Inequality in the menu: How a network of restaurants characterizes social disparities in Boston

Sirio Papa[1], Beatrice Rosi[1], Lorenzo Testa[1,2], Francesco Vaselli[1], and Giulio Rossetti[3]

[1] University of Pisa, Pisa, Italy
[2] EMbeDS, Sant'Anna School of Advanced Studies, Pisa, Italy
[3] KDD Laboratory, ISTI-CNR, Pisa, Italy

Yelp is a website collecting crowd-sourced reviews about businesses. On December 31, 2020, approximately 206.3 million reviews were available on its business listing pages. Yelp also provides data on a subset of businesses for academic purposes [1]. This data set consists of 8,635,403 reviews for a total of 160585 businesses spreading across 8 large metropolitan areas (v3, updated on April 2021). We leveraged such a data set to build a graph connecting restaurants with the aim of probing its structure and trying to predict social and economic disparities of the US census tracts where businesses are located. Data on social conditions are taken from Opportunity Insights [2].

Our analysis focuses on the Boston metropolitan area: there we selected all those businesses belonging to *at least* one category following typologies: Food, Restaurant, Cafes and Bar. Such selection generated a total of 13760 businesses, serving as the *nodes* of our network. To place an edge between two businesses, we required that at least a *same* user reviewed them both - thus assuming that the activity of a user would establish meaningful connections between the places it decided to visit (a person will visit places based on her preferences and influenced by many factors, included financial situation and other social outcomes). Due to the large numbers of reviews available for the selected business sample, we pruned the obtained graph by discarding those edges generated by less than 5 users. Such a filtering allowed us to identify the most *meaningful* connections, neglecting the spurious/noisy ones.

Communities and geographical analysis. We clustered the resulting graph using *Louvain* (see [4]), a community detection algorithm optimizing partition modularity, which returned 8 communities. We evaluated the resulting clustering from a geographical point of view to better understand the geographical distribution of businesses. Indeed, focusing on the 5 biggest communities, Louvain were able to identify meaningful geographical partitions: two of them located in Boston's downtown, one in the North, one in the West and one in the southern part of the city (Figure 1). From these results it is reasonable to postulate that restaurants' communities are identifying, at least to some extent, distinct areas of people's mobility: individuals appear to prefer going to restaurants close to each other and, presumably, within a certain radius from their home/workplace. Furthermore, it is possible to observe a consistent number of restaurants located along arterial/main roads.

Methodology and Results. As previously stated, our goal is to employ features extracted from the network to predict social disparity of the US census tracts in which restaurants are located.

To such extent, we decided to focus on the following target features: household income at age 35, number of children, median rent and employment, single parents and

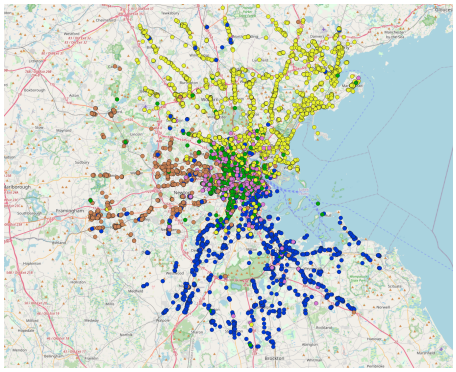

Fig. 1. Geographical visualization of businesses' community within the Boston Metropolitan Area

Index	Accuracy	ROC AUC (macro)
Household Income	0.65	0.82
Number of children	0.68	0.85
Median Rent	0.64	0.82
Employment rate	0.58	0.78
Single Parents rate	0.58	0.77
Poverty rate	0.67	0.83

Table 1. Results of the index class prediction leveraging the node2vec embeddings. The dataset being balanced, we report both the accuracy (random baseline of 33%) and the ROC AUC macro average considering all the ROCs for the three classes.

poverty rates at age 35. The data available for such indexes refer to the period 2012-2016. We assigned to each node in the network the corresponding index value for its tract. Then, we approach our objective as a classification task. First of all, noticing that all the distributions for our target variables presented a Normal - or at most right-skewed - distribution, we performed a 3 *quantiles cut*, splitting each distribution in three discrete classes (low, medium, high) each containing 33% of our nodes. Wanting to extract meaningful attributes from our network in order to perform classification, we turned to the *node2vec*. The selected parameters for the algorithm, based on various test on the validation split, were an *embedding dimension* of 128 features per node, a walk length of 170, a skip-gram context of 10, 100 epochs of training, $p = 2$ and $q = 0.25$. Interestingly, this choice of bias parameters for the random walk, with $p > q$, suggests that we obtain better classification results by favouring the embedding of the far neighborhood of a node.

The available data were split into three subsets (65% train, 17.5% validation and 17.5% for testing), and, after having performed model selection we employed 80% of the data for training and the rest for the final evaluation. The selected classification model was a *Random Forest classifier*, with *scikit-learn* (see [3]) default parameters except for *n_estimators* = 1000. Results are presented in Table 1. The obtained results underline that it is indeed possible to predict at least the broad category of a disparity index from the information given by the network alone. Having predicted each restaurant class, it would then be possible to assign the corresponding value to their tract based on majority voting.

Although no causal mechanism should be implied or supposed while dealing with this delicate task, our findings do suggest a possible relationship between behavioural processes (i.e. the choice and the review of a restaurant) and other economic and social characteristics. In particular, given the exogeneity of statistics computed on a network

structure, these become interesting and powerful instruments to be applied in regression problems tackling endogeneity issues (both in cross-sectional and panel data).

A possible and interesting extension of the previous analyses is offered by a dynamic, temporal characterization of the network. In this case, we could appreciate the behaviour of communities and of the graph itself. Social disparities may be then investigated by accounting for their trends and patterns. Moreover, given the nature of our datasets, it would also be possible to extend the previous analyses on different regions and types of activities.

Acknowledgments

L. Testa acknowledges support from the Sant'Anna School of Advanced Studies. This work is supported by the scheme 'INFRAIA-01-2018-2019: Research and Innovation action', Grant Agreement n. 871042 'SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics'

References

1. Yelp Inc., Yelp Academic Dataset, `https://www.yelp.com/dataset`, 2021
2. Opportunity Insights, "Opportunity Insights - Policy solutions to the American dream.", `https://opportunityinsights.org/`", 2016
3. Scikit-learn: Machine Learning in Python, Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E., Journal of Machine Learning Research, 12, 2825–2830, 2011
4. Fast unfolding of communities in large networks, 2008, 1742-5468, `http://dx.doi.org/10.1088/1742-5468/2008/10/P10008`, Journal of Statistical Mechanics: Theory and Experiment, IOP Publishing, Blondel, Vincent D and Guillaume, Jean-Loup and Lambiotte, Renaud and Lefebvre, Etienne, 2008, Oct, P10008