YELP FOOTPRINTS - Analysing Yelp's Filtration Mechanisms

Lara Ansari
University of Illinois at Chicago
Chicago, IL, USA
lasar2@uic.edu

Abhishek Tripathi
University of Illinois at Chicago
Chicago, IL, USA
atripa5@uic.edu

Yogeeta Kuttabadkar
University of Illinois at Chicago
Chicago, IL, USA
ykutta2@uic.edu

ABSTRACT

Social reviews are the norm in today's lifestyle. Consumers are more aware and active on social networking sites, posting their experiences of a service they have undertaken. Everyday decisions of which places to visit or which service to undertake are based directly on reading the reviews of the target service. Hence, the trustworthiness and integrity of reviews is a legitimate security issue. Fake reviews and accounts pose a threat to users, as well as businesses involved, by potentially misleading users to malicious services or products and degrading the rating of a business. Prominent sites like Yelp attempt to filter out fake reviews using specialized algorithms. We seek to explore features that are most common in accepted reviews and using this knowledge determine whether we can consistently generate reviews or construct user profiles such that they bypass these filtering algorithms. In this paper, we perform a study on Yelp restaurant data, collecting recommended reviews and user profile information. Based upon the retrieved data, we craft reviews which can bypass the filtration mechanism of YELP and be marked as "Recommended", contributing towards the overall business rating. Using created user accounts of varying characteristics (profile information, reputation, etc.), we post these reviews and monitor the reaction they receive. Our results show that yelp focuses more on the reputation built up by a user, a term we call ***"YELP FOOTPRINT"***, and that scrutiny of review filtration largely depends upon this factor. We further analyse some users and attempt to identify characteristics of these and draw conclusions.

1. INTRODUCTION

In today's world, a majority of businesses prosper and thrive off an on-line reputation. Conversely, many users depend on business reviews for their choices. The social proof contained within reviews and star ratings helps consumers short cut their research; making decisions faster and with greater confidence than ever before. A recent local consumer review survey of 2016 [1] shows that 84% of people trust on-line reviews as much as personal recommendation, and 88% look at reviews when considering any business [10]. A positive reputation is one of the most powerful marketing assets that any business can make use of to convince new customers to contact them. Added to this is the fact that there are no restrictions on who writes a review; it is merely an assumption that the reviewer is a user who has actually consumed a service from the business being reviewed. Further, there is no sure way of confirming this aspect. Thus, reviews are important, reviewers are plenty, but not all are authentic.

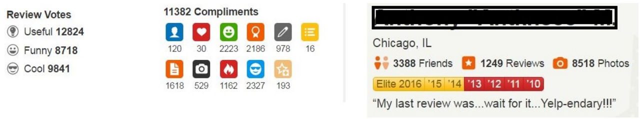
YELP - a famous on-line application is considered to be a local guide of real word-of-mouth opinions of consumers for local businesses. As of September 2016, Yelp has reported 115 million reviews, 72-75 million unique visitors over an average of 3-month period and 2.8 million claimed businesses [2]. It is reported that a one-star increase on Yelp leads to a 5-9% increase in business revenue and one negative review can cost a business 30 customers. In addition, a Yelp page will sometimes rank higher than a company's actual site on a search engine results page. Hence, the trustworthiness and integrity of reviews is a legitimate security issue.

Since its inception, Yelp has battled with multiple cases of fraudulent reviews on businesses to improve their reputation, or tarnish their competitors reputation. Further, there are in existence *"Reputation Management"* establishments that promise to raise Yelp's overall rating for a certain financial amount and in a certain period of time [4] [3]. In such cases, bots may be used to write a large number of positive reviews, however these would be easy to detect due to automatic, synchronous or other typical behaviour. More frequently, these establishments use an incentive model - **real people are paid to write fake reviews**. A 2013 study [9] found that one out of every five (20%) of all Yelp reviews are written by someone who claims to be a customer but isn't.

Since the center point of Yelp are its reviews, it combines multiple advanced filtration and detection techniques to recommend only those reviews that are of high quality, reliability and user activity. Previous work focused on detecting fake reviews in other systems like Amazon [7], using linguistic and opinion sentiment detection mechanisms. In this paper we propose to build on the existing knowledge of Yelp's filtration algorithms, craft reviews around them and using multiple user accounts post these reviews to different restaurants. In particular we look at restaurant data for major cities like Chicago, LA and Seattle. We monitor to check if our created reviews are accepted (i.e. they bypass Yelp's filtration mechanism) and remain on the site. Alternatively are flagged and rightly detected by the existing system. In addition, we also focus on identifying factors that can help determine fake user profiles from legitimate ones.

The main contributions of this paper are:

- We identified reviews that do not have corresponding text sentiment and star rating.
- We aimed to find leaks in the existing mechanisms by posting reviews with minimum characteristics that bypass checks and are termed authentic.



Review Votes
Useful 12824
Funny 8718
Cool 9841

11382 Compliments

120	30	2223	2186	978	16
1618	529	1162	2327	193	

[redacted]
Chicago, IL
3388 Friends 1249 Reviews 8518 Photos
Elite 2016 '15 '14 '13 '12 '11 '10
“My last review was...wait for it...Yelp-endary!!!”

Figure 1: YELP User Profile :Review Votes, Compliments, Elite User Status

- We identified characteristics that can differentiate between Original User Accounts and Fake User Accounts.
- We found that Yelp reviews are taken seriously by consumers as well as business owners. In order for an adversary to cause maximum harm to a business reputation or for malicious businesses to make profit by posting reviews on YELP, one would first need to build a strong reputed user account with Yelp and once their activity seems legitimate enough that their initial reviews are retained for businesses, they become a trusted user in the yelp community. Once this is achieved, one can make use of their status to begin harming other businesses. Our results support this claim as many of our golden user accounts' reviews were not filtered out whereas similar reviews posted with newly created accounts were removed in the first 12 hours.

