Network Based Profiling and Topic Clustering of Non-Emergency Calls in New York City

Kristin Cai

Abstract

New York City non-emergency calls, namely the NYC 311 calls, record detailed complaints from city residents and visitors on a variety of experience issues. While most complaint types are self-explanatory to be handled by the most relevant agencies and informational enough to trigger immediate responses, our interest is to identify patterns in the complaints (co-)occurrences that can better help understand the underlying situations.

This report briefly summarizes three analyses we build on top of the occurrence patterns we identified. 1) A Complaint Occurrence Network (CON) is constructed to understand the impacts of a complaint type. 2) Unsupervised temporal clustering of complaint co-occurrences to help identify similar situations in the history. 3) An attempt to model the latent situations (i.e. 'topics ') underlying specific event co-occurrence patterns, in a topic model setting.

Keywords: Occurrence, Network Analysis, Centrality, Features, Clustering, Topic Modeling

1. Data Input

We try to clean and prepare our data to provide the occurrence of complaints in geographic neighborhoods within a given time window. As a proof-of-concept research, we use daily aggregated occurrence of complaints ("Complaint Type") within a zip code ("Incident Zip") (as shown in Fig. 1) as the initial input data points for the following analyses and models.

Figure 1: Prepared Input Data Matrix

Occurrence #	Day 1			Day 2			Day N		
	Zipcode 1	...	Zipcode Z	Zipcode 1	...	Zipcode Z	Zipcode 1	...	Zipcode Z
Complaint Type 1	1		0	1		1	0		1
Type 2	3		1	0		0	0		0
...	...			...			...		0
Complaint Type 250	2		1	3		1	1		1

2. Complaint Occurrence - Stand-alone vs. Co-occurring

Stand-alone complaints in our analysis are defined as those complaint types that are least likely to be accompanied by the occurrences of other complaint types in the same zip code on the

same day. Similarly, we define two complaint types **co-occurring** if they both occur within the same zip code on a given day.

Given the 200+ complaint types in the data, we think that the **stand-alone** complaints are less informative and thus the least interesting to explore. Some of these complaints can occur frequently, but since they typically do not trigger other related complaint types, they are not the focus of our study.

Normalized Pointwise Mutual Information (NPMI) is calculated between each pair of the complaint types to measure the significance of co-occurrence between two complaint types. As the normalized version of pairwise mutual information(PMI), NPMI is a common association measure calculated as follows:

$$\text{PMI} = \log\left(\frac{p(x,y)}{p(x)p(y)}\right) = \log\left(\frac{p(x|y)}{p(x)}\right) = \log\left(\frac{p(y|x)}{p(y)}\right)$$
$$\text{NPMI} = \frac{\text{PMI}}{-\log(p(x,y))} = \frac{\log p(x)p(y)}{\log p(x,y)} - 1$$
$$\text{NPMI} \in (-1, 1)$$

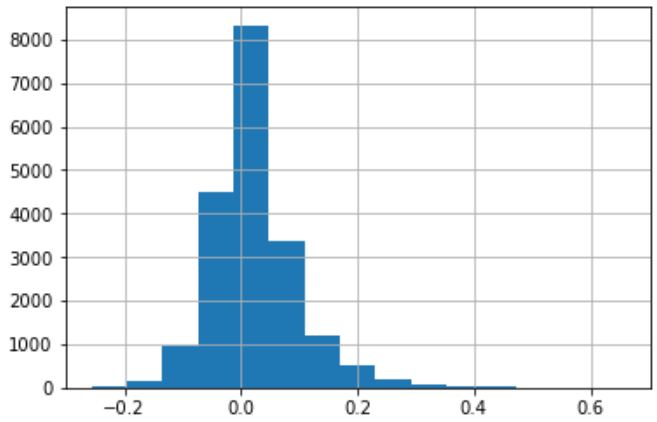
When two types only occur together, NPMI(x,y) = 1; when they occur as expected under independence, NPMI(x,y) = 0 as the numerator is 0; finally, when two words occur separately but not together, NPMI is ¬1, as it approaches this value when p(x, y) approaches 0 and p(x), p(y) are fixed. For comparison, these orientation values for PMI are respectively $\neg \log p(x,y)$, 0 and ∞

2.1. NPMI sample distribution and cutoff

As a co-occurrence association measure, NPMI can be used as the first criterion to filter out complaint types that are least likely to be accompanied by the occurrence of other complaints. These pairs are identified as those with small positive or negative NPMI values.

