Survival of Gentrification / Depreciation in Restaurants

F. M. Marsh and J. A. Clithero

February 16, 2017

Abstract

1 Introduction

Many large technology services have a large amount of data on they users. This data is used to predict consumer behavior and target advertising to maximize revenue. In order to improve their services, some companies release some of their data to the public. Often, companies that release their data to the public do so to discover new ways of using their data. For example, the business review company Yelp hosts the "Yelp Dataset Challenge", where scholars can access the data and can donate their time to help Yelp understand their users.

Participants in The Challenge have produced original research on a variety of topics.

[Alghunaim, 2015, Byers et al., 2012, Cawkwell et al., 2015, Chepurna and Makrehchi, 2015, Feng and Qian, 2013, Gutierrez, 2014, Hajas et al., 2014, Hu et al., 2014, Liu et al., 2015, Mashhadi et al., 2012, Quattrone et al., 2015]

The Yelp and Zillow public datasets have previously been combined, to produce a web tool that provides neighborhood-based restaurant information [Bonnar et al.].

Our goal is to use the correlation of two time-series:

1. The monthly median rent, as tracked by Zillow Rental Data. 2. The median restaurant review rating (stars) for each restaurant in a neighborhood.

Zillow rental data can be used to detect appreciating, and depreciating neighborhoods.

As rents rise in a given neighborhood, which types of businesses fare / worse better in the reviews? As rents fall in a given neighborhood, which types of business fare / worse better in the reviews?

We hope to present concrete suggestions to restaurant owners to improve the survivability of their businesses in times of strong appreciation / depreciation in the housing market.

2 Data

2.1 Yelp Academic Dataset

The Yelp Academic Dataset (available at `https://www.yelp.com/dataset_challenge/dataset`) contains five files. There are separate files for businesses, users, reviews, checkins.

In this study, we will use three of the files: we will use the files on businesses, reviews and users.

The Yelp Dataset Business file includes data on 77,445 businesses in the metro areas of Las Vegas, NV, Phoenix, AZ, Charlotte, NC Pittsburgh, PA Champaign, IL, Kitchener,Canada, Montreal, Canada, Edinburgh, Scotland, Karlsruhe, Germany.

The Yelp Dataset User file includes data

The Yelp Dataset Review file includes data

2.2 Zillow Public Dataset

The Zillow Public Dataset (available at `http://www.zillow.com/research/data/#bulk`) contains data on home value indices for various neighborhoods across the United States. Zillow has created a proprietary index of home value, called the "Zillow Home Value Index" (hereon ZHVI). The methodology used to calculate ZHVI can be found at `http://www.zillow.com/research/zhvi-methodology-6032/` [**?**].

Zillow divides homes into geographic "neighborhoods" with boundaries. ZHVI is reported on a monthly basis for 6,958 neighborhoods across the US. Zillow Rental Index (hereon ZRI) is reported for studio, one, two, three, four and five or more bedroom apartments are reported for a smaller set of about 300 neighborhoods. The Zillow neighborhood boundaries (initially released in 2008) can be accessed (in ESRI arcGIS shapefile format) at `http://www.zillow.com/howto/api/neighborhood-boundaries.htm`.

3 Methods

In §3.1 we describe how the restaurant locations from the Yelp Dataset and the neighborhood boundaries from the Zillow dataset were combined.

In §3.3 we describe how each users was assigned a most probable Zillow neighborhood to live in.

3.1 Combination of Datasets

In this section, we describe how we sort each Yelp business into its appropriate Zillow neighborhood. This step is necessary to attach a neighborhood Zillow Home Value Index (ZHVI) to each restaurant.

