INSEAD

The Business School for the World®

# Faculty & Research Working Paper

## Who are my Competitors? Let the Customer Decide

Jun LI
Serguei NETESSINE
2012/84/TOM

# Who are my Competitors? Let the Customer Decide

Jun Li*

Serguei Netessine**

* Assistant Professor of Technology and Operations at Stephen M. Ross School of Business University of Michigan 701 Tappan St. Ann Arbor, MI 48109-1234, USA.
Email: junwli@umich.edu

** The Timken Chaired Professor of Global Technology and Innovation, Professor of Technology and Operations Management, Research Director of the INSEAD-Wharton Alliance at INSEAD Boulevard de Constance 77305 Fontainebleau, France.
E-mail: serguei.netessine@insead.edu

# Who are My Competitors? Let the Customer Decide

Jun Li
Stephen M. Ross School of Business, University of Michigan, junwli@umich.edu
Serguei Netessine
INSEAD, serguei.netessine@insead.edu

Many competitive industries find it challenging to identify their competition set, and despite the widely accepted importance of competition set to strategy development and daily operations, few data-driven approaches have been developed to address this challenge. In this paper, we propose a simple and intuitive methodology to identify true competitors from the customer perspective using online search and click-stream data. We use data from the hotel industry to build two competition networks: one based on our customer-centric approach, and the other based on price-matching patterns from the hotelier perspective. The customer-based competition network has an average degree of six to eight, consistent with the average number of competitors found in hotel managers surveys. Our approach also reveals a property of the competition network in hotel industry — small degree of separation. A hotel is connected with another within three steps of competitive links, which has important implications for how price perturbations travel through competitive links. Comparing the two networks, we find a 50% mismatch, with hoteliers tending to ignore independent and distant hotels while over-emphasizing branded and nearby hotels. This result is robust to many alternative measures of competition. Finally, this proposed methodology can easily be applied to many other industries to aid businesses in identifying their key competitors and to enrich our understanding of networked competition.

## 1. Motivation

With whom do you compete? Many competitive and fragmented industries, such as the hotel industry, are challenged by this question. The US hotel and motel industry consists of about 40,000 companies that operate about 50,000 properties. The largest 50 companies generate only 45% of total revenue. Take New York City, for example. Here, more than 500 hotels compete for business in the city and vicinity. Besides the sheer number of potential competitors, hotels also compete on multiple dimensions, both spatially (i.e., location) and vertically (i.e., quality). Competition becomes even more complex due to increasing use of the internet as a convenient and reliable search and transaction channel — hotel used to compete mostly with other hotels in close proximity, but now they might be competing with hotels located farther away but that offer appealing services and rates.

Understanding competition structure in such a market has important theoretical and practical implications. Theoretically, we do not fully understand how firms compete in a market like

that described above. While researchers have analyzed and measured competition in duopoly and oligopoly markets with a smaller number of competitors, we currently lack sufficient understanding of networked competition on multiple dimensions including both spatial and vertical differentiation — Who competes with whom? How dense is the competition? What makes competitors competitors? How does competition network arise in equilibrium and evolve over time? How does competition network affect equilibrium prices? One can keep on asking questions, but one of the first steps to answer these questions is to define, measure and characterize networked competition.

From a practical perspective, monitoring competitor performance is a critical part of a hotel's daily operations and long-term strategy development. However, industry professionals still depend mostly on rules of thumb to define their competition set. Prevalent current practice for defining competition set in the hotel industry ranges from looking across the street to identifying properties that charge the same basic rates (appealing to customers with the same price tolerance), and to weighing and scoring property attributes. For example, Smith Travel Research(STR) advises hotels to look at the following attributes in determine competition set: price, location, restaurant and room service in hotel, meeting space, complimentary breakfast, loyalty program, full-service amenities, and brand. The lack of data-driven analytical tools may lead to inadequate reactions to market changes, especially price changes. With increasing price transparency, matching prices with the wrong competitors can generate sizable revenue losses as customers quickly punish pricing errors by flooding in at low prices or booking away.

In the present paper, we propose a new customer-centric approach to studying competition structure using online search and clickstream data. At the core of our methodology is the idea that hotels should see themselves as potential customers do, because ultimately hotels are competing for customer's demand. Thus, rather than asking themselves with whom they think they are competing, hotels should ask who their customers identify as their competition. As noted by HotelNewsNow, a professional industry website launched by STR, "According to Forrester Research, 90 percent of all travel-related purchasing decisions are made online. So, despite our best efforts to define who we think our competition is, it's really customer perception that drives purchasing decisions." [1]

Analyzing consumer footprint has been made possible by the emerging availability of online data. Which hotels did consumers click to see details? Whose websites did they visit? Answers to these and other questions can reveal which hotels customers perceive as competitors. One important advantage of using clickstream data over transaction data is its straightforward nature and its efficiency. Conventionally, we depend primarily on transaction data to understand competition by estimating price elasticities among all pairs of competitors. Imagine doing this for 500 hotels

[1] "Who are my 'true' hotel competitors?" Trevor Stuart-Hill. HotelNewsNow.com

in New York City. This is equivalent of estimating 250,000 cross price elasticities. Clickstream data, however, tell us directly which hotels have competed for a particular customer. Note that clickstream data is not new to business. In all these years organizations have been collecting this big data, we lack effective analytical tools to leverage it. Clickstream data is known for being too granular and too voluminous to be managed or analyzed. While most studies using clickstream data have focused on understanding browsing behavior and predicting clickthrough or conversion rates more effectively (e.g., Moe and Fader 2004), we unveil another potential use of clickstream data — to analyze competition structure.

Specifically, we use a combination of clickstream data containing customer page views, search data containing displayed search results even if not clicked, and hotel data from a major online travel agency (OTA). By using online data, we focus our attention on leisure and unmanaged business travelers. The rest of the market, managed business travelers, usually book through different channels under different rates, such as corporate affiliation programs. In this leisure and unmanaged business market, online travel agencies play an increasingly important role in providing both information and easy booking channels to travelers. Based on data compiled by comScore, which monitors online behavior, 81.5% of consumers perform travel-related searches, and 70% of consumers visit a travel-related site prior to booking through a supplier's website. Employing data from one particular OTA, however, still the question regarding how representative it is of the online leisure and unmanaged business market. Note that OTAs differ in the sorting algorithms they use to sort search results. Based on our data, 30% of clicks on hotel-specific pages may potentially be affected by this sorting algorithm, while the rest of the clicks that originate from other internal (e.g., a promotion page) or external (e.g., Google) pages are not affected. A question thus arises: How much of the competition observed is actually due to the sorting algorithm the particular OTA has adopted? There are two possible ways in which sorting algorithms may affect clicking and comparison patterns among hotels. First, hotels listed closer to one another may be interpreted as similar hotels and are thus more likely to be compared. Our data actually rejected this adjacency hypothesis. The average rank distance (7.12) between two hotels that are clicked together is statistically larger than the expected distance (4.21) under no adjacency effect. Second, it is a well-documented fact that higher ranks are associated with higher click-through rate (e.g., Hoque and Loshe 1999). Thus, a hotel that is ranked higher by this OTA is more likely to become a competitor to other hotels. To adjust for this possibility, we consider several co-location measures developed in the Linguistic literature. These measures control the baseline probability of one hotel being clicked when calculating the level of co-location, or in our context, competition. Furthermore, we show that ignoring heterogeneity in hotel ranks and consumer tastes can result in an aggregation bias in these measures. To address this bias, we imbed a Random Coefficient Choice Model

into constructing co-location measures. This can be done only by combining search data (which provide information on all displayed hotels) and clickstream data (which provide information on clicked hotels).

In addition to analyzing competition from the customer perspective, we also calculate a proxy of competition from hotelier's perspective. Here, we propose to measure the strength of the competitive relationship using price responsiveness (price-matching). A common practice in the hotel industry is to monitor competitors' prices on a daily basis and match prices accordingly. We discover evidence of price matching mainly along two time dimensions — travel dates and booking dates. One advantage of the pricing data recorded from consumer searches is that unlike most studies that use Average Daily Rates (ADR), we observe exact prices quoted for each travel request. Thus, we are able to measure the level of parallelism of two price series matched by each travel request rather than simply measuring the difference between two aggregated price points.

With these two measures, we visualize the local competition structures in the form of competition networks, built on the customer and hotelier perspectives, respectively. We measure the extent of mismatch between the two competition networks and identify when mismatches are likely to occur. We find that independent hotels and distant hotels are likely to be excluded from competition sets, while branded hotels and nearby hotels are more likely to be included. Moreover, hotels tend to benchmark themselves against hotels that have lower star ratings, lower price levels, lower ranks, and higher customer reviews. By showing evidence of competition set misperception among hoteliers, we illustrate how a data-driven analytical approach can bring hotel managers closer to truly understanding their competition and their competitive positions.

Notably, although the present paper uses hotel industry data, the methodology proposed here can be applied easily to other products and services. The data required for this type of co-location or co-occurrence analysis prevails in various online and even offline settings. Many merchandisers' websites, such as Verizon and BestBuy, allow consumers to choose several products (usually up to four or five) and click "compare" to review them in detail. Many websites, such as TripAdvisor and Amazon, provide recommendations to consumers by showing selections labeled "customers who viewed this product also viewed...". Consumer-generated content on social media also allows us to analyze which products customers tend to mention together (Netzer et al. 2012). Furthermore, emerging technologies such as Shopper Tracker are able to track consumers' physical movements such as touching a product in a retail store. Unlike the days when we relied mostly on transaction data to understand competition and consumer choice, we are now able to know not only a consumer's final choice but also which options he/she has considered. In other words, we can discover which products and services competed for each customer. Thus, the customer-centric approach and the view of networked competition structure that we propose in this paper are likely to benefit

many other businesses and to serve as a basis for future research as well. More questions can be examined under this framework, such as how competition set changes with rate repositioning, how competition networks evolve dynamically, and how small local perturbations may travel across a network through competitive links.

## 2. Literature Review

First and foremost, our customer-centric approach is part of the current trend of incorporating consumer-driven models into analyzing of Revenue Management problems and associated empirical applications. Discrete choice models were first introduced to the community of Revenue Management by Talluri and van Ryzin (2004) through an application to a single-leg Revenue Management problem in the airline industry. Ever since, a stream of studies have followed this line of research, including Zhang and Cooper (2005), Gallego et al. (2009), and Farias et al. (2012), to name only a few. Empirically, customer-driven demand models have also shown their values in improving Revenue Management practices in various contexts. In the airline context, Vulcano et al. (2010) show that, by accounting for customer choice behavior in inventory control optimization, airline can improve their average revenue by 1 to 5% from the current ESMR-b (Expected Marginal Seat Revenue, Version b) policy. Further, accounting for dynamic consumer choices can improve revenue in certain types of city-pair markets by 3 to 5% as shown by Li et al. (2011). In the hotel context, Bodea et al. (2009) collect data from five U.S. properties of a major hotel chain and illustrate how choice-based Revenue Management can be used with real data. Anderson and Xie (2011) estimate a nested logit model on data from firms selling hotel rooms through an opaque channel, and they optimize firms' dynamic pricing decisions with the knowledge of customer choice behavior. Using hotel transaction data supplemented with user-generated content from social media, Ghose et al. (2012) optimize the hotel ranking system based on estimated customer valuation of hotel stays.

