Information Systems Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

The Impact of Fake Reviews on Online Visibility: A Vulnerability Assessment of the Hotel Industry

Theodoros Lappas, Gaurav Sabnis, Georgios Valkanas

To cite this article:
Theodoros Lappas, Gaurav Sabnis, Georgios Valkanas (2016) The Impact of Fake Reviews on Online Visibility: A Vulnerability Assessment of the Hotel Industry. Information Systems Research

Published online in Articles in Advance 09 Nov 2016

. http://dx.doi.org/10.1287/isre.2016.0674

Information Systems Research
Articles in Advance, pp. 1–22
ISSN 1047-7047 (print) | ISSN 1526-5536 (online)

https://doi.org/10.1287/isre.2016.0674

The Impact of Fake Reviews on Online Visibility: A Vulnerability Assessment of the Hotel Industry

Theodoros Lappas, Gaurav Sabnis, Georgios Valkanas
Stevens Institute of Technology, Hoboken, New Jersey 07030
{tlappas@stevens.edu, gsabnis@stevens.edu, gvalkana@stevens.edu}

Extant research has focused on the detection of fake reviews on online review platforms, motivated by the well-documented impact of customer reviews on the users' purchase decisions. The problem is typically approached from the perspective of protecting the credibility of review platforms, as well as the reputation and revenue of the reviewed firms. However, there is little examination of the vulnerability of individual businesses to fake review attacks. This study focuses on formalizing the visibility of a business to the customer base and on evaluating its vulnerability to fake review attacks. We operationalize visibility as a function of the features that a business can cover and its position in the platform's review-based ranking. Using data from over 2.3 million reviews of 4,709 hotels from 17 cities, we study how visibility can be impacted by different attack strategies. We find that even limited injections of fake reviews can have a significant effect and explore the factors that contribute to this vulnerable state. Specifically, we find that, in certain markets, 50 fake reviews are sufficient for an attacker to surpass any of its competitors in terms of visibility. We also compare the strategy of self-injecting positive reviews with that of injecting competitors with negative reviews and find that each approach can be as much as 40% more effective than the other across different settings. We empirically explore response strategies for an attacked hotel, ranging from the enhancement of its own features to detecting and disputing fake reviews. In general, our measure of visibility and our modeling approach regarding attack and response strategies shed light on how businesses that are targeted by fake reviews can detect and tackle such attacks.

Keywords: customer reviews; fake reviews; knowledge management; vulnerability assessment; decision support systems
History: Rob Fichman, Ram Gopal, Alok Gupta, Sam Ransbotham, Senior Editors; Alok Gupta, Associate Editor. This paper was received on March 2, 2015, and was with the authors 8 months for 2 revisions. Published online in *Articles in Advance* November 9, 2016.

1. Introduction

> Mrs. Richards: When I pay for a room with a view, I expect something more interesting than that.
> Basil: That is Torquay, madam.
> Mrs. Richards: Well it's not good enough.
> Basil: Well, may I ask what you expected to see out of a Torquay hotel bedroom window? Sydney Opera House, perhaps? The Hanging Gardens of Babylon? Herds of wildebeest sweeping majestically . . . ?
> Mrs. Richards: Don't be silly. I expect to be able to see the sea.
> Basil: You can see the sea. It's over there between the land and the sky.
> Mrs. Richards: I'd need a telescope to see that!
>
> Fawlty Towers,
> Episode "Communication Problems"
> 1979, British Broadcasting Corporation

This humorous exchange between proprietor Basil Fawlty and dissatisfied customer Mrs. Richards illustrates an important phenomenon in the hospitality industry, in which customers book hotels based primarily on features that are important to them. For Mrs. Richards, the view from her window is important. For someone else, the top priority could be free breakfast or air-conditioning. Large travel websites such as TripAdvisor.com or Booking.com allow users to *filter* the hotels that they consider, based on the features that they are interested in. The popularity of such platforms is largely based on the availability of large volumes of customer reviews, which are considered to be more credible than biased promotional campaigns (Bickart and Schindler 2001, Lu et al. 2013). Relevant literature has established the impact of reviews on purchase decisions (Chatterjee 2001, Kwark et al. 2014, Ghose et al. 2014, Duan et al. 2008) and, consequently, on a firm's sales and revenue (Ghose and Ipeirotis 2011, Forman et al. 2008, Zhu and Zhang 2010). Nevertheless, online reviews are also susceptible to tampering from unscrupulous businesses who attempt to manipulate the available information by posting either fake positive reviews about themselves or fake negative reviews about their competitors, resulting in review fraud (Dellarocas 2006, Mayzlin et al. 2012, Lappas 2012, Luca and Zervas 2016). Although a considerable amount of research has focused on the identification of fake reviews (Hu et al. 2012, Lappas 2012, Xie et al. 2012, Jindal and Liu 2008, Mukherjee et al. 2013a), we still have a relatively limited understanding of the nature and nuances of the vulnerability of the businesses themselves to review fraud. To overcome such limitations, our work studies the mechanisms by which fake reviews can mislead customers and affect the targeted business. Next, we demonstrate the intuition behind our approach with a realistic example.

Consider a businesswoman planning to arrive in a new city at 11 P.M. and spend the night preparing for a 10 A.M. presentation the next day. When she searches for hotels on an online platform, she might choose free wi-fi as a filtering feature to have access to the Internet while working on her slides. She might also ask for a business center to print documents for the presentation. When the same businesswoman travels with her family for vacation, her preferences are likely to differ. For instance, she might filter her results based on the availability of a swimming pool that her children can enjoy, or limit her search to hotels that offer family rooms. In both scenarios, the hotel she eventually chooses will be among those that appear as a response to her filtered search. The website presents these competitive hotels in a ranked list, in which the position of each competitor depends on its reviews (TripAdvisor 2013, Holloway 2011a).

Previous work has repeatedly verified that the increased visibility of highly ranked items improves their chances of being considered and selected by interested users (Pan 2015, Ghose et al. 2014, Ifrach and Johari 2014, Tucker and Zhang 2011). In this setting, a hotel's visibility is determined by the set of features that it can cover, the popularity of these features among potential customers, and the hotel's position in the review-based ranking. Even though the first two elements are impervious to manipulation, the third visibility component is clearly vulnerable to fake reviews. Intuitively, the injection of a sufficient number of positive or negative fake reviews could alter the ranking and have a significant impact on the visibility of the ranked businesses. The pressing concern for such injection-based attacks motivates us to contribute to the growing literature on online reviews from a vulnerability perspective, by focusing on the following research objectives: (i) formalizing the *visibility* of a business, by defining a measure that considers its position in the review-based ranking, (ii) understanding how visibility can be impacted by review fraud, and (iii) identifying effective ways for businesses to protect their visibility from fake review attacks.

The cornerstone of our research methodology is the visibility construct. Our work identifies the components that contribute to an item's visibility within an online platform and describes the operationalization of each component. First, we use a large data set of over 2.3 million reviews of 4,709 hotels from 17 cities to estimate the popularity of the different hotel features. Our study reveals consistent user preferences across cities, and also provides specific guidelines on the amount of review data that are sufficient to obtain accurate estimates. For the review-based ranking, we evaluate the industry-standard *average rating* function, as well as TripAdvisor's Popularity Index formula, which considers the age, quantity, and quality of a hotel's reviews (TripAdvisor 2013). Our framework also accounts for the different ways in which users consider items in ranked lists. We implement and evaluate three alternative consideration models, motivated by the significant body of relevant literature. Finally, we simulate three different attack strategies to evaluate the vulnerability of visibility to fake reviews. Our evaluation reveals the factors that determine a hotel's vulnerability and leads to findings with strong implications about the ways in which platforms rank competitive hotels. We address the vulnerabilities exposed by our study by proposing a fraud-resistant method for review-based ranking, and suggesting different response strategies for businesses that want to protect their visibility from fake reviews.

The remainder of the paper is organized as follows. First, we discuss relevant work in Section 2. We describe our TripAdvisor data set in Section 3. We then describe our visibility construct in Section 4. In Section 5, we describe attack strategies based on fake reviews. Then, in Section 6, we discuss response strategies that can be used to address the exposed vulnerabilities. Finally, we conclude in Section 7 with a discussion of our findings and their implications, as well as an overview of directions for future research.

2. Background

Fake reviews have been acknowledged as a critical challenge by both the research community (Feng et al. 2012) and the e-commerce industry (Sussin and Thompson 2012, Breure 2013). Despite the commitment of review platforms to combat fraud, the percentage of fake reviews is estimated to be around 15%–30% (Sussin and Thompson 2012, Luca and Zervas 2016, Belton 2015). In a recent formal complaint filed in court by Amazon.com (Stempel 2015), the retail giant wrote: "While small in number, these reviews threaten to undermine the trust that customers, and the vast majority of sellers and manufacturers, place in Amazon, thereby tarnishing Amazon's brand." In October 2015, the company filed a lawsuit against 1,000 people who allegedly offered to hire themselves out as fake reviewers (Weise 2015). This lawsuit is a milestone in the ongoing fight against fake reviews, as it was the first legal action taken by a large corporation against individual users. In the same year, the Italian magazine *Italia a Tavola* created a fictitious restaurant on TripAdvisor. The restaurant secured the highest rank among all competitors in the region via the injection of fake positive reviews (Fenton 2015). This controversial experiment came a year after TripAdvisor was issued a $613,000 fine by the Italian Competition Authority, for failing to adopt controls against false reviews while promoting its content as "authentic and genuine" (Masoni 2014).

The prevalence of review fraud can be largely attributed to the plethora of professional review-authoring companies that submit fake reviews on major review platforms in exchange for a fee. In fact, these companies have grown so confident in their ability to continuously adapt and avoid detection that they are often willing to delay their compensation until after the fake reviews have penetrated the website's defenses.[1] This phenomenon has prompted action by government bodies worldwide. In 2015, the Attorney General of the State of New York spearheaded "Operation Clean Turf," an effort to identify and expose review-authoring firms (Schneiderman 2015). The operation resulted in the identification of 19 companies, which were forced to cancel their services and pay over $350,000 in fines. Despite such ongoing efforts, a number of similar companies are still active and even willing to inject defamatory reviews about the competitors of the business that hires them. In the same year, a study by the UK's Competition and Markets Authority (CMA) identified a growing number of companies that offer review-authoring services (CMA 2015). Additional findings from the study include the devastating effects of fake negative reviews, especially for small businesses, as well as the customer practice of using defamatory reviews to blackmail businesses into providing some concession, such as a price discount. Finally, the study provided valuable insight on the distribution of fake reviews, which we discuss in detail in Section 2.1.

2.1. Studying the Distribution of Fake Reviews

One of the main findings of the CMA study was that fake positive reviews are more common than fake negative reviews. As a possible explanation, the study hypothesized that it is easier and less risky to affect overall ratings by self-injecting positive reviews than by posting a series of negative reviews about many business rivals. This is intuitive, as culprits caught sabotaging the reviews of competitors can face serious repercussions, in terms of lawsuits and long-term damage to their brand. Therefore, a reasonable hypothesis would be that the practice of injecting negative reviews to competitors is more likely to be adopted by new or small businesses, which do not have an established brand to damage if they get caught. However, high-profile cases that expose corporate giants as solicitors of fake negative reviews suggest a much wider reach of this phenomenon. For instance, in 2013, the Taiwan Fair Trade Commission fined Samsung $340,000 for hiring two external companies to post negative reviews about HTC, one of Samsung's main competitors. The story, which was extensively covered by media outlets, demonstrated that review fraud is a widespread practice that is not limited to small businesses (Tibken 2013). Furthermore, the relevant research literature on the distribution of fake reviews is relatively limited and also provides some conflicting evidence. For instance, while positive fake reviews are generally accepted as more prevalent and easier to generate (CMA 2015, Mayzlin et al. 2012), evidence from multiple domains shows that reviews by nonverified buyers (which are more likely to be fraudulent) are significantly more negative than those by verified buyers (Anderson and Simester 2014).

Given the ambiguity of previous findings, we conduct our own study on a data set collected from Yelp.com, which applies a proprietary filter to identify fraudulent and other low-quality reviews (Kamerer 2014). Even though such reviews are not prominently displayed on the page of the reviewed business, they remain online and can be viewed by the interested user. Previous work has verified these reviews as an accurate proxy for the set of fake reviews detected by the platform (Luca and Zervas 2016). Our data set consists of about 15,000 hotel reviews. Out of these, about 56% were positive (four or five stars), 29% were negative (one or two stars), and the rest did not have a clear positive or negative rating (three stars). These percentages suggest that, even though fake positive reviews are indeed more prevalent, fake negative reviews still make up a significant part (around one-third) of all review injections. In practice, these percentages are likely to differ across platforms and domains. In addition, the estimation task is inherently problematic, as we cannot account for the undetected fake reviews that penetrated the platform's defenses. The attack-simulation framework that we present in Section 5 is the first attempt toward addressing this limitation.

2.2. Defense Mechanisms

The growing concern over fake reviews has motivated review platforms to implement different types of defense mechanisms and review filters. For instance, reviews posted on TripAdvisor are subject to review by the website's moderators (TripAdvisor 2015a). Suspicious reviews can then be placed on hold pending examination and can even be eliminated if the website's proprietary filtering process provides enough evidence. Businesses that are associated with fake reviews are penalized in the website's rankings, excluded from press releases and top-10 lists, and may even have a relevant banner placed on their page.

[1] http://realtripadvisorreviews.com/.

While such penalties serve as deterrents, they have also been exposed as a creative way for an unscrupulous business to defame its competitors. Specifically, by repeatedly injecting fake positive reviews about a competitor, a malicious business can manipulate the platform into detecting an attack and penalizing the injected competitor (TripAdvisor 2015a). In addition, given the scarcity of ground truth data and the difficulty of detecting fake reviews, filtering mechanisms are likely to lead to false positives (CMA 2015).

One of the most popular defense mechanisms against fake reviews is the "Verified Buyer" (VB) badge, which only allows reviews by users that have purchased the item or service. Verified reviews have been adopted by multiple platforms and praised as a promising solution to review fraud (Mayzlin et al. 2012, May 2011). However, as we discuss next, they also introduce limitations and can even lead to undesirable consequences.

