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 though multiple global social media platforms [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, including crime, illegal conversion and construction, tax regulation, emergencies, and many others, require large quantities of incoming data with frequent updates and on the detailed level.

This need might be met by data from social media. Focusing on records aggregated on spatial locations rather than on individuals, new approaches have been initiated different types of communication might be used for many purposes, from urban landscape description [2] [3] [4], to regional delineation [5] [6], population density estimation, land use classification [7] [8] and identification of social groups and events [9] . As many channels and platforms of communication have unique sets of features, it is crucial to develop theoretical frameworks and a real-time monitoring systems, as it is required 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 this study we use Twitter as a main source of data. Twitter was chosen for it's easy access to data through open API, rich data, including geolocation, and large user base within the city. While Twitter may represents a wide range of activities, it also provides "App signature" (credentials of application, that initiated particular tweet) along with each message. While majority of tweets is generated by it's original apps, significant fraction represents other applications, and, therefore, can characterize specific activity.

Further we provide a comparative study of twitter stream for New York City, using several approaches, such as *time series analysis*, *clustering* and *mobility networks*. The specific focus of our study is on demonstrating that multiple approaches, separately and in ansamble, can be applied in order to explore urban landscape dynamics, or used for a predictive analysis for the public good.

II. Materials and Methods of Aquisition

A. Twitter data gathering and preprocessing

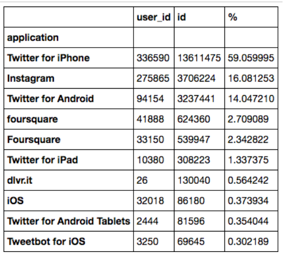
A feed of Twitter data was collected through official API using ansamble of custom scrapers. Each scraper constantly collects new tweets with geolocation mark and created within city boundaries, working 24 hours a day. Data completeness is ensured by deploying an ansamble of scrapers, working in parallel.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

B. Spam Filtering

As several applications generate tweets automatically and do not represent any human activity and they can potentially add noise to our data, 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 remove all tweets from *NYC_511* road traffic bot, *NYC job offer* bot, and a few other automated systems. As a result, roughly about .2% of tweets were removed.

C. Data overview

Final dataset consists more than 23 millions of tweets from about 6 hundred of thousands unique users, published withing the geographical boundaries of New York City, from June 2013 and until June 2016 (with few minor gaps). Tweets were generated using 603 different applications, though almost 60% were generated by *Twitter for IOS*, 16% - *Instagram* and 14% - *Twitter for Android* mobile applications.



	user_id	id	%
application			
Twitter for iPhone	336590	13611475	59.059995
Instagram	275865	3706224	16.081253
Twitter for Android	94154	3237441	14.047210
foursquare	41888	624360	2.709089
Foursquare	33150	539947	2.342822
Twitter for iPad	10380	308223	1.337375
dlvr.it	26	130040	0.564242
iOS	32018	86180	0.373934
Twitter for Android Tablets	2444	81596	0.354044
Tweetbot for iOS	3250	69645	0.302189

Figure 1. 10 top popular applications for Twitter

D. Spatial joint

Tweets were grouped and aggregated by their location, using 262 Postal Code Boundaries (PCB). Postal Code boundaries were chosen as they provide, on one hand, picture detailed enough to provide valuable and interpretable results, and, at the same time, 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* [10]

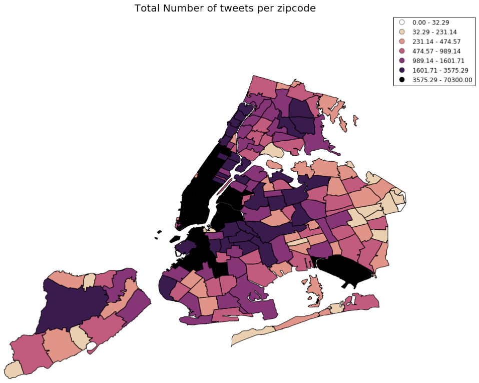


Figure 2. Replace this text with your caption

E. American Community Survey

American Community Survey (ACS) 5-year Summary tables for year 2013 were aggregated from official source [11] 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 to one, presented in [12]. 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, we generated average weekly time series - week signatures - for each group, represented by 672 (4x24x7) attributes 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.

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, it is clear that patterns for residential and business areas differs drastically: residential areas have large pikes in the evening, especially through weekday, while business districts are characterized by pikes in the first part of the day. As an example, all zipcodes in Lower Manhattan generally have more tweets during the day, and less in the evening, than New York on average. On the contrary, areas such as Flatbush (Brooklyn) and Jamaica (Queens) have well defined pikes in the evening.