We share the spirit of these papers in adopting a customer-centric approach, but focus on a different type of Revenue Management problem — determination of competition sets. This differs from the more common Revenue Management focus on inventory control and dynamic pricing problems, but it is an important yet less-understood component of Revenue Management practices. Competitive considerations interact with Revenue Management decisions through both initial strategic positioning (e.g., how to design products based on existing competitive products?) and daily pricing decisions (e.g., who to match price with? When and how to react to competitors' price changes?). Before prescribing how should revenue management decisions responde to competitors' decisions, we take a small yet important step to understand the current competitive status: with whom firms truly compete for demand, and do hotel manager have good judgement about competition. To answer these questions, we look at pre-purchase behaviors using search and clickstream

data, which is also different from what is commonly used (transaction data) in empirical studies mentioned above.

What we propose is also an innovative approach towards understanding market competition structure. In many industries, firms produce differentiated products and compete in markets that are limited in extent. Theoretical models of localized competition originates from one-dimensional spatial model, including linear (Hotelling 1929), circular (Salop 1979) and vertical (Gabszewicz and Thisse 1979), where firms only compete directly with its two nearest neighbors, and were later extended to allow for differentiations along multiple dimensions. Empirically, localized competition has seen most applications in settings such as the gasoline market (Pinkse et al. 2002) and the fast-food industry (Allon et al. 2011). Competition structures in these settings are less complex compared to the hotel industry, as gas and fast-food are to some extent homogeneous products and the main differentiation is location. Consequently, competition is frequently characterized through location-based measures such as Euclidian-distance (e.g., Allon et al. 2011, Thomadsen 2007) or common market boundary measure (e.g., Pinkse et al. 2002, Feenstra and Levinsohn 1995). However, in the hotel industry, competition structure depends not only on locations but more importantly on various service attributes.

Most empirical models measures competition using firm-level data, e.g., number of competitors in the local market (Olivares and Cachon 2009), entries and exits by competitors (Buell et al. 2011), and Herfindahl-Hirschman Index (Borenstein and Rose 1994). Alternative, empirical studies of differentiated products are cast in discrete choice model framework (McFadden 1984, Berry et al. 1995) using transaction data. Such model either estimate a full set of cross-price elasticities assuming all products compete with all others, or a local spatial model with a pre-specified competition structure based on location measures. However, these approaches are not applicable to a highly differentiated industry such as the hotel industry with numerous competitors. On one hand, it is not efficient to estimate a full set of cross-price elasticities on hundreds of hotels. On the other, location based spatial models cannot adequately characterize multi-dimensional competition in the hotel industry.

Consequently, we propose an alternative methodology taking advantage of the availability of a new type of data to capture localized competition structure. Using consumer search and clickstream data, we provide more direct and straightforward competition measures. Note that similar type of data has been used in Marketing to understand the market structure and the existence of sub-markets, i.e., customer consideration set data solicited using customer surveys. See Shocker et al. (1991) for a review. One important finding by Urban et al. (1984) is that consideration set alone already contains much competitive information as compared to more extensive data such as customer preference list. We are similar in that we also use customer comparison set to understand

market structure. Yet, we differ from this line of research in at least two aspects: 1) we aim to picture networked competition rather than just dividing a market into sub-markets; and 2) our results are based on objective data of customer consideration sets, whereas other data are usually obtained from focus groups or customer surveys.

In this paper, we visualize a networked competition structure. This attempt is a part of the burgeoning efforts to modeling economic markets as networks rather than fully connected markets, e.g., buyer-seller networks described by Kranton and Minehart (2001). While majority of these papers are theoretical models, we provide empirical characterizations of one particular economic network, i.e., competition network. In this spirit, our work is also associated with several other streams of network-related literature. Customer associative networks have been used to understand brand preferences and associations. A recent work by Netzer et al. (2012) analyze how often consumers mention two products from online user-generated content to map out the competitive market structure. Our work is similar in terms of the research goal and methodology, but it is derived from a different type of data; that is, data that captures real actions (customer views of a product page), rather than data that measures human memory and perceptions. In addition, we not only construct the competition network from the customer perspective, but also compare it against the competition network from the hotelier perspective to examine the degree of network mismatch. Although network misalignment has also been studied in other settings such as product architecture and organizational structure by Sosa et al. (2004) and Gokpinar et al. (2010), to our knowledge our paper is the first to characterize it in hotel competition.

In addition, the present work contributes to the empirical literature stemming from the increasing availability of online clickstream data (e.g., Moe and Fader 2004, Park and Fader 2004). Most papers in this stream focus on understanding customer browsing behavior and predicting clickthrough rates and conversion rates. No studies of which we are aware use the data to understand market competition structures. We contribute by highlighting one other important direction that such data can be used.

## 3. Data and Industry Overview

### 3.1. Data

We combine three data sets associated with the online search for hotels sponsored by a major OTA. The first data set, search and transaction data, contains complete histories of all product searches at the sponsoring OTA's website conducted by approximately 4,000 cookie-based users in Manhattan, New York during the first two weeks of October 2009. The travel dates actually span October 2009 to September 2010. Among these, 90.4% of requests are for dates within three months, and 96.2% are for dates within six months. This data set includes information regarding what users searched

for, what searching criteria were specified, which hotels were returned in response (with room availability, price and promotions, and reviews), which hotels visitors clicked to see details and which hotels were booked, if any. Search and transaction data are tracked internally by the OTA. The second data set, clickstream data, contains the complete clickstream history of all actions by the users identified in the search and transactions data. The clickstream dataset contains usage of all web-properties belonging to the sponsoring OTA for the same users included in the transaction data from January 2009 to October 2010. Each record describes a clickstream event such as viewing a particular webpage or submitting a form. This dataset is tracked by a third-party web analytics firm. The third dataset, hotel data, contains information associated with hotels appearing on the OTA's website, such as brand and chain name, ownership type, room capacity, star level, and so on.

In both search/transaction data and clickstream data, a unique ID is used to identify each cookie-based web user. The two datasets can be matched based on user ID and event time. Note that certain information in these two datasets overlaps – such as a click on a hotel displayed in a search result. Meanwhile, each dataset contains unique information. Since clickstream data only record clicks, those hotels displayed in search results but not clicked are only recorded by search/transaction data. In search/transaction data, we see everything that a user sees in a search result – hotels displayed and their associated prices, promotions and customer reviews. This hotel-specific information is not recorded by clickstream data. However, since clickstream data record all user *actions*, hotel page visits not resulting from standard search requests are also recorded. A visit to a specific hotel's webpage does not necessarily originate from a direct search on the OTA's website, but can also originate from external sources and other internal pages of the OTA's website. Examples of external sources include search engines such as Google or Yahoo!, other online booking sites, and email promotions. Examples of internal pages include a page promoting recent deals, flight-reservation or car-rental pages, and other hotel pages linking to similar hotels in the same destination. In fact, clicks which do not originate from direct search requests on the OTA's website account for 69.8% of all page visits. In summary, access to clickstream data gives us a more comprehensive view of a customer's actual consideration set. We could have used the 20-month clickstream data to obtain competition sets. However, in order to match the clickstream data with the search and transaction data to control for the effect of the ranking system and for heterogeneity in choice probabilities, we restrict our attention to the period in which the two datasets overlap, i.e., the first two weeks of October 2009.

As shown in Table 3, our data contains 3,514 users who searched for hotel stays in New York and vicinity during the first two weeks of October 2009. This covers 309 hotels in Manhattan[2]. In

[2] New York City and vicinity includes 255 hotels outside Manhattan (e.g., in New Jersey or Long Island). We restrict our attention to hotels in Manhattan. These hotels receive on average six times more page views than hotels outside

total, these hotels received 22,901 page views in our data, or 74.1 page views per hotel on average. Naturally, hotels may be viewed multiple times by the same user. 37.5% of users did not view any specific hotel page at all, The rest viewed 10.4 pages on average per user, or 3.5 distinct hotels. Note that a user may leave the site and return later in the day and conduct the same search again. This second search is unlikely to be considered a new search but rather an extension of the previous search. As a result, a user may visit a hotel page multiple times during a short period of time. Now a question arises — how should we define a session such that multiple visits to the same hotel during the session are considered as one visit? One attempt is to define a session based on calendar date. However, a disadvantage to this approach is that this can divide a series of actions that occur near a cut-off time, say mid-night, into two sessions. Instead, we define restarting a new search session using an inaction period of 24 hours[3]. Prices, promotions, customer reviews and hence ranks of the same hotel may change the next day. As such, the same hotel would essentially be a different hotel for a new search after 24 hours. Thus, for our purposes, multiple visits to a hotel page within a search session are counted as one visit. One can potentially weight the visits with frequencies of visits or even length of stay on the page. We consider this information secondary in this application. In sum, we consolidate 22,901 page views to 7,764 distinct page visits.

In our data, we observe price variations mostly along two dimensions: 1) day-of-travel and 2) day-of-observation. Variation along day-of-travel is mostly based on segmentation of customer types. For example, business travelers tend to stay on weekdays, while leisure travelers tend to stay on weekends. Hotels targeting different segments demonstrate different pricing strategies based on different days of the week. Variation along day-of-observation is mostly due to inventory-based Revenue Management models – how to price based on days in advance and how to dynamically change rates given remaining room inventory. An important input for price change along both dimensions is how competitors adjust price along the same two dimensions. Later, we will estimate a pricing model to examine this aspect of pricing behavior.

### 3.2. Hotel Industry Background

The hotel industry is characterized by intense competition with various types of branded and non-branded properties competing for business. Hoteliers make continuous efforts to differentiate their services from others, but they are still faced with fierce price competition. As prices becoming more and more transparent with Internet search engines, competition among hoteliers increases. In our data, 309 hotels compete for business in Manhattan, New York, offering a total of 68,584 rooms.

of Manhattan. The unique geography of Manhattan makes it convenient for our purposes of constructing competition networks without making them too sparse.

[3] We also use 8 hours and 48 hours as robustness checks and the results are very similar.

Among them, 59.2% are independent properties, representing 41.4% of room capacity. Among the branded properties, a few major chains (i.e., Hilton Worldwide, Choice, Marriott, IHG, Starwood) operate 77 properties under 32 brand names ranging from two to five stars. Competition exists not only between chains but also within a chain, since hotels in the same chain are typically operated by different individuals or management firms.

**Price Patterns.** The hotel industry exhibits significant price variation across different properties, room types, customer types, days of stay and days of booking. Our sample of 309 hotels in Manhattan offers an average room rate of $295 with a standard deviation of $200. Part of the variation is manifested through frequent promotions — an average hotel is on promotion 41% of the time in this period. Hotels sometimes reach full capacity — 2.60% on average. During the past 30 years, the hotel industry has adopted revenue management tools successfully. Rooms of a hotel are usually classified into several room types, which are further grouped into a few (typically 10 to 12) rate bands together with customer types for revenue management purposes. Rates typically start off from rack rates (which reflects no discounts; also called BAR rates) and decrease from there. Discounted rates are characterized by percentage discounts off rack rates. As room inventory depletes when the travel date approaches, the revenue management system suggests opening or closing certain rate bands. Revenue managers also retain the right to override the system when needed.

