Predictive Expert Identification within the Yelp dataset

David Greenfield (dg2815@columbia.edu), Karan Matnani (ksm2148@columbia.edu)

ABSTRACT

The HITS Algorithm as defined by Kleinberg [1] proposed a mechanism for ranking both Hubs and Authorities by leveraging a graph based solely on hyperlink structure within a focused set of websites. Since its creation the HITS algorithm has been applied more broadly to define authority within graph structures. In this project we seek to first show that we can apply the HITS algorithm to find experts within an online review network (Yelp) and additionally seek to improve the algorithm by adding weighting context derived from prior reviewer accuracy. We will validate this conclusion by comparing the experts generated through application of HITS to the manually tagged experts within the data set ("Yelp Elite") and to the broad dataset.

1. INTRODUCTION

In this project, we seek to identify expert restaurant reviewers within the Yelp dataset by examining the network of interactions between users and businesses (restaurants). Within the platform, we observe that restaurants have several phases in their business life cycle. An "infancy" phase where the restaurant has few reviews and potential for a high variance in the consensus rating over time. During the infancy phase we observe a higher review frequency rate that eventually reaches an inflection point. In testing the data, we observe this inflection point to be on average around 180 days after the first review (figure 1.c). After this inflection point, we consider the restaurant in "mature" phase where the community has reached a consensus on the overall rating and then expect that the consensus rating variance decreases as the impact of a single new review lessens.

We seek to find a set of experts that minimizes the error when comparing reviews during the infancy period and the consensus review achieved after maturity.

2. RELATED WORK

In the paper, *Improvement of HITS-based Algorithms on Web Documents [2]*, the authors explore a similar weighting modification to the original HITS algorithm. In the paper Li et al. use contextual information from the websites to establish an initial weighting. We seek to derive similar context in a review environment to establish an initial weighting.

Another paper, *Popularity Dynamics of Foursquare Micro-Reviews [3]* looks at reactions to short reviews or "tips" on the Foursquare platform. In this paper they observe that tip popularity gradually increases over time rather than diminish. In our study we see similar behavior at the restaurant level within Yelp whereas older restaurants on average receive more new ratings per month than new restaurants (Figure 1)

3. EXPERT DETECTION METHODS

In our research, we explore application and modifications of the HITS algorithm in order to identify expert nodes which more accurately predict maturity consensus ratings.

3.1 Training vs Test Data

In order to avoid bias we use all reviews prior to 01-01-2013 as our trianing set and use reviews 01-01-2013 to 01-01-2014 as our test set. We limit the restaurant nodes in both the sets to only restaurants which at the latest state of their data have at least 20 reviews in order to qualify that the restaurant has both an incubation and a mature phase within the scope of our data set.

3.2 Base Graph Model

In order to apply our algorithms we first set up a graph to represent the relationship between users and restaurants. In our graph we create nodes for users and restaurants and create an edge between a user and a restaurant when that user has reviewed the restaurant within the first 180 days after the initial review. Because we do not know the actual opening date of the restaurant we use the date of first review as a proxy. In the base model all edge weights are 1.

3.3 Accuracy Weighted Graph Model

For our enhanced model we modify the initial weight of the edges to be equal to a function of the accuracy of the user in general within the training data. We calculate the accuracy for each user given a set of user reviews u and a set of restaurant ratings in maturity r each with N items :

$$Accuracy_u = \sum_{i=1}^{N} (\frac{(u_i - r_i)}{4N})^2$$

4. RESULTS

Overall, our results (Appendix: Table 1) confirm that HITS can be used to identify experts that predict the future consensus on a restaurant more accurately than the manually labeled experts in Yelp Elite. Additionally, augmenting the model with an initial weighting based on past review accuracy in most cases improved the results. Below we outline specific conclusions and limitations as well as potential applications for the resultant expert set.