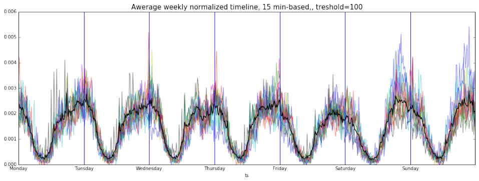


Figure 3. Averge week for each zipcode in New York. Blue vertical lines represent midnight. Black line - average week pattern for whole New York

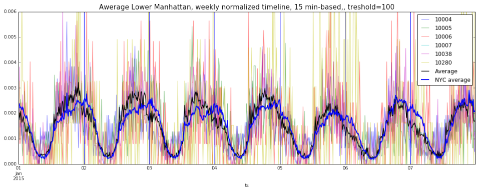


Figure 4. Average week signature for Lower Manhattan, Manhattan

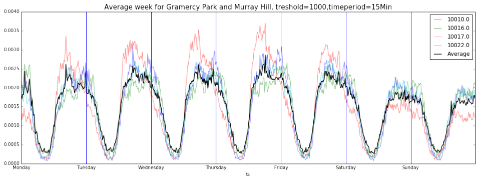


Figure 5. Average week signature for Grammercy park and Murray Hill, Manhattan

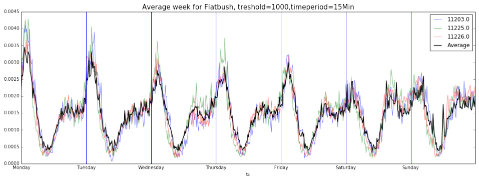


Figure 6. Average week signature for Flatbush, Brooklyn

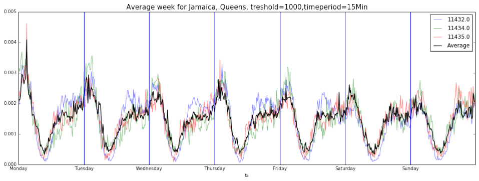


Figure 7. Average week signature for Jamaica, Queens

B. Applications

Similar week signatures were generated for most popular Apps: while data we use is generated by Twitter service, it often enough produced by other apps via API: this happens usually when people agreed to share their updates through Twitter, — many applications have this functionality, and often it is 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.

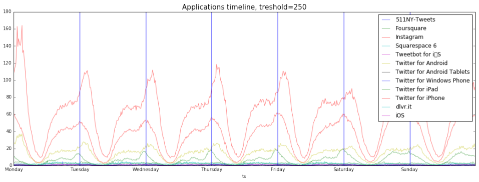


Figure 8. Average week plot for most popular applications

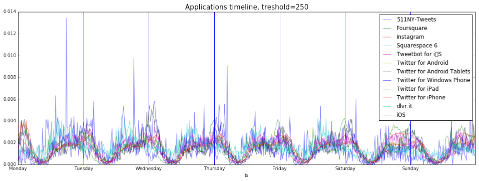


Figure 9. Average week plot for most popular applications, normalized

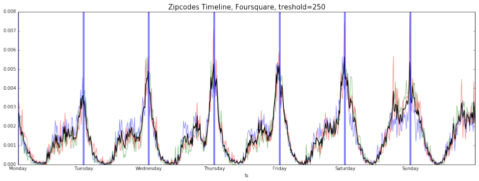


Figure 10. Awerage week plot for Foursquare app

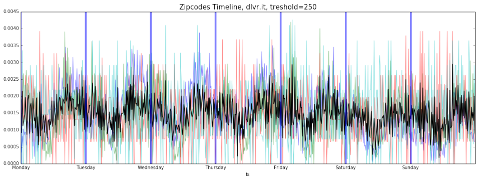


Figure 11. Awerage week plot for Dlvr.it app

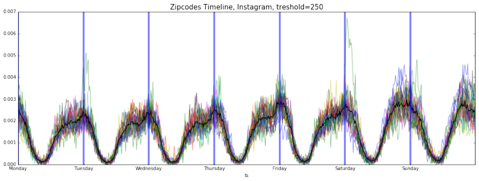


Figure 12. Average week plot for Instagram

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

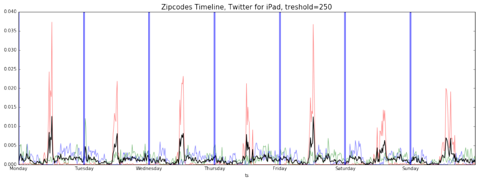


Figure 13. Average week plot for Twitter for Ipad

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. Both times, areas with total number of tweets below the threshold were dropped