Such needs usually arise due to observing changes in competitors' prices – an important input for pricing decisions in this industry. As a matter of fact, monitoring competitors' prices is part of everyday operations in the industry. "Call-around" is a common practice whereby

> Hotels engage in regular communications, typically by telephone with the hotels on their Call-Around Lists, two or three times daily, to exchange with such hotels: (i) each hotel's non-public current occupancy rate (generally expressed as a percentage of hotel rooms occupied) and (ii) the standard rate currently being charged for hotel rooms to be occupied that same day (generally expressed as the 'BAR rate' or the 'rack rate', which would not include any available discount rates)... Revenue Managers have the ability to manually override the preset grid and/or computer reservation system to adjust the applicable room rate.... In determining whether to manually override the reservation system, the Regional Revenue Manager may periodically consult various sources of information concerning competitor rates, including publicly listed rates through internet sites or other market information[4].

[4] An Agreement By and Among the Attorney General of the State of Connecticut, LQ Management L.L.C., and La Quinta Franchising, L.L.C. March, 2010

In addition to the "call-around" practice, hoteliers subscribe to automated tools, such as MarketVision, PriceTrack, and RateVIEW, to monitor competitors' rates closely, . These tools help hoteliers shop rates and availabilities of hotels in their pre-specified competition set (Cross et al. 2009). In addition, development of the online search engines exposes a wealth of readily accessible price information not only to customers but also to competitors.

**Online booking channels.** Transient business and leisure markets are captured increasingly through online reservations. 81.5% of consumers perform travel-related searches, and 70% visit a travel-related site prior to booking at the supplier's website (Withiam 2011). OTAs currently capture a large market share of online hotel reservations. Based on our analysis of a random sample of ComScore data in 2009, major OTAs — Expedia.com, Hotel.com, Hotwire, Priceline, Travelocity and Orbitz in descending order of market share— contribute 43.7% of all online observations. This percentage is even higher when we consider eye-visit market share. For example, 10.5% of customers who booked hotels online made their reservations through Expedia.com; meanwhile, 29.8% of customers visited Expedia.com before purchasing.

In an era of great price transparency enabled by online search engines, price matching with the wrong competitors has the potential to generate sizable revenue losses because any pricing error is much more visible. Customers are quicker to punish misaligned prices by booking away from hotels that have rates too high or pouncing on rates that are below market. This new reality calls for an accurate definition of competition set and a properly designed data-driven analytical approach that will allow hotel managers to benchmark prices with the right set of competitors and take the proper actions to observed price changes in the marketplace.

## 4. Methodology

In this section, we first discuss the methodology used to measure competition intensity from the customer perspective using search and clickstream data. Then, we discuss the methodology used to measure competition from the hotelier perspective based on price matching patterns. These two types of measures will later lead to the construction of two competition networks and an examination of network mismatch.

### 4.1. Customer-Based Measure of Competition

To understand which hotels compete with other hotels from the customer perspective, we use clickstream data to analyze the hotels customers have compared before making their final choice; in other words, which hotels have competed for each customer's demand. This competition pattern, as noted above, is likely to be affected by the OTA's sorting algorithms. Depending on how hotels are ranked in the results, one OTA's particular sorting algorithm may alter the consumer's click behavior and hence which hotels are compared. There are two potential ways that an OTA's ranking

result is likely to affect competition. First, customers may interpret hotels listed close to each other as similar and hence click these hotels together. For example, if customers click down the list blindly from rank one to three without careful thought, then whoever is ranked in the first three places will be "competitors". This reveals the OTA's view of competition but not necessarily the customers' view. We test this adjacency hypothesis using rank distance. Based on the rank distribution of clicked hotels in the data, we estimate the expected rank distance under no adjacency effect is 4.212 with a standard deviation of 5.889. The observed average rank distance is actually 7.119 with a standard deviation of 5.789. Obviously, no tendency exists in which consumers tend to click hotels with adjacent ranks. Second, it is well established that higher ranks induce higher click-through rates. Thus, a hotel that is ranked higher by this OTA is more likely to be clicked and compared with other hotels. We illustrate the potential bias with the following example.

**Illustrative example.** Suppose there are three hotels: $A$, $B$ and $C$, and we observe that hotel $A$ is clicked 200 times, $B$ 500 times, and $C$ 200 times. We also observe that $A$ and $B$ are clicked together (i.e., compared) 100 times, while $A$ and $C$ are only compared 50 times. Do we conclude that competition between $A$ and $B$ is more intense than that between $A$ and $C$? A simple 2 by 2 contingency table answers this question.

**Table 1 Example: 2 by 2 Contingency Table**

| | $B$ | $\bar{B}$ | **total** | | $C$ | $\bar{C}$ | **total** |
|---|---|---|---|---|---|---|---|
| $A$ | 100 | 100 | 200 | $A$ | 50 | 150 | 200 |
| $\bar{A}$ | 400 | 400 | 800 | $\bar{A}$ | 150 | 650 | 800 |
| **total** | 500 | 500 | 1000 | **total** | 200 | 800 | 1000 |
| | | | | | | | |
| **Chi-square** | 0 | | | **Chi-square** | 3.9063 | | |
| **P-value** | 1 | | | **P-value** | 0.0481 | | |

*Note: $A$ represents $A$ is clicked; $\bar{A}$ represents $A$ is not clicked.*

We observe that hotel A is compared to hotel B twice as often as it is compared to Hotel C. However, the chi-square test shows that A and B are two independent events, while A and C have a significantly positive dependence. If clicking of different hotels are completely random events, we expect to observe a frequency of 100 comparisons of A and B. However, we expect to see only 40 comparisons of A and C, which is smaller than what we actually observe, i.e., 50 comparisons.

From this simple example it is clear that we need to account for the probability of each hotel being clicked when considering the extent to which two hotels compete for demand. This is especially important when we consider potential effects of online ranking systems. Higher ranks usually lead to higher clickthough rates. In our data, the probability of clicking a hotel descends almost monotonically as one moves down along the OTA suggested ranks displayed on the page, and the

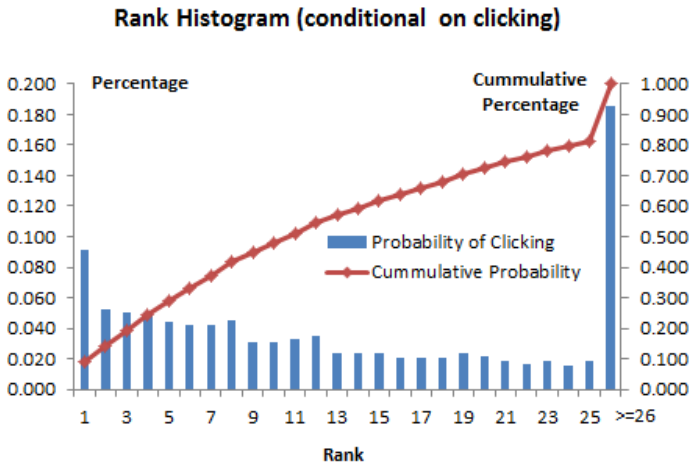


**Figure 1 Histogram of Rank conditional on Clicking**

top 10 hotels receive about half (47.7%) of all click-throughs (see Figure 1). Since higher-ranked hotels tend to receive more clicks, if we do not adjust for the probability of a hotel being clicked, we would come to a biased conclusion that there is more competition among top-ranked hotels than among lower-ranked hotels.

**Measures of co-location.** To test whether including hotel A and B into the same consideration set is a random event, we adopt four co-location measures that are widely used in both Linguistic (see Manning and Schutze 1999 for a general discussion) and Data Mining literature. The null hypothesis to be tested is that including hotel A in the consideration set is independent from including hotel B, i.e., $P(AB) = P(A)P(B)$. The four measures are 1) t statistics, 2) Chi-Square statistics , 3) Pointwise Mutual Information (PMI, also called "lift" in the Data Mining literature), and 4) and an adjusted version of PMI — normalized PMI.

The T-test uses normal approximation for the number of successes of a series of Bernoulli trials with probability $p = P(A)P(B)$ under the null hypothesis. The Chi-Square test implements the chi-square test for independence for the 2 by 2 contingency table as in Table 1. Unlike the T-test, the Chi-Square test does not rely on normal approximation to Binomial distribution.

Pointwise Mutual Information (proposed by Church and Hanks 1990) is defined as follows:

$$log\Big(\frac{P(AB)}{P(A)P(B)}\Big). \tag{1}$$

PMI is simply the ratio of A and B's actual joint probability over the expected joint probability if they were independent events. A PMI of 0 indicates that the two events are independent. A positive PMI means that the occurrence of the two events is positively correlated. One known limitation of this measure, however, is that PMI is particularly sensitive to low-frequency events (even though

all co-location measures are sensitive to low-frequency data to some extent). To see why, assume perfect dependence between clicking two hotels A and B, so we have

$$log\Big(\frac{P(AB)}{P(A)P(B)}\Big) = log\Big(\frac{1}{P(B)}\Big). \tag{2}$$

A low probability of $P(B)$ indicates a large PMI value in this case. To mitigate its impact, we introduce another measure, normalized PMI, which is defined as follows (Bouma 2007):

$$log\Big(\frac{P(AB)}{P(A)P(B)}\Big)/-log(P(AB)). \tag{3}$$

This measure is proposed to overcome two shortcomings of PMI: 1) its sensitivity to low-frequency data; and 2) the lack of a fixed upper bound. Normalized PMI is essentially a PMI scaled by joint probability, which gives lower weights to low-frequency events. When two events are completely dependent, the above measure equals the upper bound 1.

**4.1.1. Aggregation Bias.** Consider the following example of clicking probabilities which are dependent on ranks. Suppose clicking hotels A, B, and C are independent events. The probability of one hotel being clicked is completely determined by its rank – 0.7 if ranked first, 0.5 if ranked second, and 0.3 if ranked last. Also suppose all six orders appear with equal probabilities.

**Table 2 Example: Aggregation Bias**

| rank | order 1 | order 2 | order 3 | order 4 | order 5 | order 6 | Pr(click) |
|---|---|---|---|---|---|---|---|
| 1 | A | A | B | B | C | C | 0.7 |
| 2 | B | C | A | C | A | B | 0.5 |
| 3 | C | B | C | A | B | A | 0.3 |

In this example, we can calculate the following probabilities:

$$\begin{aligned} P(AB) &= \frac{1}{6}(0.35+0.21+0.35+0.21+0.15+0.15) = 0.237, \\ P(A) &= P(B) = \frac{1}{3}(0.7+0.5+0.3) = 0.5, \\ P(AB) &\neq P(A)P(B). \end{aligned}$$

If the probability of an event occurring varies with different contexts (such as different orders of results), aggregating probabilities may induce bias. The following equation shows why it is the case in general. Under the null hypothesis of independence,

$$\begin{aligned} P(AB) &= \frac{1}{N}\sum_{i=1}^{N} P_i(AB) = \frac{1}{N}\sum_{i=1}^{N} P_i(A)P_i(B) \\ &\neq (\frac{1}{N}\sum_{i=1}^{N} P_i(A))(\frac{1}{N}\sum_{i=1}^{N} P_i(B)) = P(A)P(B). \end{aligned}$$

In different types of searches and on different observation days, the same hotel may appear at different positions in the search result. If we do not account for this possibility, the co-location measure will be subject to aggregation bias.

