29	.09	.06	.03
San Diego	.37	.30	.22	.10	.00	Wilkes Barre	.25	.28	.08	.23	.16
West Palm Beach	.52	.22	.12	.08	.06	Birmingham	.63	.24	.06	.04	.02
Hartford/New Haven	.41	.21	.11	.23	.03	Providence	.48	.26	.09	.14	.04

Notes: The total segment size for each DMA is normalized to 1.

Sonata (p = 11%), Chevrolet Malibu (p = 6%), and Nissan Maxima (p = 6%). Segment 1 is present in all major markets in the United States, but it is even more prominent in Southeast and Northeast.

Consumers in Segment 2 have a strong preference for compact sedans: Honda Civic (p = 35%), Toyota Corolla (p = 33%), and Ford Focus (p = 18%), and Nissan Sentra (p = 9%). As it is evident in Table 8, this segment is popular in most of the United States with a share of 20% to 30% in most of the markets.

Fig. 7. Distributions of segment sizes in selected DMAs. Notes: From left to right, the five bars in each DMA represent segments 1, 2, 3, 4, and 5 respectively.

Segment 3 represents the customers who prefer fuel-efficient sedans: Toyota Prius (p = 41%), Mazda 3 (p = 29%), and Honda Civic (p = 12%). This segment tends to be larger in the West Coast than in the South and East Coast.

Consumers in Segment 4 are more likely to buy VW Jetta (p = 40%), Subaru Forester (p = 24%), Subaru Impreza (p = 16%), VW Passat (p = 10%), and Subaru Legacy (p = 8%). These consumers prefer sporty and powerful cars with outdoor capabilities. This segment constitutes a relatively small percentage of the market in most of the country. The exceptions are Denver, Seattle, and Portland, the mountainous DMAs where outdoor capabilities become more desirable.

The smallest segment, Segment 5, is similar to Segment 1, but is less interested in Japanese cars. This segment is primarily interested in Hyundai Sonata (p = 58%), Chevrolet Malibu (p = 36%), Chevrolet Impala (p = 11%), Honda Accord (p = 11%), and Toyota Camry (p = 11%). This segment is small in most of the country, except in the areas close to the headquarters of domestic U.S. automakers. At its peak in the DMA that includes Detroit, this segment is the largest and covers 55% of the market.

Taken together, all the above empirical evidence suggests that our approach to inferring market structure from OPQRs has strong face and predictive validity. Such an approach can offer

Table 10
The probabilities of vehicle choices for each segment.

Cars	Segments				
	1	2	3	4	5
Chevrolet Impala	.02	.00	.00	.00	.13
Chevrolet Malibu	.06	.00	.00	.00	.20
Chrysler 300	.01	.00	.00	.00	.02
Ford Focus	.00	.18	.00	.00	.00
Honda Accord	.26	.00	.00	.00	.13
Honda Civic	.00	.35	.12	.00	.00
Hyundai Sonata	.11	.00	.00	.00	.22
Kia Optima	.04	.00	.00	.00	.06
Mazda 3	.00	.01	.29	.00	.00
Mazda 6	.03	.00	.00	.00	.01
Mitsubishi Lancer	.00	.00	.05	.02	.00
Nissan Altima	.14	.00	.00	.00	.05
Nissan Maxima	.06	.00	.00	.00	.02
Nissan Sentra	.00	.09	.05	.00	.00
Subaru Forester	.00	.00	.00	.24	.00
Subaru Impreza	.00	.00	.01	.16	.00
Subaru Legacy	.00	.00	.00	.08	.00
Toyota Avalon	.01	.00	.00	.00	.01
Toyota Camry	.26	.00	.00	.00	.13
Toyota Corolla	.00	.33	.01	.00	.00
Toyota Prius	.00	.03	.41	.00	.00
VW Jetta	.00	.00	.05	.40	.00
VW Passat	.00	.00	.00	.10	.00

Notes: the sum of each column (segment) is 1.

managers several advantages. First, by embedding the product-positioning map in a sales response model, we discovered five segments in the U.S. auto market. The sizes of these segments vary across DMAs, enabling automakers to customize strategies according to regional market conditions. Second, because our approach does not require human interference to pre-define competitive sets, it allows firms to discover previously unknown patterns of competition and cannibalization. Third, our approach can be carried out on an ongoing basis; as new data arrives, the product-positioning map can be updated in real time.