Figure 2 shows the NPMI distribution from all pairs of complaint types from data. As discussed previously, we want to pre-define a cutoff value to exclude those **stand-alone** complaint types.

Figure 2: Distribution of Normalized Pairwise Mutual Information from Data



To illustrate the effectiveness of using NPMI to measure the co-occurrence likelihood, Figure 3 shows the top 20 pairs of complaint types in terms of NPMI values. These are the pairs of complaint

types that are most likely to co-occure in the data. Interestingly enough, we have observed very reasonable pairs, and some triples (e.g. PAINT-PLASTER, GENERAL CONSTRUCTION and HEATING). Triplets are the stablest connections from a network point of view.

Figure 3: Distribution of Normalized Pairwise Mutual Information from Data

Complaint_Type1	Complaint_Type2	NPMI
PAINT - PLASTER	GENERAL CONSTRUCTION	0.66
GENERAL CONSTRUCTION	HEATING	0.61
PAINT - PLASTER	HEATING	0.59
Advocate-Personal Exemptions	OEM Disabled Vehicle	0.59
Found Property	Radioactive Material	0.54
NONCONST	GENERAL CONSTRUCTION	0.54
HEATING	NONCONST	0.54
NONCONST	PAINT - PLASTER	0.54
BEST/Site Safety	Emergency Response Team (ERT)	0.48
Vacant Lot	Fire Alarm - Replacement	0.47
LinkNYC	Bus Stop Shelter Placement	0.47
Request Xmas Tree Collection	Advocate - Other	0.46
Investigations and Discipline (IAD)	Open Flame Permit	0.45
Special Projects Inspection Team (SPIT)	Emergency Response Team (ERT)	0.44
Fire Alarm - Reinspection	Legal Services Provider Complaint	0.43
Case Management Agency Complaint	Panhandling	0.43
Mosquitoes	Taxpayer Advocate Inquiry	0.43
Water Quality	Hazmat Storage/Use	0.42
Sprinkler - Mechanical	Recycling Enforcement	0.41
Bike Rack Condition	Adopt-A-Basket	0.41

We empirically choose **0.2** as a conservative cutoff value, which gives us a reduced complaint types space of size 192. The complaint types remained are thus more likely to be related to the occurrence of other complaint types in the same neighborhood (zip code) within a day.

3. Network Analysis on Selected Complaint Types - Smaller vs. Bigger Complaint Co-occurrence Networks

With stand-alone complaints filtered out, we are now trying to distinguish smaller and bigger co-occurrence networks. The rational is that complaints types in smaller networks are less complicated to handle, and thus we would like to identify reasonably big complaint occurrence networks, and further identify "informative"and "influential"complaints in the network. To this end, we are borrowing the concept of "network"and "centrality measures"from typical network analysis.

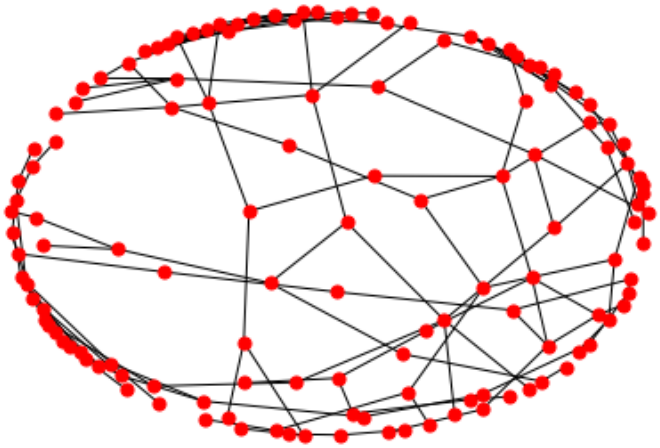
First of all, let us define a compliant occurrence network:

Definition 3.1. Complaint Occurrence Network (CON):
A Complaint Occurrence Network (CON) is a undirected network of complaint types, in which each node is a complaint type, each edge represents co-occurrence, and edge weights are normalized pairwise mutual information (NPMI) values that indicates the association strength.

The CON on selected complaint types are shown in Figure 4.
Several immediate observations are:

- The network is not fully connected. Many of subgraphs have small number of nodes, which are corresponding to the occurrences of small incidents that triggers a couple of complaints.