Suppose we know the clicking probability $P_i(A), P_i(B), i = 1, ..., n$, it is then straightforward to adjust for this potential aggregation bias. Simply substitute $P(A)P(B)$ by $\frac{1}{N}\sum_{i=1}^{N} P_i(A)P_i(B)$. We hereafter make a special note of the validity of using t-statistics in this case: under the null hypothesis of independence, we expect to observe a sequence of independent but *non-identical* Bernoulli trials with success probabilities: $p_i = P_i(A)P_i(B), i = 1, ..., N$. The total number of successes then follows a Poisson Binomial Distribution, which can also be approximated by a Normal distribution with mean $\sum_i p_i$ and variance $\sum_i p_i(1-p_i)$. Thus, we can still use a T-test of independence. Consequently, $P(AB)$ can be approximated by a normal distribution with mean $\frac{1}{N}\sum_i p_i$ and variance $\frac{1}{N^2}\sum_i p_i(1-p_i)$. The remaining question is: in t-statistics, can we still use $\frac{P(AB)(1-P(AB))}{N}$ to estimate the sample variance? Even if the null hypothesis holds, in general, $\frac{P(AB)(1-P(AB)}{N} \neq \frac{1}{N^2}\sum_i p_i(1-p_i)$. However, when $p_i$'s are small, we have $1-p_i \approx 1$. Then, we have $\frac{P(AB)(1-P(AB))}{N} \approx \frac{P(AB)}{N} = \frac{1}{N^2}\sum_i p_i \approx \frac{1}{N^2}\sum_i p_i(1-p_i)$.

Now, consider that we actually do not know $P_i(A), P_i(B), i = 1, ..., N$? These probabilities can be estimated using logistic regression. For each displayed hotel, a consumer decides to either click or not click[5], and this action is a function of both the time-variant and time-invariant characteristics of the displayed hotel: rank, room rate, promotion status, customer review, star rating and location. Other information such as amenities, transportation, and complementary services do not affect the decision of clicking as they are not part of the customer's information set yet.

One shortcoming using logistic regression is its underlying assumption of independent error terms and the Independence of Irrelevant Alternatives (IIA) property. In reality, a consumer's clicking decision can be correlated across hotels. To account for this heterogeneity in consumers' tastes and to overcome the IIA problem, we use the Random Coefficient Logit (or Mixed Logit) model (Train 2003). We predict the clicking probability $P_i(A), P_i(B), i = 1, ..., N$ based on this model while allowing for random coefficients,

$$\begin{aligned} u_{ijk} &= \beta_i X_{jk} + \epsilon_{ijk}, \beta_i \sim N(\beta_0, \Sigma), \\ u_{ijk} &= \beta_0 X_{jk} + \xi_{ijk}, \text{where } \xi_{ijk} = (\beta_i - \beta_0)X_{jk} + \epsilon_{ijk}, \end{aligned}$$

where $i$ represents an individual, $j$ represents an alternative, i.e., a hotel, and $k$ represents a search occasion. An example of a search occasion is a search request made on October 1, 2010 for a

[5] Suppose that the total number of hotels displayed is exogenously given. In reality, it is chosen endogenously, as consumers do choose whether or not to continue on to the next page or result. Should they choose to continue, more hotels will be displayed. However, in our data, almost all customers (96.4%) stayed on the first page of results and did not scroll down.

two-night stay in New York City with check-in dates on November 1, 2010. $X_{jk}$ represents the characteristics of hotel $j$ at search occasion $k$. For example, \$295 per night wws quoted for the above request. $\beta_i$ is a vector of random coefficients. $\epsilon_{ijk}$ are i.i.d. shocks across consumers, hotels and search occasions. This means that even though, at the personal level, the clicking of hotels $A$ and $B$ is independent as $\epsilon_{iAk}$ and $\epsilon_{iBk}$ are independent, the aggregate choice of the two hotels can be dependent due to heterogeneity in customers' tastes, that is, $\xi_{iAk}$ and $\xi_{iBk}$ are correlated within the same person.

### 4.2. Hotelier-Based Measure of Competition

Competitors' price is a key component when a hotel determines its own price. Each property has a list of hotels which they consider as competitors. Traditionally, they use such a list in Call-Around practice to solicit information on price and occupancy rates from competitors. Now with the proliferation of online booking tools, price becomes even more transparent. Revenue managers no longer need to call several times a day to discover competitors' prices and availability as they are just a click away. Any change in a hotel's price, such as a newly launched promotion or changes in room availability, can be immediately spotted. Such information may invoke price changes from competitors should they decide to react. As a result of price-matching practice, the prices of competing hotels will be highly correlated. We will use this as the basis for identifying competitors from the hotelier's point-of-view.

Based on theories of product differentiation, if firms can differentiate their services or products through dimensions such as quality or location, they do not need to engage in fierce price competition. Thus, a lack of price differentiation among firms indicates a lack of differentiation among other aspects, or in other words, firms perceive themselves offering very similar products. In the hotel industry in particular, 70% of variations in price, measured usually by Average Daily Rate (ADR), can be explained by a hedonic model of product characteristics (e.g., Thrane 2007), similar to housing markets. This is to say, price levels in this industry are a symbol of how similar hoteliers think their services are. Kim and Canina (2009) exploit ADR clustering patterns to identify hotel competition sets.

Rather than relying on an aggregated measure of price (ADR), one advantage of our data is that it records prices quoted for each travel occasion. Hence, for each pair of hotels, we observe a series of price pairs, which allows us to track how hotels match prices dynamically and measure the level of price parallelism using micro-level data. We use the correlation of two price series $\{p_{jk}, j = A, B, k = 1, 2, ..., K\}$ to capture competitiveness between hotel A and B from hotelier perspective. $j$ corresponds to hotels, and $k$ corresponds to search occasions. For example, Hotel A may charge \$299 per night for a two-night stay request made on a particular date for a particular

travel starting date, with or without specifications on the number of adults, number of children and type of room. Hotel B may charge $289 for the exact same request. Note that the availability of search data allows us to obtain price correlation down to the level of search requests, while most other analysis using hotel prices are based on aggregated ADR. This benefit also comes with a cost. Not all hotels are seen by a customer, due to his particular specifications of the travel request or his inaction beyond the first page of results. As a consequence, we are faced with a scarce data problem where only a few price pairs are observed for some pairs of hotels if they are rarely shown in the same search. However, the pairs for which we do not have much data for are also likely to be those pairs that adopt different price levels and hence are less likely to be competitors. For hotels that charge competitive rates, we have thousands of price pairs to calculate price correlation. Further, as price quotes are search-specific and not customer-specific, we can group the same search made by multiple customers at adjacent time, such as within the same searching day, to obtain more observations of price pairs.

**4.2.1. Decompose Price Matching.** To better understand how hotels match prices, we decompose price correlation into two components according to the following pricing model:

$$p_{jk} = \alpha_j + \beta_j X_{jk} + \epsilon_{jk}, \tag{4}$$

$\alpha_j$ is a hotel-specific intercept which represents the average price level of hotel $j$. It includes how a hotel sets prices based on its static characteristics such as location, star rating, and brand. $\beta_j$ estimates a hotel-specific factor of how hotels price characteristics of each travel request, such as length of stay, day-of-week of travel dates, days of advanced purchase, number of adults and children, etc. We suspect that price matching mostly happens along two dimensions – travel dates and booking dates, as these are the main variations of different travel requests and the two most important identifiers of travel type (i.e., business vs. leisure). A comparison of how similar $\beta_j$'s are will inform us about the level of price matching between hotel pairs. The error component, $\epsilon_{jk}$, represents idiosyncratic shocks to price, such as unpredicted changes in demand and unpredicted adjustments in competitors' prices. Applying the above model to each hotel, we are able to obtain a predicted price series $\{\hat{p}_{jk}, k = 1, 2, ..., K\}$ and residual price series $\{p_{jk} - \hat{p}_{jk}, k = 1, 2, ..., K\}$. Similar to the correlation of the original price series, we obtain correlations of predicted price series and residual price series for each hotel pair. The correlation of predicted prices is mainly based on price matching along travel dates and booking dates. The correlation of price residuals is mostly based on price matching of common demand shocks and completive matching.

One concern of using price correlation as a measure of price matching is that two price series can go in parallel with each other, but with a large discrepancy (for example, one around $100 and the

other around $500). This means that it is probably the overall demand trend that is driving the correlation rather than the hotel trying to match prices. To address this concern, we define another measure — Normalized Average Price Difference between two price series $\frac{\frac{1}{K}\sum_{k=1}^{K}|p_{1k}-p_{2k}|}{\frac{1}{2K}\sum_{k=1}^{K}(p_{1k}+p_{2k})}$. Under perfect price matching, the measure should be zero. A larger value means less price matching and less competition.

# 5. Empirical Results

## 5.1. Consumer-based Competition Measure

We compare four measures of consumer-based competition: t-statistics, Chi-Square statistics, PMI, and NPMI. We then select the most appropriate one for our application. As shown in many other cases of co-location analysis, PMI is particularly sensitive to low-frequency data. Normalized PMI is proposed to address this issue. We observe that Normalized PMI indeed adjusts this sensitivity, but only to a certain extent, as shown in Table 4 listed in descending order of NPMI. Take the first and forth rows of Table 4, for example, where both PMI and NPMI favor the low-frequency event of hotel "John Street Suites" which has been clicked only once during our period of study but was compared with "Hotel Belleclaire" at that time. The Chi-Square statistic further adjusts sensitivity to low-frequency data. Within the above example, the Chi-Square favors "Grand Union" with 36 clicks and 10 comparisons over the small-probability event of clicking "John Street Suites". However, comparing the first row and the twelfth row, the Chi-Square favors the low frequency event in this case. The t-statistic is the least sensitive to low-frequency data among all four measures. It is also one of the three measures (PMI, NPMI, and t-statistics) that can be consistently implemented when we account for heterogeneity and observe independent but *non-identical* Bernoulli trials. Due to these reasons, we focus on t-statistics as our main measure of competition[6]. The shortcoming of t-statistics is its normal approximation for Binomial distribution. A commonly applied rule of thumb is to restrict attention to cases in which $Np >= 5$. We note that, when we only choose t-statistics greater than 1.96 (i.e., $p-value=0.05$), this condition is satisfied 97.0% of time.

Based on t-statistics, the top 10 and bottom 10 pairs of competitors are listed in Table 5. On the top of this list is Park Central New York Hotel, which has been viewed 164 times, and Paramount Hotel Times Square New York, which has been viewed 93 times, and the pair have been compared 21 times. At the bottom of the list is Empire Hotel and Doubletree Guest Suites Times Square NYC. Each of them has been viewed multiple times, 70 and 191, respectively. However, they are compared only once[7].