The remainder of this paper is structured as follows: in Section 2, we offer the necessary background information about YELP, its business model and some technical aspects used in this paper. In Section 3 we offer details of our system, our proposed approach to determine the loopholes in Yelp's recommendation system and present the results of our experiments in Section 4. Further, we present some of the shortcomings of our work and provide future scope in Section 5. We discuss related work in Section 6; and conclude in Section 7.

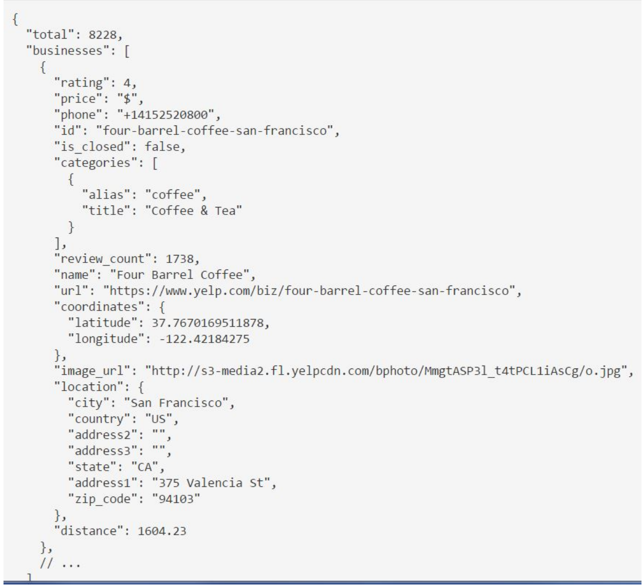
2. BACKGROUND

In this section we provide a short description of features of Yelp and an overview of our threat model.

Yelp Business Model Overview: Any user that has an account on Yelp and performs the email verification step, can search for a business and post a review. The user can check-in at the business location via the mobile phone Yelp application or even upload pictures of the food/menu/ambience of the business along with their review text. The review goes public instantly on the business page and all other users can view it. Yelp also has other social networking features like User profile data (biography, url link, favourite things, tag-line,profile picture,etc.), allows users to add friends, direct messaging to other users, vote for a review (Useful, Cool or Funny), compliment a review (Good tip, Helpful pictures, etc.), as shown in Figure 1.

Yelp Elite Squad: Yelp recognizes its users who are extremely active within the community and have posts on a wide range of subjects as marking them as *”ELITE”*, as shown in Figure 1. Elite- Worthiness is based on many factors that aren't fully disclosed. Some of them are - well written reviews, high quality tips, detailed personal profile, active voting on other reviews, complimenting records and a history of user activity [17].

Recommended Reviews: Yelp uses a filtration mechanism to distinguish between reviews that can add value to the community by marking them as *”Recommended”*. Most reviews written pass through this filter and either remain on the business page as recommended, or are moved to the end of the page tagged under *”Other Non-Recommended Reviews”*. Only recommended reviews affect a business' overall rating score. Interestingly, Yelp does not guarantee that all the reviews marked as non-recommended are necessarily fake. Some are real reviews written by legitimate users who Yelp doesn't have enough information about. For example, such a review might come from a new user with minimum profile information or the review itself might seem like an unhelpful rant or rave, lack actual details of the service, it suggests some bias (friend with the business owner) or the user just doesn't have enough trusted engagement activity on the Yelp community [18]. We term such an aspect as a **”YELP FOOTPRINT”**. The more a user writes reviews on different categories of businesses like -Restaurants, Shopping Services, Stores, or makes friends with other users, compliments other reviews, etc. the more this footprint grows. This is to say that, a user with a larger Yelp footprint, will have higher chances of their review remaining on the page as recommended than another user having a smaller footprint.



```
{
  "total": 8228,
  "businesses": [
    {
      "rating": 4,
      "price": "$",
      "phone": "+14152520800",
      "id": "four-barrel-coffee-san-francisco",
      "is_closed": false,
      "categories": [
        {
          "alias": "coffee",
          "title": "Coffee & Tea"
        }
      ],
      "review_count": 1738,
      "name": "Four Barrel Coffee",
      "url": "https://www.yelp.com/biz/four-barrel-coffee-san-francisco",
      "coordinates": {
        "latitude": 37.7670169511878,
        "longitude": -122.42184275
      },
      "image_url": "http://s3-media2.fl.yelpcdn.com/bphoto/MmgtASP3l_t4tPCL1iAsCg/o.jpg",
      "location": {
        "city": "San Francisco",
        "country": "US",
        "address2": "",
        "address3": "",
        "state": "CA",
        "address1": "375 Valencia St",
        "zip_code": "94103"
      },
      "distance": 1604.23
    },
    // ...
  ]
```

Figure 2: YELP Search API returned JSON object response

YELP API V2 - SEARCH API: Yelp has restricted developer access to their data via established API's. In this paper we only make use of the Yelp SEARCH API, which takes a location and category as parameters and returns a JSON object of available business information.

For Example,

Request: https://api.yelp.com/v2/search?location=Seattle&categories=restaurants

Response: Described in Figure 2

The Yelp Dataset Challenge: Every year Yelp releases a dataset with a large number of users, reviews, businesses, check-ins data for academic research purposes. We considered using this data for our analysis, but, the links between the users' and their written reviews for particular businesses were obfuscated. Hence, we only scanned this data to gain additional business URL's, apart from the ones returned by the Yelp Search API.

2.1 Threat Model

Our study is geared towards an attacker who is a real user writing fake reviews for a business, and to understand characteristics that makes such reviews sustain. In this paper we seek to explore features of Yelp that can help any attacker in two ways:

1. Construct reviews which can bypass the Yelp's filtration mechanisms.
2. Construct User Profiles and User Activity such that the reviews posted by such users are accepted by Yelp.