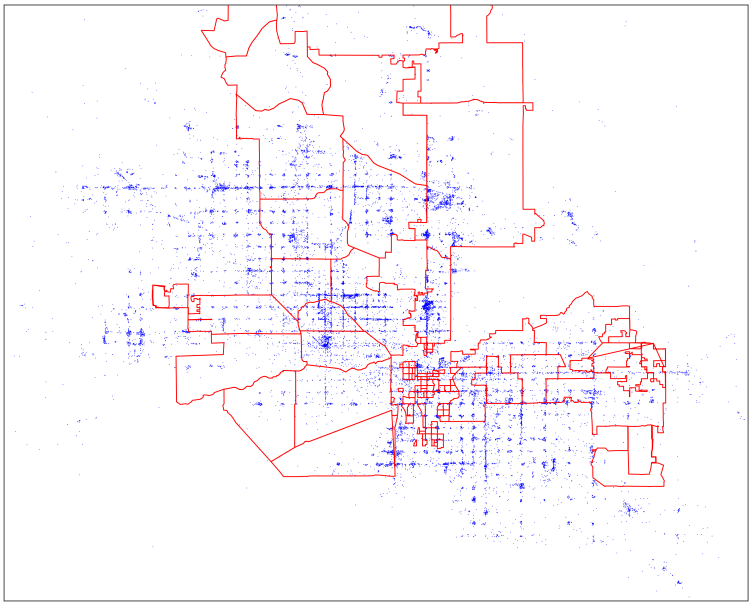

Figure 1: Yelp Businesses (points in blue) and Zillow neighborhood boundaries (lines in red) for the Phoenix, AZ metro area. In §**??**, we describe how we sort each Yelp Business into its appropriate Zillow neighborhood.

Each Yelp business is tagged with a geographic (latitude, longitude) coordinate, and each Zillow neighborhood is reported as an ArcGIS Shape (`.shp`) file. The shapefile includes a set of (typically 100 to 200) (latitude, longitude) points that describe the boundaries of each neighborhood. This file also includes a set of four points that describe the four corners of the neighborhood's *bounding box*: the smallest box in latitude and longitude that includes the entire neighborhood polygon.

To perform the sorting of Yelp businesses into Zillow neighborhoods, we employ a two-step approach. In the first step, we test every Yelp business for inclusion in the bounding box of every Zillow neighborhood. In the second step, we test every Yelp business for polygon inclusion in the neighborhoods which bounding boxes it lies within. We use this two-step approach because the first step can rule out all but two or three of the 6,958 possible Zillow neighborhoods.

We test each Yelp business for inclusion in the set of 6,958 bounding boxes. In Fig. 2, we see a randomly selected Yelp business, displayed as a red point. We see that this

business is included in the bounding boxes of two Zillow neighborhoods.

We then test for point-in-polygon inclusion using an implementation of a ray-casting method in `Python` [**?**]. For each Yelp business, we only test the Zillow neighborhoods whose bounding boxes it lies within.

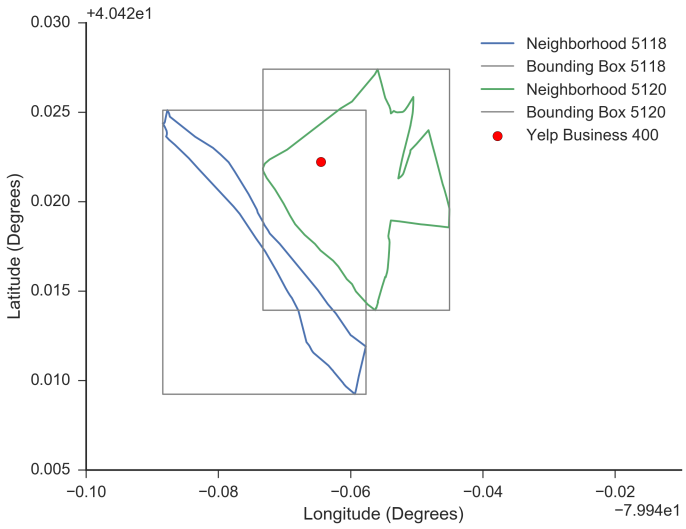

Figure 2: Example of neighborhood and neighborhood bounding box inclusion method. Yelp business 400 (the red point) is included in the bounding boxes of two Zillow neighborhoods. It is only included in one neighborhood polygon, however.