First, a motivated attacker or professional review-authoring firm can still buy a competitor's product or service (especially in low-cost markets), earn the VB badge, and proceed to inject fake negative reviews. The task of self-injecting fake positive reviews would then be even easier, since the purchase cost would essentially return to the culprit. Second, only the person who used their information to make the purchase would be able to submit a review. In the hotels domain, this excludes friends and family members who stayed in the same room but would not have the chance to share their insights. Similarly, verification excludes customers who do not book through a website and prefer to pay cash or go through a travel agency. This limits the reviewer demographic and could introduce an unwanted bias to the ratings. This is a legitimate concern, especially because only an estimated 43% of all hotel bookings are done online (StatisticBrain 2015). In fact, online payments are also not the norm for other businesses in the service industry, such as restaurants, making verification a challenging task. Finally, given that online reviews are public, a website that allows only verified customers will unavoidably reveal information about its users. This could raise privacy concerns and discourage reviewers who do not want to openly share their purchase decisions. While such concerns could be overcome by allowing reviewers to hide their identities, this would also take away their ability to build their reputation within the online community, which has been shown to be one of the principal motivations for review authoring (Wang 2010).

The above concerns about the VB badge have discouraged many platforms, including TripAdvisor, the largest travel website in the world. In a recent 2015 statement, the platform defended its choice to allow nonverified reviews (Schaal 2015b): "We have considered all of the verification options out there, and have elected to use our current model for one simple reason: The volume of opinions provides for the most in-depth coverage of consumer experience, and it is the model that consumers prefer." A hybrid policy that allows both verified and unverified reviews has enabled TripAdvisor to accumulate over 250 million reviews, a number that far surpasses that of its competitors (Schaal 2015a, Olery 2015, TripAdvisor 2015b). The website also boasts the largest market share among all travel websites in the hotel industry in terms of traffic, more than twice its nearest competitor Booking.com (Tnooz 2013), which only allows verified reviews. TripAdvisor's unwillingness to sacrifice volume for verification is consistent with relevant research, which identifies the number of reviews as the factor with the highest influence on the users' trust in the platform, higher even than that of the quality and detail of the reviews (Breure 2013). The hybrid approach has also been adopted by other major travel websites, including Orbitz.com and InsiderPages.com, as well as by market leaders in other industries, such as Amazon.com.

In conclusion, while buyer verification can support the effort against fake reviews, it also has a number of limitations and undesirable effects. This has motivated large review platforms to opt for a hybrid policy that allows both verified and unverified reviewers, while bolstering their effort to identify and eliminate fake reviews. Our goal is to make a decisive contribution to this effort, by quantifying the vulnerability of review-based visibility to different types of review fraud, and presenting informed response strategies for vigilant businesses and review platforms.

3. The TripAdvisor Data Set

We use a large data set of reviews from TripAdvisor.com, one of the largest review platforms and travel portals. The data were collected during the last week of January 2015, and include over 2.3 million reviews on 4,709 hotels from 17 cities. Each hotel is mapped to a space of 18 features, for which we adopt the representation presented by Ding et al. (2008) in their seminal work on lexicon-based opinion mining. Each feature $f \in \mathcal{F}$ is represented via a finite set of words or phrases W_f, which includes synonyms and other terms with a strong semantic connection to the feature. Table 1 shows the set of words for each feature. For example, any occurrence of the words "Internet" or "wifi" is mapped to the "Internet" feature. In addition, we use the opinion-mining algorithm of Ding et al. (2008) to analyze the sentiment (positive/negative) and the strength of the opinions expressed about each feature within a review. The algorithm assigns a value in the $[-1, +1]$ range to each reviewed feature, with -1 and $+1$ indicating a highly negative and

Table 1 **Hotel Features and Their Respective Sets of Relevant Words**

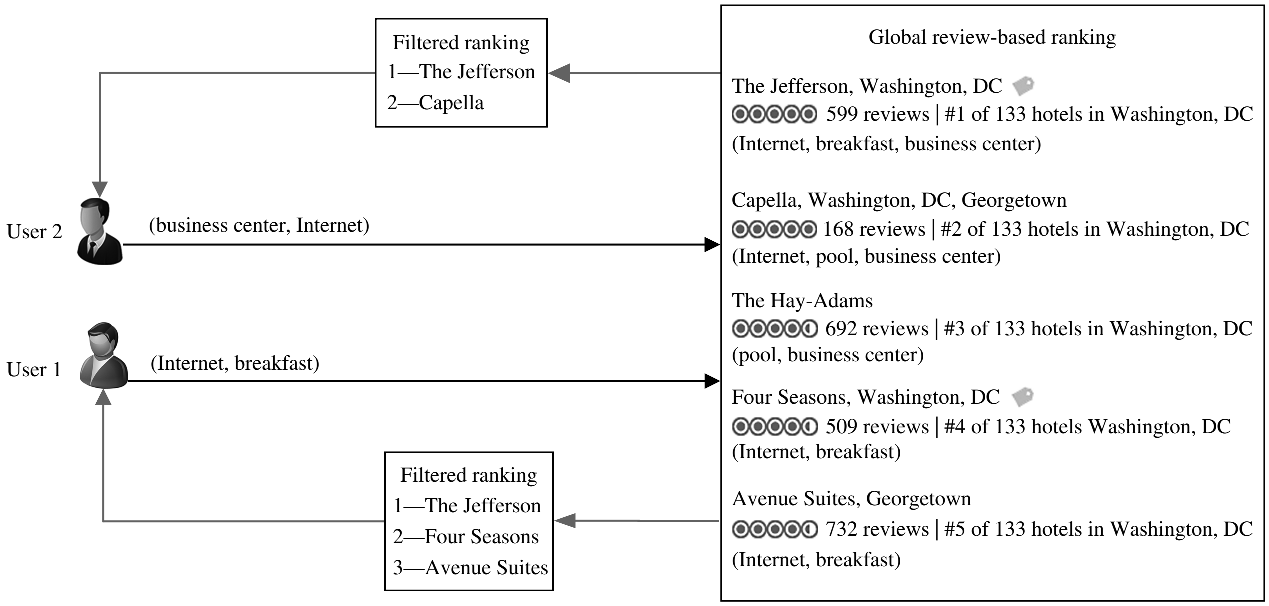
Feature f	Set of relevant keywords W_f
1. Air-conditioning	a/c, ac, air condition(ing, ed), room temperature
2. Airport transportation	airport transportation, airport transport, airport shuttle, airport ride, airport bus
3. Bar/Lounge	bar, lounge, sitting area, cafeteria, cafe
4. Business services	business center, business area, business service(s), conference room, conference area
5. Concierge	concierge, doorman, attendant
6. Fitness	fitness, workout, gym, exercise, athletics center
7. Breakfast	breakfast, morning meal
8. Parking	parking, park (any word) car
9. Internet	Internet, wi-fi, wifi, wireless, network, ethernet
10. Kitchenette	kitchenette, kitchen, cooking area, oven, stove
11. Nonsmoking	nonsmoking, smoking, smoker, smoke, cigarette(s)
12. Pets	pet(s), cat(s), dog(s), bird(s), parrot(s), pet friendly, animal(s)
13. Pool	pool, swimming pool
14. Reduced mobility	reduced mobility, limited mobility, disabled, disability, disabilities, handicapped, wheelchair, ramp
15. Restaurant	restaurant, buffet, food
16. Room service	room service
17. Spa	spa, sauna, massage
18. Suites	suite, suites

highly positive sentiment, respectively. The sentiment is then aggregated to the hotel level via averaging. We utilize the opinions mined from our data set for (i) the robustness check of our method for estimating the importance of queries, presented in Section 4.1.1, and (ii) the feature enhancement strategy described in Section 6.1. Additional details on the data are provided in Table 1 of the online appendix (available as supplemental material at https://doi.org/10.1287/isre.2016.0674).

4. Measuring Visibility

Previous studies have focused on evaluating the effects of customer reviews on booking intentions (Mauri and Minazzi 2013, Sparks and Browning 2011, Vermeulen and Seegers 2009) and sales (Ye et al. 2009, 2011). While the consensus is that reviews can significantly affect user decisions, the precise mechanism by which this effect occurs is typically overlooked. Instead, the corpus of reviews about an item is simply represented by measures that capture aggregate valence, such as the average star rating and the number of positive or negative reviews. In practice, however, the utilization of reviews by online platforms goes beyond the computation of such simple measures. Specifically, the set of reviews about each item serves as the input to a ranking function. The function then translates the reviews into a score that is used to rank the item among its competitors. As we discuss in detail in Section 4.2, ranking functions are typically proprietary and tend to differ across platforms. However, the underlying theme is that items with better reviews are ranked higher and receive increased *visibility* within the platform. Consider the illustration presented in Figure 1, which demonstrates the typical session of the average user on a review platform, such as TripAdvisor.com

Figure 1 **An Illustration of a Typical User Session: Given the Features Required by the User, the Review-Based Ranking Is Filtered to Eliminate Hotels That Cannot Cover the Requirements**


Filtered ranking
1—The Jefferson
2—Capella
Global review-based ranking
The Jefferson, Washington, DC
599 reviews | #1 of 133 hotels in Washington, DC
(Internet, breakfast, business center)
Capella, Washington, DC, Georgetown
168 reviews | #2 of 133 hotels in Washington, DC
(Internet, pool, business center)
User 2
(business center, Internet)
The Hay-Adams
692 reviews | #3 of 133 hotels in Washington, DC
(pool, business center)
User 1
(Internet, breakfast)
Four Seasons, Washington, DC
509 reviews | #4 of 133 hotels Washington, DC
(Internet, breakfast)
Avenue Suites, Georgetown
732 reviews | #5 of 133 hotels in Washington, DC
(Internet, breakfast)
Filtered ranking
1—The Jefferson
2—Four Seasons
3—Avenue Suites

Note. The filtered ranking is then shown to the user, who considers it prior to making a purchase decision.

or Booking.com. First, the user specifies her requirements with respect to different hotel features, such as Internet access, breakfast, a business center, and a pool. The platform then filters the ranking by eliminating hotels that cannot cover the specified requirements and returns a filtered list of qualified candidates. The user then browses through this ranked list and considers a subset of the ranked items by clicking on them to obtain further information and, ultimately, make a purchase decision. We refer to this subset as the user's *consideration set*.

As illustrated by this example, a hotel's visibility is determined by two factors:

1. *The hotel's position in the review-based ranking*. Intuitively, hotels with a higher ranking are more likely to be considered by the user. This intuition has been verified by numerous studies that have demonstrated the significant impact of ranking on visibility. In Section 4.3, we provide a detailed literature review and describe how the users' tendency to favor hotels with higher rankings is taken into account by our framework.

2. *The set of features that the hotel can cover*. By providing more amenities, a hotel is able to cover the requirements of a larger part of the customer base. The popularity of the provided amenities is also crucial: a hotel that cannot cover a popular requirement (e.g., breakfast) will be eliminated by a large number of users. On the other hand, the lack of a less popular amenity (e.g., a kitchenette or a spa) is less detrimental to a hotel's visibility.

Our work combines these two factors to deliver a fully operationalized definition of visibility with a probabilistic interpretation. Specifically, we define an item's visibility as the probability that it is included in the consideration set of a random user. Formally, the user specifies the query q of features that she requires, where q is a subset of the universe of all possible features $\mathcal{F}$. A query q is sampled with probability $p(q)$ from a categorical distribution with $|2^{\mathcal{F}}|$ possible outcomes, such that $\sum_{q \in 2^{\mathcal{F}}} p(q) = 1$. As illustrated in Figure 1, the website maintains a global review-based ranking $\mathcal{G}_H$ of all of the hotels in the city. Given the user's query q, the platform returns a filtered ranking $\mathcal{G}_H^q$, including only the hotels that can cover all of the features in q. A hotel h is included in the user's consideration set with probability $\Pr(h \mid q)$, which depends on its rank in $\mathcal{G}_H^q$. Finally, we define the visibility of any hotel $h \in H$ as follows:

$$v(h) = \sum_{q \in 2^{\mathcal{F}}} p(q) \times \Pr(h \mid q). \tag{1}$$

The following three sections describe the components of our visibility framework in detail. Section 4.1 describes two methods for estimating the probability $p(q)$ for every possible query of features $q \in 2^{\mathcal{F}}$. Section 4.2 describes two alternative functions for computing the global review-based ranking $\mathcal{G}_H$ of a given set of competitive items H. Finally, Section 4.3 describes three different simulation models for the consideration probability $\Pr(h \mid q)$.

4.1. Using Big Data to Estimate User Preferences

Our definition of visibility considers the probability $p(q)$ that a random customer is interested in a specific query of features q, for every possible $q \in 2^{\mathcal{F}}$. Next, we describe how we can estimate these probabilities from real data. Ideally, an algorithm with access to extensive logs of user queries could be used to learn the required probabilities (Baeza-Yates et al. 2005). In practice however, the sensitive and proprietary nature of such information makes it very hard for firms to share it publicly. This is a typical limitation for research on search behavior (Korolova et al. 2009). Cognizant of this limitation, we present an estimation process based on a resource that is already available for the evaluation of review-based visibility: *customer reviews*.