We first aim to understand the most common features in accepted recommended reviews, using this knowledge determine whether we can consistently generate reviews and develop user profiles such that they bypass Yelp's filtering algorithm. Knowing these common features of *"recommended"* reviews by Yelp's standards illuminates some of the inner workings of their algorithm, which in turn reveals potential loopholes that a malicious user or business could exploit. Further, if a user initially builds a good reputation within Yelp's community by performing activities as described above, one could easily fall off the filtration radar of Yelp. Then this user could later exploit their built reputation to write fake reviews and continue to be undetected, thereby misguiding other consumers and manipulating overall business rating.

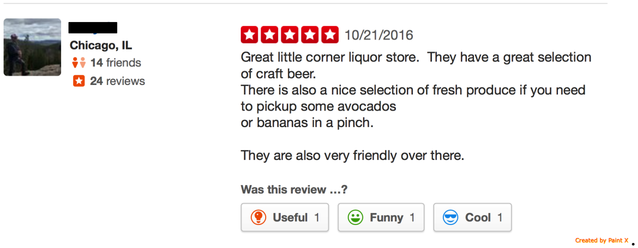
We developed user profiles and worked towards building a decent footprint of these profiles within the community (for some accounts). This included posting a number of real reviews from users with completed profiles, check-ins at multiple restaurants, frequent user activity by liking other reviews and adding friends. Simultaneously, we also collected recommended reviews posted on many restaurant businesses in and around Chicago, Los Angeles and Seattle. Using this data we constructed different combinations of reviews (as discussed in Section 3) and posted them using our created user profiles. We performed this experiment with user profiles that had a good footprint as well as users with minimum footprint and compared the results. Further, we began looking at user profiles on Yelp, to identify suspicious user accounts writing illegitimate or fake reviews and gain a perspective on the characteristics of such accounts.

3. SYSTEM OVERVIEW

This section gives a detailed explanation of the various modules of our project. Most of the modules are implemented in Python. Our project mainly constituted of the following modules:

3.1 User Account Creation

Each Yelp user account needs to be registered with a valid email id and confirm enrollment with the link sent to this email. We created 75 fake user profiles, each with varying differences in types of accounts and from IP addresses of different regions. This was done to ensure that we had accounts that are not similar in characteristics. Accounts were created incrementally between September 20th,2016 - October 26th, 2016.We created 30 accounts with user profiles pictures (celebrities, random individuals found on the internet, elite users of Yelp), 15 profiles with complete profile details filled (What you love, Find me In, Home-town, Favourite Website, Favourite Movie, Best Book,etc.). Out of these 15 accounts we concentrated on building a potential Yelp footprint for 5 *(Golden User Accounts)* of them, using our own mobile devices. We checked-in at restaurant locations and left an almost equal number of positive and negative reviews. A point worth mentioning here is that user account reputation building is a time consuming process, hence, we could not gain a larger number of such accounts in practice within a restricted amount of time. This was done to establish a reputation for these accounts and to differentiate the behaviour of automatic recommendation detection of these same reviews posted from such users. Additionally, 2 user accounts were our personal accounts which were created in February 2016 and April 2016, we used these accounts to validate if the date of account creation plays a vital role in the detection mechanism of Yelp. The remaining created accounts were blank user accounts with basic information such as name and location.



[redacted]
Chicago, IL
14 friends
24 reviews

★★★★★ 10/21/2016

Great little corner liquor store. They have a great selection
of craft beer.
There is also a nice selection of fresh produce if you need
to pickup some avocados
or bananas in a pinch.

They are also very friendly over there.

Was this review ...?

Figure 3: YELP Sample Review on a business page

Location	#Businesses	#Reviews	#User Profiles
Aurora	1031	68,659	NA
Chicago	1034	269,682	3724
Elgin	1022	64,820	2068
Joliet	1027	61,578	2702
Los Angeles	1000	84,247	NA
Naperville	1014	105,458	2154
New York	1010	90,000	NA
Schaumburg	1006	56,911	NA
Seattle	1000	86,900	NA
TOTAL	**9144**	**888,255**	**10,648**

Table 1: Scrapped Data Statistics

3.2 Data Collection

In this part we focused on collecting multiple aspects of data from the Yelp system. Mainly:

1. Using the Search API of Yelp in combination with continuous updation of page offset parameters we collected restaurant information for multiple locations.
2. For each of these businesses we developed web scrapers that retrieved details of review and reviewer (user that posted the review) present on the first 10 pages. The statistics of data collected is shown in Table 1, the characteristics collected from each page is described in Figure 4.
3. Finally, we scrapped user profiles, choosing those user id's that seemed to have a significant difference between their review text and the rating of the review.

Data Collection Restrictions: Yelp has many restrictions on the method and amount of data one can access off their system. Some of the major concerns in this project were:

1. Yelp doesn't work with TOR network. It blocks almost all requests originating from a tor browser and displays a 503 - Service Unavailable Response.
2. Per IP address, we can scrape reviews for roughly 100 businesses (First 10 pages of each business), but beyond this limit the IP is blocked for roughly 12 hours.
3. Per IP Address, we can scrape roughly 100 user profiles after which the IP is blocked.
4. Yelp Search API returns page wise details of a business, restricting the results to 20 items in one request. Even with manipulating a page offset parameter, the API returns duplicate businesses in each page and can go up to a maximum of 980 pages, post which it times out. Hence, even though the total count of businesses for the requested parameters seem higher, one cannot obtain all the businesses within the set page offset restriction.
5. CAPTCHA Challenges - After every 5th consecutive user login from the same IP address, a CAPTCHA challenge is provided to proceed.