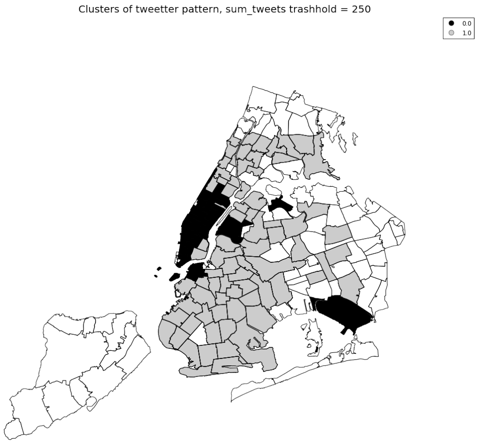


Figure 14. K-mean clustering, n=2, treshold=250

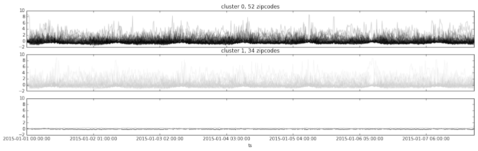


Figure 15. K-mean cluster, n=2

B. Affinity propagation

The second technique, Affinity propagation (AP), is particularly popular for 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, three clusters of residential areas, and several unique PCBs, including JFK, La Guardia, Bay Ridge, and few others. As with K-mean, areas with number of tweets below the threshold were dropped.

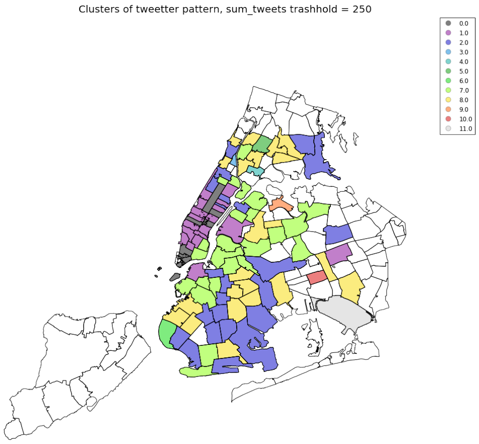


Figure 16. Affinity Propagation clustering outcomes: 12 clusters, including singleton clusters for JFK and La Guardia Airports

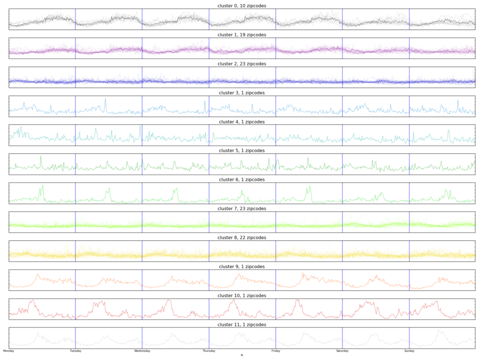


Figure 17. Affinity Propagation clustering outcomes: time Series of postal codes, grouped into 12 clusters

Another partition algorithm applied was Combo partition, described in [13], which returned 3 groups of areas with modularity metric of 0.20311

V. Socioeconomic properties of the partition

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.

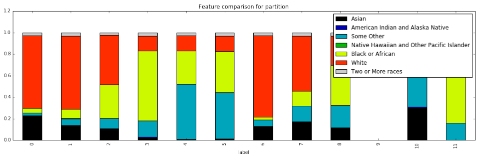


Figure 18. Racial decomposition of 12 time series clusters, defined by AP

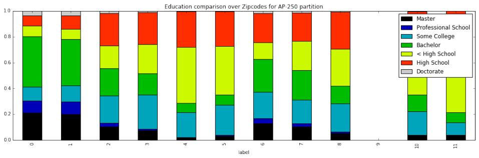


Figure 19. Highter education decomposition of 12 time series clusters, defined by AP

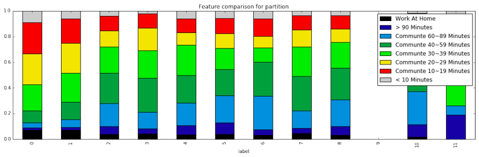


Figure 20. Commute decomposition of 12 time series clusters, defined by AP

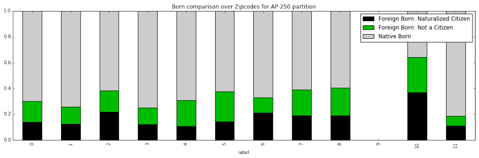


Figure 21. Foreign population decomposition of 12 time series clusters, defined by AP

VI. Event detection: 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.

VII. Discussion

In this study, we explored twitter data in many ways: as a network of mobility, time series area signatures, ap-

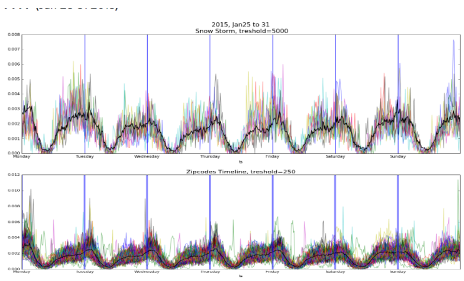


Figure 22. Snow storm week (above) versus average week (below).