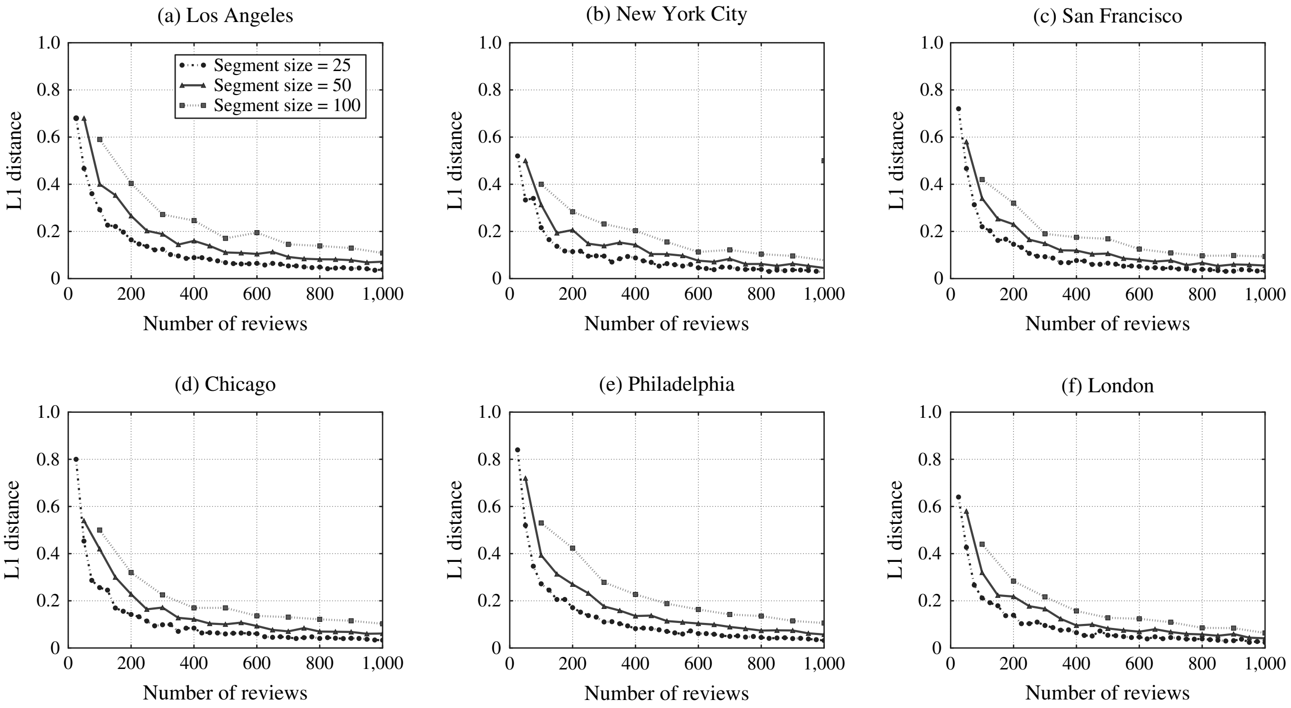
Extant research has validated the process of using reviews to estimate user preferences in multiple domains (Marrese-Taylor et al. 2013, Ghose et al. 2012, Decker and Trusov 2010, Leung et al. 2011). An intuitive approach would be to estimate the demand for each feature separately, and then aggregate the individual estimates at the query level. However, this approach assumes that the occurrence of a feature in a query q is not affected by the other features in q. To avoid this assumption and capture possible correlations, we consider the set of features mentioned in each review as a single query. We then compute the probability of a query q by computing its frequency in our review corpus R, and dividing it by the sum of all query frequencies. Formally

$$p(q) = \frac{\text{freq}(q, R)}{\sum_{q \in 2^{\mathcal{F}}} \text{freq}(q', R)}. \tag{2}$$

Ideally, we would have access to the set of requirements of all possible customers. The maximum likelihood estimate of Equation (2) would then deliver the true probability of any query q. While this type of global access is unrealistic, Equation (2) can still deliver accurate estimates if the number of reviews in R is large enough to accurately represent the customer population. The usefulness of the estimator is thus determined by a simple question: How many reviews do we need to achieve accurate estimates? As we demonstrate next, the unprecedented availability of reviews allows us to answer this question and validate the estimated probabilities.

We conduct our study as follows: first, we merge the reviews from all of the hotels in a city into a single large

Figure 2 Evaluating the Convergence of Query Probability Estimates with Respect to the Number of Reviews



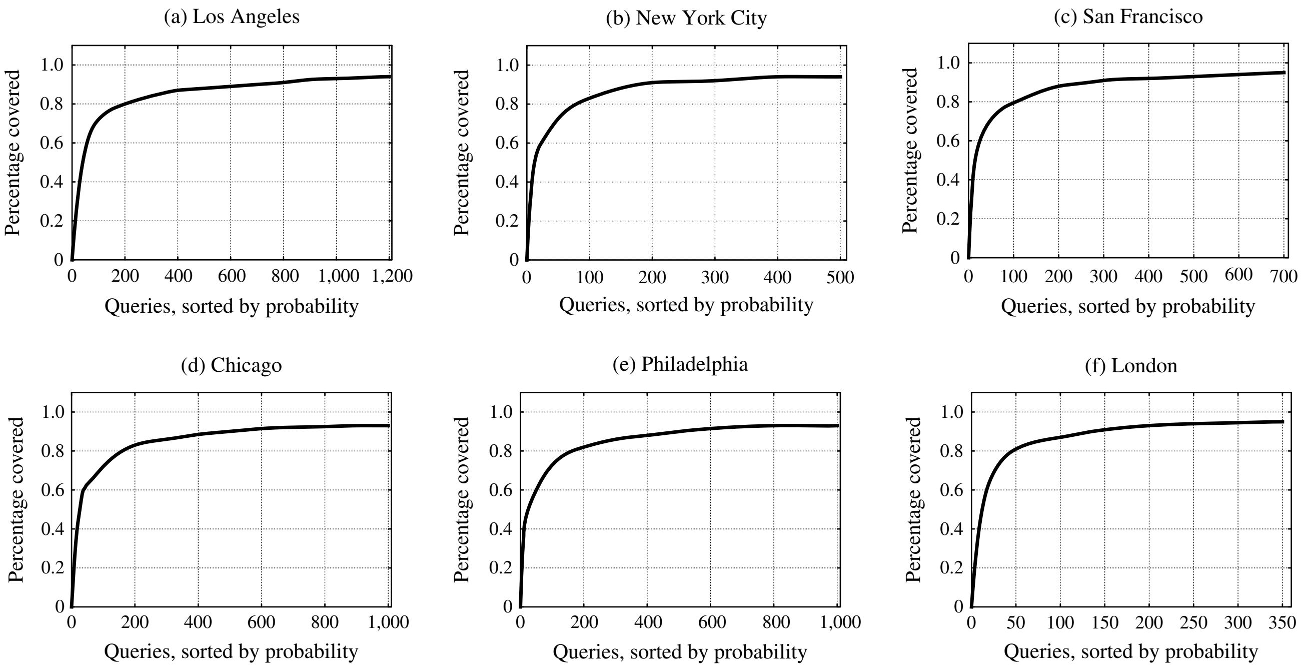
set, sort them by submission date, and split the sorted sequence into fixed-size segments. We then iteratively append segments to the review corpus R considered by Equation (2) and recompute the probability of each query. The vector of probabilities from the ith iteration is compared with that from the $(i-1)$th iteration via the L_1 distance: the sum of the absolute differences of corresponding queries. We repeat the process for segments of 25, 50, and 100 reviews. Figure 2 shows the results for 6 cities. The results for the remaining cities were nearly identical and are omitted for lack of space.

First, we observe near-identical trends for all cities. This is an encouraging finding, which suggests that our conclusions will generalize beyond the cities in our data set. Second, the figures clearly demonstrate the rapid convergence of the probabilities, with the reported L_1 distance dropping to trivial levels below 0.1 after the consideration of less than 1,000 reviews. In fact, 600 reviews were generally sufficient to achieve a distance of around 0.1, for all cities. This trend is consistent for all three segment sizes. As anticipated, adding larger segments is more likely to delay the convergence. However, even for segments of 100 reviews, the observed trend was the same in terms of both the rapid decay and the convergence to a low of around 0.1 after around 1,000 reviews. The rapid convergence of the probabilities is an especially encouraging finding that (i) reveals a stable categorical distribution for the preferences of the users over the various feature queries, and (ii) demonstrates that 1,000 reviews are sufficient to converge to this distribution, a number that is orders of magnitude smaller than the hundreds of thousands of reviews available in our data set.

The consistent underlying distribution motivates us to study its properties and consider possible implications for our work. We begin by plotting the cumulative distribution function (CDF) for each city in Figure 3, in which the queries on the x-axis are sorted in descending order of their respective probabilities. The results reveal that the 500 most popular queries in each city cover about 90% of the distribution. This is an intriguing observation, given that these 500 queries account for only 0.2% of the $2^{18} = 262{,}144$ possible feature queries. This finding has strong implications for the efficiency of our computation: rather than iterating over the exponential number of all possible queries, the results suggest that we can accurately approximate visibility by focusing on a small subset of the most popular queries. As we discuss in detail in Section 6, these queries can also help a business strategically select which features to improve to bolster its visibility. Further investigation reveals a near-normal distribution of the query size, with around 70%–80% of the queries including three–six features. This finding can inform methods for the estimation of query popularity, as it constrains the space of sizes that have to be considered.

4.1.1. Evaluating the Robustness of Our Estimation Method. Our estimation method evaluates features in *sets*, based on their occurrence in customer

Figure 3 The Cumulative Distribution Function of the Categorical Distribution Over the Queries



reviews. As a robustness check, we study an alternative approach that estimates the probabilities of *individual* features and then aggregates them at the query level. For this study, we adopt the utility model proposed by Li et al. (2011), which uses regression analysis to estimate the sensitivity of a random user to each feature.[2] Formally, given N hotels with K features, let D be the $N \times 1$ vector of demands (i.e., hotel bookings), X be a $N \times K$ feature matrix, β be a $N \times 1$ vector of coefficients, and ϵ an i.i.d. random error term. The model can then be written as follows:

$$\ln(D) = X\beta + \epsilon. \tag{3}$$

We take the logarithm of the dependent variable to address skewness in the demand. Given that we do not have access to demand data, we utilize the number of reviews on each hotel as a proxy for D. We populate the feature matrix X with the opinions that the reviewers of each business express on each feature, as described in Section 3. Then, we estimate the coefficient vector β via the standard ordinary least squares (OLS) approach. Conceptually, each coefficient gives us the importance of the corresponding feature for a random user, as captured by the demand proxy. Given that our goal is to estimate query probabilities, we normalize the coefficients in β into a pseudoprobability distribution ξ, such that $\sum_{f \in \mathcal{F}} \xi_f = 1$. Finally, we need to define a formal way of computing the probability $p(q)$ of a given query q, based on the probabilities of the features that it includes. A trivial approach would be to compute the probability of a query q as $p(q) = \prod_{f \in q} \xi_f$. However, this would falsely penalize large queries. Thus, we adopt the following alternative formula:

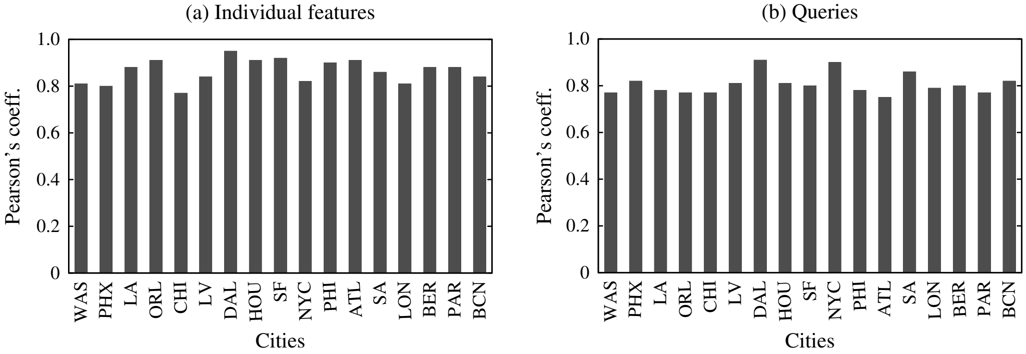
$$p(q) = \psi_{|q|} \times \prod_{f \in q} \xi_f, \tag{4}$$

where $\psi_{|q|}$ is the probability that a random user is interested in exactly $|q|$ features. In other words, ψ is the probability distribution of all of the possible subset sizes (i.e., $\sum_s \psi_s = 1$), which we estimate via the distribution of the number of features that appear in the reviews of our data set.

Let v_1 be the vector of probabilities of the top 500 queries for a city, as ranked by Equation (2) of the review-based method described in the beginning of this section, which considers all of the features in the review as a single query. Then, let v_2 be the corresponding vector computed via Equation (4). We compare the two vectors via Pearson's correlation coefficient (Pearson 1895), a standard measure of the linear correlation between two variables. The measure returns +1 for a perfect positive correlation, 0 for no correlation, and −1 for a perfect negative correlation. We focus on the top 500 queries because, as we discussed earlier in the section, they account for more than 90% of the entire probability distribution. We report the results in Figure 4(b). For completeness, we report the results of the same process at the feature level in Figure 4(a). For this second test, we compare the probabilities of individual

[2] We thank one of the reviewers for suggesting this robustness check.

Figure 4 Comparing the Probabilities Learned by the Two Estimation Methods for Queries and Individual Features



features (rather than queries), as estimated by the two methods.

The figures demonstrate a consistently high similarity between the results of the two methods. Specifically, the reported coefficients were around 0.8 for both the query-level and feature-level comparisons, for all 17 cities in our data set. The correlation was even higher for the feature-level comparison, with most cities reporting near-perfect correlations. Higher values are anticipated for this comparison, as the size of the compared vectors is considerably smaller (i.e., 18 features versus 500 queries for the query-level comparison). A secondary analysis that compares the two vectors via their L_1 distance produced similar results, with the reported values being consistently around 0.2 and 0.15 for the query-level and feature-level comparisons, respectively. Overall, the study validates the results of our estimation method and demonstrates that large review volumes are a valuable source for mining user preferences. This is an encouraging finding with positive implications for the ongoing research efforts in this area, which are often hindered by the absence of query logs and similar data sets from large platforms. Finally, it is important to stress that our visibility framework is compatible with *any* method that can assign a probability to a given query of features.