Data Collection Workarounds: We made use of the following to work around the restrictions mentioned above:

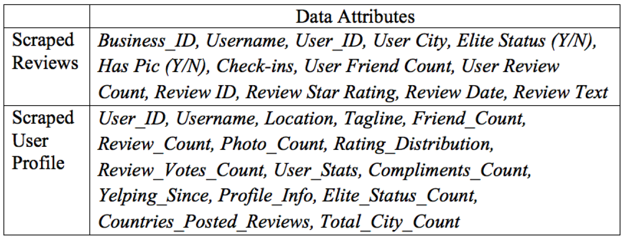
1. **WINDSCRIBE**: Using this free VPN client that provides a wide range of region based IP addresses, we scrapped the Yelp system for reviews and user profile data [15].
2. **HMA PRO:** We exhausted all the free IP addresses provided by Windscribe. Hence, we switched to HMA Pro which is a paid VPN service with a colossal number of IP addresses from all over the world [14].
3. Using a python script and manipulating the ”Page Offset” and ”Limit” parameters, we were able to access roughly 1000 unique Chicago restaurant url's from 980 returned request pages. We performed the same to collect restaurant information from suburbs of Chicago like Elgin, Joliet, Schaumburg,etc. as well as from Los Angeles and Seattle.
4. To overcome the CAPTCHA challenge restriction, we added a random wait time between each user login request within our automatic selenium scripts. Additionally, if a CAPTCHA challenge was encountered we solved it manually and proceeded with execution.

3.3 Crafting Fake Reviews

From the reviews we collected in our data collection phase, we created particular reviews as various test cases to answer specific questions. An explanation of various test cases is described below.

Test Case # 1: Natural User Reviews

Hypothesis: Does Yelp have a limit to the number of reviews posted per user per day?



	Data Attributes
Scraped Reviews	*Business_ID, Username, User_ID, User City, Elite Status (Y/N), Has Pic (Y/N), Check-ins, User Friend Count, User Review Count, Review ID, Review Star Rating, Review Date, Review Text*
Scraped User Profile	*User_ID, Username, Location, Tagline, Friend_Count, Review_Count, Photo_Count, Rating_Distribution, Review_Votes_Count, User_Stats, Compliments_Count, Yelping_Since, Profile_Info, Elite_Status_Count, Countries_Posted_Reviews, Total_City_Count*

Figure 4: Data Attributes Collected for Reviews & User Profiles

Procedure: We picked a random sample of 10 reviews (without replacement) from all reviews of the same businesses, split the reviews into a list of sentences, combined a random number of sentences to form a text. We cleaned the text to ensure each sentence started with a capital letter, there are more than 5 sentences in each review, and the length of the text is substantial. The rating of these constructed reviews is the average rating of all the combined reviews.

Test Case # 2: Positive Review text with Minimum Rating, vice versa

Hypothesis: Does Yelp check the correspondence between the sentiment present in the text and the star rating given to it?

Procedure: From the same business we picked all positive (4,5) and negative (1,2) rating reviews and using the same method described above, we split the review with full stops and combine random sentences to form a review. These reviews are then posted with opposite ratings positive(1,2) and negative (4,5).

Test Case # 3: Reposting the same Older Elite user review

Hypothesis: Does Yelp check if the same reviews from the past are re-posted for the same business?

Procedure: For each business, we picked an older positive (rating: 4,5) and negative (rating: 1,2) review from the previous years' of 2014 or earlier and repost it on the same business.

Test Case # 4: Reviews from businesses of the same categories

Hypothesis: Does Yelp filtration detect if the text is closely relevant to the business but doesn't make sense?

Procedure: We picked businesses of the same categories (Chinese, Sandwiches, Irish bars,etc.) and, taking random sentences from reviews of each category, we prepared reviews for other businesses.

3.4 Review Posting

These crafted reviews are posted using the different user accounts we create in Section 3.1. The results of this is discussed in Section 4.

3.5 Sentiment Analysis

For all the reviews we collected, we ran sentiment analysis to analyse if the review text and the rating correspond correctly. We ran different types of sentiment analysis on subsets of data, to speed up the process of gaining results. The NLTK Vader (Valence Aware Dictionary and Sentiment Reasoner) Library is a lexicon and rule-based tool specifically used to determine sentiments used in social media [6]

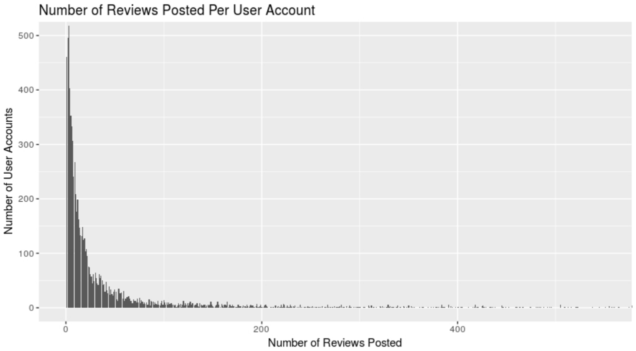


Figure 5: Total number of reviews posted per user, taken since the user's first appearance on Yelp. Ideally, distinctly high review counts might indicate suspicious user behaviour. However, upon manual inspection of the top 50 profiles with the highest review counts, this was not necessarily the case.

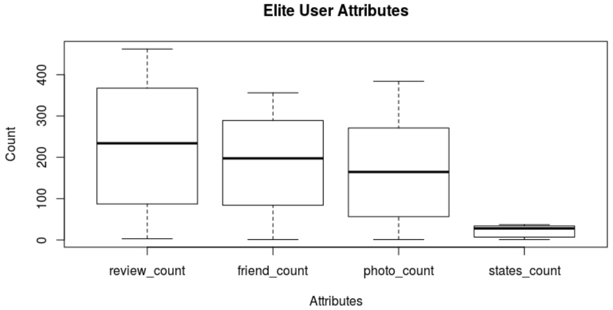


Figure 6: Select user attribute and behavior counts for Elite User. Number of photos and friends appear to most effectively characterize Elite users in comparison to non-Elite users.