[6] We also applied NPMI with a cut-off value where the event has to occur at least 3 times, suggested by (Manning and Schutze 1999). Results are consistent.

[7] For the purpose of illustration, we only list those pairs of hotels which were compared at least once. Many other pairs have no comparisons at all.

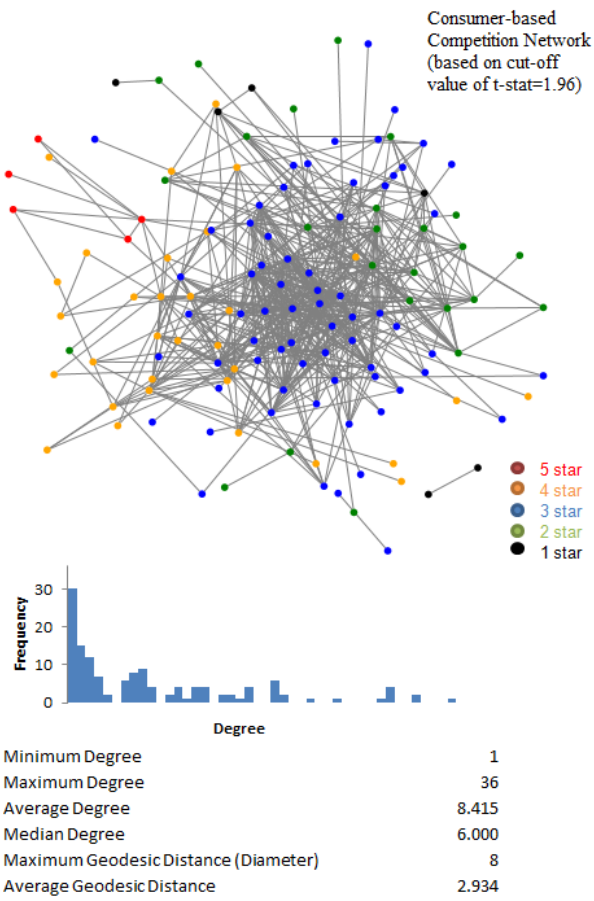


**Figure 2 Visualize Customer-based Competition Network (t-stat$>1.96$)**

Using a t-statistics cut-off value of 1.96, we build an undirected competition network of Manhattan hotels in Figure 2. As we expected, there is a strong clustering pattern within hotels of the same star level, although hotels also occasionally compete with hotels of adjacent star levels. The degree distribution of the network has a mean of 8.42, median of 6.00, and minimum of 1 and maximum of 36. That is, on average a hotel has 6 to 8 competitors. This is also consistent with a survey conducted by Hotel Compete, where they report that the average size of Benchmark Competition Set is 6.87 to 7.83 by directly asking 2,833 hotels to name their competitors.[8]. The average geodesic distance in this graph is 2.93, with a maximum (i.e., diameter) of 8. That is, on average, a hotel can be linked to another hotel through three steps of competitive relationships, or two competitors in between. This small degree of separation has interesting implications, as it means that shocks to a local hotel (such as promotions resulting from demand or supply shifts) may spread quickly to other parts of the network through price matching.

One caveat of using a universal cut-off is that some hotels with only weak competition links (usually those which do not appear often and are not compared often with others) may be left out

[8] "Hotel Comp Set Analysis – Untapped Opportunity #1: Market Dynamics." Hotel Compete. May 16, 2012.

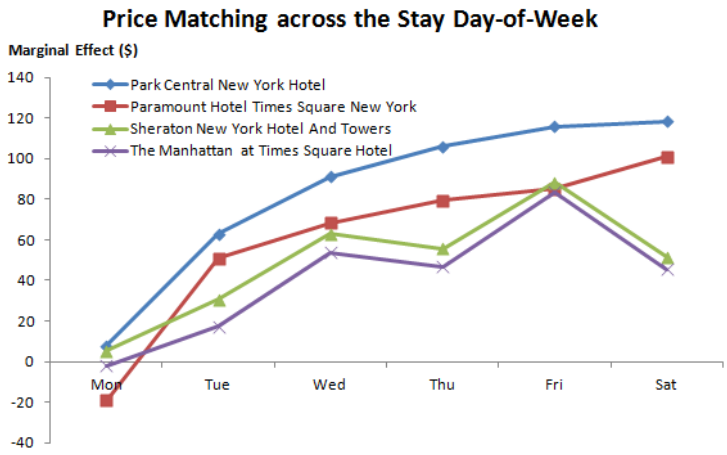


**Figure 3 Price Matching across Day of Week (Sunday is the benchmark)**

of this graph. If we move down the ladder of t-statistics and choose lower cut-off values, we would visualize more competitive links.

### 5.2. Evidence of Price Matching

To illustrate price matching among competitive hotels, we take the top two pairs of competitive hotels from Table 5 as an example. The basic information for each hotel is listed in the table below:

| | Hotel Name | Star | Brand | Location | Room Capacity |
|---|---|---|---|---|---|
| **Pair1** | **Park Central New York Hotel** | 3.5 | Independent | Midtown | 935 |
| | **Paramount Hotel Times Square New York** | 3 | Independent | Times Square | 567 |
| **Pair2** | **Sheraton New York Hotel And Towers** | 3.5 | Sheraton Starwood | Midtown | 1750 |
| | **The Manhattan at Times Square Hotel** | 3 | Starwood other brand | Midtown | 665 |

We analyze the pricing strategy of each hotel using Equation 4, the results of which are shown in Table 7. The graph reveals consistent competitive relationship as that revealed from consumer clickstream data. The estimated coefficients are very similar within each pair, but quite different across pairs. For example, days in advance have positive coefficients for both hotels in Pair 1 and negative coefficients for both hotels in Pair 2. The two hotels from Pair 1 increase prices linearly from Monday to Saturday, with the highest room rate on Saturday nights. Hotels of Pair 2 tend to charge the highest rate on Friday nights, with the second highest on Wednesday nights. These patterns can be seen clearly in Figures 3 and 4. It is obvious that Park Central and Paramount adopt similar pricing strategies, and that Sheraton and Manhattan adopt similar pricing strategies. This evidence suggests that hoteliers match prices with their competitors, which supports our approach of using similarity in price to represent who hotel managers identify as their close competitors.

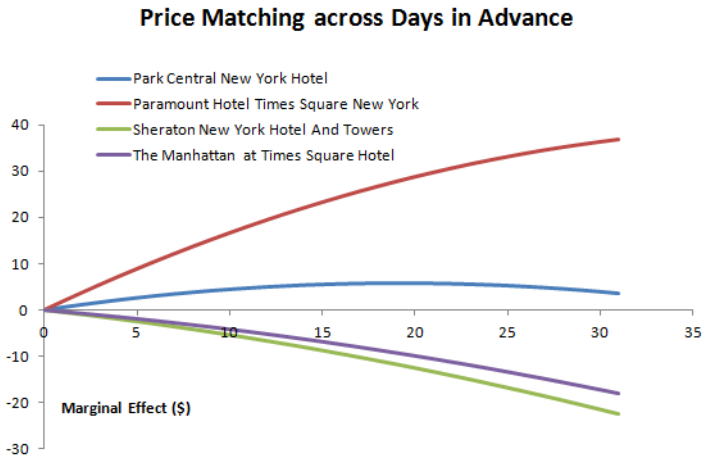


**Figure 4 Price Matching across Days in Advance**

Table 8 lists 4 measures of price matching for the top 10 and bottom 10 pairs displayed in Table 5. Generally, higher t-statistics are associated with higher price correlation scores and lower normalized average price difference, even though the order might not be exactly the same. The association is even stronger once we combine correlation and difference in a weighted manner to determine competitiveness.

### 5.3. Network Mismatch

After obtaining both consumer-based and hotelier-based competition measures, we are ready to evaluate the mismatch of the two networks. As some hotels, especially one- and two-star hotels, do not often appear to have a robust measure of price correlation, we restrict our attention to the sub-network of hotels with three stars or higher, which gives us around 193 hotels. There is not a sufficient number of comparisons or price pairs for some pairs of hotels (which most likely are not strong competitors in any way). We further restrict our attention to those pairs with sufficient number of comparisons from consumer data (i.e., greater than 3 comparisons) and sufficient number of price pairs (i.e, at least 30 price pairs). In this way, our networks are finally constructed on 89 hotels, which actually represent 64.1% of the total number of rooms offered collectively by hotels with a minimum star level of three.

We note that, when it comes to price matching, different hotels match prices on different levels. Some hotels tend to have a higher correlation with most other hotels, while some others have a lower price correlation with others in general. For example, five-star hotels usually do not engage in much price competition. Therefore, using a universal cut-off of price correlation may seem unfair. Instead, we identify up to top five competitors for each hotel based on the price correlations of this hotel with all other hotels. Similar logic also applies to consumer-based networks: we choose

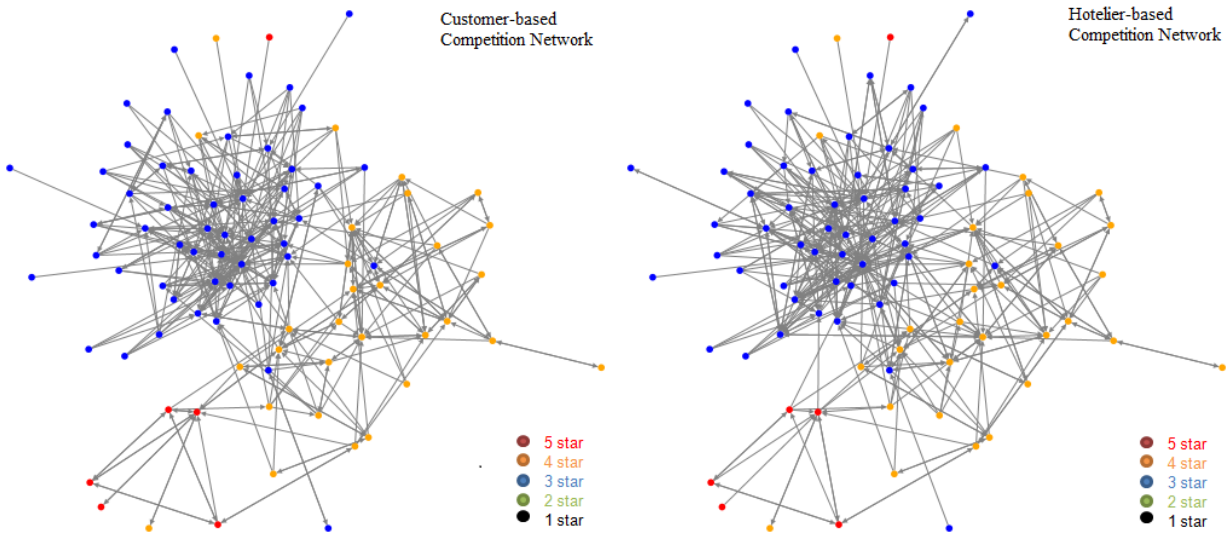


**Figure 5 Top 5 Competitors Networks of Hotels with 3 Stars and Up**

up to the top five hotels based on the t-statistics as competitors of the focal hotel[9]. In this way, we derive two directed networks, one each from the customer and the hotelier perspective, as shown in Figure 5. Directed links in this figure point from a hotel to its competitors.

