

# Social Media of Quantified Communities

*Philipp Kats*  
*New York University (NYU)*

## 1 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!!!

## 2 Materials and Methods of Aquisition

### 2.1 Twitter stream

A stream of twitter data was collected through official API using custom scrapers. Each scraper constantly scrapes new geolocated tweets from within New York City Boundaries, working 24 hours a day. Data completeness was ensured by deploying several scrapers to work at the same time. Same scrapers are responsible for 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.

As several Applications generate tweets automatically and do not represent any real user activity, we considered them as “spam”. In order to remove most of those “spam” tweets, 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 311 Agency twitter application, and few other automated systems. As a result, about 600 tweets was removed.

Final dataset consists more than 23 millions of tweets from more than 6 hundred of thousands of unique users, published within the geographical boundaries of New York City, from June 2013 and until June 2016 (with few minor gaps)

## 2.2 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.

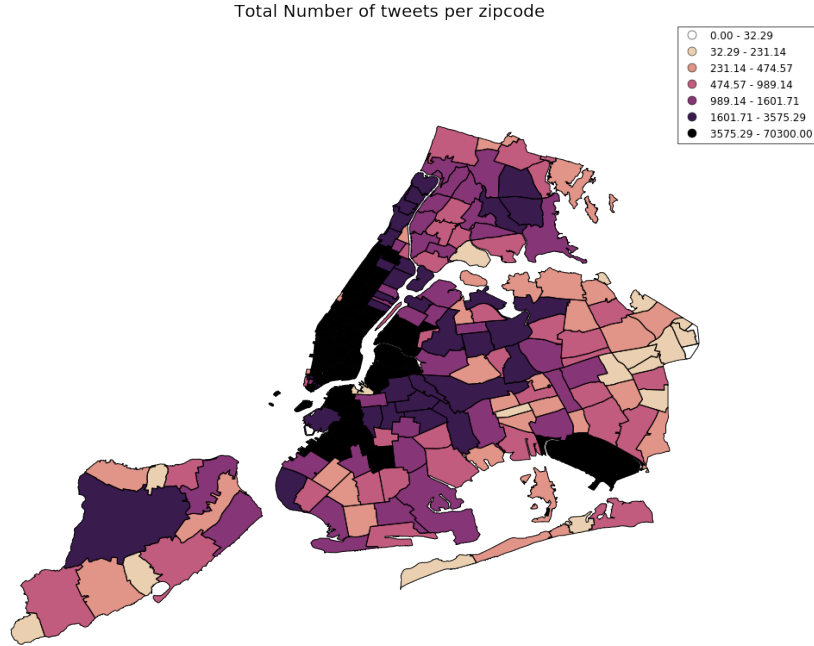


Figure 1: Replace this text with your caption

## 2.3 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.

## 3 Time series

In order to understand basic patterns of twitter sharing, time series analysis, based on average week pattern was chosen as our main method. Procedure is similar to one, presented in [3]

In order to apply time series-based analysis, few additional processing steps were done. 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 each, and normalized by dividing each attribute by their total sum.

Few additional processing steps were done in order to prepare time series. First, as we need frequency levels, dataset was aggregated to the total number of tweets for each 15-minutes range. As there are few gaps in our data collection, every 15-minute sample for dates with zero tweets per day was removed. An aggregation was made for multiple tweet partitions, — city in total, each PCB, and every of the top popular applications. Next, Average weak was generated for each group, represented by 672 attributes each, and normalized by dividing each attribute by their total sum.

By doing this, we revealed both local and universal city-wise 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 higher. It would be reasonable to expect seasonal and weather-related dependencies.

### 3.1 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

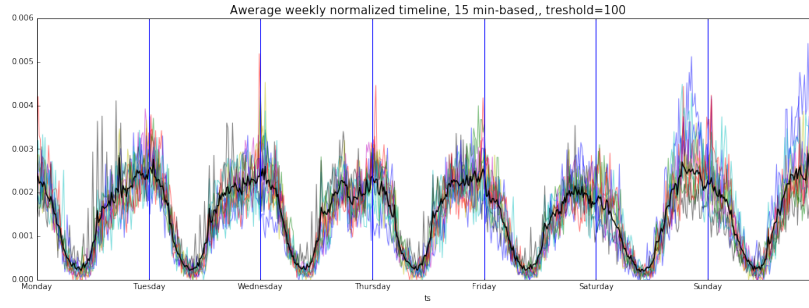


Figure 2: Average week for each zipcode in New York. Blue vertical lines represent midnight. Black line - average week pattern for whole New York

York on average - this definitely correlates with the business functionality of the area, as Lower Manhattan represents New York's Downtown.