4.2. Review-Based Ranking

In this section we describe two alternative functions for the computation of the global review-based ranking $\mathcal{G}_H$ of the set of hotels H in a given city. The first function is the standard *average stars* rating, employed by a majority of online portals. Formally, let R_h be the complete set of reviews on a given hotel h. The `Average Rating` formula is then defined as follows:

$$g_{\mathtt{AV}}(h) = \frac{\sum_{r \in R_h} \text{rating}(r)}{|R_h|}. \tag{5}$$

The principal drawback of the $g_{\mathtt{AV}}(\cdot)$ function is that it assigns an equal importance to all of the reviews in R_h. As a result, it maintains the effect of old and possibly outdated reviews. In addition, the function favors older items that have more time to accumulate reviews. In combination with the users' well-documented tendency to prefer items that are already popular (see Section 4.3 for a detailed literature review), this bias leads to rich-get-richer effects and makes it harder for new, high-quality items to rise in the rankings. Motivated by such shortcomings, leading review platforms such as Yelp and TripAdvisor employ proprietary ranking functions that consider the "freshness" of the reviews, in addition to their quantity and quality. A characteristic example is TripAdvisor's *Popularity Index*, which considers the *Quality, Quantity*, and *Age of the reviews* (TripAdvisor 2013). To assess the practical implications of our work in realistic settings, we use the large review data set described in Section 3 to estimate TripAdvisor's Popularity Index. Next, we describe the estimation process and verify the validity of the estimated function via a prediction task.

While we do not have access to the score assigned by TripAdvisor's function to each hotel, our data include the official ranking of all of the hotels in each city, as computed based on their scores. Given the full set of reviews R_h on a hotel h, we first sort the reviews in R_h by their date of submission, from newest to oldest. We then split the sorted sequence into semesters, and define R_h^t as the subset of reviews submitted during the t-th (most recent) semester. The use of semesters delivered the best results in our experimental evaluation, which also considered months, quarters, and years. For each semester R_h^t, we consider the number of reviews that it includes, as well as the corresponding average rating $\bar{R}_h^t$. The `PopularityIndex` of an item h is then computed as follows:

$$\begin{aligned} g_{\mathtt{PI}}(h) = {} & \mathbf{w}_1 \cdot (|R_h^1|, |R_h^2|, \ldots, |R_h^T|, |R_h^{>T}|) \\ & + \mathbf{w}_2 \cdot (\bar{R}_h^1, \bar{R}_h^2, \ldots, \bar{R}_h^T, \bar{R}_h^{>T}), \end{aligned} \tag{6}$$

where $R_h^{>T}$ includes the reviews submitted before the T-th semester. Each interval is mapped to two coefficients: one for the number of reviews and one for the average rating. This formulation allows us to control the number of coefficients that need to be estimated by tuning the value of T. The coefficients in

Table 2 **Computed Coefficients for the $g_{\mathtt{PI}}(\cdot)$ Ranking Function**

First semester		Second semester		Third semester		Older reviews	
$\lvert R_h^1\rvert$	$\bar{R}_h^1$	$\lvert R_h^2\rvert$	$\bar{R}_h^2$	$\lvert R_h^3\rvert$	$\bar{R}_h^3$	$\lvert R_h^{>3}\rvert$	$\bar{R}_h^{>3}$
−0.0003	−2.647	−0.0004	−0.532	−0.0004	−0.043	−0.0002	−0.008

the $\mathbf{w}_1$ vector represent the importance of the number of reviews submitted during each semester. Similarly, the coefficients in $\mathbf{w}_2$ represent the importance of the corresponding average rating. From an optimization perspective, $\mathbf{w}_1$ and $\mathbf{w}_2$ constitute a single vector of coefficients that needs to be estimated to minimize the difference between the ranking of our $g_{\mathtt{PI}}(\cdot)$ function and the official TripAdvisor ranking. Given two sets of rankings of hotels H in a city, a pair of hotels (h_i, h_j) is *concordant* if both rankings agree in the relative order of h_i and h_j. Given the corresponding hotel sets $H_1, H_2, \ldots, H_m$ from the m cities in our data set, the goal is to estimate a ranking function $g_{\mathtt{PI}}(\cdot)$ that maximizes the average percentage of concordant pairs per city. As shown by Joachims (2002), this problem can be formulated as a constrained optimization task and solved by a linear support vector machine (SVM). We complete the optimization and obtain the coefficients using `SVM-rank`,[3] the state of the art for training ranking SVMs. We set the number of semesters T via a 10-fold cross validation. On each of the 10 folds, 90% of the cities in our data set were used to learn the coefficients, and predict the rankings for the remaining 10%. The best results were achieved for $T = 3$, for which 91% of all possible hotel pairs were reported as concordant. Increasing the value of T led to near-zero values for the additional coefficients and did not improve the results. Table 2 holds the coefficients for $T = 3$.

The high percentage of concordant pairs (91%) verifies that our `PopularityIndex` function can accurately reproduce TripAdvisor's ranking. All of the computed coefficients were negative, as anticipated, because lower rankings are desirable. We also observe a declining importance of the average rating for older semesters, a finding that verifies that the function favors recent reviews. Finally, the coefficients of the number of reviews in each semester were consistently around −0.0003, which represents the marginal contribution of every additional review to the function.

4.3. Computing the Consideration Probabilities

In this section, we describe three alternative models for the consideration probability $\Pr(h \mid q)$, the probability that a user with a query q includes a specific business h in her consideration set. Our models are motivated by the extensive work on the impact of ranking on visibility. In the hotels domain, Ghose et al. (2014) reported an average click-through rate (CTR) increase of 10.07% for a one-position improvement in rank. In the Web apps domain, Carare (2012) found that an app's status as a best seller is a highly significant determinant of demand. In the same domain, Ifrach and Johari (2014) report that the top-ranked app received 90% more downloads than the one ranked in the 20th position. Best sellers were also considered by Tucker and Zhang (2011), who found that sales-based rankings have a significant impact on user behavior, and lead to a rich-get-richer effect.

Furthermore, ranking effects have been extensively studied in the context of search engines. Pan et al. (2007) found that users strongly favor high-ranked results, even if they appear to be less relevant than lower-ranked alternatives. In a later study, they found that the CTR for hotels follows a power-law distribution as a function of their rank on popular search engines (Pan 2015). Brooks (2004) observed that an ad's click potential and conversion rate can drop significantly as its search engine rank deteriorates. He reported a nearly 90% drop between the first and 10th position, although the reductions between consecutive positions did not follow a predictable pattern. Henzinger (2007) found that the majority of search-engine users do not look beyond the first three pages of the ranked results. In their eye-tracking studies, Guan and Cutrell (2007) found that users browse through ranked items in a linear fashion, from top to bottom. They also found that users typically look at the first three to four options, even if they do not ultimately select one of them. Similar findings were reported by Joachims et al. (2005) and Lorigo et al. (2008).

The extensive amount of relevant work reveals three interesting findings: (i) there is a strong causal relationship between an item's position in a ranked list and its visibility; (ii) users tend to focus on a small set of top-ranked items that monopolize their attention; and (iii) the diversity in user behavior makes flexibility essential for modeling the connection between rank and visibility, as it emerges in different settings. These three observations inform our models for the computation of the consideration probabilities, which we describe in the remainder of this section.

Let $\mathcal{G}_H$ be the review-based ranking of a set of hotels H, and let $\mathcal{G}_H^q$ be a filtered version of the ranking, which includes only the hotels that cover all of the features in a query q. The user browses through this ranked list and considers a subset of the ranked items by clicking on them to obtain further information. The inclusion of a hotel into the user's consideration set is modeled via a random Bernoulli variable, which takes the value 1 with probability $\Pr(h \mid q)$, computed as a function of the hotel's rank in $\mathcal{G}_H^q$. We consider three alternative consideration models, which we refer to as `Linear`, `Exponential`, and `Stretched Exponential`.

[3] http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html.

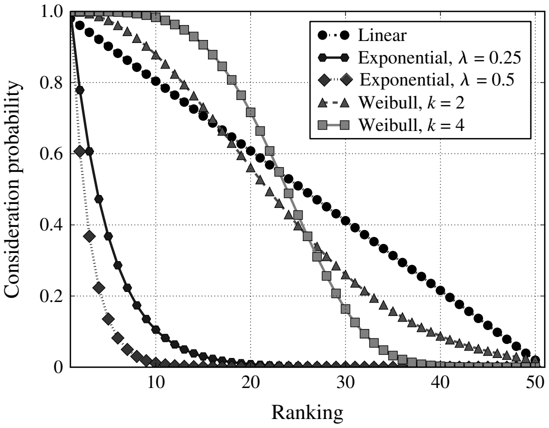
Table 3 **Alternative Models for the Consideration Probability as a Function of an Item's Position x in a Ranked List with N Items**

Title	Formula
`Linear`	$\Pr(x) = 1 - \dfrac{x-1}{N+1}$
`Exponential`	$\Pr(x) = e^{-\lambda x}$
`Stretched Exponential`	$\Pr(x) = e^{-(2(x-1)/N)^k}$

Table 3 includes the formula for the consideration probability of an item ranked in the xth position of a list of N items, for all three models. Figure 5 then illustrates the computed probabilities for a ranked list of 50 items.

The `Linear` model is a parameter-free model that assigns a consideration probability proportional to the item's rank. The `Exponential` model introduces the standard rate parameter λ of the exponential distribution and applies an exponential decay to the consideration probability as the rank increases (becomes worse). This is a strict model that greatly favors top-ranked items. This pattern has been repeatedly reported by previous work, as we described in detail earlier in this section. Figure 5 demonstrates that the strictness of the model can be tuned by increasing the value of the rate parameter to expedite the decay. Finally, the `Stretched Exponential` model follows a reverse sigmoid shape, controlled by a parameter k. As shown in the figure, increasing the value of k from 2 to 4 delays the decay and assigns a high probability to a larger number of top-k items. In practice, the `Stretched Exponential` model allows us to expand the bias that favors top-ranked items, while maintaining an exponential decay for the others. This is a particularly useful property because, as discussed earlier in this section, the exact number of top-ranked items that monopolize the users' interest has been shown to differ across studies and domains.

Figure 5 **Consideration Probability for the Top 50 Positions, for Each of Our Three Models**



The `Exponential` and `Stretched Exponential` models inherit the parameters of the distributions that they are based on. The interested practitioner can intuitively tune these parameters to simulate different types of consideration patterns. In practice, however, review platforms have access to extensive user logs of search-and-click data, that can be used to *learn* the most appropriate consideration model for each context. Previous work has presented a number of effective algorithms for this task (Agichtein et al. 2006, Joachims 2002, Chapelle and Zhang 2009). It is important to stress that our framework has no dependency on distribution-based models and their parameters. Instead, it is compatible with *any* consideration model that assigns a probability to each position in the ranking, including models that can be learned from real search data.

Finally, while higher rankings are typically connected with increased consideration, it is impossible to exclude scenarios that are not significantly influenced by rankings. For instance, consider a market with a very small number of competitive items, or a highly demanding consumer whose multiple constraints limit the consideration set to a trivial size. In such cases, the user is likely to consider all available items, regardless of their rank. While we acknowledge such possibilities, our goal is to formalize the effect of ranking in scenarios that include a nontrivial number of alternatives and do not allow for an exhaustive evaluation. The prevalence of such scenarios in modern markets has been acknowledged by relevant literature, as well as by virtually all major review platforms, which develop and utilize ranking systems to improve their users' experience.

5. Attack Strategies

Based on our definition of visibility, hotels that are ranked higher in the global review-based ranking have a higher probability of being added to the consideration set of a potential customer. Therefore, by injecting fake reviews and manipulating the ranking, a competitor can directly affect a hotel's visibility. To achieve this, the attacker can follow one of three strategies:

- `Self-Pos`: self-inject positive reviews into its own review set.
- `Target-Neg`: inject negative reviews into the review set of a competitor.
- `Mixed`: a mixed strategy that combines both positive and negative injections.

The common goal of all three strategies is to increase the attacker's visibility at the expense of the visibility of its competitors. This raises the issue of measuring a hotel's vulnerability to such attacks and motivates the following question:

Research Question. Let H be a set of competitive hotels. Then, given any two hotels $h, h' \in H$, how vulnerable is h to a review-injection attack by h'?

If a competitor could inject an infinite number of reviews, then a hotel's visibility could be manipulated at will. In practice, however, the probability of being detected increases with the volume of injected reviews. This is because of the resulting traffic patterns (Mukherjee et al. 2012, Xie et al. 2012), as well as due to the increasing complexity of creating fake reviews that are both realistic and dissimilar enough to avoid detection (Jindal et al. 2010, Lappas 2012). Therefore, the attacker would reasonably want to surpass a competitor with the *minimum* number of injections. In the context of our research question, we thus have to compute the smallest possible set of injections that enables a hotel h' to surpass a competitor h, for each of the three strategies. An injection attack can be described in terms of the two alternative review-based ranking functions described in Section 4.2: the Average Rating function $g_{\mathtt{AV}}(\cdot)$ and the PopularityIndex function $g_{\mathtt{PI}}(\cdot)$. Formally, let R_h be the set of reviews on a given hotel h and let $\mathcal{I}$ be a set of injected reviews. Then, considering Equation (5), the updated Average Rating $g^*_{\mathtt{AV}}(h)$ can be computed as

$$g^*_{\mathtt{AV}}(h) = \frac{\sum_{r \in R_h \cup \mathcal{I}} \text{rating}(r)}{|R_h| + |\mathcal{I}|}. \tag{7}$$

With respect to the PopularityIndex function defined in Equation (6), the key observation is that all injected reviews will be more recent than all preexisting reviews for the hotel. Therefore, all of the injections will be effectively applied to R_h^1, which is the first (most recent) semester of reviews considered by the function. Formally, the updated $g^*_{\mathtt{PI}}(h)$ can be computed as follows:

$$\begin{aligned} g^*_{\mathtt{PI}}(h) = {}& \mathbf{w}_1 \cdot (|R_h^1| + |\mathcal{I}|, |R_h^2|, \ldots, |R_h^T|, |R_h^{>T}|) \\ & + \mathbf{w}_2 \cdot (\bar{R}_h^1 \cup \mathcal{I}, \bar{R}_h^2, \ldots, \bar{R}_h^T, \bar{R}_h^{>T}). \end{aligned} \tag{8}$$

Given Equations (7) and (8), an attacker h' can optimally implement the Self-Pos strategy by greedily self-injecting positive reviews until its visibility surpasses that of the target h. Similarly, the practice of greedily injecting negative reviews to h, until its visibility falls below that of h', is optimal for the Target-Neg strategy. It can be trivially shown that the greedy approach is not optimal for the Mixed strategy. In the online appendix, we include an algorithm for computing an optimal solution that is much faster than naively evaluating all possible combinations of positive and negative injections. We use this algorithm for the experiments of this section.