Concluding Remarks

In this research, we use a case study to demonstrate how one can leverage the spatiotemporal granularity and massive scale of online consumer activity data for affinity analysis, even without a unique individual identifier. We use multiple approaches to investigate the validity of our results and find consistent evidence that the granularity of the data (at five-minute interval and by Zip code) makes it possible to infer patterns of vehicle co-considerations from the patterns of OPQR co-occurrences. The similarities between the resulting brand level product-positioning map and the map reported by Netzer et al. (2012), along with the fact that our vehicle model level map outperforms the feature- or shared-based map in a sale response model, suggest spatiotemporal patterns embedded in the data contain genuine information about the underlying competitive relationships among the vehicles. Furthermore, our proposed map appears to be particularly useful for understanding products that are similar in both features and market shares (thus indiscernible on either a feature- or share-based map).

Sharing consumer activity data from multiple sources (e.g., different websites or organizations) can make the combined data much more useful (de Montjoye et al. 2015; Lazer et al. 2014). Doing so, however, typically requires unique individual identifiers that can tie records from different sources together, raising privacy concerns due to the possibility of re-identification (de Montjoye et al. 2015; Li and Sarkar 2011, 2014; Menon, Sarkar, and Mukherjee 2005). Our case study illustrates that consumer activity data that does not contain any unique individual identifier may still be leveraged for affinity analysis, with minimum privacy concerns about sharing information across sources.

More broadly, our research demonstrates that creative use of not-so-perfect but massive online consumer activity data can be rewarding. Some academic and industry experts have argued for a shift in perspective when working with big data (e.g., Einav and Levin 2014; Mayer-Schönberger and Cukier 2013), given the intrinsic differences between big data and traditional data. In the past, researchers were used to working with data with relatively few high-quality observations gathered using carefully designed measurement instruments. Nowadays, with wide adoptions of sophisticated information systems, big data common in businesses are results of automatic data collections not optimized for high quality measurements. As technological advances have removed many of the barriers for analyzing as much data as we wish (Varian 2013), the challenge lies in that these data are often collected without specific research purposes and with reduced measurement quality, making them not so ideal for conventional analyses. They can be messy and missing what is deemed as a crucial piece of information (Mayer-Schönberger and Cukier 2013). On the other hand, they are massive in size and tend to be very granular. The hope is that the richness and size of the data allow certain real signals to emerge from the noisy and messy data (Fan, Han, and Liu 2014). In our case, as an example of embracing this shift in perspective, we take advantage of the granular spatiotemporal information to compensate for the lack of unique individual identifiers. Our empirical results show that we can have a valid measurement despite the deficiencies of the data, as long as we use proper procedures to establish validity of the measurement.

Rather than making a broad claim on the generalizability of our approach, we acknowledge that limitations exist. For example, in a different empirical context there is no guarantee that one can infer co-considerations by leveraging spatiotemporal information like we did in our empirical context. As our robustness analysis indicates, our approach relies on data granularity and relative sparsity. In cases where unrelated events co-occur with a higher density and/or raw records are aggregated to a longer time span or over a larger area, random co-occurrences can become dominant enough to prevent the reliable detection of genuine co-consideration signals. Thus, the applicability of our approach in other empirical contexts should be examined on a case-by-case basis.

For future research, we see several directions to extend our study. First, we use a static product-positioning map created from inferred co-considerations. One can imagine that as the market evolves, competition can shift as well. Although we do not expect competition to change every day, it is certainly possible to extend the current approach to create a dynamic product-positioning map using methods such as dynamic MDS. As a result, managers can receive frequent updates on the competitive landscape. Second, we focused on one auto body type — sedans. A straightforward extension is to examine all body types. Such a study shall help discover cross-class competitions (e.g., between a sedan and an SUV) that may exist in the auto industry. Third, the location information in our data makes it possible to extend the research scope to study contagions in co-considerations. When people in one location start to co-consider a pair of products with higher frequency, it is conceivable that through various social contacts and observations, people in neighboring locations may increasingly co-consider the same pair as well. Different from contagion of individual products, contagion of co-considerations, if exists, would suggest competitions can spread through multiple markets following a similar diffusion process. Finally, OPQRs happen at the lower purchase funnel. Although they are closer to the final purchases and may be more accurate in predicting actual purchases, they tend to cover a consideration set that has narrowed down considerably from the upper purchase funnel. At this stage, firms have fewer levers to influence consumer choices. If one can combine this data with upper funnel co-consideration activities, one should be able to identify competitive relationships at different stages of the purchase funnel.

