Gentrification Index Using Yelp Data

Gene Burinskiy, Jianwei Li, Tianyi Wang, Xiaoxian Wu

December 10, 2015

1 Introduction

1.1 Motivation

Gentrification is a contentious topic as it has the potential to substantially raise the living standards and aspirations of a community whilst displacing the residents who would most benefit from the improved services and increased wealth. Consider San Francisco's well known story. With the influx of prosperous tech companies and the beneficiaries of their wealth, San Francisco's property prices have sky rocketed thereby pricing out less affluent individuals who work in or around the city. Driven by the incentive of high property prices, property owners favor the more affluent and either cater new development to them or invoke forces to vacate the rentals of those who cannot pay the higher prices. As a result, many long-time residents, particularly the less affluent ones, feel pressure to move out of the area in search of more affordable housing options. Of course, with the increased wealth and development, local services improve, businesses also improve, and foster an overall improved living standard. However, the poor who would see the most incremental benefit of improved schools and services can no longer reside in these communities and thus do not benefit from these improvements. Consequently, the gentrifying process can cause justified alarm among the present residents.

Examining the above process, we note that its focus is property prices and so the only way to characterize a neighborhood as one being gentrified is by looking at the change in the property value. If they grow substantially for a sustained period then it is frequently considered as a neighborhood that is being gentrified. Yet, this type of analysis ignores many factors, including the business movement aspect. The gentrifying process typically first initiates through an influx of individuals with higher disposable income moving to an affordable area and that, often through a concerted effort, lures in businesses that cater to their tastes. These new businesses tend to be more expensive than existing ones and so can be a signal of ongoing gentrification. Hitherto, this process could not be examined because there was no data on this process. We gained a full access to San Diego Yelp on businesses and its users, including its users' consumption history and are thus able to, for the first time, examine the issue of gentrification from a business perspective. Our goal is to identify communities that are likely being or will be gentrified using only the user and business data from Yelp. At the present moment, in favor of simplicity we ignore more nuanced processes of time, space, business types, and coordination with housing data though we wish to incorporate them in future work.

2 Method

2.1 Data and Motives

The largest headache entailed the tying of seemingly unrelated user and business data to gentrification. On the business side of the Yelp data, we utilize the following fields: location of the

businesses, their price range, and the self-assigned business category. In addition, we also have operation hours, the text and date from all of the businesses' reviews, and miscellaneous data such as whether they accept credit cards though these remain unused at the moment. On the user side, we know all of the businesses that the users have reviewed on Yelp and the business data on those reviewed. Additionally, we are able to construct a user network based on either on a list of their peers or the percent of commonly reviewed businesses but we do not use these elements yet.

As is readily apparent, we do not know which businesses are gentrifiers or who are the users who tend to gentrify and, overall, how can we construct an index on the likelihood of a community's gentrification. The final process boils down to five steps: (0) generate business communities (1) identify gentrifying businesses, (2) categorize users based on their consumption history using both prices and business types, (3) identify which category of users tends to consume gentrifying businesses most frequently, (4) approximate the relative composition of the each group type in a community, (5) relate a user category's tendency to gentrify and that group's prevalence in a community to an overall likelihood of being gentrified.

2.2 Communities

A gentrifying index only makes sense over an area and is vacuous when assigned to each business individually so we created business communities based on their tendency to cluster together. The city of San Diego officially designates 40+ communities through the city bounds but does not publish the community bounds so we had to approximate these communities somehow. Though numerous unsupervised clustering algorithms exist, a simple k-means classifier seemed to do a good job in approximating the community designations so we opted for that. This stage of the index-building could be honed through the use of alternative unsupervised clustering algorithms as it is unlikely that each cluster is of the same size, as k-means creates. Consequently, an algorithm, such as spectral clustering, that allows of unequal group sizes may be better at approximating communities but that remains unimplemented at the moment.

