Social Media of Quantiffied Communities

Philipp Kats
New York University (NYU)

I. Introduction

As digital technologies are becoming more and more widespread, big data created by recording the digital traces left behind human activities become a powerful mean to study various aspects of human behavior. Many of those aspects can be described with social media feeds - data, generated and shared by people globally [1]. At the same time, the increasing urbanization of the world's population and great diversity of new urban population deeply affect urban environment. Solving many challenges of modern cities, from crime to illegal conversion, to tax regulation, etc, require large quantities of incoming data with frequent updates and on the detailed level. This need might be met by social media data, as multiple studies show [2] [3]. However, It is crucial to develop theoretical frameworks as well as real-time monitoring systems beforehand, to understand how the individual dynamics shape the structure of our cities in order to make better tactical decisions and general strategies for city governance [1].

In a now-famous paper, Eagle and Pentland [2] showed that it was possible to decompose mobile phone activity patterns of university students into regular daily routines, and that these routines were linked to each student's major and also to employment levels. Building upon this work TOFINISHS!!!

II. Materials and Methods of Aquisition

A. Twitter data gathering and preprocessing

A feed of Twitter data was collected through official API using ansamble of custom scrapers. Each scraper constantly scrapes new tweets with geolocation and created within New York City Boundaries, working 24 hours a day. Data completeness is ensured by deploying an ansamble of scrapers working in parallel. Scrapers also responsible for major data processing, saving only a subset of attributes — user id, location (latitude and longitude), time, text, amounts of retweets and favorites, application signature for each tweet, and timestamp, adjusted to EST.

B. Spam Filtering

As several Applications generate tweets automatically and do not represent any real user activity, and may skew our time series due to their automatic nature, we considered them as "spam". In order to filter them, we drop all the tweets, generated by all applications, for which more than 10% of total tweets belongs to one user. As such, we are dropping all tweets from NYC 511 Road traffic bot, and few other automated systems. As a result, roughly about .2% of tweets were removed.

C. Final Dataset

Final dataset consists more than 23 millions of tweets from more than 6 hundred of thousands of unique users, published withing the geographical boundaries of New York City, from June 2013 and until June 2016 (with few minor gaps). Tweets were created using 603 different applications, though more than 60% were generated by Twitter for Instagram

	user_id	tweets
application		
Instagram	6120	9164
Foursquare	701	1313
Squarespace	174	252
Twitter for iPhone	119	267
Twitter for Android	82	341
Tweetbot for iOS	36	145
Untappd	20	27
Twitter for Windows Phone	18	60
Twitter for Android Tablets	17	88
iOS	15	15

Figure 1. Top 10 Applications. Numbers to be updated

D. Spatial joint

Tweets were grouped and aggregated by spatial location, using 262 Postal Code Boundaries (PCB). Postal Code boundaries were chosen as they provide, on one hand, image detailed enough to provide valuable and interpretable results. At the same time, they are large enough to have enough tweets aggregated within each code to receive statistically significant results. As PCB were designed for service purposes, they provide roughly similar number of population per each district, which is useful for normalization purposes and general modeling. Each tweet was given a Postal Code, basing on its geographical location through spatial joint. Postal codes were later interpreted, where applicable, to neighborhoods, basing on *ZIP Code Definitions of New York City Neighborhoods* [3]

E. American Community Survey

American Community Survey (ACS) 5-year Summary tables for year 2013 were aggregated from official source and used to compare clusters and train our prediction models. ACS contains hundreds of questions on multiple topics, from population and demographics, to household median income, commute, health and insurance, education, rent, access to the internet, and many others. It is given on many geographical levels, from the whole country and states, and down to block group level, and include postal codes as well.

III. Time series

In order to understand basic patterns data was converted to average week time series. Data processing procedure is similar

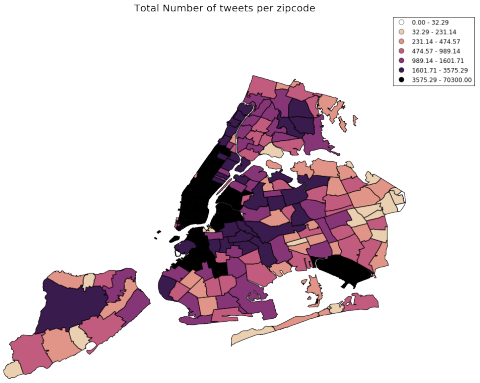

Figure 2. Replace this text with your caption

to one, presented in [4]. First, all tweets were aggregated to the total number of tweets per each 15-minutes range. As there are several gaps in data collection, for days with zero tweets in total all 15-minute samples were removed. An aggregation was made separately for every group of tweets: — city in total, each Postal Code, and every of the top popular applications. Using those aggregated and filtered time series, we generated average weekly time series for each group, represented by 672 attributes (4*2*4*7*) each, and normalized by dividing each attribute by their total sum.