Conclusions:

1. We are able to get the top N experts, unlike the Yelp Elite, which do not give any information about rankings.
2. We saw a consistent positive bias in overall reviews among all groups with the mean rating for all groups being between 3.75 and 3.85 out of 5.
3. As the number of experts increases, we observe a normal distribution of average ratings provided by them. Elite and Non-Elite general populations also fall into a normal curve rating distribution.
4. As noted in the results table, the weighted HITS algorithm performs better than the baseline implementation consistently.
5. The mean percentage error of the Yelp Elite is greater than that of the experts we have found, for up to the top 500-1000 but degrades below Elite or even normal beyond 1000 experts in the weighted HITS algorithm implementation.

Limitations:

1. This method does not perform as well as manual selection when the number of top N experts is increased significantly.
2. Given the constraints in measurement requiring users to make reviews in the infancy period, many users may fall in the rankings if they do not generally make reviews early. This likely causes the drop we see of expert groups 1000-5000 below the general population as users with poor accuracy data would be ranked above users with no data.
3. This system is simple, and therefore gameable. If the system was open and being an expert was incentivized, a user could place reviews near the mean near the end of the incubation period. A modification to weight accuracy early in the incubation period could alleviate that weakness.

Applications:

1. Yelp Elite users are selected manually. We have automated the process, with nearly the same accuracy as that of a human selecting an expert user.
2. After an expert predictor set has been identified, geographic data from the reviews from experts could be used to show geographic regions with a high frequency of positive reviews on new restaurants (Figure 3 and Figure 4)
3. Because our experts are derived based on activity in the incubation period, they would also be useful in other applications that rely on collecting data from early-adopters. One example of this is to identify restaurants with an abmormally large number of expert reviews during the incubation period to identify new restaurants that are popular with trendsetters.
4. The method used in this paper could extend to other datasets & graph problems where past predictive accuracy can be measured.As it works with high accuracy to pick the first N users, where $N < 100$, the method will be useful where a small set of experts can provide useful insights.

5. ADDITIONAL AREAS FOR RESEARCH

While the results from our test showed that we were able to identify a set of predictive experts within the dataset that are more accurate than both the general population and the Yelp Elite experts, we believe that the model could be enhanced even more to identify more accurate predictors.

5.1 Use of Rating Feedback

One way to explore would be to create a second factor for weighting users based on the feedback from the community on their reviews. After a review is placed in Yelp, users who read the review can rate the review as: Useful, Funny or Cool. Using the useful votes on a users reviews could be used as an additional factor to the accuracy of reviews in the test data in the initial weights of the graph may improve the accuracy of the overall model.

5.2 Combination with Multinomial Model

In addition to the social graph based approach outlined in this paper it may be possible to create an expert ranking based on a multinomial regression on characteristics of the user such as: number of reviews, accuracy, number of years elite, number of friends, average star rating. This model would work well with the social graph based model.

5.3 Location Specific Modeling

Much of our modeling is done using population means and trends to determine infancy/maturity. A model with individual location cutoffs would likely be more accurate but require more processing.

6. APPENDICES

6.1 Yelp Elite Qualifications

Description of Qualifications of Yelp Elite from Yelp.com:

The Elite Squad wants you! Or do we...? Here are some of the things we look for:

Authenticity Real people. Real reviews.® That means use your real name, a real (and clear!) profile photo, and an honest, unbiased opinion. You've got a lot of sway in the community; let everyone see your face so they know that you stand by what you say.

Contribution Your reviews for everything from bars to boutiques, and even your trusted accountant, add to the local experience. Reviews of your band, dog, or mother's Pecan Surprise? Not so much. Also share happenings on the Events tab, create lists, and use the mobile app on the go to write meaningful tips, upload photos, and check in.

Connection Vote (Useful, Funny, or Cool) on your favorite reviews and send compliments. Welcome new members and watch out for your community. Choose diplomacy and intelligent wit over crassness and mean-spiritedness; Yelp's a big bowl of cherries, but nobody likes the pits.

Finally, we look for a certain je ne sais quoi when reviewing Elite candidates. Like Supreme CourtJustice Potter Stewart, we know it when we see it.

