

# Introduction to Dataset for Detecting Collective Anomalies from Multiple Data Source

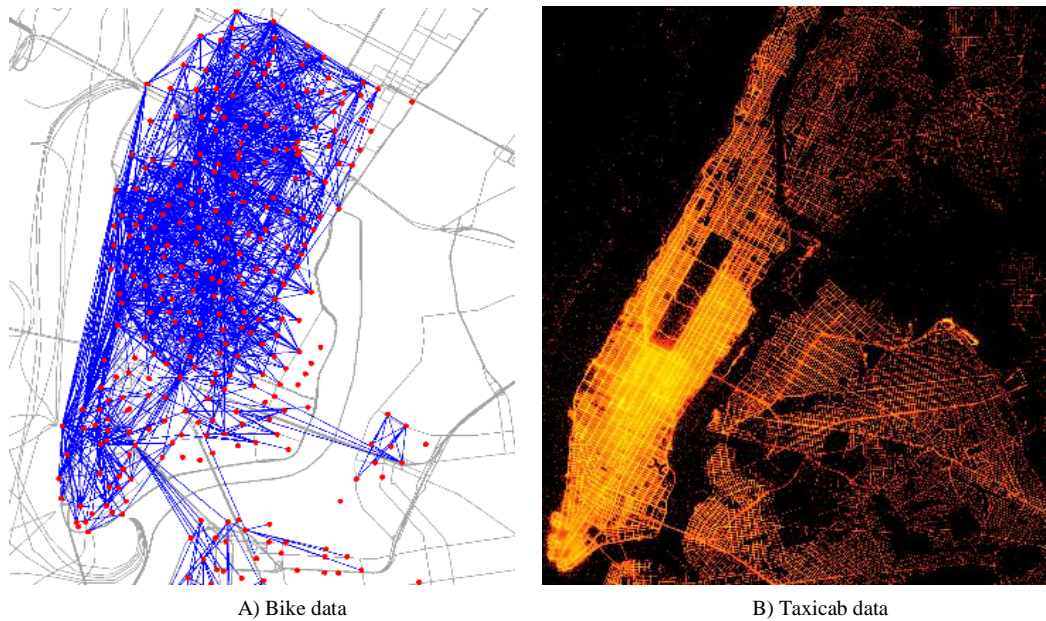
## Introduction:

This document introduces dataset used in paper [1]. It is comprised of five parts of data, named Taxi Trip Data, Bike Trip Data, 311 Service Data, POI data and Road Network data. The first three datasets are either FOIL-able or open data which anyone can get access to. Method of acquiring these datasets will be detailed below. Table 1 is an overall statistics about the dataset.

Data sources	Properties	values
<b>Taxi trip data</b> <b>1/1/2014-12/31/2014</b>	number of trips	165M
	total duration (hour)	36.5M
	total distances (km)	5,671M
<b>Bike trip Data</b> <b>1/1/2014-12/31/2014</b>	number of stations	344
	number of bikes	6,811
	number of trips	8,081,216
	total duration (hour)	1.9M
<b>311 Service Data</b> <b>1/1/2014-12/31/2014</b>	number of categories	4
	number of instances	292,811
<b>Road network</b> <b>2013</b>	number of nodes	79,315
	number of road segments (level $\leq$ 5)	32,210
	number of road segments (level $>$ 5)	83,655
	number of regions	862
<b>POIs</b> <b>2013</b>	number of categories	14
	number of instances	24,031

**Table 1**

Figure 1 presents the geographical distributions of the taxicab and bike on a digital map. As shown in Figure 1 A), each red point stands for a bike station and a blue edge denotes the aggregation of bike commutes between two stations. To generate a clear graph of stations, we remove the edges with the number of commutes smaller 700 in the time interval 1/1/2014-12/31/2014. Figure 1 B) is a heat map of the drop-off and pickup points of all the taxi trips from 1/1/2014 to 12/31/2014. The lighter the denser.



**Figure 1. Visualization of the data source**

Please cite the following two papers when using the codes.

- [1] Yu Zheng, Huichu Zhang, Yong Yu. Detecting Collective Anomalies from Multiple Spatio-Temporal Datasets across Different Domains. In the Proceeding of the 23rd ACM International Conference on Advances in Geographical Information Systems (ACM SIGSPATIAL 2015).
- [2] Yu Zheng, Licia Capra, Ouri Wolfson, Hai Yang. Urban Computing: concepts, methodologies, and applications. ACM Transaction on Intelligent Systems and Technology (ACM TIST). 5(3), 38, 2014

## **1. Taxi Trip Data**

### *Description:*

This dataset contains all taxi trip data from 2014-01-01 to 2014-12-31. Because the data is too large to release, we will provide a taste of the dataset containing 20000 trip records in “taxi.csv”. Also we will release some statistics of the dataset. Anyone who are interested in the full dataset can submit a FOIL request to NYC Taxi & Limousine Commission.

### *Schema:*

Each row stands for a single trip record. Recorded attributes are as follows

vendor\_id, **pickup\_datetime**, **dropoff\_datetime**, **passenger\_count**, trip\_distance,  
**pickup\_longitude**, **pickup\_latitude**, rate\_code, store\_and\_fwd\_flag, **dropoff\_longitude**,  
**dropoff\_latitude**, payment\_type, fare\_amount, surcharge, mta\_tax, tip\_amount, tolls\_amount,  
total\_amount

Bold red formatted attributes are those we are interested in.