2.3 Identifying the Gentrifiers

We define a business as a gentrifier if it is more expensive than the businesses near it as that suggests that it may not cater to the predominant local tastes. Numerous caveats abound with this definition. For one, suppose an expensive restaurant has been in a community for over 50 years while the community around it changed to reflect a more affordable composition. In our static picture, this restaurant will be classified as a gentrifier despite its established belonging to the community; however, without a time element we cannot nuance our approach to identify these types of cases. Another potential issue is the level of price discrepancy that is sufficient to classify a business as a gentrifier. Most businesses are inherently \$\$ and the distribution of price ranges is skewed for pricier businesses. In such a case, if we map the price range set $\{\$, \$\$, \$\$\$, \$\$\$\$\} \to \{1, 2, 3, 4\}$ and let $p \in \{1, 2, 3, 4\}$, and use numerical averages to classify businesses as gentrifiers, a $p \in (1, 2)$ is ambiguous in regards to the businesses' gentrifying status. As a result, we omit any business that has a price range of 2 and use only inexpensive and expensive categories to identify gentrifiers. Namely, we define any businesses with price category $p = 1$ is an inexpensive business whereas any business with price range $p > 2$. Although this results in a loss of nuance and sample size, we could not find a manner of classification that would do well without this omission.

Following the labelling of businesses as either expensive or not, we focus on labeling businesses as gentrifiers using classification error under the presumption that a gentrifier should be different from the predominant local price ranges. To achieve this, we opt to classify a business

based on its neighbors. Let $i \in B$ be a business, where B is our set of businesses. If the majority of $j \in B(i) = \{j \in B : d(i,j) < r\}$ are inexpensive then $\hat{p_i} = 1$ and if the majority of $j \in B(i)$ are expensive then $\hat{p} > 2$. Because we know the true value of p, we can then compare $p = \hat{p}$ and if $p > \hat{p}$ then we classify business i as a gentrifier.

Though we have a plethora of supervised classification learning algorithms, we opted to try SVM with the radial kernel and the radial nearest neighbor. One issue is that both algorithms require a subjective input for the range of search which proved difficult to assess. As a result, we picked the radius that seemed to classify a reasonable proportion of expensive businesses as gentrifiers. After speaking with city planners and those in the industry, it was suggested we use search at maximum over block. In downtown San Diego, the approximate diameter of a block is .001 degrees so we used a half of that or $r = .0005$ to avoid reaching too far into other blocks while still searching over the area. This returned about 10% of the expensive businesses as gentrifiers. For the SVM, the model tuning was a matter of trial and error but the parameterization followed a non-linear form so we opted for one that classified expensive areas in closed shapes. In the end, the SVM returned 40% of businesses as gentrifiers which to us seemed too high. Nevertheless, in later stages, the nearest neighbor approach did not yield a differentiable index whereas the SVM managed to differentiate the user categories so we used the predictions of the SVM for our final output.

2.4 Categorizing Users

To cluster users by their consumption patterns we used a Latent Dirichlet Allocation (LDA) model for topic modeling over the user prices and business categories. Namely, let U be the set of users and for each user $u \in U$ let b_u be the set of businesses that user reviewed in San Diego. For a user u who reviewed n businesses, let $b_u = \{(p_i, c_i)\}_i^n$ where p_i is the price of business i and c_i is the category set for that business (ie, {cafe, coffee, pastry, tea}). If we let $m = |U|$ be the number of users for which we have data, we wish to create categories based on $\{b_k\}_k = 1^m$. Initial conceptions for this step entailed a simple categorization of users into price brackets based on some type of a summary statistic over their price history; however, this approach ignores the users' preferences for business types. For example, user i and user j could both consume businesses in the expensive business bracket but if one only reviews country clubs in rich areas whereas the other prefers $40 cupcakes in downtown San Diego then it is unreasonable to categorize the two together for predictive purposes. Alternatively, we considered topic modeling based on the business types the users have consumed but that ignores the price factor of the consumption history. Consequently, in a matter of, "why not?" we combined business categories with prices and let LDA do its magic. We tried two forms of string concatenation. Namely, if user i consumed a business with price category 2 and category set of {cafe, coffee, pastry, tea} then we tried "2cafe,coffee,pastry,tea" and {2cafe,2coffee,2pastry,2tea}. Because the former did not yield interpretable results whereas the latter gave consistent and intuitively coherent user categorizations, we use the latter scheme for out final results. We then experimented with the number of topics and decided that 9 topics gave categories with a large variety in user consumption business types and prices.