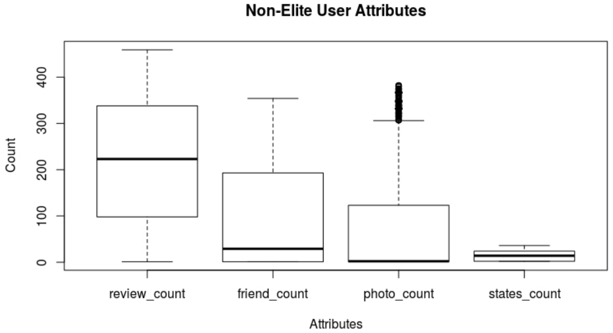


Figure 7: Select user attribute and behavior counts for Non-Elite users.

[13]. The other tool used consists of AFINN [12] (gives a score of -5 to 5 for each word) and Bing Liu's [8] dictionary of most positive and negative words. We ran each review against this dictionary and the higher affinity is considered as overall affinity. Lastly, we also used the Stanford NLTK core which performs sentence level sentiment calculation by categorizing each sentence into very positive, positive, neutral, negative or very negative [16].We found that the Stanford NLTK library gives the most accurate results, but it is a little slower. The results of our sentiment analysis is discussed in Section 4.

4. EXPERIMENTAL EVALUATION

The below section contains our results of each module described in Section 3. We begin with our results of Data Analysis of scrapped reviews and user profile data. Followed by Review Posting, explaining the results and observations made from each experiment. Next, we record our overall findings which we considered worth mentioning, followed by preliminary results of user profile scanning and some typical characteristics of suspicious users we encountered.

4.1 Data Analysis

From the user profile data, we considered the number of reviews posted per account, to identify any potential posting-volume anomalies. Ideally, the concept was to identify high review numbers as suspicious. For example, even for a user who has existed on Yelp since its launch 12 years ago, having a 2,000+ review count would insinuate a posting rate of one review every other day. For casual users, who are not externally motivated by money or other gains, this seems excessive- since reviews are necessarily tied to physical business visits and experiences. Figure 5 is a truncated bar plot of the number of reviews posted per user account. As anticipated, a dominant majority of users have review counts below 50, and there is a long tail of review counts greater than this. However, this tail turns out to be extremely long, with the maximum review count of our set comprising 7,355 reviews posted by a sole user. The next-highest review count was 2,515, and from this the representative review count values attenuated more gradually. We manually inspected the top 50 user IDs with the highest review counts in hopes of identifying at least one suspicious profile. However, some of these (including the user with 7,355 reviews) were multi-year Elite users, and the rest appeared legitimate based on the consistent quality and subject of their photos, as well as the overall completeness of their profile.

Yelp has Elite users, which are selected each year by a designated Yelp Elite Council. These Elite users are selected from a pool of nominated users submitted by Yelp Community Managers and regular users alike. While there is no specified official criteria for becoming an Elite user, the official Yelp page states that Elite-worthy users are generally selected based on quality of reviews and photos, an active voting and complimenting record, and generally *”playing well”* with other users. Elite users must also appear *”real”* by providing a convincing name and photo, as well as a detailed personal profile. Since the Elite user account represents a human-recognized pinnacle of approved site behaviour, we elected to interpret them as proxies for optimally trusted site users.

We compared select profile attributes and star-rating behaviours of Elite vs. non-Elite users, to identify any significant features a fake reviewer could hone to optimize his trustworthiness. Figure 6 and Figure 7 compare select user attributes for Elite versus non-Elite users. Review count refers to the total number of reviews a user has posted. The Friend count and photo count refer to the total number of friends and photos of a user, respectively. Finally, the States count is the total number of unique states (US) and/or countries a user has posted reviews for. There is clearly a sta-

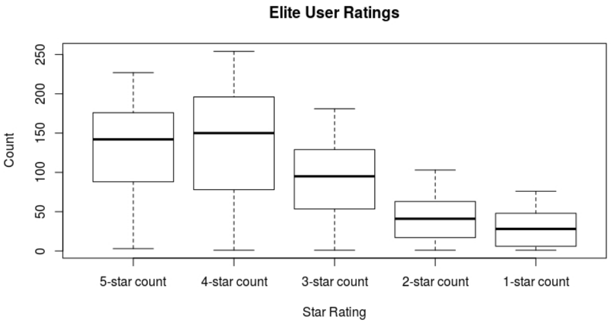


Figure 8: Per-category counts of user star rating for Elite users. Elite users appear to exhibit more diverse star-rating distributions than non-Elite users, which may reflect a more nuanced or high-quality review.

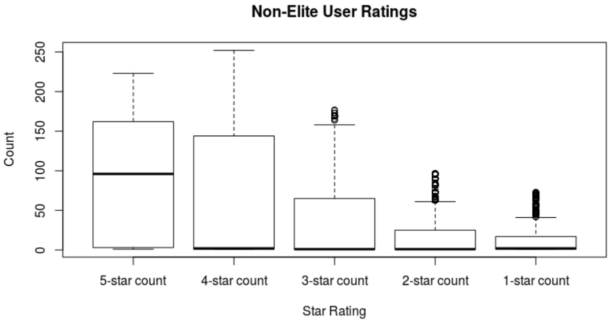


Figure 9: Per-category counts of user star rating for Non-Elite users

tistically significant difference between the friend and photo counts of Elite and non-Elite users. In particular, the mean and median friend counts for Elites are 187.5 and 197.5, respectively, and 100.7 and 29.0 for non-Elites, respectively. If a fake reviewer wished to increase his trustworthiness on the site, garnering friendships and posting photos would likely be the most effective attributes to consider.

Figure 8 and Figure 9 compare the star-rating behaviour of Elite vs. non-Elite users. Non-Elite rating behaviour appears disproportionately skewed toward giving five stars, whereas Elite rating behaviour exhibits a more diverse distribution of ratings, particularly in giving less than five stars. This may be an indication of review quality, in that a more diverse distribution of star ratings insinuates a more nuanced approach to writing reviews âĂŤ since a user with more sensitive criteria for judging businesses is more likely to consider more star-rating categories.