By doing this, we revealed both local and universal city-vise twitter usage patterns: on average, people used to tweet the most around midnight. This pattern changes slightly through the week, with pike going late on Fridays and earlier on Sundays and Mondays. At the same time, variance for the weekend is much highter. It would be reasonable to expect seasonal and weather-related dependencies.

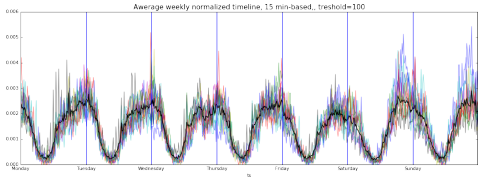

Figure 3. Averge week for each zipcode in New York. Blue vertical lines represent midnight. Black line - average week pattern for whole New York

A. *Postal code Boundaries and Neighborhoods*

Each Postal code has its unique week signature. However, many of signatures are similar and present spatial correlation. For example, All Zipcodes in Lower Manhattan generally have more tweets during the day, and less in the evening, than New York on average - this definitely correlates with the business functionality of the area, as Lower Manhattan represents New York's Downtown.

????

B. *Applications*

Similar signatures were generated for most popular Apps: while data we use is generated by Twitter service, it ofthen enough produced by other apps via API: this happens usually

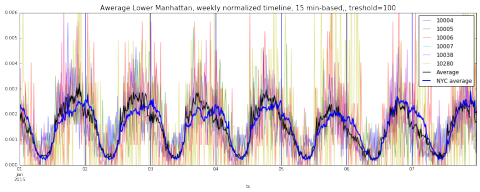

Figure 4. Average week signature for Lower Manhattan (postcodes 10004, 10005, 10006, 10007, 10038, 10280)

when people would prefer to share their app-related updates through Twitter, — many applications have this functionality, and many of them have it turned on by default. Having the application signature for each tweet, we were able to create a detailed city-wise timeline per each of the top 15 popular applications. Timeline for each application is depends heavily on its particular application. For example, **Dlvr.it** app, which is used to deliver food and grocceries, have a smooth and balansed timeline with pikes during the day, On the contrary, **Foursquare** app has pikes in the evening. For some behavioral reason, **Twitter for iPad** has a narrow pike during the lunch time for every workday.

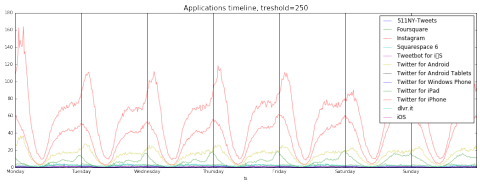

Figure 5. Average week plot for most popular applications

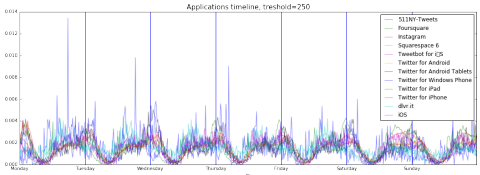

Figure 6. Average week plot for most popular applications, normalized

Average weak plot for most popular applications, normalized

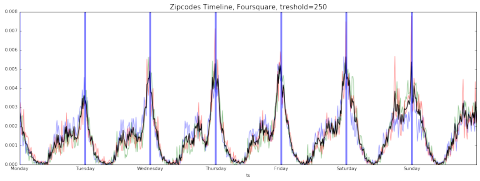

Figure 7. Average week plot for Foursquare app

IV. Clustering

While each time series is unique, many of them represent similar patterns. To determine those main patterns and interpret them, certain clustering techniques were introduced. Two clustering algorithms were used: k-mean and affinity propagation. All PCBs with a total number of tweets lower

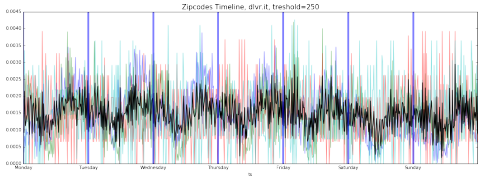

Figure 8. Awerage week plot for Dlvr.it app

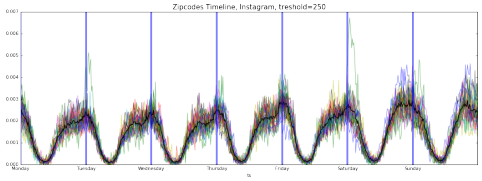

Figure 9. Awerage week plot for Instagram

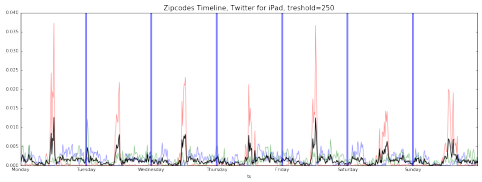

Figure 10. Awerage week plot for Twitter for Ipad

than the threshold — 250 tweets — were removed in order to ensure data consistency.