First, we make a note on the average amount of overlap between the two networks. The amount of mismatch is visualized in Figure 6, and the numbers are shown in Table 9. Among 386 competitors identified by customer-based measures, 191 or 49.5% are also recognized by hoteliers as competitors. 160 of the 386 competitive links in the hotelier-based competition network are also reciprocal – that is, both hotels consider each other among the top five competitors. A similar number of links, i.e., 172, are reciprocal in the customer-based competition network. To further examine when mismatch is likely to occur, we predict chances of true competitors being left out or wrong competitors being included using characteristics of hotels relative to the hotel in consideration.

Results of when mismatch is likely to occur are shown in Table 10. The coefficients in Column 1 are the marginal effects obtained from a logit model for predicting a competitor being left out from competition sets by hoteliers *conditional* on customers considering it as a competitor. The coefficients in Column 2 are the marginal effects obtained from a logit model for predicting a competitor being included in competition sets by hoteliers *conditional* on it is not a competitor from the customer perspective. We find that independent hotels are likely to be left out by hoteliers from their competition sets, and branded hotels are often wrongly included in the competition set. Reading from Column 1 and 2, an independent hotel has a 20% higher chance of being left out of a competition set than a branded hotel, and a branded hotel has a 10.5% higher chance of being

[9] Top five is a conservative choice, as the average number of competitors of a hotel is usually 6 to 8. We also tried the top ten competitors, and the results are similar.

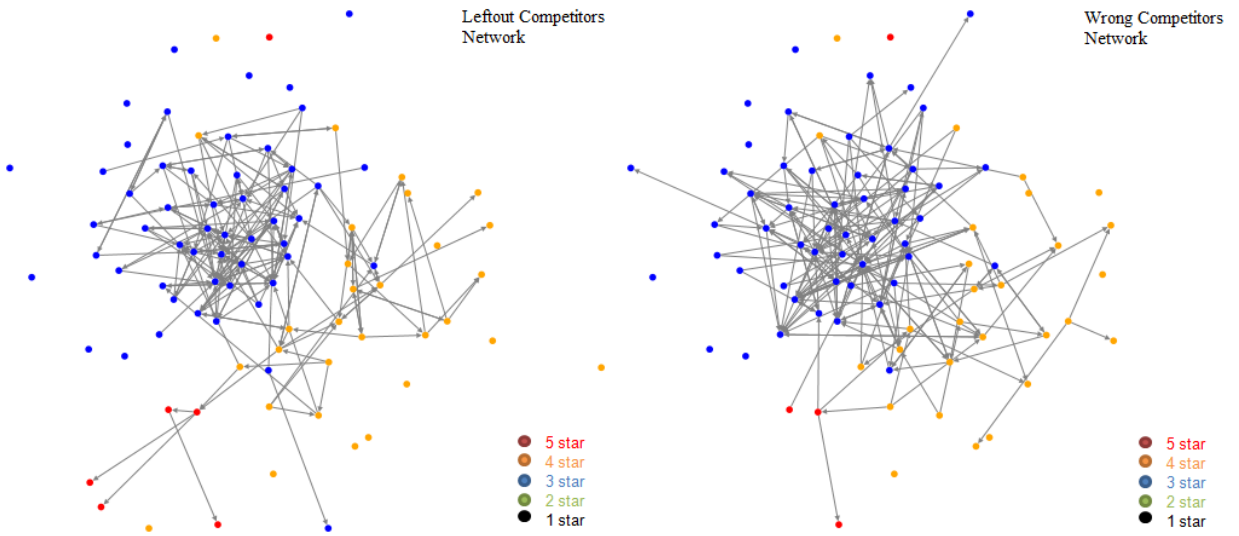


**Figure 6 Network Mismatch**

wrongly included in a competition set. In forming their competition sets, hoteliers also have a bias for hotels in the same district[10]. A hotel in a different district has a 10.9% higher chance of being left out from a competition set[11]. Hotels with lower traveler reviews are likely to be left out from a competition set as well. It is possible that hotels set up a high benchmark or mistakenly think they are equivalent to higher rated hotels. Finally, we also notice that hotels tend to benchmark themselves with other hotels with lower star ratings, lower price levels and lower ranks.

To check the robustness of mismatch patterns discussed above, we use alternative measures of customer-based and hotelier-based competition. For customer-based competition, we adjust the aggregation bias using Logit and Mixed Logit models. The estimation results of the two logit models are displayed in Table 11. Mismatch patterns based on t-statistics resulting from the Mixed Logit model are shown in the third and fourth columns of Table 9. For the hotelier-based competition measure, we present results based on the normalized average price difference in columns 5 and 6. We first note that the two alternative measures yield similar levels of overall overlapping between the two networks. The amounts of overlap are 45.1% and 48.4% respectively, as compared to 49.5% previously. Second, these results confirm the findings with regard to when mismatch is likely to occur, though with minor changes in significance levels in some cases.

[10] Districts are defined according to the OTA's definition as follows: Broadway – Times Square, Upper East Side, Upper West Side, Financial District / Downtown, Grand Central, Lower, Lower East Side, Lower West Side, Mid-town East, Mid-town West, North Manhattan, and Uptown East.

[11] To test the location effect, we also used distance rather than the same district indicator. The signs are exactly the same but sometimes not statistically significant.

## 6. Conclusions and Discussions

In this paper, we propose a methodology to identify competition sets from the customer perspective. Our main contributions are threefold: 1) We develop a customer-centric, data-driven approach to tackle the question that challenges many businesses: who are we really competing with? 2) We provide a network view to understand market competition structure; and 3) We point out the potential of using online data to study competition among firms. We apply this methodology to the hotel industry, using a combination of search, clickstream and hotel data from a major OTA. Using these data, we also identify price-matching behavior among hoteliers and use it as a basis to develop competition sets from the hotelier perspective. We contrast the two competition networks on a subset of hotels in Manhattan. We find a 50% mismatch of the two competition networks. We also find that independent and distant hotels are more likely to be left out from competition sets. Meanwhile, hotels have a tendency to compare themselves to other hotels with lower star ratings, lower price levels, lower ranks and higher customers reviews.

A shortcoming with our application is that the data is from a single OTA. That is, we do not observe customer actions on other websites, such as other OTAs or hotels' own websites.This concern is mitigated, though, by the amount of overlap of listed hotels among multiple OTA sites (especially in the major destination that we study in this paper - Manhattan), and by the fact that 70% of the customers visit a travel-related site before booking on the supplier's own website. Certainly, a more comprehensive dataset would allow for more precise estimation.

Our approach can easily be applied to many other industries. We note that there exist various formats of data which can support the type of analysis that we conduct in this paper, including "customers who viewed this product also viewed ..." data, product comparison data tracked by sites providing "compare" options for online shoppers, and even offline technologies tracking customer movements in supermarkets. We would like to highlight that the business analytics opportunities of using this type of data are vast and promising.

We see this study as opening up many future research avenues on network competition as well. In the competition network of Manhattan hotels, we find that, on average, a hotel has 6 to 8 close competitors, and the average number of steps needed to reach from one hotel to another is 3. The small degree of separation potentially means that a small perturbation, such as a price promotion, may quickly spread to other parts of the market through competitive links. This would be an interesting phenomenon to study. Additionally, a competition network is an endogenous network. How is it formed in the marketplace, and how firms choose to position themselves in such a network is also an interesting question to answer. Other topics like dynamic evolution of competition networks and overlap of competition networks with other networks are also worth addressing.

## References

Allon, G., A. Federgruen, M. Pierson. 2011. How much is a reduction of your customers' wait worth? An empirical study of the fast-food drive-fhru industry based on structural estimation methods. *Manufaturing and Service Operations Management* **13**(4) 489–507.

Anderson, C.K., X.Q. Xie. 2011. A choice-based dynamic programming approach for setting opaque prices. *Production and Operations Management* **21**(3) 590–605.

Berry, S., J. Levinsohn, A. Pakes. 1995. Automobile prices in market equilibrium. *Econometrica* **63**(4) 841–890.

Bodea, T., M. Ferguson, L. Garrow. 2009. Choice-based revenue management: Data from a major hotel chain. *Manufacturing & Service Operations Management* **11**(2) 356–361.

Borenstein, S., N.L. Rose. 1994. Competition and price dispersion in the u.s. airline industry. *The Journal of Political Economy* **102**(4) 653–683.

Bouma, G. 2007. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL* 31–40.

Buell, R.W., D. Campbell, F.X. Frei. 2011. How do incumbents fare in the face of increased service competition? Working Paper, Harvard University, Cambridge, MA.

Church, K.W., P. Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* **16**(1) 22–29.

Cross, R.G., J.A. Higbie, D.Q. Cross. 2009. Revenue management's renaissance: A rebirth of the art and science of profitable revenue generation. *Cornell Hospitality Quarterly* **50**(1) 56–81.

Farias, V. F., S. Jagabathula, D. Shah. 2012. A new approach to modeling choice with limited data. *Management Science* Forthcoming.

Feenstra, R.C., J.A. Levinsohn. 1995. Estimating markups and market conduct with multidimensional product attributes. *Review of Economic Studies* **62**(1) 19–52.

Gabszewicz, J.J., J.F. Thisse. 1979. Price competition, quality, and income disparities. *Journal of Economic Theory* **20**(3) 340–359.

Gallego, G., L. Lin, R. Ratliff. 2009. Choice-based EMSR methods for single-leg revenue management with demand dependencies. *Journal of Revenue & Pricing Management* **8** 207–240.

Ghose, A., P. Ipeirotis, B. Li. 2012. Designing ranking systems for hotels on travel search engines by mining user-generated and crowd-sourced content. *Marketing Science* **31**(3).

Gokpinar, B., W.J. Hopp, S.M.R. Iravani. 2010. The impact of misalignment of organizational structure and product architecture on quality in complex product development. *Management Science* **56**(3) 468–484.

Hoque, A.Y., G.L. Loshe. 1999. An information search cost perspective for designing interfaces for electronic commerce. *Journal of Marketing Research* **36**(3) 387–394.

Hotelling, H. 1929. Stability in competition. *Economic Journal* **39**(153) 41–57.

Kim, J.Y., L. Canina. 2009. Product tiers and ADR clusters: Integrating two methods for determining hotel competitive sets. *Cornell Hospitality Report* **9**(14) 207–240.

Kranton, R.E., D.F. Minehart. 2001. A theory of buyer-seller network. *The American Economic Review* **91**(3) 485–508.

Li, J., N. F. Granados, S. Netessine. 2011. Are consumers strategic? Structural estimation from the air-travel industry. Working paper, University of Pennsylvania, Philadelphia, PA.

Manning, C., H. Schutze. 1999. *Foundations of Statistical Natural Language Processing.*. MIT Press, Cambridge, MA.

McFadden, D.L. 1984. Econometric analysis of qualitative response models. Z. Griliches, M. D. Intriligator, eds., *Handbook of Econometrics*, vol. 2, chap. 24. Elsevier, 1395–1457.