We design our experiments as follows. First, we rank the hotels in each city in descending order of their individual visibilities. For every hotel h' in the city, we compute the *target set* $\{h: v(h) \geq v(h')\}$ of all hotels with an equal or higher visibility. Then, for each attack strategy, we compute the minimum number of injections that h' needs to become more visible than each target. We repeat the process for all cities, all of the hotels in each city, the two ranking functions described in Section 4.2, and the three consideration models described in Section 4.3. For the Exponential and Stretched Exponential consideration models, we set $\lambda = 1$ and $k = 2$, respectively. We omit the results for the Linear model, as they were similar to those of Stretched Exponential and did not deliver additional insights. Finally, we focus our discussion on the results from New York City, San Francisco, and Los Angeles. The results with different parameter settings and the remaining cities were also similar and are omitted for lack of space.

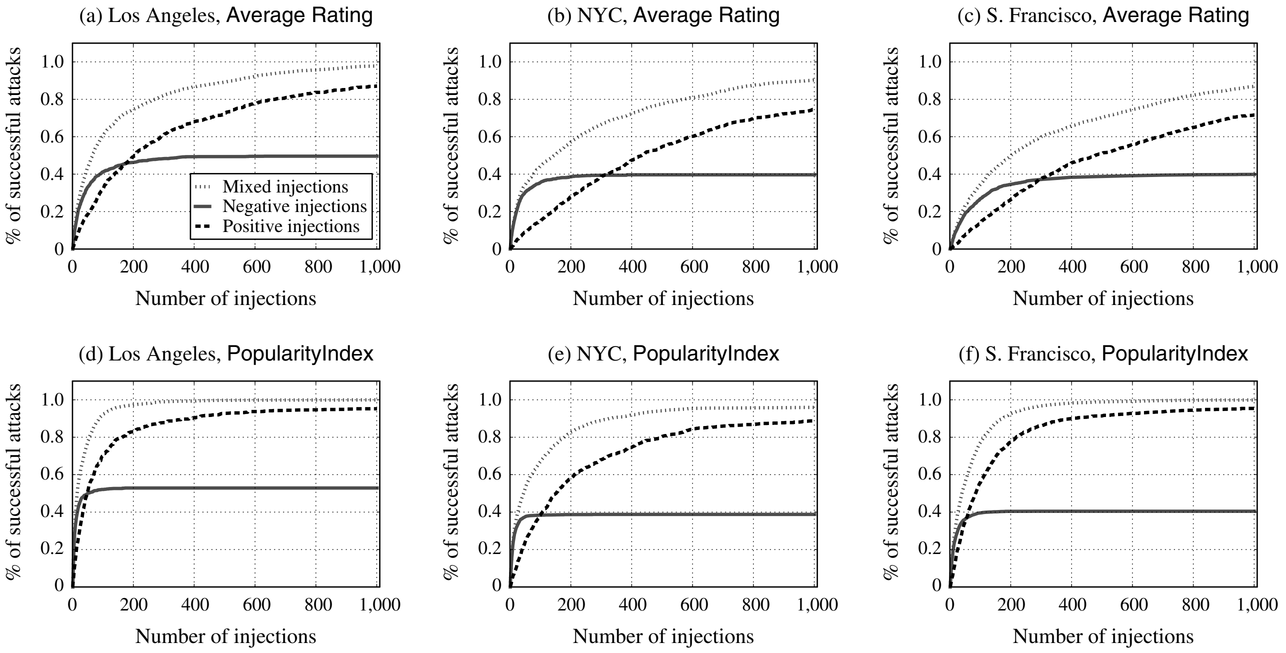
5.1. Measuring the Effectiveness of the Three Strategies

First, we evaluate the percentage of all possible attacks (for all hotels and their respective targets) that can be completed for increasing numbers of review injections. Figure 6 shows the results for the Average Rating and PopularityIndex ranking functions, under the Exponential consideration model. Figure 7 shows the respective results for Stretched Exponential.

The figures reveal several findings. First, we observe similar injection effects for all three cities, within each combination of ranking functions and consideration models. This demonstrates the influence of these two factors and exposes a generalized vulnerability to attacks that manipulate an item's ranking and consideration probability. Furthermore, a series of noteworthy findings and respective implications comes from the differences observed among the three attack strategies and between the two ranking functions. Next, we discuss these findings in more detail.

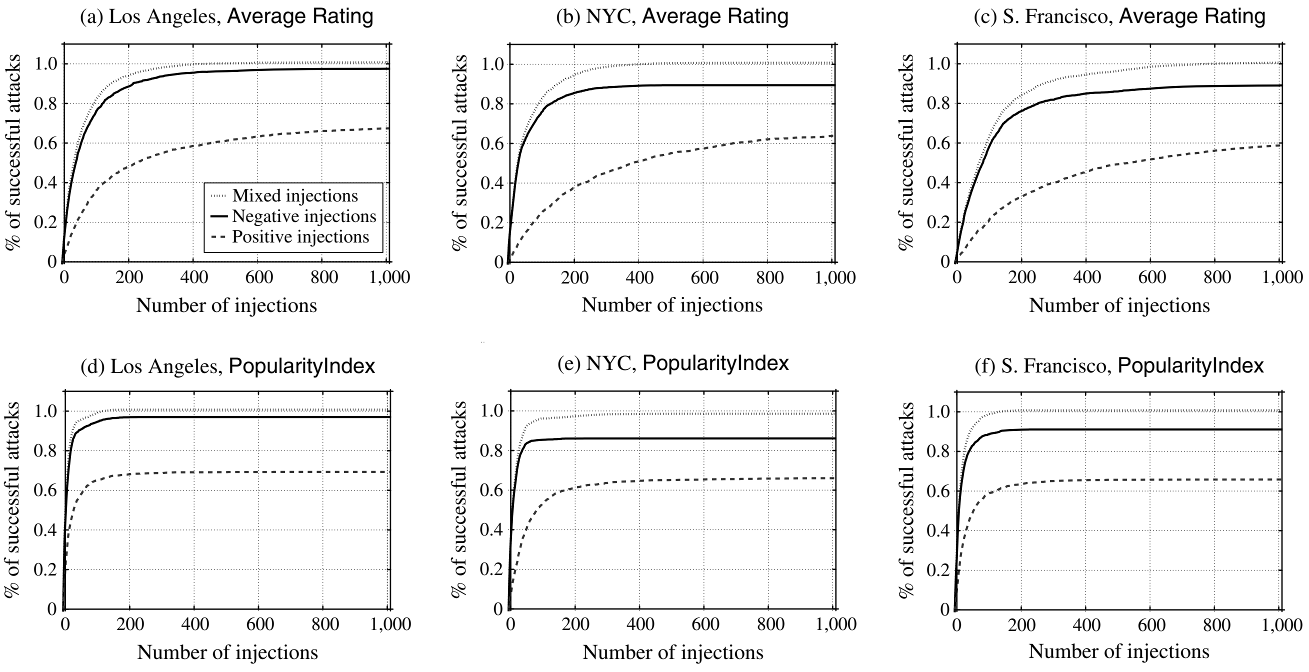
Are hotels more vulnerable to positive or negative reviews? As anticipated, the Mixed strategy consistently demonstrates its ability to successfully complete an attack with less injections than the other two approaches. A less anticipated observation is the role reversal of the Self-Pos and Target-Neg strategies with respect to the two consideration models. For the Exponential model, the Target-Neg strategy fails to complete more than 40% of the attacks, even after 1,000 injections. On the other hand, the percentage of Self-Pos increases consistently, reaching a value around 70% for 1,000 injections. Surprisingly, the roles are reversed for the Stretched Exponential model: Self-Pos converges after completing about 60% of the attacks, while Target-Neg achieves percentages as high as 90%. In addition, the reversal is observed for both the Average Rating and PopularityIndex ranking functions. As we discuss next, this finding is the result of the different way in which the two models distribute user consideration.

Figure 6 **Percentage of Successful Injection Attacks for the** `Average Rating` **and** `PopularityIndex` **Ranking Functions, Under the** `Exponential` **Consideration Model with** $\lambda = 1$



The `Exponential` model limits the user's consideration to a very small group of top-ranked hotels. The self-injection of positive reviews is effective in this setting, as it allows the hotel to eventually enter the top listings and dramatically increase its consideration probability. On the other hand, `Target-Neg` is only effective in a minority of cases when both the attacker and the target are ranked in top positions with a nontrivial consideration probability. Conceptually, an attacker cannot achieve significant visibility earnings by attacking a target that already has a zero or near-zero visibility. On the other hand, the stretched decay of the `Stretched Exponential` model allocates a nontrivial consideration probability to a much larger

Figure 7 **Percentage of Successful Injection Attacks for the** `Average Rating` **and** `PopularityIndex` **Ranking Functions, Under the** `Stretched Exponential` **Consideration Model with** $k = 2$



number of hotels. This favors negative injections, as they can benefit the attacker by both increasing its own visibility and reducing that of the target, as long as the number of injections is enough to surpass the target in the review-based ranking. We revisit these findings and their implications for businesses in Section 7.

Which ranking function is more vulnerable to fake reviews? Without loss of generality, we focus our study on the Mixed strategy, which computes the optimal (smallest) number of injections required for each attack and thus allows us to objectively compare the vulnerability of the two functions. First, we examine the results for the Exponential model, shown in Figure 6. For New York City, 200 injections were enough to manipulate the PopularityIndex ranking and complete 80% of all possible attacks. On the other hand, for the Average Rating ranking, the same number of injections was sufficient for only 60% of the attacks. For Los Angeles, just 50 reviews were sufficient for 80% of the attacks for the PopularityIndex. On the other hand, four times as many reviews were required to achieve the same percentage under the Average Rating.

Furthermore, for the PopularityIndex, 450–500 injections were enough to eventually complete nearly 100% of the attacks across cities. The number was even smaller for San Francisco and Los Angeles. By contrast, about 700 injections were required to reach 100% in Los Angeles under the Average Rating. In fact, this percentage was unattainable even after 1,000 injections for San Francisco and New York City. Finally, as demonstrated by the steep curves for the PopularityIndex, its vulnerability is even larger for small injection numbers (<50). This is a critical observation, as small attacks are more likely to occur and avoid detection in realistic settings (Xie et al. 2012, Fei et al. 2013). Figure 7 motivates similar observations for the Stretched Exponential consideration model, with the curves of the Mixed strategy exhibiting a steeper rise for the PopularityIndex than for the Average Rating. In addition, less than 50 reviews were enough to manipulate the PopularityIndex and complete nearly 100% of the attacks. On the other hand, four times as many reviews were required to achieve similar results for the Average Rating.

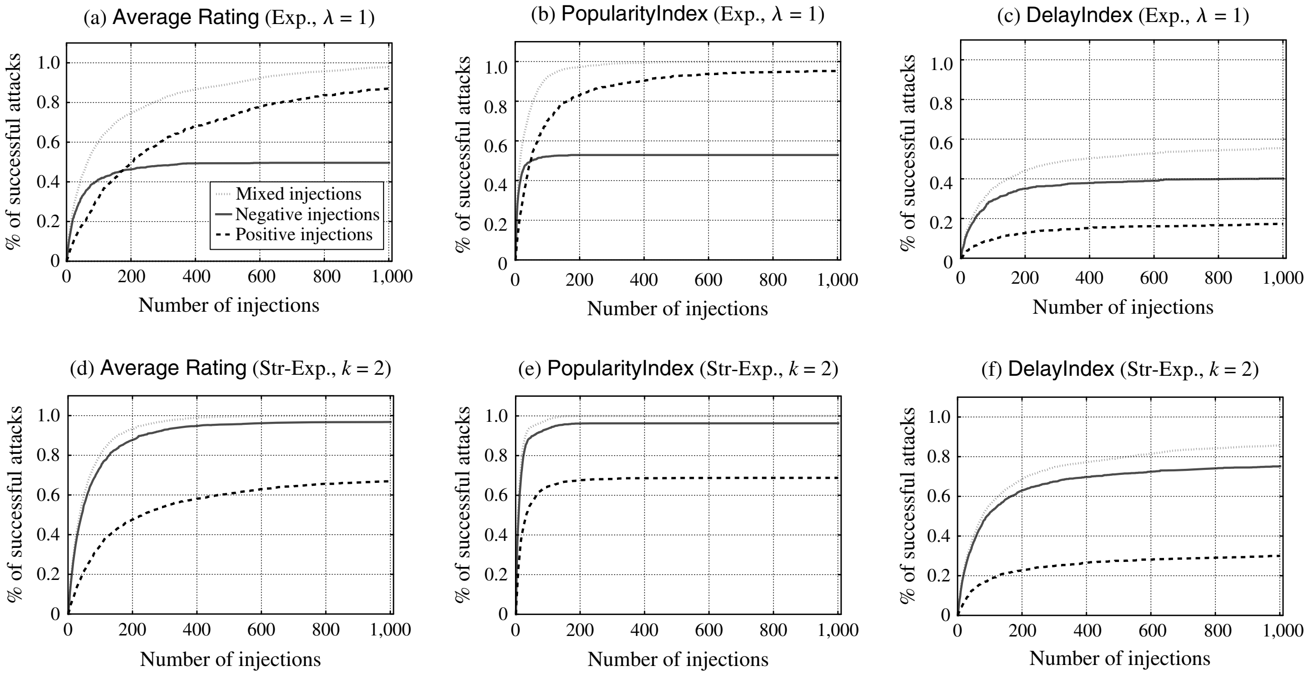
The results demonstrate that the PopularityIndex is consistently more vulnerable to fake reviews than the Average Rating. This is a surprising finding, as the PopularityIndex was designed to address the shortcomings of the Average Rating, as discussed in Section 4.2. However, a careful examination of the function reveals the cause of its increased vulnerability. By definition, the PopularityIndex assigns increased importance to the average rating of the most recent batch of reviews. In fact, as shown in Table 2, the influence of each review declines with time. While this practice can eliminate outdated information, it also favors fake reviews. This bias has two causes. First, fake reviews can only be injected in the present and, therefore, have the highest possible contemporary influence among the reviews of the injected hotel. Second, by reducing the importance of older reviews, the PopularityIndex effectively reduces the number of reviews that contribute to the final score and makes the score more vulnerable to manipulation. As an example, consider a hotel with an average rating of four stars and 1,000 reviews. To reduce the rating to three stars, we would need to inject 500 one-star reviews. Now, suppose that we consider only the 100 most recent hotel reviews, and the average of these 100 reviews is again four stars. In this setting, just 50 one-star reviews are sufficient to reduce the rating to three stars.