6.2 Resources and Methods

- Dataset used : The Yelp Dataset provided for the "Yelp Dataset Challenge".
- Database: MongoDb
- Programming Language: Python
- We downloaded the data from the above source, and created the following tables in the database:
  1. business : Records for all businesses. We focussed on restaurants.
  2. reviews : Records for all reviews.
  3. user : Holds user objects from the original data as it is.
  4. $review\_starts$: This is a table we created to get the date of the first review of every business that has receivedd at least one review. These were used in calculating restaurant age, and to get the frequency of reviews for every restaurant over every month.
  5. $reviews\_rest$: This table was created to hold reviews specific to restaurants (among all other businesses) only.
  6. $reviews\_rest\_min20$: This table holds records of only those restaurants, that have a minimum of 20 reviews.
  7. Full Repo available at https://bitbucket.org/SocialNetworksProject/socialfinalproject/overview

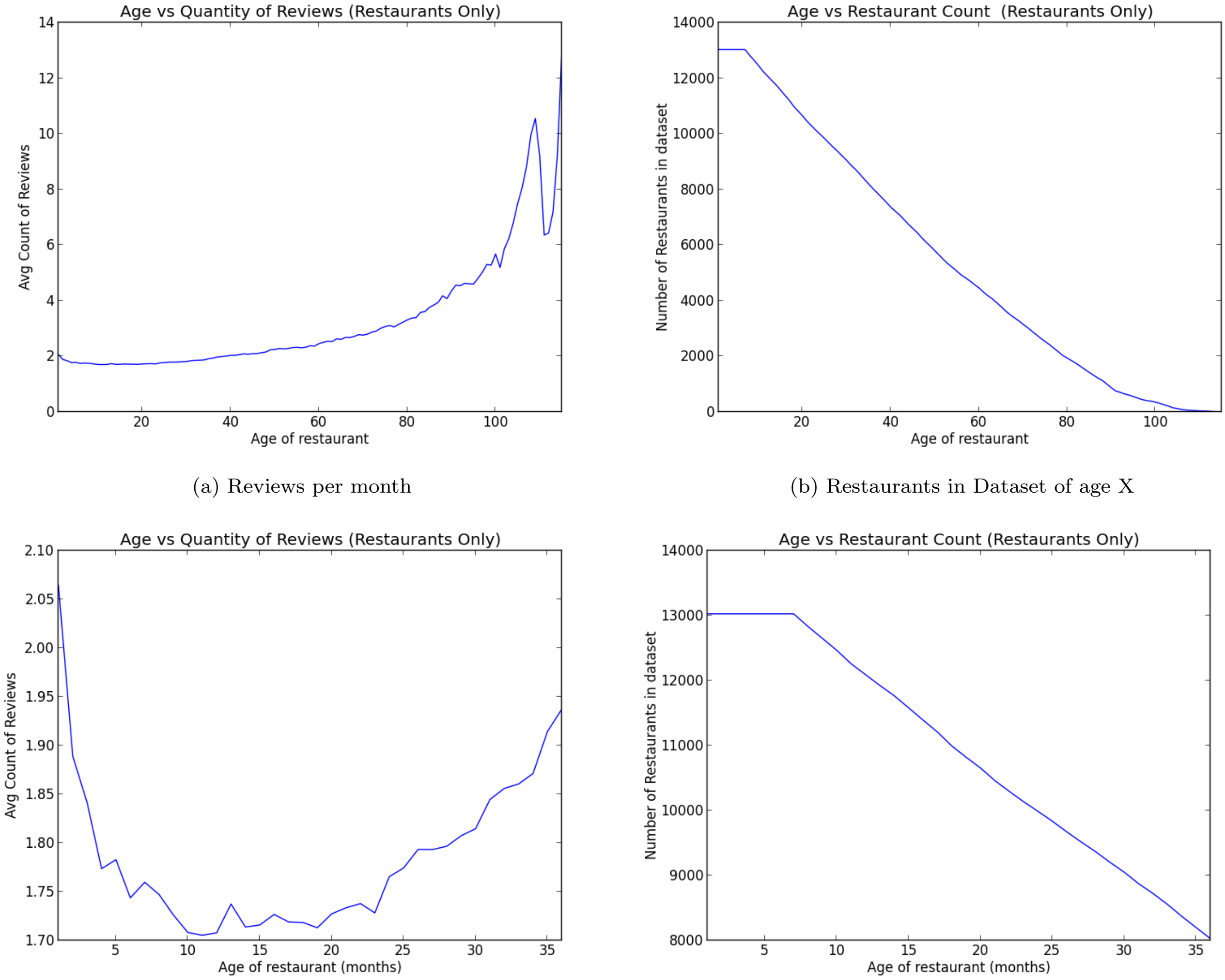
7. REFERENCES

[1]Jon Kleinberg . *Authoritative Sources in a Hyperlinked Environment.* 1999

[2]Longzhuang Li, Yi Shang, Wei Zhang. *Improvement of HITS-based Algorithms on Web Documents.* 2002.

[3]Marisa Vasconcelos, Jussara Almeida, Marcos Gonçalves, Daniel Souza, Guilherme Gomes. *Popularity Dynamics of Foursquare Micro-Reviews.* 2014.