4.2 Review Posting Results

The main objective of this experiment was to analyse the Yelp review filter and identify loopholes, using which an adversary can continue to misuse the system, post fake reviews, and affect business rating. Using the test cases described in Section 3, we proceeded to build automatic Selenium scripts to post reviews on multiple businesses using our user accounts. We maintained random time gaps between each posting to ensure we do not overwhelm Yelp with requests. Further, we used the text from already written reviews for each business to ensure no harm was caused to the various businesses we targeted. The users from which these reviews were posted all belong to our created accounts or our personal accounts and no other account was compromised. The results of each test case is recorded below.

Test Case #0: Identifying the threshold of reviews posted in a day that does not get detected

This experiment was a base case to understand the limits of the system and boundaries within which we could continue all other tests without being detected due to unusual or overloaded behaviour. Since Yelp is a user experience rating application, it is unlikely that users would rate a large number of services everyday, as doing so would be considered abnormal behaviour. This gave us an understanding to be able to isolate that, reviews are removed due to other characteristics, as well as confirm that Yelp checks user activity for normal timing behaviour.

User Profile Characteristic: Newly Created Users with basic profiles (Name & Location)

Procedure:We posted reviews in increments of 10, 20, 30, 40, 50 and 60 from different user accounts, to different businesses and with different IP addresses. A wait time of 5-10 minutes between each post was added to depict more natural behaviour. We also added noise in between these requests, by posting authentic reviews from restaurants with check-ins and completely unrelated topics. For instance, we picked a review of a Chocolate box from Amazon and posted it as a review under a restaurant.

Results/Observations: Our observations consists of the following:

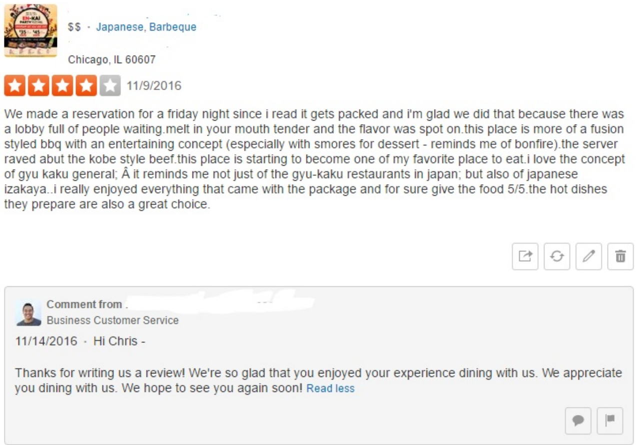
1. Most of reviews were filtered out as "Not Recommended" within the first 24 hours. Beyond this time period, the filtration was slower. Reviews which remained unfiltered after 24 hours tended to remain for a longer time on the business page. Eventually every review was filtered within a period of 10 days

2. Posting as little as 10 reviews a day does not guarantee that the reviews would stay on the site and eventually affect a business rating.

3. It is interesting to note that the unrelated review of the chocolate box remained on the restaurant's page(name not disclosed) for 4 days, where some natural looking reviews were taken down within a day. This tells us that the filtration mechanism is not instant nor accurately checks for relationship between business and text.

4. Some reviews were voted as cool, funny or useful by other users as well as some of our created users. For many cases, the business representatives replied thanking the our users for the review, as shown in Figure 10. These reviews were also filtered out by Yelp. This gives us an insight that reviews voted as funny or useful, or which received a reply from the owner doesn't assure that the review will remain.

5. Even though many of the reviews were removed from the business page itself, the same review remained publicly viewable under particular food menu items tags. For instance, one of our reviews contained the item "fried bacon mac", which was removed from the business page, but was visible on the food page.



$$ · Japanese, Barbeque

Chicago, IL 60607

★★★★☆ 11/9/2016

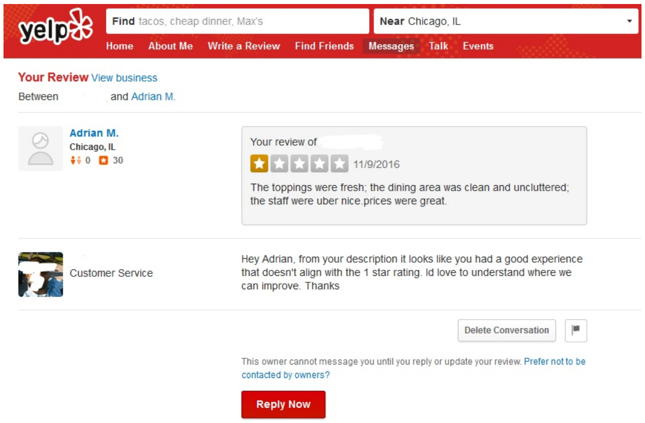
We made a reservation for a friday night since i read it gets packed and i'm glad we did that because there was a lobby full of people waiting.melt in your mouth tender and the flavor was spot on.this place is more of a fusion styled bbq with an entertaining concept (especially with smores for dessert - reminds me of bonfire).the server raved abut the kobe style beef.this place is starting to become one of my favorite place to eat.i love the concept of gyu kaku general; Â it reminds me not just of the gyu-kaku restaurants in japan; but also of japanese izakaya..i really enjoyed everything that came with the package and for sure give the food 5/5.the hot dishes they prepare are also a great choice.

Comment from [redacted]
Business Customer Service

11/14/2016 · Hi Chris -

Thanks for writing us a review! We're so glad that you enjoyed your experience dining with us. We appreciate you dining with us. We hope to see you again soon! Read less

Figure 10: Business Reply to posted Review



Your Review View business

Between [redacted] and Adrian M.

Adrian M.
Chicago, IL
0 30

Your review of [redacted]

★☆☆☆☆ 11/9/2016

The toppings were fresh; the dining area was clean and uncluttered; the staff were uber nice.prices were great.