This process assigns a Zillow neighborhood ID `z_hood` to each Yelp business which resides in a Zillow neighborhood.

3.2 Area Computation

We compute the geographic area of each neighborhood in square miles. We retrieve the (longitude, latitude) points that describe the boundary of each neighborhood and use

3.3 Yelp User Description

We would like to determine where each Yelp user lives, in order to estimate their income. A Yelp user can write reviews of different restaurants. It is common for one Yelp user to write reviews of more than one restaurant. We can see a list of the restaurants that each user has reviewed.

We assume that the user resides in the neighborhood which contains the their *medoid* restaurant.

3.4 Description of Combined Dataset

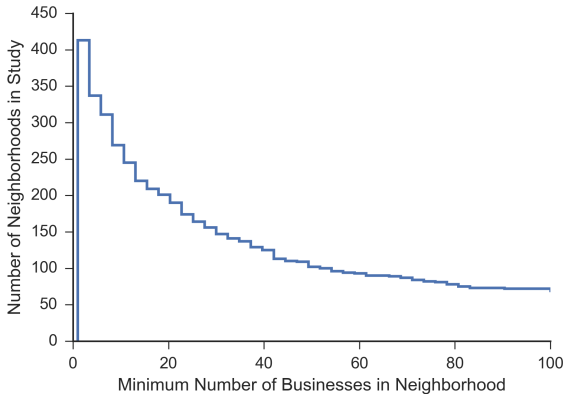

Figure 3:

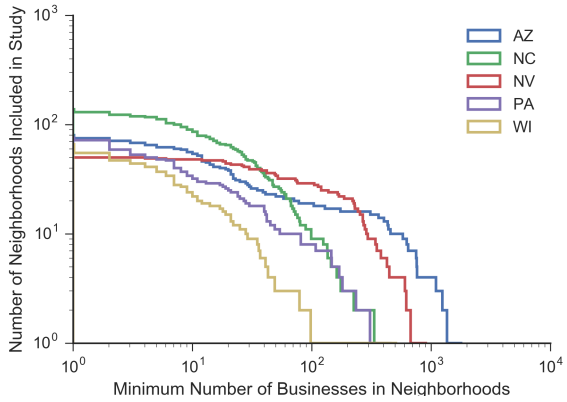

Figure 4:

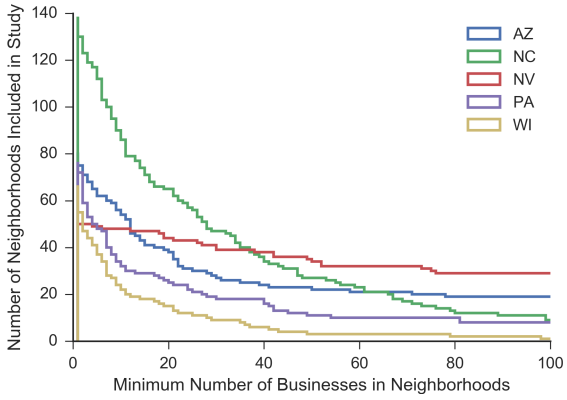

Figure 5:

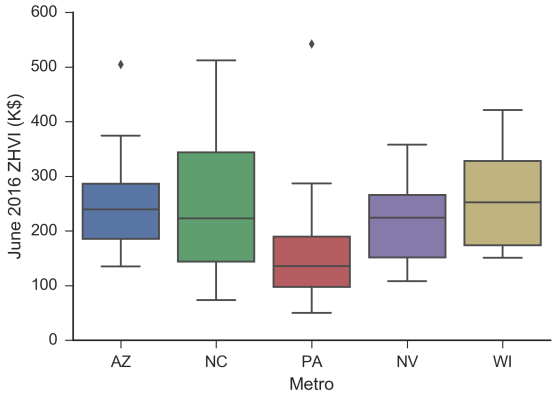

Figure 6:

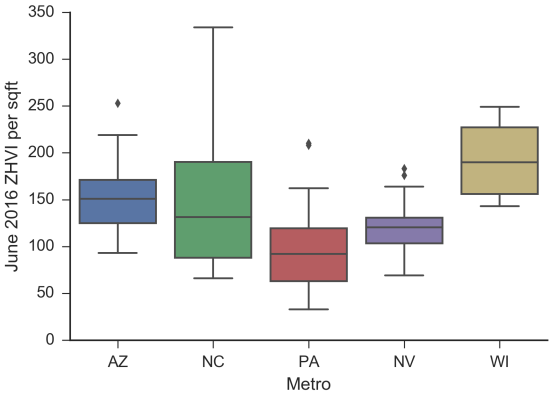

Figure 7:

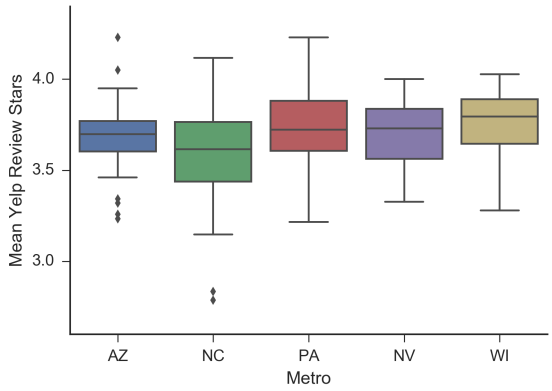

Figure 8:

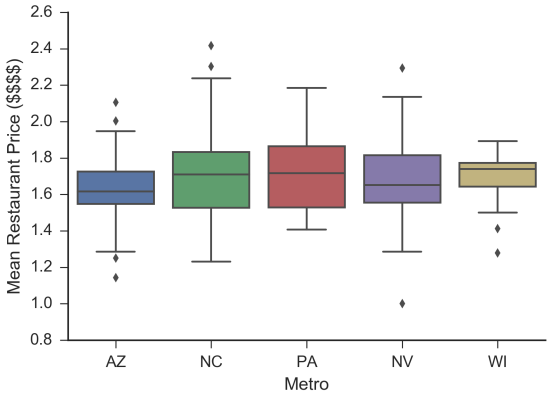

Figure 9:

Metro	Original	Retained	Rate
Phoenix, AZ	32,615	19,958	61.2%
Charlotte, NC	6,162	4,853	78.7%
Las Vegas, NV	21,233	9,694	45.7%
Pittsburgh, PA	3,754	2,749	73.2%
Madison, WI	2,802	1,355	48.3%
Champaign, IL	x	0	0%
Kitchener, Canada	x	0	0%
Montreal, Canada	x	0	0%
Edinburgh, Scotland	x	0	0%
Karlsruhe, Germany	x	0	0%

Table 1: min n = 0

The combined Zillow and Yelp Dataset contains

Metro	median (B/N)	mean (B/N)	Total B	Total N
Phoenix, AZ	116.5	467.6	19,640	42
Charlotte, NC	43.0	66.1	4,296	65
Las Vegas, NV	178.0	218.6	9,619	44
Pittsburgh, PA	43.5	94.8	2,464	26
Madison, WI	36.0	71.4	1,071	15

Table 2: min n = 20

state	median	mean	sum	len
AZ	31.609536	62.769432	2636.316160	42
NC	2.776164	3.466144	225.299334	65
NV	11.553754	22.659192	997.004433	44
PA	1.192682	1.586050	41.237288	26
WI	0.057833	0.066647	0.999710	15

Table 4: n = 20

City	median (B/N)	mean (B/N)	Total B	Total N
Charlotte	43.0	66.1	4296	65
Henderson	105.0	153.0	2907	19
Las Vegas	241.5	242.0	5809	24
Madison	36.0	71.4	1071	15
Mesa	491.5	509.5	3057	6
North Las Vegas	903.0	903.0	903	1
Phoenix	625.0	763.30	11450	15
Pittsburgh	43.5	94.8	2464	26
Scottsdale	2100.0	1531.7	4595	3
Tempe	26.5	29.9	538	18