8. DATA AND CHARTS



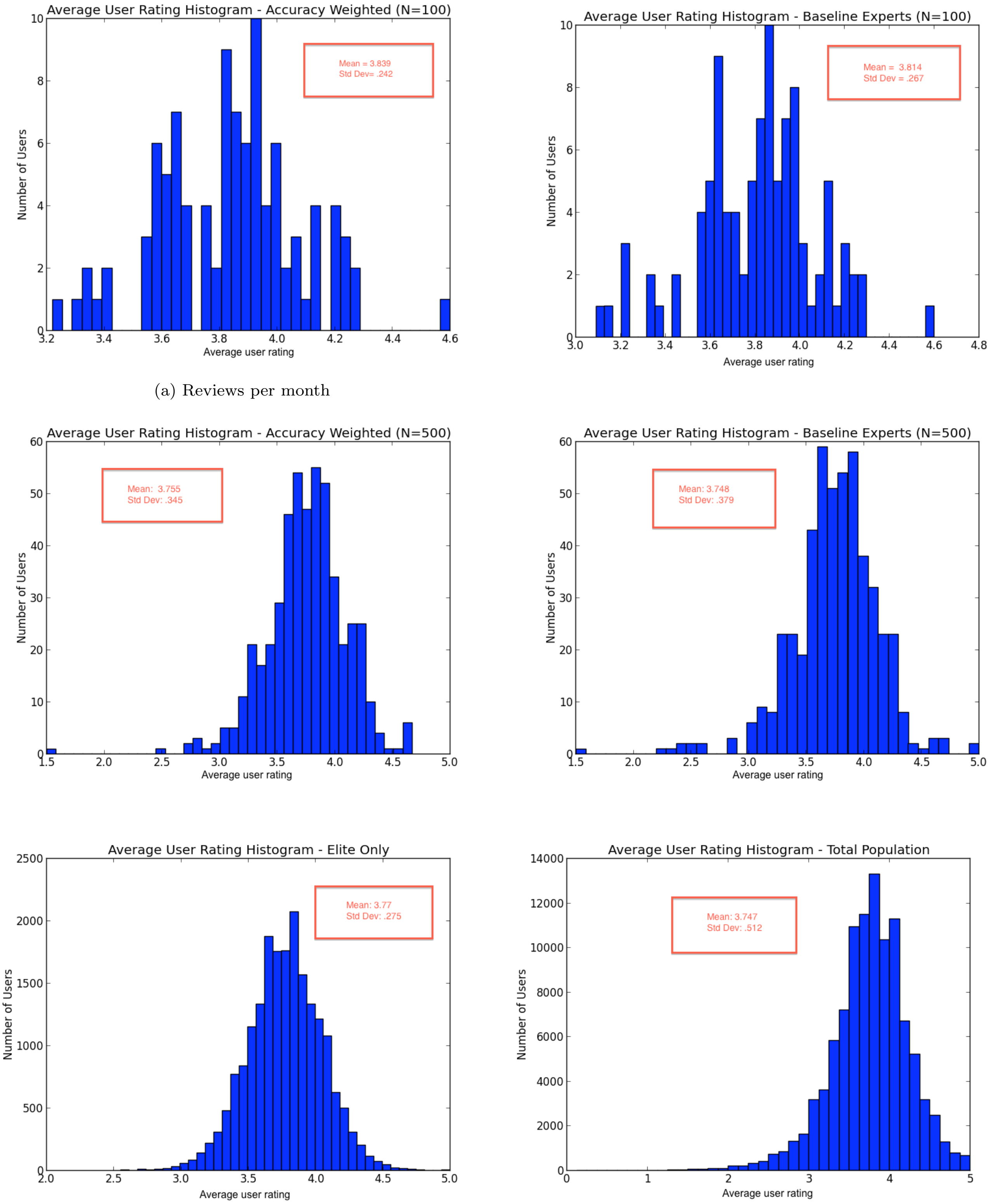
(a) Reviews per month

(b) Restaurants in Dataset of age X

(c) Reviews per month

(d) Reviews per month

Figure 1: Review Rate Graphs



(a) Reviews per month

Figure 2: Review Rate Graphs

User Subset	Mean % Error	#Reviews(Test)	P-Value vs Elite
Systemic Error	6.2	19907	NA
Total Population	23.5	19907	NA
Yelp Elite	18.4	3997	NA
Baseline HITS Top 10	14.4	188	<.01
Weighted HITS Top 10	14.8	170	<.01
Baseline HITS Top 50	16.1	589	<.01
Weighted HITS Top 50	15.6	413	<.01
Baseline HITS Top 100	17.2	963	.07
Weighted HITS Top 100	15.8	597	<.01
Baseline HITS Top 200	17.8	1311	.61
Weighted HITS Top 200	15.8	720	<.01
Baseline HITS Top 300	18.3	1553	.50
Weighted HITS Top 300	16.8	1087	.01
Baseline HITS Top 400	18.6	1706	.16
Weighted HITS Top 400	17.0	1255	.02
Baseline HITS Top 500	18.7	1772	.09
Weighted HITS Top 500	17.2	1317	.07
Baseline HITS Top 1000	19.3	2247	<.01
Weighted HITS Top 1000	18.3	1742	.47
Baseline HITS Top 2000	20.3	3227	<.01
Weighted HITS Top 2000	18.4	2459	.33
Baseline HITS Top 5000	20.76	4691	<.01
Weighted HITS Top 5000	19.86	3760	<.01

Table 1: Results

Key Definitions

Systemic Error: The systemic error is the error inherant to the system due to the inability of users to rate more granularly than an integer star rating. For example, a restaurant with a mean rating of 4.5 would cause an error of .5 for a user inputing a 4 or 5 star rating. We expect that systemic error would approach 6.25 percent in as the system grows large as that is halfway between the largest unavoidable error of .5 per review and the minimum unavoidable error of 0.

Mean Pecentage Error: This key figure is defined as the average difference between the review star rating in infancy r_i and the latest date consensus star rating of the restaurant in maturity r_m. We grade error as a percentage of the maximum 4 rating error representing a consensus rating of 1 and a review of 5 stars or vice versa.