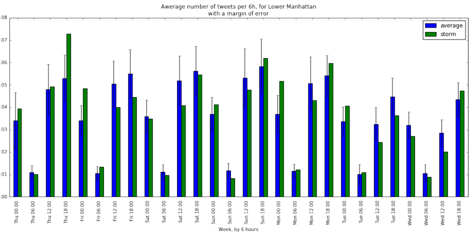


Figure 23. Average week vs Snow Storm week for Lower Manhattan

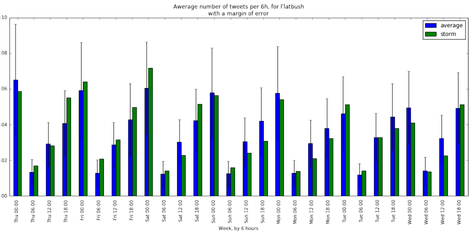


Figure 24. Average week vs Snow Storm week for Flatbush area

plying network communities detection and time series based clustering techniques. As study shows, after cleaning and processing, the stream of social media may be used to define functionality, define and predict socioeconomic characteristics, and represent human mobility. Our results showed, that average week signature for residential postal codes differs clearly from the office/retail areas. As such, over continuous monitoring, we might outline shifts in area functionality and demographics, detect events and predict how they will influence mobility. Using clustering techniques and average week signature of the areas as an input, we were able to group areas by their social behavior. This partition, in its term, effectively represent both functional and socioeconomic properties of the neighborhoods, and, therefore, might be used to control changes in those properties at the detailed level. This prediction model might be improved in the future by adding signatures for specific applications.

Through modeling mobility network, we were able to get inter-connections and transportation-based partition of the areas. This model represents human mobility, might be specified to exact subsets, and monitored in near-real time.

Significant events can be spotted through their twitter "fingerprints". As such, snow storm of 2015 was presented by behavioral changes both in business areas residential neighborhoods, most likely due to the problems with transportation.

While there is a room for advancements for any techniques, the results are already useful. Any of the approaches might be turned into practical implementation, helping to make decisions, allocate resources and understand shifting urban landscapes in near-real time

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] V. Frias-Martinez, V. Soto, H. Hohwald, and E. Frias-Martinez, "Characterizing urban landscapes using geolocated tweets," in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. IEEE, pp. 239–248. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6406289

[3] C. G. W. Jacobs-Crisioni and E. Koomen, "Linking urban structure and activity dynamics using cell phone usage data," in *Workshop on complexity modeling for urban structure and dynamics, 15th AGILE international conference on Geographic Information Science, Avignon*. [Online]. Available: http://www.feweb.vu.nl/gis/publications/docs/Linking_urban_structure_and_activity_dynamics_using_cell_phone_usage_data_abstract%20AGILE2012.pdf

[4] C. Ratti, D. Frenchman, R. M. Pulselli, and S. Williams, "Mobile landscapes: using location data from cell phones for urban analysis," vol. 33, no. 5, pp. 727–748. [Online]. Available: http://epb.sagepub.com/content/33/5/727.short

[5] A. Amini, K. Kung, C. Kang, S. Sobolevsky, and C. Ratti, "The Impact of Social Segregation on Human Mobility in Developing and Urbanized Regions." [Online]. Available: http://arxiv.org/abs/1401.5743

[6] K. S. Kung, K. Greco, S. Sobolevsky, and C. Ratti, "Exploring universal patterns in human home-work commuting from mobile phone data," vol. 9, no. 6, p. e96180. [Online]. Available: http://arxiv.org/abs/1311.2911

[7] T. Pei, S. Sobolevsky, C. Ratti, S.-L. Shaw, T. Li, and C. Zhou, "A new insight into land use classification based on aggregated mobile phone data," vol. 28, no. 9, pp. 1988–2007. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/13658816.2014.913794

[8] 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," in *Computational Approaches for Urban Environments*, ser. Geotechnologies and the Environment, M. Helbich, J. J. Arsanjani, and M. Leitner, Eds. Springer International Publishing, no. 13, pp. 363–387, DOI: 10.1007/978-3-319-11469-9_15. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-319-11469-9_15

[9] J. Reades, F. Calabrese, and C. Ratti, "Eigenplaces: analysing cities using the space–time structure of the mobile phone network," vol. 36, no. 5, pp. 824–836. [Online]. Available: http://epb.sagepub.com/content/36/5/824.short

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

[11] U. C. Bureau. American Community Survey (ACS). [Online]. Available: https://www.census.gov/programs-surveys/acs/

[12] 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

[13] 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