References

Anderson, Chris (2006), *The Long Tail why the Future of Business Is Selling Less of More*. NY: Hyperion.

Besanko, David, Jean-Pierre Dube, and Sachin Gupta (2003), "Competitive Price Discrimination Strategies in a Vertical Channel Using Aggregate Retail Data," *Management Science*, 49, 9, 1121–238.

Bloch, Peter H. (1995), "Seeking the Ideal Form: Product Design and Consumer Response," *Journal of Marketing*, 59, 3, 16–29.

Brijs, Tom, Gilbert Swinnen, Koen Vanhoof, and Geert Wets (2004), "Building an Association Rules Framework," *Data Mining and Knowledge Discovery*, 8, 1, 7–23.

Chen, Hsinchun, Roger H.L. Chiang, and Veda C. Storey (2012), "Business Intelligence and Analytics: From Big Data to Big Impact," *MIS Quarterly*, 36, 4, 1165–88.

Choi, S. Chan (1991), "Price Competition in a Channel Structure with a Common Retailer," *Marketing Science*, 10, 4, 271–96.

Cooper, Lee G. and Akihiro Inoue (1996), "Building Market Structures from Consumer Preferences," *Journal of Marketing Research*, 33, 3, 293–306.

Day, George S., Allan D. Shocker, and Rajendra K. Srivastava (1979), "Customer-oriented Approaches to Identifying Product-markets," *Journal of Marketing*, 43, 4, 8–20.

de Montjoye, Yves-Alexandre, Laura Radaelli, Vivek Kumar Singh, and Alex "Sandy" Pentland (2015), "Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata," *Science*, 347, 6221, 536–9.

Dillon, William R. and Narendra Mulani (1989), "LADI: A Latent Discriminant Model for Analyzing Marketing Research Data," *Journal of Marketing Research*, 26, 1, 15–29.

Du, Rex Yuxing, Ye Hu, and Sina Damangir (2015), "Leveraging Trends in Online Searches for Product Features in Market Response Modeling," *Journal of Marketing*, 79, 1, 29–43.

Einav, Liran and Jonathan Levin (2014), "Economics in the Age of Big Data," *Science*, 346, 6210, 715–21.

Elrod, Terry (1991), "Internal Analysis of Market Structure: Recent Developments and Future Prospects," *Marketing Letters*, 2, 3, 253–66.

Fan, Jianqing, Fang Han, and Han Liu (2014), "Challenges of Big Data Analysis," *National Science Review*, 1, 2, 293–314.

Gao, Guodong (Gordon), Brad N. Greenwood, Ritu Agarwal, and Jeffrey McCulllough (2015), "Vocal Minority and Silent Majority: How Do Online Ratings Reflect Population Perceptions of Quality?" *MIS Quarterly*, 39, 3, 565–89.

Gardial, Sarah Fisher, D. Scott Clemons, Robert B. Woodruff, David W. Schumann, and Mary Jane Burns (1994), "Comparing Consumers' Recall of Prepurchase and Postpurchase Product Evaluation Experiences," *Journal of Consumer Research*, 20, 4, 548–60.

Grover, Rajiv and William R. Dillon (1985), "A Probabilistic Model for Testing Hypothesized Hierarchical Market Structures," *Marketing Science*, 4, 4, 312–35.

Hunt, Shelby D., Richard D. Sparkman Jr., and James B. Wilcox (1982), "The Pretest in Survey Research: Issues and Preliminary Findings," *Journal of Marketing Research*, 19, 2, 269–73.

Kamakura, Wagner A. and Gary J. Russell (1989), "A Probabilistic Choice Model for Market Segmentation and Elasticity Structure," *Journal of Marketing Research*, 26, 4, 379–90.

Kim, Jun B., Paulo Albuquerque, and Bart J. Bronnenberg (2011), "Mapping Online Consumer Search," *Journal of Marketing Research*, 48, 1, 13–27.