5.2. Designing Fraud-Resistant Ranking Functions

The policy of favoring recent reviews in review-based rankings is adopted by industry leaders such as TripAdvisor and Yelp (Kamerer 2014, Mukherjee et al. 2013b). Therefore, the vulnerability of the PopularityIndex has major implications for these platforms and their users. Even though the goal of this policy is to eliminate outdated information, it is also contrary to standard quality principles for user-generated content, such as those found in large crowdsourcing communities. For instance, on Wikipedia, new contributions to a page are monitored and evaluated by users who have placed the page on their *watch list* (Viégas et al. 2004). The community can thus promptly identify and eliminate false or malicious edits before they can be seen by a large number of users.

The watch list paradigm motivates us to propose the DelayIndex ranking function, which maintains the benefits of the PopularityIndex while significantly reducing its vulnerability to fake reviews. As described in detail in Section 4.2, the PopularityIndex first aggregates a hotel's reviews into groups (e.g., semesters) according to their date of submission. It then assigns increasingly high coefficients to more recent groups. The DelayIndex function takes the same approach, except that it swaps the coefficients of the most recent group with those of the second most recent. This reduces the effect of freshly injected reviews by assigning them to the second level of importance. In addition, this introduces a delay period before a newly injected fake review reaches the semester with the highest coefficient. This gives the platform time to identify and eliminate fake reviews and allows the attacked business to respond to such reviews before they can significantly impact its visibility. We show the results of the DelayIndex function for the Exponential and Stretched Exponential consideration models in Figure 8. For lack of space, we present the results for Los Angeles. The results for the other cities were similar and did not lead to additional findings.

Figure 8 **Percentage of Attacks that can be Successfully Completed for Increasing Numbers of Review Injections, for the** Average Rating, PopularityIndex, **and** DelayIndex **Ranking Functions Under the** Exponential **(a–c) and** Stretched Exponential **(d–f) Consideration Models (LA Data)**



The figures show that the DelayIndex is consistently less vulnerable than the other two functions, for both consideration models. In fact, under this function, the Mixed strategy completed only 50% and 70% of the possible attacks for the Exponential and Stretched Exponential consideration models, respectively, even after 1,000 injections. These promising results provide insights on how more robust ranking functions can be designed to protect visibility from fake review injections. We discuss possible research directions for further improvement in Section 7.

5.3. Accounting for the Effects of Defense Mechanisms

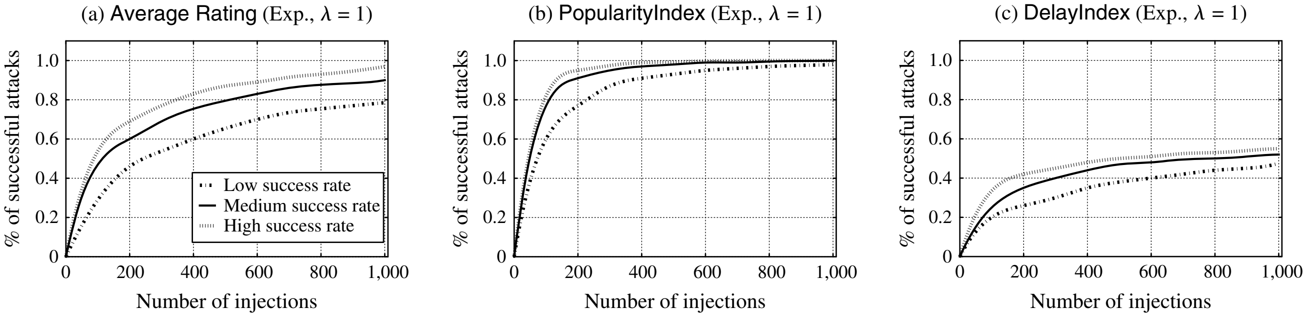
As we discussed in detail in Section 2, review platforms utilize different types of defenses against fake reviews. The results that we presented in Section 5.2 do not account for the effects of such mechanisms, which could prevent at least a percentage of the attempted injections. Given an accurate estimate of this percentage, it would be trivial to adjust the expected number of successful attacks in our experiments and update our figures accordingly. However, as we discuss next, this estimation task is ambiguous and far from trivial. In fact, the correct percentage is likely to differ across platforms. Therefore, rather than trying to find a percentage that would only be meaningful in the context of our data set, we present an analysis of the estimation task and the challenges that it presents, as well as of the scenario where business managers personally dispute fake reviews.

In theory, we can compute the percentage of identified fake reviews as the True Positives/(True Positives + False Negatives) ratio. However, even in the most optimistic scenario, we can only have access to the set of reviews that have already been marked as fake by the platform. Unfortunately, the true positives (i.e., fake reviews marked by the platform) in such a data set would be mixed with false positives (i.e., real reviews that were falsely marked as fake) and could not be untangled. In addition, the set of false negatives is unattainable, given that the platform cannot be aware of fake reviews that were missed by its filters. The only way to obtain such information would be to execute a strategically designed barrage of injection attacks on the website. However, this would be a direct violation of the platform's terms and is not an acceptable option.

5.3.1. Simulating Defense Levels. To address this limitation, we consider hypothetical scenarios involving low, medium, and high injection success rates (ISRs) as a result of screening mechanisms used by platforms.[4] Specifically, we evaluate the cases where only 25% (low), 50% (medium), and 75% (high) of all injections are completed, while the rest are prevented. We then compute the percentage of successful attacks that are possible given each ISR, for different numbers of attempted injections. We present the results

[4] We thank one of the reviewers for suggesting this experiment.

Figure 9 Percentage of Successful Injection Attacks with Low, Medium, and High Injection Success Rates Under the `Exponential` Consideration Model (LA Data)

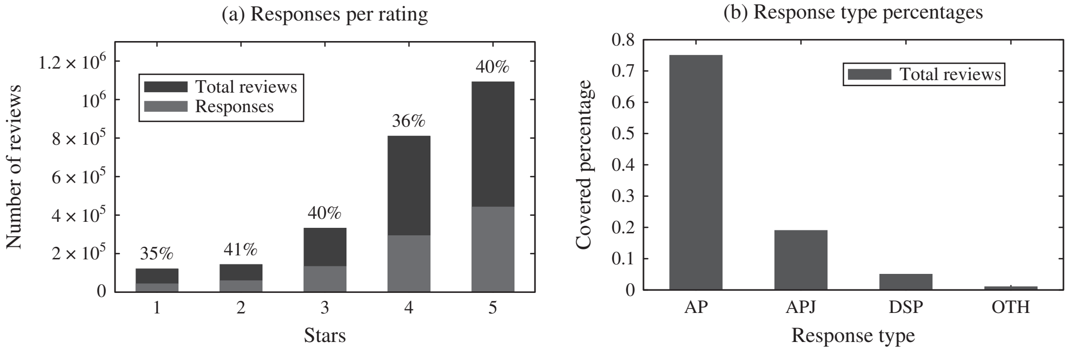


for all three ranking functions in Figure 9. For lack of space, we only include the results of the superior `Mixed` attack strategy under the `Exponential` consideration model on the Los Angeles data set. The results for the other models and cities lead to similar findings, consistent with the trends that we observed in our previous experiments.

The figure verifies the vulnerability of the `PopularityIndex`: even for a low success rate, 50 injection attempts were enough to complete 40% of the attacks, while 100 injections led to a 60% completion rate. These rates were about two times those yielded by the `Average Rating` function and almost three times those yielded by the `DelayIndex` for the same ISR and number of attempts. In fact, 400 attempts were enough to complete more than 90% of *all* possible attacks under the `PopularityIndex`, regardless of the ISR. Another interesting finding is that, while lower ISRs predictably reduce the effect of attacks, it remains significant. For instance, at a 50% ISR, 50 attempts were enough to complete 50% and 30% of all possible attacks for the `PopularityIndex` and `Average Rating` functions, respectively. This highlights the need for robust ranking functions and vigilant policies against fake reviews. A platform's effort to strengthen its defenses and reduce the success rate of injection attacks is likely to require valuable resources. Therefore, this type of simulation can be used to evaluate different options and conduct the cost-benefit analysis.

5.3.2. Responding to Reviews. Because of the inherent difficulties in estimating the percentage of detected fake reviews, we focus on an alternative measure of the platform's vigilance. On large review platforms, such as Yelp or TripAdvisor, businesses can respond to reviews posted by their customers. Relevant research has consistently verified the benefits of this practice (Cheng and Loi 2014, Ye et al. 2008, Barsky and Frame 2009, Avant 2013, Xie et al. 2014). Our TripAdvisor data set includes 969,469 such responses. Figure 10(a) shows the distribution of the reviews over the possible star ratings, as well as the number of reviews in each rating that received a response. We observe that the number of reviews that received a response is consistently between 35% and 41%, for all five ratings. Given that we are interested in the vigilance of businesses with respect to defamatory comments, we focus on the 101,759 one-star and two-star reviews in our data. Figure 10(a) suggests that around two-thirds of all negative reviews do not receive a response. While this could be an alarming finding, we cannot confidently attribute it to lack of vigilance, as the true percentage of fake reviews is unattainable. To put this

Figure 10 Information on the Types and Distribution of Responses Submitted by Businesses to Online Reviews



finding into context and gain a deeper understanding of the types of responses, we adopt a topic-modeling approach, based on the highly cited Latent Dirichlet Allocation (LDA) (Blei et al. 2003) model. LDA models the generation of each term in a response d as follows: first, the responder samples a topic i from a document-specific distribution θ_d. A term is then sampled from a topic-specific distribution ϕ_i. The document is thus modeled as a mixture of topics, which, in turn, are modeled as term distributions. Given the number of topics to be learned k, the ϕ and θ distributions are estimated via Gibbs sampling (Blei et al. 2003). We set k via a 10-fold cross validation with the perplexity measure, which has been consistently used for the evaluation of topic models (Blei et al. 2003, Asuncion et al. 2009). After experimenting with $k \in \{10, 20, 30, 40, 50, 60, 70\}$, we found that $k = 50$ delivered the lowest perplexity.

The next step is to determine the context of each of the 50 learned topics. Given that each response is modeled as a mixture of topics, we achieve this by focusing on *pure* responses that include a single dominant topic in their distribution. Formally, let θ_{di} be the probability of topic i in the mixture of response d. Then, i is the dominant topic of d if $\theta_{di} > 0.8$ or, equivalently, if the response is 80% about topic T. Then, for each topic i, we sample 20 responses with $\theta_{di} > 0.8$. We choose 80% as the threshold, as it is the highest value that delivers at least 20 responses for all 50 topics. An examination of the data reveals at least 3 response types: (i) apologies without a justification, posted simply to placate the reviewer; (ii) apologies with a justification, in which the responder offers an explanation for the customer's unpleasant experience; and (iii) unapologetic responses, posted to challenge the customer's opinions. Table 4 includes an example of each type.

To verify this composition and measure the prevalence of each type in our corpus, we asked five annotators to assign one of the following four labels to each topic, based on its 20 associated responses: AP: *apology without justification*, APJ: *apology with justification*, DSP: *dispute*, or OTH: *other*. Note that the annotation task is trivial, due to the purity of the sampled responses. Therefore the number of disagreements was minimal and a label with at least four votes was reported for all 50 topics. The process revealed 35, 11, 3, and 1 topics of type AP, APJ, DSP, and OTH, respectively. Simply reporting the number of topics for each type would be incomplete, since some topics are more prevalent than others in the corpus. Therefore, we also compute the percentage $p(i) = (\sum_d \theta_{di})/(\sum_d \sum_{i'} \theta_{di'})$ covered by each topic i and report the cumulative percentage for each of the four response types in Figure 10(b). We observe that 75% of all responses are simply apologetic and make no effort to challenge negative reviews. An additional 19% combined an apology with a justification, while only 5% disputed the reviews directly.

Our study is inherently limited, given that we cannot know which reviews are fake and should have been disputed. Nevertheless, the results reveal that even businesses that are willing to monitor and respond to their reviews are unlikely to be confrontational and dispute reviews. We identify two alternative explanations for this finding. First, a confrontational response might demonstrate that the business is insensitive to the comments of its customers and unwilling to acknowledge its faults, thus damaging its brand. Second, given that reviews are inherently subjective, it is easy for an attacker to write reviews that are hard to dispute as fake. For instance, a business can only respond to generic statements such as "I really didn't like their breakfast" or "The rooms were too small, I would never go back again" with apologies, justifications, or promises for improvement.

In conclusion, we find that (i) around two-thirds of all negative reviews do not receive responses from businesses and (ii) only 5% of the responses are actually disputes. Thus, despite the well-documented benefits of engaging the customers and responding to criticism, this practice is rarely used for disputing fake reviews. As we discuss in Section 7, the detailed interpretation of these findings suggests a promising direction for future research.