Customer Service

Hey Adrian, from your description it looks like you had a good experience that doesn't align with the 1 star rating. Id love to understand where we can improve. Thanks

This owner cannot message you until you reply or update your review. Prefer not to be contacted by owners?

Figure 11: Business Reply to posted Review

Test Case # 1: Posting Natural User Reviews, Results detected

User Profile Characteristic: New User created with basic profile information (Name & Location)

Results: With this user we were able to post over 518 reviews in a time span of 12 hours overall. Our automatic script continued to run after this recorded number, although, Yelp blocked this user account without any communication. We were not able to log into this account any more (message: User Account Closed) and all the reviews posted were removed entirely from all the businesses.

Test Case # 2: Positive Review text with Minimum Rating, Results

User Profile Characteristic: Moderate footprint user, with some profile information

- User 1 posted 5 extremely positive reviews with 1 star.
- User 2 posted 5 dismissive reviews with 5 star ratings.
- User 3 (With profile picture) posted 3 natural reviews along with one review of user 1, user 2 each.

Observation/Results:

1. All the reviews for User 1 and User 2 were taken down within 48 hours.
2. Business representatives replied to some reviews pointing out the contradiction between the review and the rating as shown in Figure 11.
3. For user 3, the two reviews with opposite ratings were taken down within 24 hours but the natural reviews stayed for almost 12 days.
4. We can conclude from the above observations that Yelp definitely performs a sentiment analysis on the reviews and it does filter the reviews which demonstrate conflicting sentiments.

Test Case # 3: Reposting the same Older Elite user review

User Profile Characteristic: Golden User Account, Newly Created User Account, Personal User Account created in February

Procedure:Reposted reviews from past elite users on the same business page with same rating. Important point to note here is that even though the review is the same text, their rendering on the browser is different and the line endings were not identical. Hence, this case does not coincide with the same hash value test case.

Results: All reviews posted with the golder user account and the personal account remained on the website whereas, all reviews posted from the new user account were removed in the first 12 hours.

Test Case # 4: Reviews from businesses of same categories

User Profile Characteristic: Golden User Account, Newly Created User Account, Personal User Account created in February

Procedure:As described in test case 4 description in Section 3, the reviews of same categories of businesses were jumbled and posted.

Results: 10 out of 12 reviews posted with the golden user account and all posted from the personal account remained on the website whereas, all reviews posted from the new user account were removed in the first 12 hours.

4.3 Additional Test Results

1. ***Does a profile picture and complete bio for a user profile have relevance in reviews?***

   We wanted to test if an account which has a profile picture and a complete bio filled out is given any preference against the accounts with none/incompletely filled information and no profile picture. To explore this, we conducted the following experiments:

   User Characteristic: We created four different personas; A, B, C and D with the following attributes:

   - A: Most Basic, No details filled; No Profile Picture
   - B: Profile Picture but No other details filled
   - C: No Profile Picture but details filled completely
   - D: Complete Profile

   Procedure: These accounts were created at the same time using different IP addresses. We posted 3 & 5 (total:8) reviews, in two different runs, from each user manually. The reviews posted by each user were identical and were posted to different branches of a chain of restaurant such as The Cheesecake factory, Starbucks, etc. This was done to ensure indistinguishable behaviour of the users except for the biography information.

Observations:

- A: 1 review after 5 days, 1 remained
- B: 4 reviews after 5 days, 2 review after 2 weeks
- C: 4 review after 5 days, 1 remained
- D: 7 reviews after 5 days, 7 reviews after 2 weeks

These observations indicate that having a profile picture and a completed profile definitely helps in building a better Yelp footprint and in most cases result in the review remaining as recommended.

2. ***Does Yelp keep a hash value of all the reviews posted under a business, such that if two identical reviews are posted it catches them?***

   It is essential for Yelp to detect identical reviews posted to the same business multiple times. If this is not implemented efficiently, any business with negative competition could be affected if a user reposts multiple legitimate negative reviews (remaining over-time on the business page).

   User Characteristic:

   - User A: Golden Account with good Yelp footprint (10 accepted reviews)
   - User B: Moderate Account (4 accepted reviews)
   - User C: Golder Account with high acceptance (12 accepted reviews)

   Procedure: Posted the same review, ensuring line endings, text and fonts are identical using user accounts with a gap of 3 days. User A, User B, User C in the order of posting.

   Observations:

   - User A: Review remained on the business page
   - User B: Review was removed within 12 hours, User A review continued to remain
   - User C: Review remained, User A review removed within 24 hours

   From these findings, we can conclude that Yelp keeps a hash of all the reviews. But from the identical reviews the one removed depends upon the user's reputation. As seen above when the same review was posted with User C, the review removed was of User A (which did not have as good a reputation as C). Hence, user's Yelp footprint undoubtedly comes into play while deciding which review would be recommended on the website.

4.4 Other Interesting Findings

While working closely with Yelp and its filtration mechanism we made the following general observations that are not necessarily tagged to any particular experiments.

1. All reviews become live as soon as a user posts them and the review remains on the site for at least a few hours before Yelp filters it out.
2. A user is not notified when their review has been marked as non-recommended. Surprisingly, when a user visits a business page, which they have personally reviewed, from their own profile, the review continues to appear as if it is on the public business page.
3. A user's profile will always show the total count of reviews written by that user posted regardless if they are present on the business page or not.
4. More than 10 business representatives replied to our reviews, some thanking the user or asking for improvement feedback, a few others even mentioned that we have posted the review to a wrong business. Surprisingly, many out of these 10 reviews did not make complete sense as they were abstract sentences put together. The business representatives, in some cases, were even faster than the Yelp filter, since they replied even before Yelp acted on our reviews and moved them into the non-recommended section. This shows that businesses take their reputation on Yelp seriously, and might also have some automatic script that thanks users for their reviews regardless of the content.