Moe, W., P. Fader. 2004. Dynamic conversion behavior at e-commerce sites. *Management Science* **50**(3) 326–335.

Netzer, O., R. Feldman, J. Goldenberg, M. Fresko. 2012. Mine your own business: Market-structure surveillance through text mining. *Marketing Science* **31**(3).

Olivares, M., G.P. Cachon. 2009. Competing retailers and inventory: An empirical investigation of general motors dealerships in isolated U.S. markets. *Management Science* **55**(9) 1586–1604.

Park, Y.H., P. Fader. 2004. Modeling browsing behavior at multiple websites. *Marketing Science* **23**(3) 280–303.

Pinkse, J., M. E. Slade, C. Brett. 2002. Spatial price competition: A semiparametric approach. *Econometrica* **70**(3) 1111–1153.

Salop, S. 1979. Monopolistic competition with outside goods. *Bell Journal of Economics* **10**(1) 141–156.

Shocker, A.D., M. Ben-Akiva, B. Boccara, P. Nedungadi. 1991. Consideration set influences on consumer decision-making and choice: Issues, models, and suggestions. *Marketing Letters* **2**(3) 181–197.

Sosa, M.E., S.D. Eppinger, C.M. Rowles. 2004. The misalignment of product architecture and organizational structure in complex product development. *Management Science* **50**(12) 1674–1689.

Talluri, K.T., G. J. van Ryzin. 2004. Revenue management under a general discrete choice model of consumer behavior. *Management Science* **50**(1) 15–33.

Thomadsen, R. 2007. Product opsitioning and competition: The role of location in the fast food industry. *Marketing Science* **26**(6) 792–804.

Thrane, C. 2007. Examining the determinants of room rates for hotels in capital cities: The Oslo experience. *Journal of Revenue & Pricing Management* **5**(4) 315–323.

Train, K. 2003. *Discrete choice methods with simulation*. Cambridge University Press, Cambridge, UK.

Urban, G.L., P.L. Johnson, J.R. Hauser. 1984. Testing competitive market structures. *Marketing Science* **3**(2) 83–112.

Vulcano, G., G. van Ryzin, W. Chaar. 2010. Choice-based revenue management: An empirical study of estimation and optimization. *Manufacturing and Service Operations Management* **12**(3) 371–392.

Withiam, G. 2011. Brave new world: Online hotel distribution. *Cornell Hospitality Roundtable & Conference Proceedings* **3**(4).

Zhang, D., W.L. Cooper. 2005. Revenue management for parallel flights with customer-choice behavior. *Operations Research* **53**(3) 415–431.

**Table 3 Summary Statistics**

| | Manhattan |
|---|---|
| **# hotels** | 309 |
| **# rooms** | 68,584 |
| **Independent Properties** | 59.2% |
| **major chains** | 24.9% |
| | |
| **# users** | 3,514 |
| **# hotel page views** | 22,901 |
| **# hotel page views/hotel** | 74.1 |
| **% no-click users** | 37.5% |
| **# page views/user**† | 10.4 |
| **# distinct hotels/user**† | 3.5 |
| | |
| **average price** | $295 |
| **std. dev. of price** | $200 |
| **% on promotion** | 41% |
| **full capacity** | 2.60% |

†: *conditional on clicking at least once.*

**Table 4 Sensitivity to Low Frequency Data**

| **hotel1** | **hotel2** | **# click1** | **# click2** | **# compare** | **PMI** | **NPMI** | **chi square** | **t** |
|---|---|---|---|---|---|---|---|---|
| Hotel Belleclaire | John Street Suites | 204 | 1 | 1 | 4.813 | 0.385 | 27.117 | 0.965 |
| Hotel Belleclaire | Brownstone Bed and Breakfast | 204 | 3 | 2 | 4.228 | 0.368 | 34.848 | 1.339 |
| Hotel Belleclaire | Oxbridge Carnegie Hill Apartments | 204 | 7 | 3 | 3.591 | 0.329 | 31.553 | 1.589 |
| Hotel Belleclaire | Grand Union | 204 | 36 | 10 | 2.965 | 0.324 | 61.950 | 2.760 |
| Hotel Belleclaire | Holiday Inn New York City - Wall Street | 204 | 12 | 4 | 3.228 | 0.308 | 31.079 | 1.787 |
| Hotel Belleclaire | Hotel 373 Fifth Avenue | 204 | 21 | 6 | 3.006 | 0.304 | 38.444 | 2.146 |
| Hotel Belleclaire | The Ritz-Carlton New York, Battery Park | 204 | 14 | 4 | 3.006 | 0.287 | 25.598 | 1.752 |
| Hotel Belleclaire | Comfort Inn Manhattan Bridge | 204 | 10 | 3 | 3.076 | 0.282 | 20.418 | 1.527 |
| Hotel Belleclaire | Thirty Thirty Hotel New York | 204 | 75 | 14 | 2.392 | 0.276 | 50.573 | 3.032 |
| Hotel Belleclaire | Hampton Inn Manhattan Soho | 204 | 11 | 3 | 2.939 | 0.270 | 18.069 | 1.507 |
| Hotel Belleclaire | Belnord Hotel | 204 | 59 | 11 | 2.390 | 0.265 | 39.553 | 2.686 |
| Hotel Belleclaire | 254 East Vacation | 204 | 7 | 2 | 3.006 | 0.262 | 12.783 | 1.238 |
| Hotel Belleclaire | Washington Jefferson Hotel | 204 | 54 | 10 | 2.380 | 0.260 | 35.570 | 2.557 |
| Hotel Belleclaire | Sutton Place Suites | 204 | 3 | 1 | 3.228 | 0.259 | 7.757 | 0.893 |
| Hotel Belleclaire | Best Western Hospitality House | 204 | 3 | 1 | 3.228 | 0.259 | 7.757 | 0.893 |

**Table 5 Intensity of Competition between Hotels – sorted by t-statistics**

| hotel1 | hotel2 | # click1 | # click2 | # compare | PMI | NPMI | chi-square | t |
|---|---|---|---|---|---|---|---|---|
| **top 10** | | | | | | | | |
| Park Central New York Hotel | Paramount Hotel Times Square New York | 164 | 93 | 21 | 2.981 | 0.368 | 132.352 | 4.010 |
| Sheraton New York Hotel And Towers | The Manhattan at Times Square Hotel | 141 | 89 | 19 | 3.118 | 0.379 | 134.511 | 3.863 |
| Hilton Garden Inn Times Square | Hampton Inn Times Square North | 146 | 68 | 18 | 3.378 | 0.406 | 158.769 | 3.841 |
| Hilton New York | Roosevelt Hotel New York | 133 | 79 | 18 | 3.296 | 0.396 | 148.108 | 3.817 |
| Hilton Garden Inn Times Square | The New Yorker Hotel | 146 | 150 | 21 | 2.459 | 0.304 | 81.454 | 3.756 |
| The Edison Hotel | Paramount Hotel Times Square New York | 180 | 93 | 19 | 2.702 | 0.328 | 92.973 | 3.695 |
| Sheraton New York Hotel And Towers | Hilton Garden Inn Times Square | 141 | 146 | 20 | 2.478 | 0.304 | 78.925 | 3.676 |
| Hilton Garden Inn Times Square | Roosevelt Hotel New York | 146 | 79 | 17 | 3.079 | 0.367 | 116.229 | 3.641 |
| The Edison Hotel | Salisbury Hotel | 180 | 127 | 20 | 2.327 | 0.285 | 67.925 | 3.587 |
| Hilton Garden Inn Times Square | The Belvedere Hotel | 146 | 114 | 18 | 2.633 | 0.317 | 82.229 | 3.564 |
| **bottom 10** | | | | | | | | |
| Hotel Pennsylvania | Doubletree Guest Suites Times Square NYC | 46 | 191 | 1 | -0.615 | -0.049 | 0.193 | -0.532 |
| Doubletree Guest Suites Times Square NYC | Ink48 Hotel, a Kimpton Hotel | 191 | 47 | 1 | -0.646 | -0.052 | 0.213 | -0.565 |
| St. Giles - The Court New York | Hotel Belleclaire | 45 | 204 | 1 | -0.679 | -0.054 | 0.236 | -0.601 |
| Hilton New York | Latham Hotel | 133 | 73 | 1 | -0.760 | -0.061 | 0.294 | -0.693 |
| Doubletree Guest Suites Times Square NYC | The Pierre, A Taj Hotel | 191 | 51 | 1 | -0.764 | -0.061 | 0.300 | -0.699 |
| Sheraton New York Hotel And Towers | Latham Hotel | 141 | 73 | 1 | -0.844 | -0.068 | 0.365 | -0.795 |
| The Edison Hotel | Sofitel New York | 180 | 65 | 1 | -1.029 | -0.082 | 0.554 | -1.040 |
| Doubletree Guest Suites Times Square NYC | The Waldorf Astoria | 191 | 105 | 2 | -0.806 | -0.070 | 0.675 | -1.059 |
| The Plaza | Hotel Belleclaire | 59 | 204 | 1 | -1.069 | -0.086 | 0.603 | -1.099 |
| Empire Hotel | Doubletree Guest Suites Times Square NYC | 70 | 191 | 1 | -1.221 | -0.098 | 0.796 | -1.331 |

**Table 6 T-statistics Accounting for Aggregation Bias**

| hotel1 | hotel2 | basic | logit | mixed logit |
|---|---|---|---|---|
| **top 10** | | | | |
| Park Central New York Hotel | Paramount Hotel Times Square New York | 4.010 | 0.378 | 0.157 |
| Sheraton New York Hotel And Towers | The Manhattan at Times Square Hotel | 3.863 | 1.641 | 1.474 |
| Hilton Garden Inn Times Square | Hampton Inn Times Square North | 3.841 | 0.920 | 0.911 |
| Hilton New York | Roosevelt Hotel New York | 3.817 | 2.257 | 2.206 |
| Hilton Garden Inn Times Square | The New Yorker Hotel | 3.756 | 2.051 | 1.968 |
| The Edison Hotel | Paramount Hotel Times Square New York | 3.695 | 0.638 | 0.464 |
| Sheraton New York Hotel And Towers | Hilton Garden Inn Times Square | 3.676 | 2.406 | 2.342 |
| Hilton Garden Inn Times Square | Roosevelt Hotel New York | 3.641 | 2.772 | 2.753 |
| The Edison Hotel | Salisbury Hotel | 3.587 | 1.249 | 1.092 |
| Hilton Garden Inn Times Square | The Belvedere Hotel | 3.564 | 1.098 | 1.026 |
| **bottom 10** | | | | |
| Hotel Pennsylvania | Doubletree Guest Suites Times Square NYC | -0.532 | 0.284 | 0.198 |
| Doubletree Guest Suites Times Square NYC | Ink48 Hotel, a Kimpton Hotel | -0.565 | -0.359 | -0.386 |
| St. Giles - The Court New York | Hotel Belleclaire | -0.601 | 0.816 | 0.790 |
| Hilton New York | Latham Hotel | -0.693 | -0.776 | -0.995 |
| Doubletree Guest Suites Times Square NYC | The Pierre, A Taj Hotel | -0.699 | -1.161 | -1.218 |
| Sheraton New York Hotel And Towers | Latham Hotel | -0.795 | -1.103 | -1.413 |
| The Edison Hotel | Sofitel New York | -1.040 | -0.173 | -0.271 |
| Doubletree Guest Suites Times Square NYC | The Waldorf Astoria | -1.059 | -0.578 | -0.563 |
| The Plaza | Hotel Belleclaire | -1.099 | -0.021 | -0.086 |
| Empire Hotel | Doubletree Guest Suites Times Square NYC | -1.331 | -1.024 | -1.141 |