Table 4 **Examples of Response Types for Negative Reviews**

Type	Example response
AP	*Thank you for your feedback. We were sorry for the impression your visit left you with, your comments have been forwarded to management for improvement. We are confident that they will work diligently to prevent a future recurrence. We truly appreciate your business and hope to see you on a visit in the future.*
APJ	*Upon check-in the guest was advised that housekeeping was finishing up with cleaning the rooms due to being sold out the previous night, and was welcomed to wait in the lobby and housekeeping would let them know as soon as the room was ready. We do apologize about this inconvenience.*
DSP	*After looking into these comments we can find no trace or evidence from staff or records kept that this is a genuine review. It is also deeply suspicious as all of our singles, doubles, and twins have brand new carpets and curtains throughout. According to the vast majority of reviews we are noted for our cleanliness and charm and that speaks volumes.*

6. Response Strategies

In this section, we focus on response strategies that address the following question: How can a business protect its visibility from fake-review attacks? We identify two different types of strategies: those based on *enhancement* and those based on *confrontation*. Enhancement strategies focus on improving the attacked hotel's qualities, to increase its visibility and overcome the losses resulting from an attack. On the other hand, confrontational strategies focus on reversing the results of an attack, by identifying and disputing the injected reviews.

6.1. Enhancement Strategies

An enhancement strategy focuses on improving a hotel, to increase its visibility and compensate for the losses caused by an attack. Such a strategy can also be applied preemptively, to safeguard the hotel's visibility or gain an advantage over its competitors. Next, we discuss how a hotel can increase its visibility by (i) improving its position in the review-based ranking, (ii) covering additional features, and (iii) improving the quality of existing features.

6.1.1. Rising Through the Ranks. To improve its position in the ranking, a business needs to attract positive reviews. If the firm is confident in its own quality, then the need to attract positive reviews translates into a marketing effort. In fact, the connection between promotional campaigns and the consequent arrival of new reviews has been verified by previous work (Byers et al. 2012a, b). By using our framework to simulate a positive-injection attack, as described in Section 5, a business can estimate the number of positive reviews required to surpass any competitor and set its marketing goals accordingly. We note that a firm's rank would also improve if its competitors receive enough negative reviews to drop below the firm in the ranking. However, given that the firm cannot influence such a development without violating ethical principles and the terms and conditions of the review platform, we do not consider this approach.

6.1.2. Covering More Features. The features that a hotel can cover determine an upper bound for its visibility. If a firm had infinite resources, it could maximize its coverage by simply enhancing the hotel to cover all possible features. In practice, however, the process of improving a hotel by adding new features requires significant financial and human capital. In addition, it is reasonable to anticipate a variance in the costs required to cover additional features. For example, adding free wi-fi is arguably much cheaper and easier than adding a pool. The task of evaluating and incorporating new features is a nontrivial process that needs to consider multiple factors in addition to cost, such as the nature of the market, and the firm's reputation and pricing policy (Nowlis and Simonson 1996, Tholke et al. 2001). Our analysis in Section 4.1 revealed that 0.2% of all possible queries make up 90% of the entire probability distribution. By focusing on features that frequently occur in such influential queries, a firm can strategically allocate its resources and improve its visibility. In addition to adding well-known features, an innovative firm can gain *a pioneering advantage* by introducing a new feature that is not covered by any of its competitors (Carpenter and Nakamoto 1989). As an example, consider the first hotel to offer free parking in a densely populated metropolitan area with limited parking options. This will provide a first-mover advantage (Kerin et al. 1992) proportional to the number of customers that include this new feature in their requirements. Previous work has verified that such moves require careful consideration and high implementation costs, especially in mature markets (Tholke et al. 2001).

In conclusion, after considering all contributing factors, a business may choose between *differentiation*, which requires the introduction of a new feature, and *imitation*, which calls for the adoption of a known feature from its competitors (Narasimhan and Turut 2013). Our framework can inform such efforts by computing visibility before and after the addition of each candidate feature, and suggesting candidates that maximize the expected visibility gain. The fact that our analysis found similar user preferences and visibility distributions across cities is encouraging, since it suggests that a hotel's efforts can be informed by the success or failure of similar ventures in other cities.

6.1.3. Improving the Quality of Existing Features. Feature coverage may go beyond simple availability. For instance, while a hotel might have a pool, it might be too small or poorly maintained, leading to negative comments. To further inform enhancement strategies, we study the correlation between visibility and customer sentiment on different features. Given that visibility takes values in $[0, 1]$, we perform a beta regression that is appropriate for continuous dependent variables with such values. Formally, the visibility of y_i of the ith business is modeled as a random variable that follows the beta distribution, as parametrized by Ferrari and Cribari-Neto (2004). The regression model is then obtained by writing the mean μ_i of y_i as $g(\mu_i) = \sum_{j=1}^{K} X_{ij}\beta_j$, where β is a $K \times 1$ vector of coefficients and X_{ij} is the aggregate sentiment of the reviewers of the ith business on the jth feature, as described in Section 3. Finally, $g(\cdot)$ is a link function that maps $[0, 1]$ to $\mathbb{R}$. After experimenting with different alternatives (probit, cloglog), we adopt the logit function based on log-likelihood. The regression reveals multiple features with positive coefficients that are significant at a level of 0.01: *Business services* (0.69), *Kitchenette* (0.25), *Non-smoking* (0.36), *Pool* (0.24), *Restaurant* (0.22), and *Room service* (0.54). We provide the full regression output in Table 2 of the online appendix. This type of analysis can help a business estimate the visibility gains after improving specific features. The interested analyst could extend our methodology and control for possible confounding variables, such as the price range or size of the hotel.

6.2. Confrontational Strategies

A confrontational strategy focuses on locating and addressing the injected fake reviews. The importance of responding to negative comments has been stressed by both relevant research and review platforms. For instance, TripAdvisor has published an article on handling consumer reviews (Barsky and Frame 2009), which highlights the importance of review-monitoring and encourages business managers to directly respond to reviews. Most platforms allow firms to maintain a verified business account, through which they can connect with their customers and directly respond to criticism. The importance of this communication mechanism has recently received increased attention from the research community. For instance, Avant (2013) found that the practice of systematically responding to negative reviews improves a hotel's image and enhances the customer's intent to return. Xie et al. (2014) found that the number of responses is significantly correlated with the hotel's performance, formalized as revenue per available room. Furthermore, an increasing number of research efforts (Xie et al. 2011, Min et al. 2015, Sparks and Bradley 2014, Park and Allen 2013) are focusing on designing effective response strategies, to help firms manage their reputation. If the responder is convinced that the review is fake or malicious, then she can post a dispute, or even report the review to the platform for further examination. In either case, vigilance and a commitment to reputation management are necessary for a business in a competitive market.

For review platforms, the effort to identify and eliminate fake reviews has been established as a top priority (Holloway 2011b, TripAdvisor 2015c). In addition, previous work suggests that defensive mechanisms against review fraud can demonstrate the portal's commitment to content quality and enhance its credibility (Lowe 2010). Our framework can be used to complement such mechanisms by monitoring the visibility of each business and identifying bursts of reviews that lead to significant changes. These reviews can then be cross-referenced with those reported by other traffic-based (Xie et al. 2012, Fei et al. 2013) or content-based techniques (Agichtein et al. 2006). Previous work has focused on detecting fake reviews by looking for bursts in the number of reviews submitted for a single item or by a single reviewer (Xie et al. 2012, Fei et al. 2013). By applying similar techniques on the sequence of a firm's visibility scores, we can identify influential activity that cannot be detected by simply considering the rate of submission, such as a small set of strategically injected reviews. Finally, as we showed in Section 5, a review platform can use our framework to simulate different types of injection attacks and compose a vulnerability report for businesses that are interested in managing their reputation and protecting their visibility.

7. Implications, Limitations, and Directions for Future Research

Our work studies the vulnerability of businesses to fake review attacks. We propose a formal measure of the visibility of a business on a review platform, based on the probability that a random user chooses to include it in her consideration set. Our operationalization of visibility takes into account the popularity of the features that the business can cover, its position in the platform's review-based ranking, and the way in which users consider rankings. Our work introduces a number of artifacts that can support research on online reviews, including (i) a methodology for estimating feature popularity, (ii) two review-based ranking functions, (iii) three models that simulate alternative behavioral patterns of how users process rankings, and (iv) three different attack strategies that can be used to estimate the vulnerability of a business in different settings.

Even though our framework is designed to measure vulnerability to *fake* reviews, it can also be applied as is to help a firm monitor its visibility and inform its marketing and enhancement strategies. In addition, a review platform with access to the reviews of all of the players in a market can use our framework to generate reports with valuable information, such as the distribution of visibility in the market, up-and-coming competitors that are threatening the firm's visibility, and the features that the firm needs to cover or improve to maximize its own visibility.

Our analysis on a large data set of hotel reviews from TripAdvisor.com revealed multiple findings with implications for platforms, businesses, and relevant research:

- A mixed strategy of self-injecting fake positive and injecting fake negative reviews about competitors is the most effective way for attackers to overtake their competitors in visibility. While the exact number of injections required to complete an attack varies across scenarios, our study revealed that just 50 injections were enough to complete 80% of all possible attacks when using TripAdvisor's `PopularityIndex` ranking function.
- Positive injections have a stronger impact than negative injections in markets where users focus only on a small set of top-ranked items. On the other hand, fake injections become increasingly effective as user

consideration expands to a larger set of items. In addition to informing a platform's detection efforts, our analysis can inform a firm of its visibility status and potential. For instance, a firm in an oligopolistic market can use our framework to determine the number of positive reviews that it needs to enter the "winner's circle" and enjoy vast visibility gains. It can then calibrate its marketing and improvement efforts accordingly, as we describe in Section 6.1.

- While ranking functions that consider the "freshness" of the reviews reduce the effects of outdated information, they are also more vulnerable to the injection of fake reviews than simple alternatives, such as the average star rating. This finding has strong implications for the vulnerability of leading platforms that have adopted such functions, such as TripAdvisor and Yelp. Despite the well-documented impact of review-based rankings, findings on ways to improve ranking functions have been extremely limited. We hope that the evidence that we provide in our work will motivate relevant research in this area. We make our own contribution in Section 5.2, where we introduce a new function that accounts for review fraud, while eliminating outdated reviews.
- In the absence of extensive query logs that can be used to directly learn user preferences over a set of features, customer reviews can serve as an effective proxy. In Section 4.1, we describe a methodology for estimating the number of reviews required to converge to confident estimations. For the hotels domain, 1,000 reviews were sufficient for all of the cities in our data set.
- Our study on the TripAdvisor data set revealed that 0.2% of all possible feature queries cover around 90% of the entire distribution of query popularity. This is a promising finding that can inform popularity-estimation efforts, as well as help businesses identify influential features with a high impact on their visibility, even beyond the context of fake reviews.

The primary limitation of our work is the absence of a large data set of injection attacks, including the fake reviews that were successfully injected, as well as those that were blocked by the attacked platform's defenses. This is a standard limitation for research on fake reviews. As we described in Section 5, our solution to this was to emulate the environment of a review platform, simulate different types of attacks, and evaluate their success ratio. Additional limitations include the lack of extensive search-and-click logs that can be used to learn user preferences, as well as to estimate the way in which users process rankings. These are typical issues for research on online reviews, caused by the sensitive and proprietary nature of the required information. We describe the way in which we deal with these limitations in Sections 4.1 and 4.3, respectively.

Finally, our work provides evidence that we hope will motivate novel research on online reviews. Our findings can support the development of defense mechanisms and fraud-resistant ranking functions, frameworks for attack simulation, user consideration models, and strategies that help businesses manage their visibility. We studied such a strategy in Section 5.3, which focused on the responses that businesses submit to reviews. Despite the well-documented benefits of this practice, we found strong evidence that this mechanism is rarely utilized to dispute fake reviews. While we provided our own interpretation of this finding after studying a large corpus of responses, more extensive research is required to help businesses optimize their response strategy. The possibilities for further research are manifold and exciting. We hope that, with this paper, we further the boundaries of our knowledge in this domain and help foster rigorous and thoughtful research.

Supplemental Material

Supplemental material to this paper is available at https://doi.org/10.1287/isre.2016.0674.

References

Agichtein E, Brill E, Dumais S (2006) Improving Web search ranking by incorporating user behavior information. *Proc. 29th Annual ACM SIGIR Conf.* (ACM, New York), 19–26.

Anderson ET, Simester DI (2014) Reviews without a purchase: Low ratings, loyal customers, and deception. *J. Marketing Res.* 51(3):249–269.

Asuncion A, Welling M, Smyth P, Teh YW (2009) On smoothing and inference for topic models. *Proc. 25th Conf. Uncertainty Artificial Intelligence* (AUAI Press, Corvallis, OR), 27–34.

Avant T (2013) Responding to TripAdvisor: How hotel responses to negative online reviews effect hotel image, intent to stay, and intent to return. Unpublished Master's thesis, University of South Carolina, Columbia.

Baeza-Yates R, Hurtado C, Mendoza M (2005) Query recommendation using query logs in search engines. *Proc. 2004 Internat. Conf. Current Trends Database Tech.* (Springer-Verlag, Berlin Heidelberg), 588–596.

Barsky J, Frame C (2009) Handling online reviews: Best practices. *Hospitalitynet* (August 28), http://www.hospitalitynet.org/news/4043169.html.

Belton P (2015) Navigating the potentially murky world of online reviews. *BBC* (June 23), http://www.bbc.com/news/business-33205905.

Bickart B, Schindler RM (2001) Internet forums as influential sources of consumer information. *J. Interactive Marketing* 15(3):31–40.

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. *J. Machine Learn. Res.* 3:993–1022.

Breure E (2013) Infographic: Hotel reviews, can we trust them? *Olery* (April 24), http://www.olery.com/blog/infographic-hotel-reviews-can-we-trust-them/.

Brooks N (2004) The Atlas Rank report: How search engine rank impacts traffic. *Insights, Atlas Institute Digital Marketing*, http://www.inesting.org/ad2006/adminsc1/app/marketingtecnologico/uploads/Estudos/atlas%20onepoint%20-%20how%20search%20engine%20rank%20impacts%20traffic.pdf.