*Example:*

VTS,2014-12-12 18:16:00,2014-12-12 18:35:00,3,4.0300000000000002,-74.014049999999997,4  
0.7117069999999997,1,, -73.995626999999999,40.759461999999999,CSH,16,1,0.5,0,0,17.5  
CMT,2014-12-25 22:37:45,2014-12-25 22:47:43,1,4.799999999999998,-73.872994000000006,  
40.774079,1,N,-73.846035000000001,40.779392000000001,CRD,15,0.5,0.5,4,0,20

VTS,2014-06-20 13:41:00,2014-06-20 14:19:00,1,6.8099999999999996,-73.940678000000005,4  
0.8141719999999999,1,, -73.979681999999997,40.744644999999998,CRD,28.5,0,0.5,5.70000000  
00000002,0,34.700000000000003

## 2. Bike Trip Data

*Description:*

This dataset contains data from <http://www.citibikenyc.com/system-data>, which records bike rental information of NYC citibike. Because the data is too large to release, we will provide a taste of the dataset 20000 trip records in "bike.csv". Also we will release some statistics of the dataset. Detailed information of this dataset is listed online, also anyone can get access to this dataset.

*Schema:*

Each row stands for a single trip record. Recorded attributes are as follows

tripduration, **starttime**, **stoptime**, start station id, start station name, **start station latitude**, **start station longitude**, end station id, end station name, **end station latitude**, **end station longitude**, bikeid, usertype, birth year, gender"

Bold red formatted attributes are those we are interested in.

*Example:*

1017,2014-03-01 09:45:06,2014-03-01 10:02:03,444,Broadway & W 24 St,40.7423543,-73.98  
915076,519,E 42 St & Vanderbilt Ave,40.752416,-73.97837,18008,Customer,N,0

606,9/1/2014 09:32:39,9/1/2014 09:42:45,375,Mercer St & Bleecker St,40.72679454,-73.996  
95094,511,E 14 St & Avenue B,40.72938685,-73.97772429,17003,Subscriber,1960,1

### \*\*\*Taxi and Bike Data Statistics

We separate one day into 48 time slots.

"holiday" includes weekends and official holidays

"workingday" includes days that are not holiday

*TaxiIn\_holiday.txt:*

We calculate the cumulative count of incoming records during holidays from 2014-01-01 to 2015-01-01 of each region at each time slot.

*TaxiIn\_workingday.txt:*

We calculate the cumulative count of incoming records during workingday from 2014-01-01 to 2015-01-01 of each region at each time slot.

*BikeIn\_holiday.txt* and *BikeIn\_workingday.txt* are similar to the above.

*TaxiOut\_holiday.txt*, *TaxiOut\_workingday.txt*, *BikeOut\_holiday.txt*, *BikeOut\_workingday.txt* are those calculating number of outgoing records.

*TaxiInDay.txt*:

Each line stands for a region, which is the number of incoming records of each timeslot in each day from 2014-01-01 to 2015-01-01. So each line contains 365\*48 values.

*TaxiOutDay.txt*

Each line stands for a region, which is the number of outgoing records of each timeslot in each day from 2014-01-01 to 2015-01-01. So each line contains 365\*48 values.

*BikeInDay.txt* and *BikeOutDay.txt* are the similar to the above.

### **3. 311 Service Data**

*Description:*

311 Services Request are part of the NYC open data. You can find details about and download the dataset on "<https://nycopendata.socrata.com/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9?>". This dataset contains all kinds of 311 Services Request Data from 2010 to Present. We are only interested in 311 Service Request of four kinds "Noise", "Blocked Driveway", "Building/Use", "Illegal Parking". Also, we eliminated those with invalid information, which result in the file "pure\_Four.csv".

*Schema:*

There are many attributes contained in the dataset. Only few of the attributes are used in our project, as listed below. Detailed structure can be found online.

Created Date, Complaint Type, Descriptor, Latitude, Longitude

*Example:*

Too long to display here.

Please find examples in "pure\_Four.csv"

### **4. POI Feature**

*POI.txt:*

*Description:* contains number of different POI categories of each region. There are 14 categories in all.

*Format:* each line stands for a region, with numbers of 14 categories separated by space. Order of categories is "Arts & Entertainment", "Automotive & Vehicles", "Business to Business", "Computers & Technology", "Education", "Food & Dining", "Government & Community", "Health

& Beauty", "Home & Family", "Legal & Finance", "Real Estate & Construction", "Shopping", "Sports & Recreation", "other".

## **5. Road Network Feature**

*RN.txt:*

*Description:* contains road network information of each region

*Format:* each line stands for a region. Order of the features is "intersection count", "level 1 road length", "level 2 road length", "level 3 road length", "level 4 road length", "level 5 road length", "level 6 road length"

## **6. Region Segmentation File**

*NYC\_862.txt:*

*Description:* we treat region bounded by the latitude longitude box (40.918, -74.259, 40.486, -73.7) as a 2400 \* 2400 matrix. Any lat/lng coordinate can be mapped to a unit in the matrix. The value of matrix unit is the index of region where the corresponding lat/lng coordinate lies. Zero stands for "not a region".

*Format:* a 2400\*2400 matrix

*Usage:* A translator that get a lat/lng as input and give the region index as output is contained in the released code.

## **Contact:**

Dr. Yu Zheng, [yuzheng@microsoft.com](mailto:yuzheng@microsoft.com)

Lead Researcher at Microsoft Research

<http://research.microsoft.com/en-us/people/yuzheng/>