Figure 4: Complaint Occurrence Network(CON)



- Some nodes (i.e. complaint types) are more informative and important than others, either directly connecting to many other complaints (degree centrality), closely connected to other complaints (closeness centrality) or serving as bridges (betweenness centrality).

3.1. Further Complaint Type Selection by Selecting Subgraphs and Complaint Porfiling

We decompose the network of 192 selected complaint types and identify all connected component subgraphs. There are 32 subgraphs, with the single biggest one of size of 64. Others are relatively smaller suggesting relatively less complicated situations.

$$[64, 7, 3, 7, 6, 2, 11, 3, 3, 6, 12, 11, 2, 2, 2, 2, 4, 2, 4, 2, 2, 3, 2, 2, 5, 3, 4, 4, 4, 3, 3, 2]$$

We therefore select the 64 complaint types to focus on - they are more closely connected complaints than those in other subgraphs. Figure 5 shows the fully connected network structure. And Figure 6 shows the top bridging complaints (as measured by betweenness centrality) and complaints most closely related to other complaints (as measured by closeness centrality).

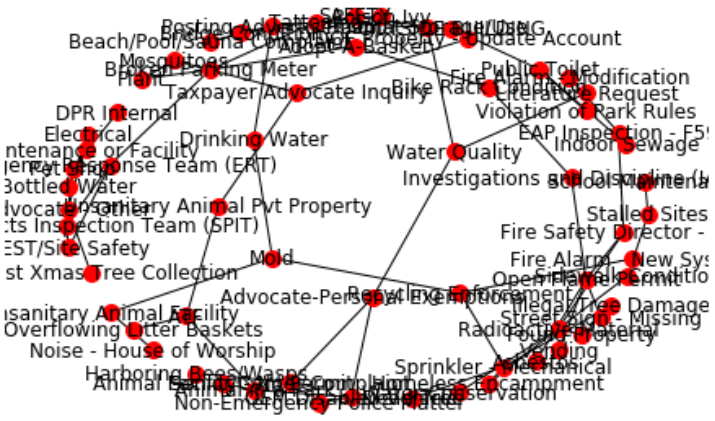
4. Temporal Clustering

With the selected 64 complaint types, we are ready to do temporal clustering on the input data shown in Figure 1. The goal, as stated previously, is to apply unsupervised temporal clustering of complaint co-occurrences to help identify similar situations in the history.

In Figure 7, if we view each column as a document (imagine as a "daily incident log"in a NYC sheriff's office) and view rows as bags of "words"in the corpus, we are facing a clustering task very similar to document clustering based on bag of words models, which are common in **topic modeling**.

Therefore, TF-IDF (Term Frequency - Invserse Docuemnt Frequency) based document clustering algorithms can be applied here to unsupervised learn temporal clusters. TF-IDF is combined with kmeans to cluster complaints occurrence patterns by day and zip codes.

Figure 5: Biggest Sub - Complaint Occurrence Network(CON)



80% of input data is used to derive the clusters, and the most recent 20% data are assigned to the learned clusters, which can be used to find historical days with similar situations/complaint patterns. Number of clusters is chosen to be 40 based on the elbow method on percentage of variance explained (as shown in Figure 8).

Figure 9 and Figure 10 shows two examples of clusters, where x-axis are the "daily document in a zipcode", and each row (y-axis) is one of the selected complaint types. Again, the first 80% are in-sample clustering assignment results, and the last 20% are out of sample cluster assignment, which show pretty good out-of-sample cluster label results.

5. Topic Modeling - Latent Dirichlet Allocation

Along the same line as viewing this as a topic modeling question, we can apply delicate latent topic models on the input data in Figure 1. The assumption is now that daily complaint patterns in zip codes are from a mixture of "latent topics"(or situations), with each "topic"having probability distribution of generating relevant complaint types. The observed "bag of complaint types"are thus from the mixture latent topics on that day.

Identifying the latent topics and daily topic patterns can better help understand the actual situation on that day. Without further introduction to the Latent Dirichlet Allocation (LDA) model, we applied LDA to the input data with NPMI filtered subset of 192 complaint types. 80 topics are used in the model, which are determined based on log-perplexity in 20% cross-validation set. The latent topics from the model are saved (LDA_topics.xlxs) and a subset of the topics are shown in Figure 11.