Byers JW, Mitzenmacher M, Zervas G (2012a) Daily deals: Prediction, social diffusion, and reputational ramifications. *Proc. Fifth ACM Internat. Conf. Web Search Data Mining* (ACM, New York), 543–552.

Byers JW, Mitzenmacher M, Zervas G (2012b) The Groupon effect on Yelp ratings: A root cause analysis. *Proc. 13th ACM Conf. Electronic Commerce* (ACM, New York), 248–265.

Carare O (2012) The impact of bestseller rank on demand: Evidence from the app market. *Internat. Econom. Rev.* 53(3):717–742.

Carpenter GS, Nakamoto K (1989) Consumer preference formation and pioneering advantage. *J. Marketing Res.* 26(3):285–298.

Chapelle O, Zhang Y (2009) A dynamic Bayesian network click model for Web search ranking. *Proc. 18th Internat. Conf. World Wide Web* (ACM, New York), 1–10.

Chatterjee P (2001) Online reviews—Do consumers use them? Gilly MC, Meyers-Levy J, eds. *Advances in Consumer Research* , Vol. 28 (Association for Consumer Research, Valdosta, GA), 129–133.

Cheng VT, Loi MK (2014) Handling negative online customer reviews: The effects of elaboration likelihood model and distributive justice. *J. Travel Tourism Marketing* 31(1):1–15.

CMA (2015) Online reviews and endorsements. Competition and Markets Authority (February 26), https://goo.gl/GxZ4J7.

Decker R, Trusov M (2010) Estimating aggregate consumer preferences from online product reviews. *Internat. J. Res. Marketing* 27(4):293–307.

Dellarocas C (2006) Strategic manipulation of Internet opinion forums: Implications for consumers and firms. *Management Sci.* 52(10):1577–1593.

Ding X, Liu B, Yu PS (2008) A holistic lexicon-based approach to opinion mining. *Proc. 2008 Internat. Conf. Web Search Data Mining* (ACM, New York), 231–240.

Duan W, Gu B, Whinston AB (2008) Do online reviews matter? An empirical investigation of panel data. *Decision Support Systems* 45(4):1007–1016.

Fei G, Mukherjee A, Bing L, Hsu M, Castellanos M, Ghosh R (2013) Exploiting burstiness in reviews for review spammer detection. *Proc. 7th Internat. Conf. Weblogs Social Media*, 175–184.

Feng S, Xing L, Gogar A, Choi Y (2012) Distributional footprints of deceptive product reviews. *Proc. 6th Internat. Conf. Weblogs Social Media*, Vol. 12, 98–105.

Fenton S (2015) TripAdvisor denies rating system is flawed, after fake restaurant tops rankings in Italy. *Independent* (June 30), http://goo.gl/NtSKpi.

Ferrari S, Cribari-Neto F (2004) Beta regression for modelling rates and proportions. *J. Appl. Statist.* 31(7):799–815.

Forman C, Ghose A, Wiesenfeld B (2008) Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Inform. Systems Res.* 19(3): 291–313.

Ghose A, Ipeirotis PG (2011) Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *Knowledge Data Engrg., IEEE Trans.* 23(10): 1498–1512.

Ghose A, Ipeirotis PG, Li B (2012) Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Marketing Sci.* 31(3):493–520.

Ghose A, Ipeirotis PG, Li B (2014) Examining the impact of ranking on consumer behavior and search engine revenue. *Management Sci.* 60(7):1632–1654.

Guan Z, Cutrell E (2007) An eye tracking study of the effect of target rank on Web search. *Proc. SIGCHI Conf. Human Factors Comput. Systems* (ACM, New York), 417–420.

Henzinger M (2007) Search technologies for the Internet. *Science* 317(5837):468–471.

Holloway D (2011a) How are Yelp's search results ranked? *Yelp* (June 21), https://biz.yelp.com/blog/how-are-yelps-search-results-ranked.

Holloway D (2011b) Just another reason why we have a review filter. *Yelp* (October 3), http://officialblog.yelp.com/2011/10/just-another-reason-why-we-have-a-review-filter.html.

Hu N, Bose I, Koh NS, Liu L (2012) Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decision Support Systems* 52(3):674–684.

Ifrach B, Johari R (2014) The impact of visibility on demand in the market for mobile apps. Working paper, Stanford University, Stanford, CA.

Jindal N, Liu B (2008) Opinion spam and analysis. *Proc. 2008 Internat. Conf. Web Search Data Mining* (ACM, New York), 219–230.

Jindal N, Liu B, Lim E (2010) Finding unusual review patterns using unexpected rules. *Proc. 19th ACM Internat. Conf. Inform. Knowledge Management* (ACM, New York), 1549–1552.

Joachims T (2002) Optimizing search engines using clickthrough data. *Proc. Eighth ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM, New York), 133–142.

Joachims T, Granka L, Pan B, Hembrooke H, Gay G (2005) Accurately interpreting clickthrough data as implicit feedback. *Proc. 28th Annual ACM SIGIR Conf.* (ACM, New York), 154–161.

Kamerer D (2014) Understanding the Yelp review filter: An exploratory study. *First Monday* 19(9). http://firstmonday.org/article/view/5436/4111.

Kerin RA, Varadarajan PR, Peterson RA (1992) First-mover advantage: A synthesis, conceptual framework, and research propositions. *J. Marketing* 56(4):33–52.

Korolova A, Kenthapadi K, Mishra N, Ntoulas A (2009) Releasing search queries and clicks privately. *Proc. 18th Internat. Conf. World Wide Web* (ACM, New York), 171–180.

Kwark Y, Chen J, Raghunathan S (2014) Online product reviews: Implications for retailers and competing manufacturers. *Inform. Systems Res.* 25(1):93–110.

Lappas T (2012) Fake reviews: The malicious perspective. Bouma G, Ittoo A, Metais E, Wortmann H, eds. *17th Internat. Conf. Applications Natural Language Processing Inform. Systems*, Lecture Notes Comput. Sci., Vol. 7337 (Springer-Verlag, Berlin Heidelberg), 23–34.

Leung CWK, Chan SCF, Chung FL, Ngai G (2011) A probabilistic rating inference framework for mining user preferences from reviews. *World Wide Web* 14(2):187–215.

Li B, Ghose A, Ipeirotis PG (2011) Towards a theory model for product search. *Proc. 20th Internat. Conf. World Wide Web* (ACM, New York), 327–336.

Lorigo L, Haridasan M, Brynjarsdóttir H, Xia L, Joachims T, Gay G, Granka L, Pellacini F, Pan B (2008) Eye tracking and online search: Lessons learned and challenges ahead. *J. Amer. Soc. Inform. Sci. Tech.* 59(7):1041–1052.

Lowe L (2010) Yelp's review filter explained. *Yelp* (March 18), http://officialblog.yelp.com/2010/03/yelp-review-filter-explained.html.

Lu X, Ba S, Huang L, Feng Y (2013) Promotional marketing or word-of-mouth? Evidence from online restaurant reviews. *Inform. Systems Res.* 24(3):596–612.

Luca M, Zervas G (2016) Fake it till you make it: Reputation, competition, and yelp review fraud. *Management Sci.*, ePub ahead of print January 28, http://dx.doi.org/10.1287/mnsc.2015.2304.

Marrese-Taylor E, Velásquez JD, Bravo-Marquez F, Matsuo Y (2013) Identifying customer preferences about tourism products using an aspect-based opinion mining approach. *Procedia Comput. Sci.* 22:182–191.

Masoni D (2014) TripAdvisor fined in Italy for misleading reviews. *Reuters* (December 22), http://www.reuters.com/article/tripadvisor-italy-fine-idUSL6N0U62MW20141222.

Mauri AG, Minazzi R (2013) Web reviews influence on expectations and purchasing intentions of hotel potential customers. *Internat. J. Hospitality Management* 34:99–107.

May K (2011) Goodbye TripAdvisor, welcome to verified reviews on Expedia. *tnooz* (December 28), http://goo.gl/A3dgpQ.

Mayzlin D, Dover Y, Chevalier JA (2012) Promotional reviews: An empirical investigation of online review manipulation. Technical report, National Bureau of Economic Research, Cambridge, MA.

Min H, Lim Y, Magnini V (2015) Factors affecting customer satisfaction in responses to negative online hotel reviews: The impact of empathy, paraphrasing, and speed. *Cornell Hospitality Quart.* 56(2):223–231.

Mukherjee A, Liu B, Glance N (2012) Spotting fake reviewer groups in consumer reviews. *Proc. 21st Internat. Conf. World Wide Web* (ACM, New York), 191–200.

Mukherjee A, Venkataraman V, Liu B, Glance N (2013a) Fake review detection: Classification and analysis of real and pseudo reviews. Technical report, UIC-CS-2013-03, University of Illinois at Chicago, Chicago.

Mukherjee A, Venkataraman V, Liu B, Glance NS (2013b) What Yelp fake review filter might be doing? *Proc. 7th Internat. Conf. Weblogs Social Media*.

Narasimhan C, Turut Ö (2013) Differentiate or imitate? The role of context-dependent preferences. *Marketing Sci.* 32(3): 393–410.

Nowlis SM, Simonson I (1996) The effect of new product features on brand choice. *J. Marketing Res.* 33(1):36–46.

Olery (2015) Hotel review sites. www.olery.com/reviewsites/.

Pan B (2015) The power of search engine ranking for tourist destinations. *Tourism Management* 47:79–87.

Pan B, Hembrooke H, Joachims T, Lorigo L, Gay G, Granka L (2007) In Google we trust: Users' decisions on rank, position, and relevance. *J. Comput.-Mediated Comm.* 12(3):801–823.

Park SY, Allen JP (2013) Responding to online reviews problem solving and engagement in hotels. *Cornell Hospitality Quart.* 54(1):64–73.

Pearson K (1895) Note on regression and inheritance in the case of two parents. *Proc. Roy. Soc. London* 58:240–242.

Schaal D (2015a) TripAdvisor's hotel review collection is leaving rivals in the dust. *Skift* (May 4), http://goo.gl/vrOxbX.

Schaal D (2015b) TripAdvisor's reviews won't change even as it becomes the next big booking site. *Skift* (July 28), https://skift.com/2015/07/28/tripadvisors-reviews-wont-change-even-as-it-becomes-the-next-big-booking-site/.

Schneiderman ET (2015) Operation clean turf. http://goo.gl/WZfq5L.

Sparks BA, Bradley GL (2014) A "triple a" typology of responding to negative consumer-generated online reviews. *J. Hospitality Tourism Res.*, ePub ahead of print July 2, http://dx.doi.org/10.1177/1096348014538052.

Sparks BA, Browning V (2011) The impact of online reviews on hotel booking intentions and perception of trust. *Tourism Management* 32(6):1310–1323.

StatisticBrain (2015) Internet travel hotel booking statistics. http://www.statisticbrain.com/internet-travel-hotel-booking-statistics/.

Stempel J (2015) Amazon sues to block alleged fake reviews on its website. *Reuters* (April 9), http://goo.gl/oz13iZ.

Sussin J, Thompson E (2012) The consequences of fake fans, likes, and reviews on social networks. Gartner, Inc., https://www.gartner.com/doc/2091515/consequences-fake-fans-likes-reviews.

Tholke JM, Hultinka EJ, Robben HSJ (2001) Launching new product features: A multiple case examination. *J. Product Innovation Management* 18(1):3–14.

Tibken S (2013) Taiwan fines Samsung $340,000 for bashing HTC. *CNET* (October 24), http://www.cnet.com/news/taiwan-fines-samsung-340000-for-bashing-htc/.

Tnooz (2013) Top U.S. travel websites. http://goo.gl/qJgOcu.

TripAdvisor (2013) TripAdvisor popularity ranking: Key factors and how to improve. https://goo.gl/TzpHUS.

TripAdvisor (2015a) Review moderation and guidlines. http://www.tripadvisor.com/vpages/review_mod_fraud_detect.html.

TripAdvisor (2015b) TripAdvisor fact sheet. http://www.tripadvisor.com/PressCenterc4Fact_Sheet.html.

TripAdvisor (2015c) We have zero tolerance for fake reviews! https://www.tripadvisor.com/pages/fraud.html.

Tucker C, Zhang J (2011) How does popularity information affect choices? A field experiment. *Management Sci.* 57(5):828–842.

Vermeulen IE, Seegers D (2009) Tried and tested: The impact of online hotel reviews on consumer consideration. *Tourism Management* 30(1):123–127.

Viégas FB, Wattenberg M, Dave K (2004) Studying cooperation and conflict between authors with history flow visualizations. *Proc. SIGCHI Conf. Human Factors Comput. Systems* (ACM, New York), 575–582.

Wang Z (2010) Anonymity, social image, and the competition for volunteers: A case study of the online market for reviews. *BE J. Econom. Anal. Policy* 10(1):1–35.

Weise E (2015) Amazon cracks down on fake reviews. *USA Today* (October 19), http://www.usatoday.com/story/tech/2015/10/19/amazon-cracks-down-fake-reviews/74213892/.

Xie HJ, Miao L, Kuo P, Lee B (2011) Consumers responses to ambivalent online hotel reviews: The role of perceived source credibility and pre-decisional disposition. *Internat. J. Hospitality Management* 30(1):178–183.

Xie KL, Zhang Z, Zhang Z (2014) The business value of online consumer reviews and management response to hotel performance. *Internat. J. Hospitality Management* 43:1–12.

Xie S, Wang G, Lin S, Yu PS (2012) Review spam detection via temporal pattern discovery. *Proc.* 18*th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM, New York).

Ye Q, Law R, Gu B (2009) The impact of online user reviews on hotel room sales. *Internat. J. Hospitality Management* 28(1): 180–182.

Ye Q, Gu B, Chen W, Law R (2008) Measuring the value of managerial responses to online reviews—A natural experiment of two online travel agencies. *ICIS* 2008 *Proc.*, 115.

Ye Q, Law R, Gu B, Chen W (2011) The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. *Comput. Human Behav.* 27(2):634–639.

Zhu F, Zhang X (2010) Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *J. Marketing* 74(2):133–148.