A. K-mean

K-mean is a widely used and relatively easy algorithm, that require the number of clusters to be set manually. While silhouette score may be used in order to determine the best number of clusters in our case it might make sense to start with a fairly small number of clusters in order to be able to interpret them.

As such, we started with 2 clusters, receiving expected partition of working and residential areas. Adding more clusters, we received a complete model with partitions of (as interpreted) three main areas: central business district (lower Manhattan), upper Manhattan and downtown Brooklyn, airports and all others, mostly residential areas.

B. Affinity propagation

The second technique, Affinity propagation (AP), works particularly good with time series. It does not require a predefined number of clusters and returns representative examples same way as k-medoids. It is also more capable of defining unique time series as a unique singleton clusters. For tweets time series, AP detected 12 clusters, detecting (as interpreted) clusters of Central Business District, Downtown and around, Residential areas, and several unique PCBs, — JFK, La Guardia, Bay Ridge, and few others.

V. Naive Network

Another approach facilitated was to create a *naive* full network of PCBs with weighted edges, where edge weight

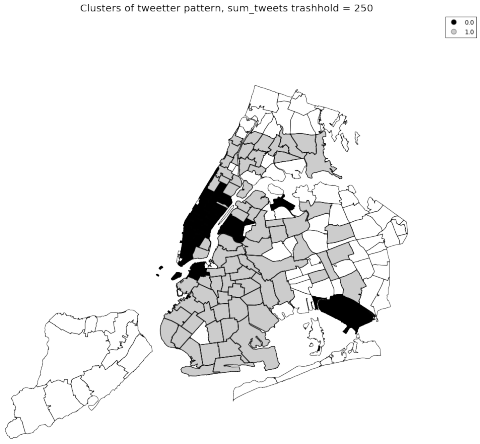

Figure 11. K-mean clustering, n=2, treshold=250

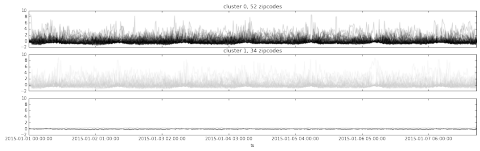

Figure 12. K-mean cluster, n=2

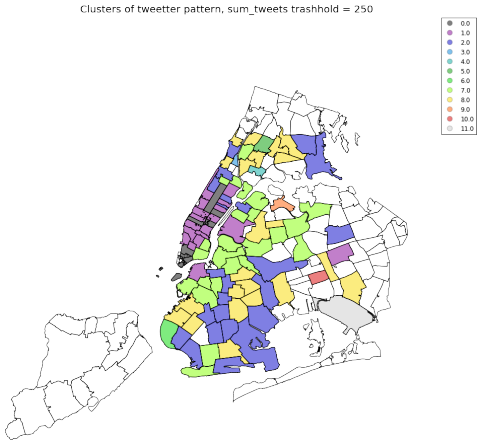

Figure 13. Affinity Propagation clustering outcomes: 12 clusters, including singleton clusters for JFK and La Guardia Airports

is measured by the number of tweets created within any of two postal codes and by any user who posted at least once in both postal codes. After that, edges are normalized by the total number of tweets in the dataset. Having this network, we can perform certain partitioning techniques, receiving certain areas, connected to each other the most.

First, we use widely known Luvain partitioning algorithm, which returned partition of 4 groups with modularity of 0.203

Another partition algorithm applied was Combo partition, described in [5], which returned 3 groups of areas with modularity metric of 0.20311

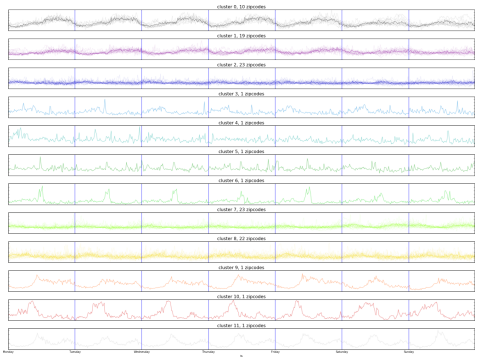

Figure 14. Affinity Propagation clustering outcomes: time Series of postal codes, grouped into 12 clusters

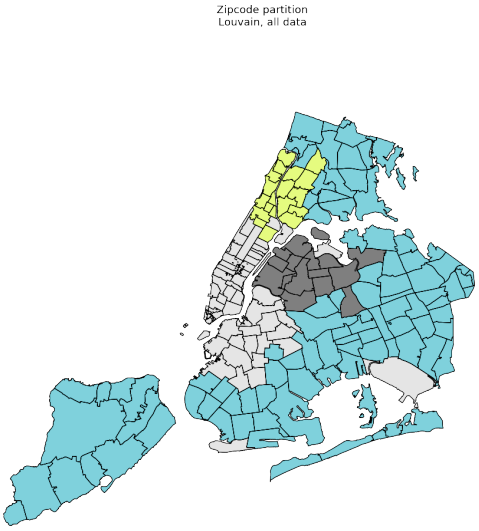

Figure 15. Luvain partition over naive tweets network, modularity = 0.203319362677