2.5 Assigning Likelihoods

At this stage, our aim is to combine the gentrifier classifications with user categories and then associate the proclivity of a category to gentrify the business communities. For this task, we first assigned each user u to a category based on her highest posterior probability of belonging to that topic. To tie the groups to gentrification, we used a logistic regression. Namely, let $y = [l_1, l_2, \ldots, l_n]$ be a vector of gentrification labels for our n businesses where $l_i = 1$ if business

i is a gentrifier and $l_i = 0$ if the business was not labeled as a gentrifier. Let X be a $n \times k$ matrix where k is the number of user categories and X_{ij} is the number of times business i shows up in topic j. Then in $g(y) = X\beta$, the β gives us the likelihood of the users in that group to consume a gentrifying business. We then used a contingency table to establish the relative composition of each community, namely, how many businesses in group j are in community c. Let these proportions be in a vector γ, to calculate the index we plugged the proportions into the linear equation for the likelihood of gentrification, $g(v) = \gamma\hat{\beta}$.

3 Results

For brevity, we only display the main output in our results section and encourage the reader to examine the poster for more detailed visual information.

As can be seen in the maps displayed in **Table 2**, the surprising result is the fact that the location of gentrifying businesses does not necessarily correspond with the likelihood of a community being gentrified. This suggests that the main drivers of the likelihood are the people whom that community attracts which in turn implies that gentrifying businesses follow the consumers.

Table 1: Sample output of user categories/topics

Topic 0:	mexican1.0 sandwiches1.0 burgers1.0 hotdogs1.0 fastfood1.0 seafood1.0 delis1.0 mattresses2.0 furniture2.0 furniturestores2.0 homedecor2.0 chickenwings1.0 grocery1.0 pizza1.0
Topic 1:	breakfast2.0 breakfastbrunch2.0 brunch2.0 newamerican2.0 american(new)2.0 bars2.0 italian2.0 winebars2.0 sandwiches2.0 american(traditional)2.0 tradamerican2.0 cafes2.0 pizza2.0 french2.0
Topic 2:	divebars1.0 pubs1.0 grocery2.0 bars1.0 barbers2.0 barbers1.0 sportsbars1.0 spirits2.0 wine2.0 beerandwine2.0 beer2.0 drugstores2.0 pizza1.0 musicvenues1.0
Topic 3:	japanese2.0 sushi2.0 sushibars2.0 seafood2.0 asianfusion2.0 mexican2.0 chinese2.0 japanese3.0 bars2.0 lounges2.0 juicebars2.0 sushi3.0 sushibars3.0 thai2.0
Topic 4:	coffee1.0 tea1.0 sandwiches1.0 cafes1.0 bakeries1.0 breakfast1.0 brunch1.0 breakfastbrunch1.0 juicebars1.0 bagels1.0 delis1.0 mexican1.0 pizza1.0 donuts1.0
Topic 5:	bars2.0 sportsbars2.0 burgers2.0 tradamerican2.0 american(traditional)2.0 mexican2.0 pubs2.0 seafood2.0 pizza2.0 american(new)2.0 newamerican2.0 mexican1.0 italian2.0 cocktailbars2.0
Topic 6:	pizza1.0 italian1.0 sandwiches1.0 seafood3.0 salad1.0 steak3.0 steakhouses3.0 delis1.0 tobaccoshops2.0 italian3.0 italian2.0 seafood2.0 vapeshops2.0 winebars3.0
Topic 7:	coffee2.0 icecream1.0 chinese1.0 bakeries2.0 vietnamese1.0 desserts2.0 korean2.0 tea2.0 frozenyogurt1.0 desserts1.0 japanese1.0 chinese2.0 asianfusion1.0 barbeque2.0
Topic 8:	italian2.0 vegetarian2.0 pizza2.0 vegetarian1.0 vegan2.0 vegan1.0 mexican1.0 salad2.0 thai1.0 mediterranean1.0 glutenfree2.0 gluten-free2.0 buffets2.0 juicebars1.0

Table 2: Location of gentrifying business, our likelihood of gentrification, census-based gentrification