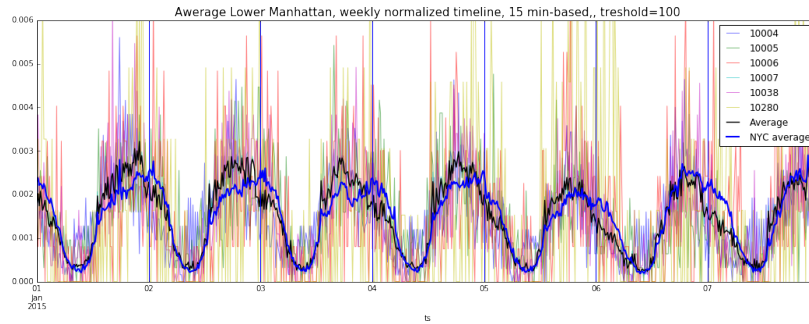


Figure 3: Replace this text with your caption

????

### 3.2 Applications

Similar 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 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 groceries, have a smooth and balanced 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.

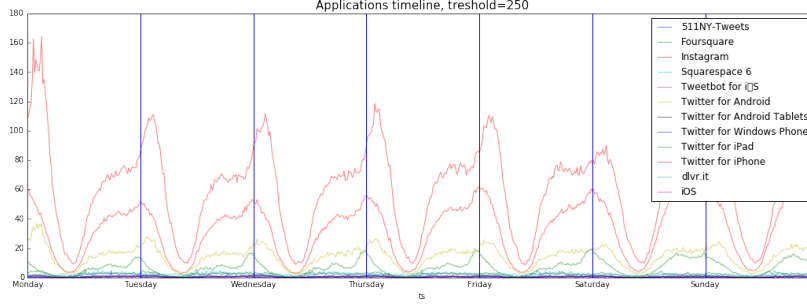


Figure 4: Average week plot for most popular applications

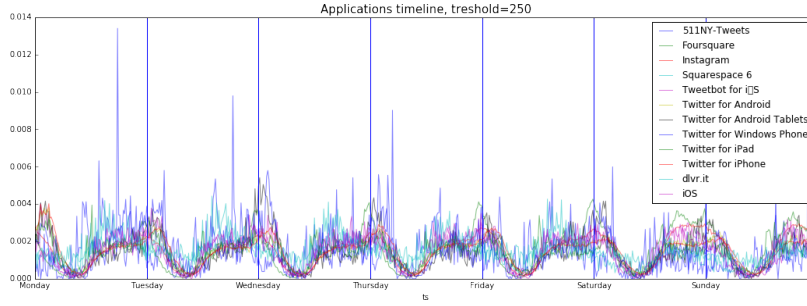


Figure 5: Replace this text with your caption

Average week plot for most popular applications, normalized

## 4 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.

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.

The second technique, Affinity propagation, 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. On our dataset, it 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 others.

## 5 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.

## 6 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. 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.

????

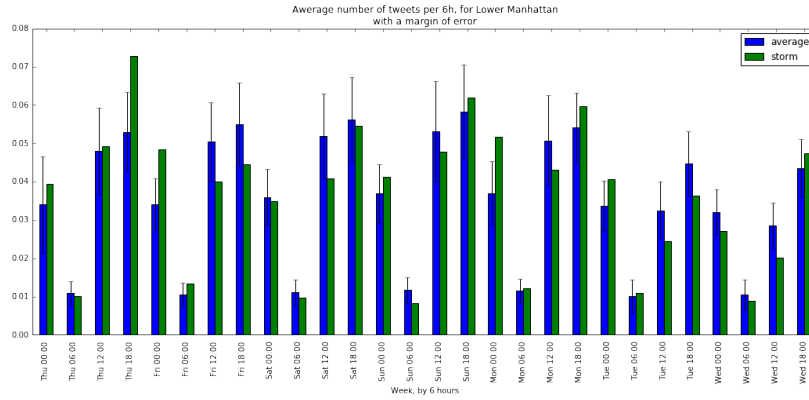


Figure 6: Replace this text with your caption

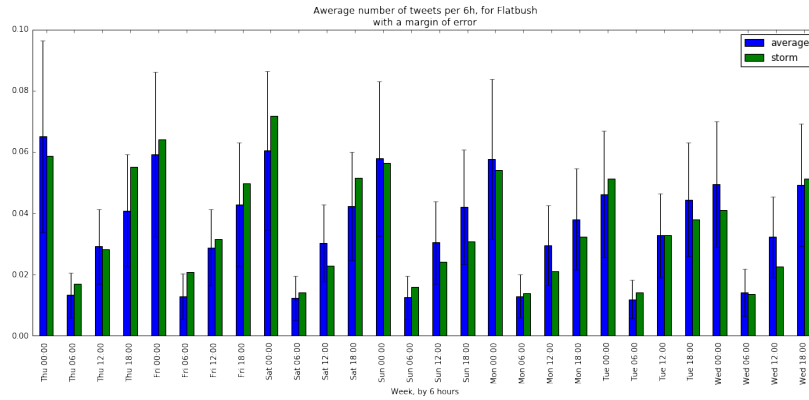


Figure 7: Replace this text with your caption

## 7 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.
- [2] Nathan Eagle and Alex Pentland. Reality mining: sensing complex social systems. 10(4):255–268.

- [3] Jonathan Reades, Francesco Calabrese, Andres Sevtsuk, and Carlo Ratti. Cellular census: Explorations in urban data collection. (3):30–38.