Table 3:

3.5 Determination of Most Common Yelp Categories and Chains

Each Yelp business has user-generated tags, that allow other users to determine what genre the business is. In the full Yelp Dataset, there are 892 distinct categories. In the full Yelp Dataset (n = 77,445)

Each Yelp business has a user-generated name. Many prominent chains have locations in neighborhoods across the country

category	counts
Restaurants	25071
Shopping	11233
Food	9250
Beauty & Spas	6583
Health & Medical	5121
Nightlife	5088
Home Services	4785
Bars	4328
Automotive	4208
Local Services	3468
Active Life	3103
Fashion	3078
Event Planning & Services	2975
Fast Food	2851
Pizza	2657
Mexican	2515
Hotels & Travel	2495
American (Traditional)	2416
Sandwiches	2364
Arts & Entertainment	2271

Table 5: The 20 most common Yelp tags in the Full Dataset (n = 77,445).

counts	name
483	Starbucks
365	Subway
345	McDonald's
200	Walgreens
180	Taco Bell
155	Pizza Hut
147	Burger King
144	Wendy's
134	The UPS Store
120	Panda Express
119	Dunkin' Donuts
118	Bank of America
114	Great Clips
108	Wells Fargo Bank
107	Circle K
97	Domino's Pizza
95	Chipotle Mexican Grill
95	Jimmy John's
93	KFC
88	US Post Office

Table 6: The 20 most common Yelp restaurant names in the Full Dataset (n = 77,445).

3.6 Yelp Review Description

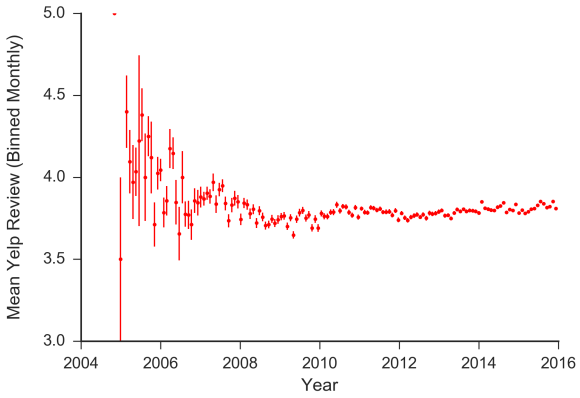

Figure 10:

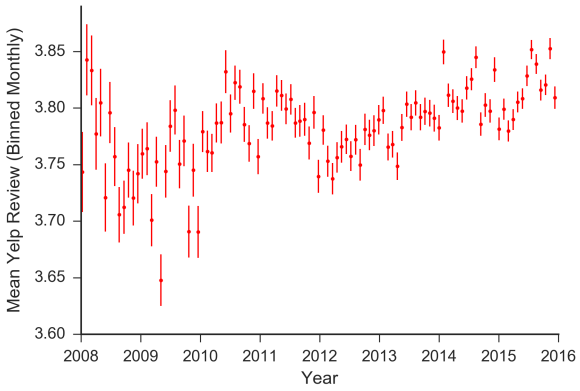

Figure 11:

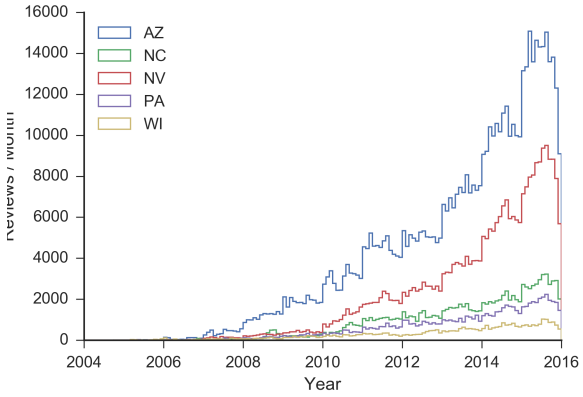

Figure 12:

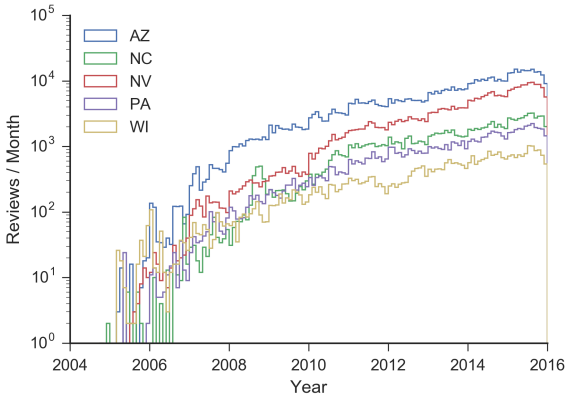

Figure 13:

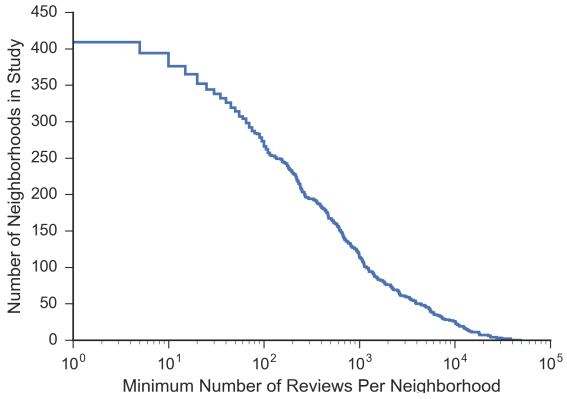

Figure 14:

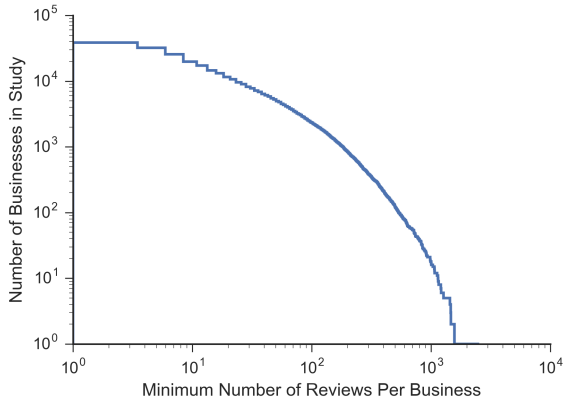

Figure 15:

3.7 Computation of Price Trend

4 Results

Ideas:

1. For each chain, how does chain density (spatial, and fractional) relate to ZHVI.
2. How does "chainy-ness" ($\frac{N_{\text{chain}}}{N_{\text{local}}}$) relate to neighborhood ZHVI / size.

3. For each chain, how does chain review correlate with neighborhood ZHVI / size.

4.1 Statics Results

In this section, we compute static results for each neighborhood.

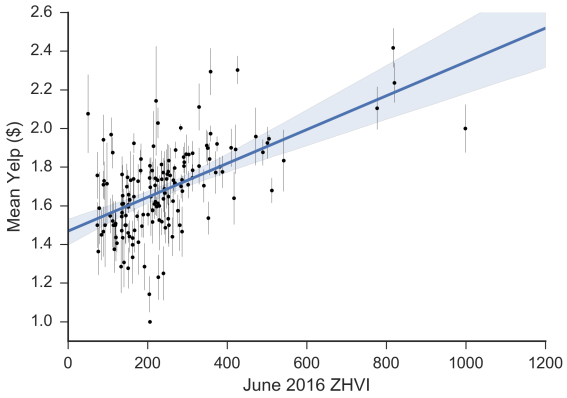

Figure 16:

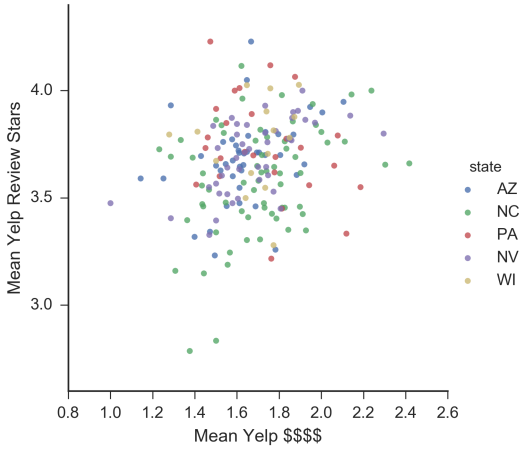

Figure 17:

5 Correlation of Yelp Rating and Zillow Housing Time-Series

Zillow reports the ZHVI on at the end of each month from 1996 to 2016. Yelp reports reviews by date.

We first normalize the ZHVI time-series for each neighborhood, $Z(t)$. For each time t:

$$Z'_h(t) = \frac{n(Z_h(t))}{\sum_{h=1}^{n} Z_h(t)} \tag{1}$$

Then for each neighborhood h:

$$Z''_h(t) = \frac{m(Z'_h(t))}{\sum_{t=1}^{m} Z'_h(t)} - 1 \tag{2}$$

We perform an identical normalization to the Yelp rating time-series $Y(t)$. For each time t:

$$Y'_h(t) = \frac{n(Y_h(t))}{\sum_{h=1}^{n} Y_h(t)} \tag{3}$$

Then for each neighborhood h:

$$Y''_h(t) = \frac{m(Y'_h(t))}{\sum_{t=1}^{m} Y'_h(t)} - 1 \tag{4}$$

$$\tau_{\text{delay}} = \overset{\arg\max}{t} \Big((f \star g)(t) \Big) \tag{5}$$

For each neighborhood h, we wish to compute

References

A. Alghunaim. *A Vector Space Approach for Aspect-Based Sentiment Analysis*. PhD thesis, Massachusetts Institute of Technology, 2015.

C. Bonnar, F. Cordeiro, and J. Michelman. With a little help from yelp.

J. W. Byers, M. Mitzenmacher, and G. Zervas. The groupon effect on yelp ratings: a root cause analysis. In *Proceedings of the 13th ACM conference on electronic commerce*, pages 248–265. ACM, 2012.

P. B. Cawkwell, L. Lee, M. Weitzman, and S. E. Sherman. Tracking hookah bars in new york: Utilizing yelp as a powerful public health tool. *JMIR public health and surveillance*, 1(2), 2015.

I. Chepurna and M. Makrehchi. Exploiting class bias for discovery of topical experts in social media. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 64–71. IEEE, 2015.

H. Feng and X. Qian. Recommendation via user's personality and social contextual. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1521–1524. ACM, 2013.

L. A. Gutierrez. *Noise reduction in user generated datasets*. PhD thesis, RENSSELAER POLYTECHNIC INSTITUTE, 2014.

P. Hajas, L. Gutierrez, and M. S. Krishnamoorthy. Analysis of yelp reviews. *arXiv preprint arXiv:1407.1443*, 2014.

L. Hu, A. Sun, and Y. Liu. Your neighbors affect your ratings: on geographical neighborhood influence to rating prediction. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 345–354. ACM, 2014.

J. Liu, J. Shang, C. Wang, X. Ren, and J. Han. Mining quality phrases from massive text corpora. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 1729–1744. ACM, 2015.

A. Mashhadi, G. Quattrone, L. Capra, and P. Mooney. On the accuracy of urban crowdsourcing for maintaining large-scale geospatial databases. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, page 15. ACM, 2012.

G. Quattrone, L. Capra, and P. De Meo. There's no such thing as the perfect map: Quantifying bias in spatial crowd-sourcing datasets. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1021–1032. ACM, 2015.