Kotsiantis, Sotiris and Dimitris Kanellopoulos (2006), "Association Rules Mining: A Recent Overview," *GESTS International Transactions on Computer Science and Engineering*, 32, 1, 71–82.

Kruskal, Joseph B. (1964), "Nonmetric Multidimensional Scaling: A Numerical Method," *Psychometrika*, 29, 115–29.

Lattin, James M. and Leigh McAlister (1985), "Using a Variety-seeking Model to Identify Substitute and Complementary Relationships among Competing Products," *Journal of Marketing Research*, 22, 3, 330–9.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani (2014), "The Parable of Google Flu: Traps in Big Data Analysis," *Science*, 343, 6167, 1203–5.

Lee, Thomas Y. and Eric T. Bradlow (2011), "Automated Marketing Research Using Online Customer Reviews," *Journal of Marketing Research*, 48, 5, 881–94.

Li, Xiao-Bai and Sumit Sarkar (2011), "Protecting Privacy Against Record Linkage Disclosure: A Bounded Swapping Approach for Numeric Data," *Information Systems Research*, 22, 4, 774–89.

——— and ——— (2014), "Digression and Value Concatenation to Enable Privacy-preserving Regression," *MIS Quarterly*, 38, 3, 679–98.

Li, Xinxin and Lorin M. Hitt (2008), "Self-selection and Information Role of Online Product Reviews," *Information Systems Research*, 19, 4, 456–74.

MacKay, David B., Robert F. Easley, and Joseph L. Zinnes (1995), "Single Structure Ideal Point Model for Market Analysis," *Journal of Marketing Research*, 32, 4, 433–43.

Mayer-Schönberger, Victor and Kenneth Cukier (2013), *Big Data: A Revolution that Will Transform how We Live, Work, and Think*. Eamon Dolan/Houghton Mifflin Harcourt.

Menon, Syam, Sumit Sarkar, and Shibnath Mukherjee (2005), "Maximizing Accuracy of Shared Databases when Concealing Sensitive Patterns," *Information Systems Research*, 16, 3, 256–70.

Netzer, Oded, Ronen Feldman, Jacob Goldenberg, and Moshe Fresko (2012), "Mine your Own Business: Market-structure Surveillance Through Text Mining," *Marketing Science*, 31, 3, 521–43.

Park, Young-Hoon and Peter S. Fader (2004), "Modeling Browsing Behavior at Multiple Websites," *Marketing Science*, 23, 3, 280–303.

Ratchford, Brian T., Myung-Soo Lee, and Debabrata Talukdar (2003), "The Impact of the Internet on Information Search for Automobiles," *Journal of Marketing Research*, 40, 2, 193–220.

Ringel, Daniel M. and Bernd Skiera (2016), "Visualizing Asymmetric Competition among over 1000 Products Using Big Search Data," *Marketing Science*, 35, 3, 511–34.

Russell, Gary J. and Ruth N. Bolton (1988), "Implications of Market Structure for Elasticity Structure," *Journal of Marketing*, 25, 3, 229–41.

Shocker, Allan D., Moshe Ben-Akiva, Bruno Boccara, and Prakash Nedungadi (1991), "Consideration Set Influences on Consumer Decision-making and Choices: Issues, Models, and Suggestions," *Marketing Letters*, 2, 3, 181–97.

Shugan, Steven M. (2014), "Market Structure Research," in *The History of Marketing Science.* R.S. Winer, S.A. Neslin, editors. Hanover, MA: World Scientific Publishing Co., 129–64.

Urban, Glen L., Philip L. Johnson, and John R. Hauser (1984), "Testing Competitive Market Structures," *Marketing Science*, 3, 2, 83–112.

Varian, Hal R. (2013), "Beyond Big Data," Retrieved September 7, 2017: http://people.ischool.berkeley.edu/~hal/Papers/2013/BeyondBigDataPaperFINAL.pdf.

Zheng, Zhiqiang Eric, Peter Fader, and Balaji Padmanabhan (2012), "From Business Intelligence to Competitive Intelligence: Inferring Competitive Measures Using Augmented Site-centric Data," *Information Systems Research*, 23, 3, 698–720.