Mixture of topics on new data: When new data come in, it will be fitted to the learned LDA model and get assigned to a mixture of topics - which will greatly help us find out the latent situation on that day.

Clustering on latent topics: Although not covered in this summary, clusters can be formed on the latent topic level, which in turn will help derive clusters on the input daily data.

Figure 6: Top Subgraph Complaints w/ Different Centrality Measures

Top Complaints by Betweenness	Top Complaints by Closeness
Broken Parking Meter	Broken Parking Meter
Open Flame Permit	Adopt-A-Basket
Adopt-A-Basket	Bike Rack Condition
Bike Rack Condition	Investigations and Discipline (IAD)
Investigations and Discipline (IAD)	Open Flame Permit
Fire Safety Director - F58	Plant
Radioactive Material	Emergency Response Team (ERT)
Plant	Fire Safety Director - F58
Mosquitoes	Tattooing
Taxpayer Advocate Inquiry	Mosquitoes
Homeless Encampment	Radioactive Material
Unsanitary Animal Pvt Property	Taxpayer Advocate Inquiry
Sprinkler - Mechanical	Asbestos
ATF	Street Sign - Missing
Animal in a Park	Vending
Recycling Enforcement	Special Projects Inspection Team (SPIT)
Emergency Response Team (ERT)	BEST/Site Safety
OEM Disabled Vehicle	Homeless Encampment
Mold	EAP Inspection - F59
Special Projects Inspection Team (SPIT)	OUTSIDE BUILDING

Figure 7: Prepared Input Data Matrix

Occurrence #	Day 1			Day 2			Day N		
	Zipcode 1	...	Zipcode Z	Zipcode 1	...	Zipcode Z	Zipcode 1	...	Zipcode Z
Complaint Type 1	1		0	1		1	0		1
Type 2	3		1	0		0	0		0
...	...			...			...		0
Complaint Type 250	2		1	3		1	1		1