$$MeanError_u = \sum_{i=1}^{N} \frac{|(r_i - r_m)|}{4N}$$

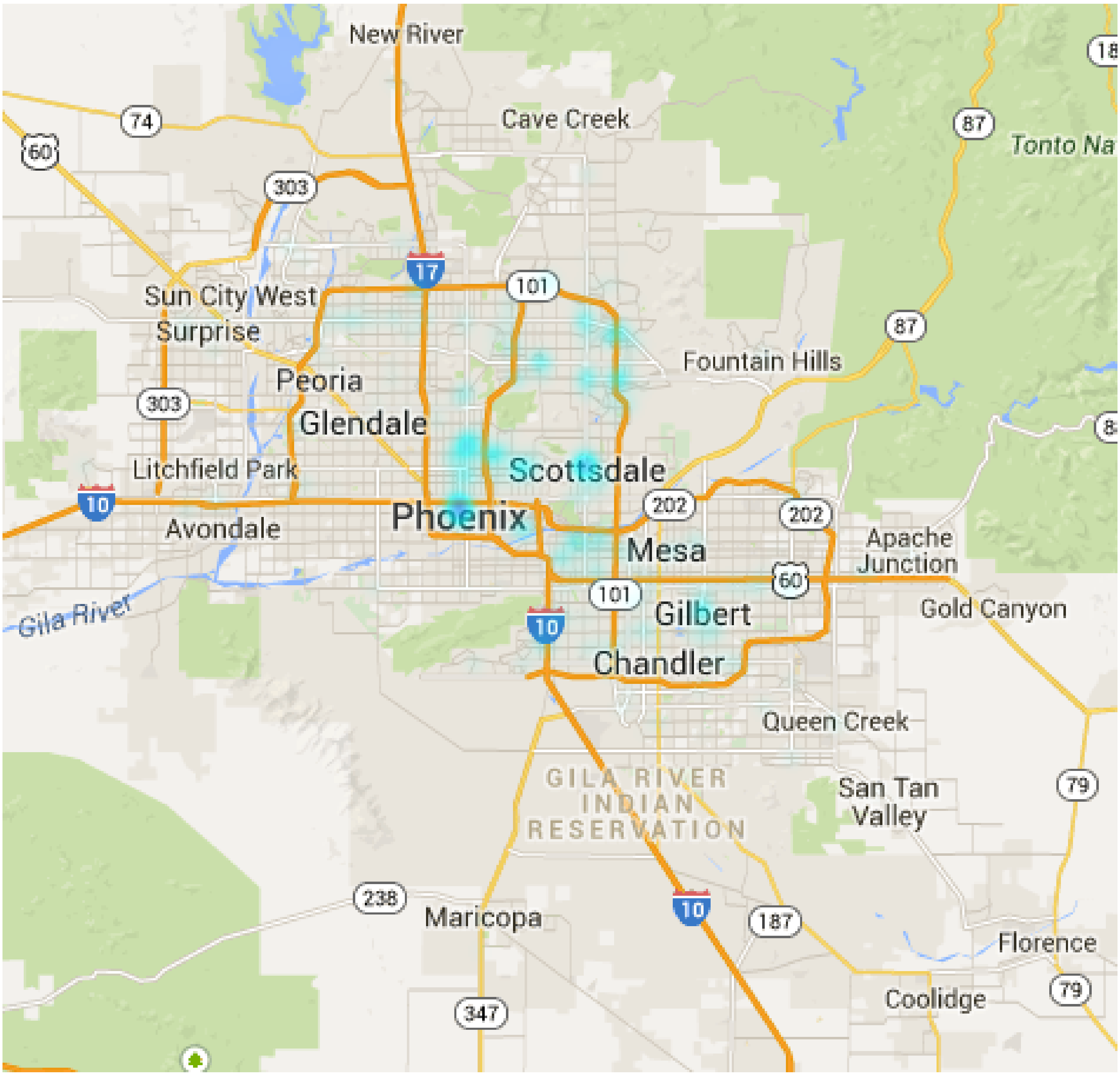


Figure 3: Heatmap of 4 or 5 Star Reviews at incubation period restaurants by experts in Pheonix during 2013 Dynamic Version at: http://www.dgreenfield.com/Heatmap/map1000.html