5. DISCUSSION

Here we can discuss shortcomings of existing approaches, limitations of our approach, and potential ways to strengthen existing systems.

A major limitation of our project is the lack of time to further investigate our potential sources. For example, from our collected user data we have certain cases of user accounts wherein the number of reviews are less than 15, and such users have written reviews in more than 5 states. Such accounts could be suspicious as they have few reviews, scattered in multiple locations. We manually inspected a few of these accounts and found some typical characteristics, but these are not enough to make any concrete claims. A further analysis could be to consider the time-line of reviews posted for such users to determine if there are short gaps or longer gaps and conclude the account activity is normal/abnormal.

Further, we noticed that Yelp stores the reviews that are marked "Not Recommended" for a business towards the end of a business page. Hence, we could improve our findings by scraping such reviews and drawing comparisons between "Recommended" and "Non Recommended" to understand characteristics of both. We also noticed some users with the following characteristics:

- *User A : [Photos:0, Reviews:178, Friends:41, # 5Star-Rating: 3]*
- *User B : [Photos:0, Reviews:72, Friends:0, # 5Star-Rating: 22]*
- *User C : [Photos:369, Reviews:95, Friends:150, # 5Star-Rating: 58]*

We manually checked if these reviews were accepted by the businesses but we found that all their reviews were marked as "Not Recommended". This seems worth investigating further, since if these users have photos and friends and are a part of the Yelp community for a long time with reviews that seem legitimate, it is strange that Yelp filters out such users as well.

6. RELATED WORK

Earlier attempts to detect overall ratings looked at extracting opinion features present in customer reviews using data mining and natural language processing, and classifying them as positive or negative [5]. In recent years, Deceptive opinion spam was studied using supervised learning methods with linguistic and behavioural features. Jindal and Liu, 2008 [7] made use of Amazon data to train models based on review text, reviewer, and product, to distinguish between duplicate opinions (considered deceptive spam) and non-duplicate opinions (considered truthful). Li and Hitt (2008) show that the distribution of social reviews for many products tends to be bimodal, either tending towards the low extreme 1 star- or high extreme 5-stars and relatively little in the middle. They also argue that this is because humans tend to leave reviews only if they have either a really positive or negative experience and not so much for regular services.

In more relevant depth, similar techniques were applied on Yelp to identify the intricate details of its filtration algorithm, and it was noted that Yelp relies on a more abnormal behaviour (of reviewer) based detection mechanism [11] rather than linguistic characteristics. Lastly, Yelp.com themselves had launched a sting operation to catch businesses who indulge in fraudulent reviews. Businesses identified soliciting in such activates are marked with a ***"Consumer Alert"*** flag for a period of 3 months.

7. CONCLUSIONS

Summarizing our work, we started off with a basic understanding of YELP and its filtration system and a goal to understand how an adversary could bypass this profound system and continue to remain undetected and cause harm. We created multiple user accounts, collected a lot of recommended review data from businesses and looked into some user profiles. On analysis, we drew conclusions that Yelp does not just use factors like IP address checking, Timing between posting of reviews, location of business versus location of user, but it focuses more on the overall User Credibility. As the Yelp slogan goes - *'REAL PEOPLE, REAL REVIEWS'*, they look for a user who's activity they can account for. Our results of posting multiple fake reviews from different types of accounts confirmed this aspect of their system. Hence, if one was to be able to build a trusted reputation (say Elite) on Yelp with initial legitimate activity, then these accounts could easily be used to perform malicious behaviour later on. We also found existing user accounts with questionable characteristics that we would like to further investigate before we make certain claims on their nature.

8. REFERENCES

[1] B. S. B. Local Consumer Review Survey. In *BrightLocal Survey Blog '16: https://www.brightlocal.com/learn/local-consumer-review-survey.*

[2] BlogSpot. An Introduction to Yelp Metrics as of September 30, 2016. In *https://www.yelp.com/factsheet.*

[3] C. Crum. Companies Pay Up For Fake Yelp Reviews. In *http://www.webpronews.com/companies-pay-up-for-fake-yelp-reviews-2016-02.*

[4] C. Flatt. "Reputation Management" Firm Will Get You Positive Reviews on Yelp for 995495. In *http://philly.eater.com/2012/3/28/6601099/reputation-management-firm-will-get-you-positive-reviews-on-yelp-for.*

[5] M. Hu and B. Liu. Mining and Summarizing Customer Reviews. In *KDD '04.*

[6] C. Hutto and E. Gilbert. VADER: A Parsimonious Rule based Model for Sentiment Analysis of Social Media Text. In *AAAI '14.*

[7] N. Jindal and B. Liu. Opinion Spam and Analysis. In *WSDM '08.*

[8] B. Liu. https://www.cs.uic.edu/ liub/FBS/sentiment-analysis.html. In *Positive and Negative Word Dictionary.*

[9] M. Luca and G. Zervas. Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud. In *SSRN Jan '13.*

[10] K. McCormick. The Power of Online Customer Reviews. In *http://propelmarketing.com/blog/2016/power-online-customer-reviews.*

[11] A. Mukherjee, V. Venkatraman, B. Liu, and N. Glance. What Yelp Fake Review Filter Might Be Doing? In *AAAI '13.*

[12] F. A. Nielsen. AFFIN. In *IMM '11.*

[13] M. S. NLTK. http://www.nltk.org/api/nltk.sentiment.html. In *NLTK Sentiment Analysis.*

[14] H.-P. M. Site. https://www.hidemyass.com. In *HMA.com.*

[15] W. M. Site. https://www.windscribe.com/home. In *Windscribe.com.*

[16] C. N. StandFord. http://stanfordnlp.github.io/CoreNLP/. In *Standford Core NLP.*

[17] F. S. Yelp. What is YelpâĂŹs Elite Squad. In *https://www.yelp-support.com/article/What-is-Yelps-Elite-Squad.*

[18] M. S. Yelp. www.yelp.com. In *Yelp FAQs.*