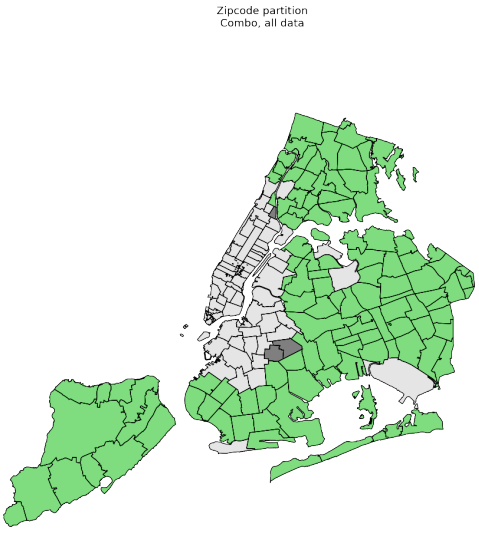

Figure 16. Combo partition using for naive tweets network

VI. Demographics

As most of the clusters were formed by neighbor areas, it leads to the question whatever they represent similar socio-economic properties of neighborhoods. To test that, we used data from last American Community Survey, which provides this particular level of geographies - ACS 2013 Summary table. For each cluster, a distribution of most important parameters was created, covering topics as race, poverty, commute time, median income and median household rent, and others. Comparing those distributions, we were able to establish a set of parameters, that has a significant difference in values from cluster to cluster. Thus, our tweets time series dataset might be used to predict any of those features.

VII. Snow storm

In January 2015, 23-31, a powerful blizzard affected the Eastern United States. On Monday, January 25, Storm approached New York. Next day, many roads were closed. As Twitter stream represents human behavior, it would be correct to assume that certain changes from the general pattern would occur within this period. Here, as data is sparse, we switch to the 6-hour range for each day, and compare differences between average week and week of the storm. Indeed, there is a shift in patterns for both January 25 and 26. The way pattern changes, however, differs depending on the area: for Lower Manhattan, twitter activity was significantly higher on Monday evening, and significantly lower during Tuesday work hours. On the contrary, for Flatbush neighborhood twitter activity was bigger during work hours on Monday and Tuesday evening, and had a clean spike on Tuesday morning. While average weeks model was able to represent those shifts, other models, such as networks, might perform better, visualizing the shift in human mobility through severe weather , closed roads, and public transport.

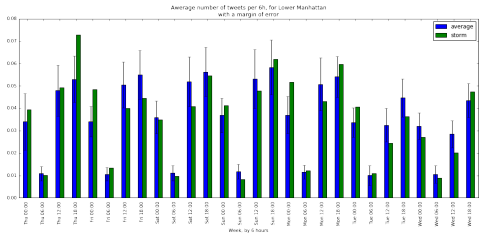

Figure 17. Average week vs Snow Storm week for Lower Manhattan

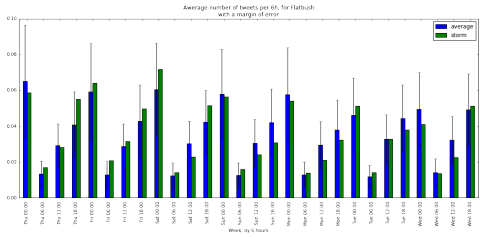

Figure 18. Average week vs Snow Storm week for Flatbush area

????

VIII. Discussion

????????

REFERENCES

[1] S. Grauwin, S. Sobolevsky, S. Moritz, I. Gódor, and C. Ratti, "Towards a comparative science of cities: using mobile traffic records in New York, London and Hong Kong." [Online]. Available: http://arxiv.org/abs/1406.4400

[2] N. Eagle and A. Pentland, "Reality mining: sensing complex social systems," vol. 10, no. 4, pp. 255–268. [Online]. Available: http://dl.acm.org/citation.cfm?id=1122745

[3] New York State. NYC Neighborhood ZIP Code Definitions. [Online]. Available: https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoods.htm

[4] J. Reades, F. Calabrese, A. Sevtsuk, and C. Ratti, "Cellular census: Explorations in urban data collection," no. 3, pp. 30–38. [Online]. Available: http://www.computer.org/csdl/mags/pc/2007/03/b3030-abs.html

[5] S. Sobolevsky, R. Campari, A. Belyi, and C. Ratti, "General optimization technique for high-quality community detection in complex networks," vol. 90, no. 1, p. 012811. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevE.90.012811