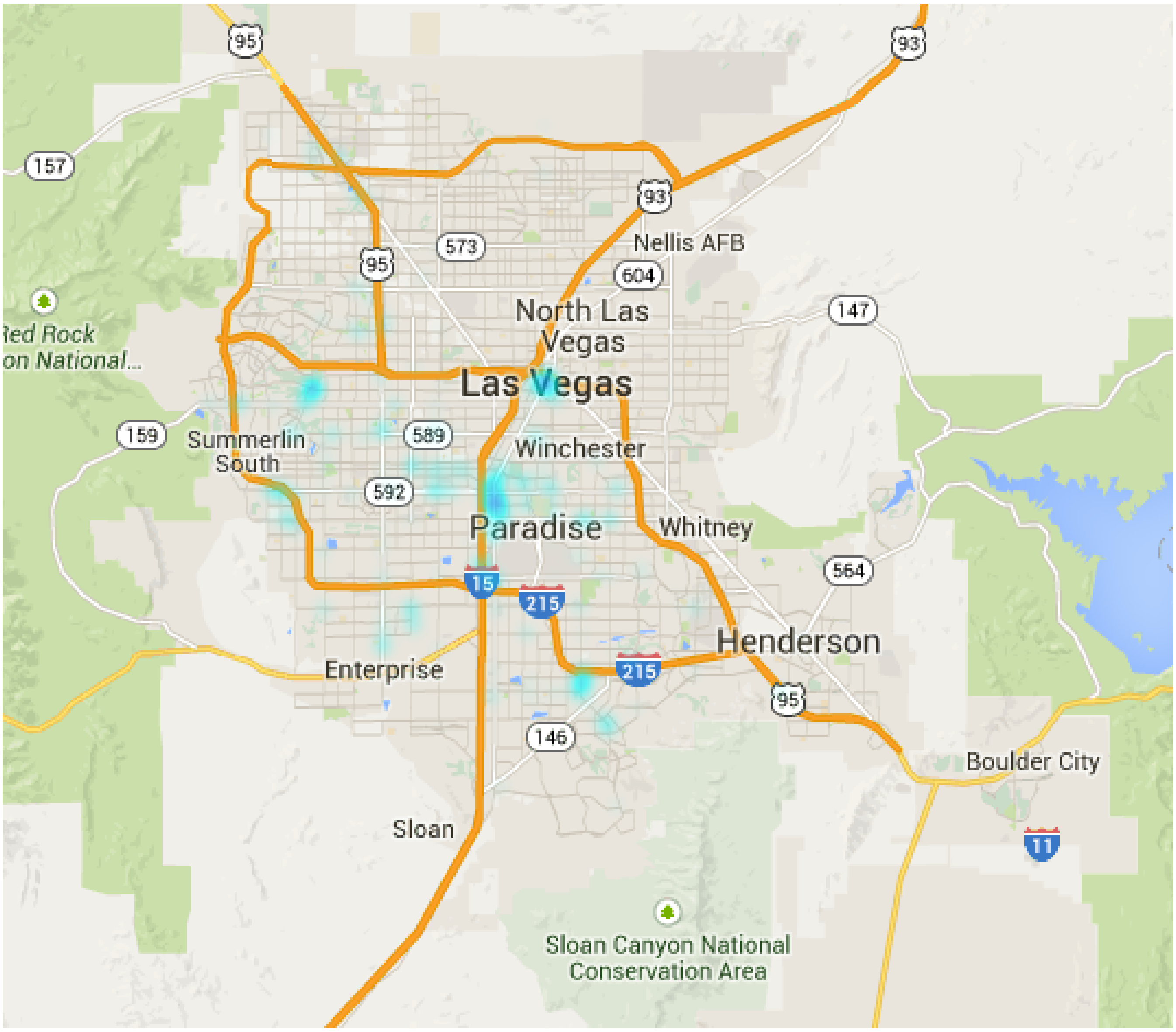


Figure 4: Heatmap of 4 or 5 Star Reviews at incubation period restaurants by experts in Las Vegas during 2013 Dynamic Version at: http://www.dgreenfield.com/Heatmap/map1000.html