**Table 7 Price-Matching Patterns**

| | Park Central New York Hotel | Paramount Hotel Times Square New York | Sheraton New York Hotel And Towers | The Manhattan at Times Square Hotel |
|---|---|---|---|---|
| **days in advance** | 0.59*** (0.14) | 1.89*** (0.14) | -0.45*** (0.13) | -0.34*** (0.13) |
| **days in advance squared** | -0.02*** (0.00) | -0.02*** (0.00) | -0.01*** (0.00) | -0.01*** (0.00) |
| **length of stay** | -9.42*** (0.64) | -10.88*** (0.63) | -4.85*** (0.69) | -6.30*** (0.62) |
| **Mon** | 7.82 (10.68) | -19.23 (9.60) | 5.27 (9.11) | -1.94 (10.13) |
| **Tue** | 62.86*** (9.35) | 50.82*** (8.43) | 30.75*** (8.04) | 17.16** (8.64) |
| **Wed** | 91.00*** (8.73) | 68.20*** (7.72) | 63.09*** (7.53) | 53.59*** (8.06) |
| **Thu** | 106.12*** (8.36) | 79.57*** (7.51) | 55.50*** (7.60) | 46.85*** (8.07) |
| **Fri** | 115.73*** (7.91) | 84.98*** (7.04) | 88.32*** (6.75) | 83.51*** (7.45) |
| **Sat** | 118.52*** (9.56) | 100.94*** (10.16) | 51.53*** (7.01) | 45.65*** (7.68) |
| **Christmas-New Year** | 95.22*** (5.18) | 43.67*** (4.89) | 3.82 (6.39) | -11.06* (6.17) |
| **# adults per room** | 15.40*** (1.96) | 49.16*** (1.86) | 4.98*** (1.88) | 1.63 (2.05) |
| **# children per room** | 29.43*** (3.27) | 87.59*** (3.01) | 8.82*** (3.40) | 2.95 (5.00) |
| **const** | 225.61*** (7.99) | 161.23*** (7.06) | 308.76*** (6.93) | 313.28 (7.69) |
| **Adj R_sq** | 0.3267 | 0.4781 | 0.3888 | 0.3691 |
| **# obs** | 3945 | 3690 | 3231 | 3306 |

**Table 8 Measures of Price Competitiveness**

| hotel1 | hotel2 | t | Price Correlation total $\rho(p)$ | predicted $\rho(\hat{p})$ | unpredicted $\rho(p_{res})$ | Price Difference $ |
|---|---|---|---|---|---|---|
| **top 10** | | | | | | |
| Park Central New York Hotel | Paramount Hotel Times Square New York | 4.010 | 0.777 | 0.814 | 0.739 | 42.3 |
| Sheraton New York Hotel And Towers | The Manhattan at Times Square Hotel | 3.863 | 0.985 | 0.993 | 0.973 | 9.9 |
| Hilton Garden Inn Times Square | Hampton Inn Times Square North | 3.841 | 0.885 | 0.842 | 0.784 | 33.2 |
| Hilton New York | Roosevelt Hotel New York | 3.817 | 0.773 | 0.878 | 0.726 | 43.1 |
| Hilton Garden Inn Times Square | The New Yorker Hotel | 3.756 | 0.671 | 0.811 | 0.559 | 54.7 |
| The Edison Hotel | Paramount Hotel Times Square New York | 3.695 | 0.602 | 0.664 | 0.518 | 90.1 |
| Sheraton New York Hotel And Towers | Hilton Garden Inn Times Square | 3.676 | 0.842 | 0.916 | 0.774 | 37.4 |
| Hilton Garden Inn Times Square | Roosevelt Hotel New York | 3.641 | 0.827 | 0.872 | 0.803 | 33.1 |
| The Edison Hotel | Salisbury Hotel | 3.587 | 0.724 | 0.955 | 0.522 | 54.9 |
| Hilton Garden Inn Times Square | The Belvedere Hotel | 3.564 | 0.851 | 0.971 | 0.794 | 39.8 |
| **bottom 10** | | | | | | |
| Hotel Pennsylvania | Doubletree Guest Suites Times Square NYC | -0.532 | 0.538 | 0.504 | 0.641 | 509.2 |
| Doubletree Guest Suites Times Square NYC | Ink48 Hotel, a Kimpton Hotel | -0.565 | 0.475 | 0.410 | 0.497 | 66.6 |
| St. Giles - The Court New York | Hotel Belleclaire | -0.601 | 0.481 | 0.353 | 0.270 | 77.4 |
| Hilton New York | Latham Hotel | -0.693 | 0.551 | 0.686 | 0.572 | 210.4 |
| Doubletree Guest Suites Times Square NYC | The Pierre, A Taj Hotel | -0.699 | 0.122 | -0.001 | 0.484 | 262.4 |
| Sheraton New York Hotel And Towers | Latham Hotel | -0.795 | 0.641 | 0.844 | 0.551 | 179.5 |
| The Edison Hotel | Sofitel New York | -1.040 | 0.735 | 0.902 | 0.559 | 153.2 |
| Doubletree Guest Suites Times Square NYC | The Waldorf Astoria | -1.059 | 0.170 | -0.060 | 0.238 | 278.7 |
| The Plaza | Hotel Belleclaire | -1.099 | 0.457 | 0.821 | 0.259 | 527.2 |
| Empire Hotel | Doubletree Guest Suites Times Square NYC | -1.331 | 0.569 | 0.474 | 0.575 | 502.5 |

**Table 9 Network Overlap and Mismatch among Hotels with 3 Stars and Up**

| customer-based competitors | hotelier-based competitors 0 | 1 | Total |
|---|---|---|---|
| 0 | 859 | 195 | 1,054 |
| 1 | 195 | 191 | 386 |
| Total | 1,054 | 386 | 1,440 |

**Table 10 When Network Mismatch Is Likely to Occur**

| | Basic measures Leftout | Wrong | Alternative measure of t-stat Leftout | Wrong | Alternative measure of price matching Leftout | Wrong |
|---|---|---|---|---|---|---|
| **Higher Star Rating** | 0.012 (0.008) | -0.010*** (0.002) | 0.011 (0.007) | -0.011*** (0.003) | -0.007 (0.008) | -0.007** (0.003) |
| **Larger Hotel** | 0.008 (0.032) | 0.019 (0.013) | -0.006 (0.029) | 0.008 (0.013) | 0.003 (0.030) | 0.006 (0.013) |
| **Independent Hotel** | 0.203*** (0.053) | -0.105*** (0.025) | 0.065 (0.055) | -0.146*** (0.025) | 0.102* (0.053) | -0.078*** (0.025) |
| **More Expensive Hotel** | 0.001 (0.000) | 0.000** (0.000) | 0.002 (0.004) | -0.005*** (0.001) | 0.005 (0.004) | -0.001 (0.001) |
| **Higher Customer Review** | -0.076** (0.036) | -0.022 (0.019) | -0.074** (0.035) | -0.014 (0.020) | -0.067** (0.034) | -0.012 (0.021) |
| **Higher Avg. Rank** | 0.008*** (0.002) | 0.002* (0.001) | 0.009*** (0.002) | 0.002 (0.001) | 0.004* (0.002) | 0.000 (0.001) |
| **Different sub-market** | 0.109* (0.060) | -0.032 (0.028) | 0.117* (0.061) | -0.044 (0.029) | 0.066 (0.058) | -0.061** (0.030) |
| **Log Likelihood** | -244.8 | -472.3 | -248.9 | -486.8 | -256.4 | -496.2 |
| **# Obs** | 386 | 1054 | 386 | 1054 | 386 | 1054 |

*Note: Coefficients are marginal effects predicted from logit models.*

**Table 11 Choice Models Predicting Probability of Clicking**

| | Logit | Mixed Logit | |
|---|---|---|---|
| | | Mean | S.D. |
| **rank** | -0.045*** (0.004) | -0.063*** (0.004) | 0.045*** (0.003) |
| **total # of displayed hotels** | -0.073*** (0.004) | -0.086*** (0.005) | 0.016*** (0.002) |
| **no available room** | -0.180*** (0.022) | -0.549*** (0.130) | 0.039 (0.245) |
| **price** | -0.122 (0.106) | -0.244*** (0.028) | 0.051** (0.023) |
| **promotion** | 0.530*** (0.049) | 0.392*** (0.065) | 0.892*** (0.065) |
| **star rating** | 0.066** (0.028) | -0.019 (0.043) | 0.006 (0.013) |
| **customer review** | -0.013 (0.039) | 0.061** (0.029) | 0.006 (0.011) |
| **sub-market dummies** | | | |
| **financial** | -0.092 (0.592) | -0.084 (0.621) | 0.347 (0.332) |
| **central** | -0.709 (0.609) | -1.505 (0.957) | 1.610** (0.699) |
| **lower** | -0.349 (0.608) | -0.330 (0.651) | 0.290 (0.623) |
| **lowereast** | -0.061 (0.621) | -0.117 (0.673) | 0.440 (0.675) |
| **lowerwest** | -0.210 (0.575) | -0.160 (0.600) | 0.411* (0.234) |
| **mideast** | -0.003 (0.573) | 0.056 (0.595) | 0.302* (0.163) |
| **midwest** | -0.019 (0.573) | 0.062 (0.594) | 0.139 (0.121) |
| **timessquare** | 0.021 (0.573) | 0.082 (0.596) | 0.312** (0.159) |
| **uppereast** | -0.301 (0.584) | -0.354 (0.621) | 0.584* (0.351) |
| **upperwest** | 0.206 (0.578) | 0.346 (0.600) | 0.191 (0.352) |
| **uptowneast** | 1.357 (0.986) | 1.221 (1.196) | 0.886 (1.559) |
| **const** | -0.197 (0.588) | 0.317 (0.610) | 0.011 (0.046) |
| **Log likelihood** | -7800.5209 | -7542.7694 | |
| **# obs** | 32223 | 32223 | |

## Acknowledgments

The authors would like to thank the corporate sponsor for making the data available and Wharton Customer Analytics Initiative for the research opportunity.

Europe Campus
Boulevard de Constance
77305 Fontainebleau Cedex, France
Tel: +33 (0)1 60 72 40 00
Fax: +33 (0)1 60 74 55 00/01

Asia Campus
1 Ayer Rajah Avenue, Singapore 138676
Tel: +65 67 99 53 88
Fax: +65 67 99 53 99

Abu Dhabi Campus
Muroor Road - Street No 4
P.O. Box 48049
Abu Dhabi, United Arab Emirates
Tel: +971 2 651 5200
Fax: +971 2 443 9461

www.insead.edu

INSEAD

The Business School
for the World®

Printed by INSEAD