Figure 8: Choose Number of Clustering in TF-IDF KMeans Clustering

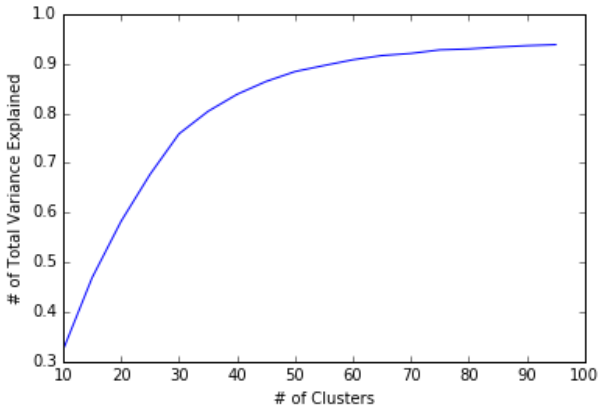


Figure 9: Illustration of Temporal Clustering - I: Cluster #33

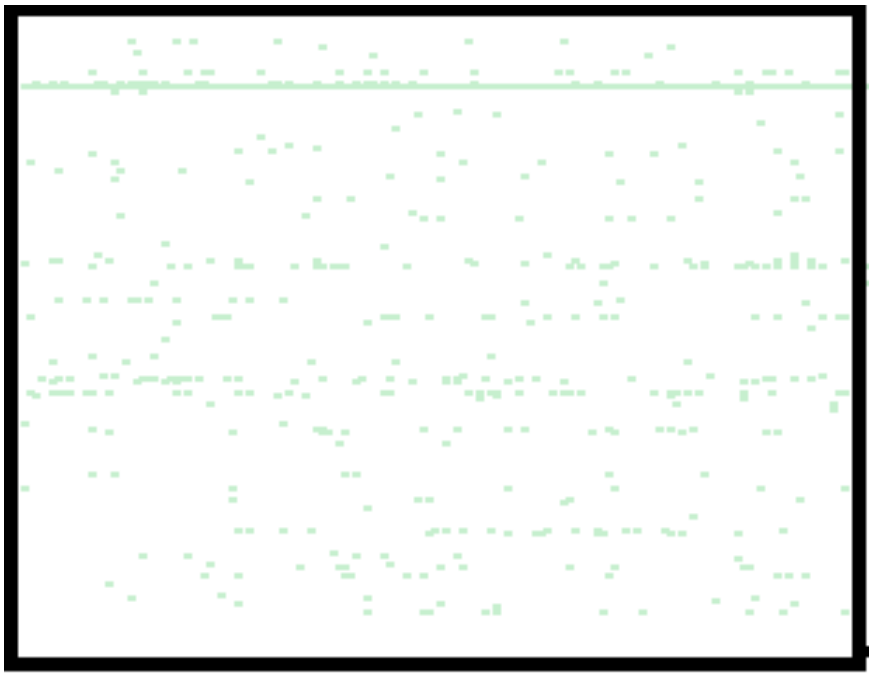


Figure 10: Illustration of Temporal Clustering - II: Cluster #20

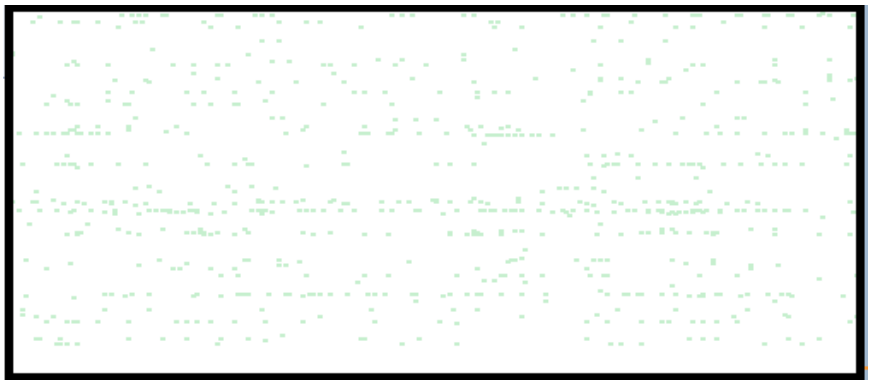


Figure 11: Selected Topics from Fitting LDA Topic Model

Latent Topic ID	Topic Top Three Words
0	0.985*"For Hire Vehicle Report" + 0.000*"PLUMBING" + 0.000*"Special Projects Inspection Team (SPIT)"
1	0.977*"PAINT/PLASTER" + 0.016*"Posting Advertisement" + 0.007*"PLUMBING"
2	**0.493*"Sewer" + 0.493*"DOF Property - Update Account" + 0.003*"Electronics Waste"**
3	0.963*"DOF Property - RPIE Issue" + 0.001*"GENERAL CONSTRUCTION" + 0.001*"HEATING"
4	0.917*"Illegal Tree Damage" + 0.081*"Senior Center Complaint" + 0.000*"Electronics Waste"
5	**0.722*"Plumbing" + 0.276*"Smoking" + 0.000*"PLUMBING"**
6	0.994*"SAFETY" + 0.005*"PLUMBING" + 0.000*"PAINT/PLASTER"
7	1.000*"Homeless Person Assistance" + 0.000*"PLUMBING" + 0.000*"HEATING"
8	0.802*"Homeless Encampment" + 0.197*"DPR Internal" + 0.000*"Special Projects Inspection Team (SPIT)"
9	0.803*"Literature Request" + 0.003*"Graffiti" + 0.002*"GENERAL CONSTRUCTION"
10	0.999*"Sidewalk Condition" + 0.000*"PLUMBING" + 0.000*"Radioactive Material"
11	0.461*"Bike/Roller/Skate Chronic" + 0.399*"Fire Alarm - Reinspection" + 0.132*"Plant"
12	**0.773*"WATER LEAK" + 0.188*"PAINT/PLASTER" + 0.034*"OUTSIDE BUILDING"**
13	0.997*"Violation of Park Rules" + 0.000*"PLUMBING" + 0.000*"HEATING"
14	0.842*"Panhandling" + 0.148*"Snow" + 0.000*"Plumbing"
15	0.940*"Noise - Vehicle" + 0.060*"Water Quality" + 0.000*"UNSANITARY CONDITION"
16	0.997*"Sewer" + 0.003*"Unsanitary Animal Facility" + 0.000*"PLUMBING"
17	0.999*"Standing Water" + 0.000*"UNSANITARY CONDITION" + 0.000*"Electronics Waste"
18	**0.704*"Electrical" + 0.248*"Overflowing Litter Baskets" + 0.046*"Drinking Water"**
19	0.947*"Noise - Helicopter" + 0.046*"Bottled Water" + 0.000*"SAFETY"
20	0.994*"DOF Property - Update Account" + 0.000*"PAINT - PLASTER" + 0.000*"GENERAL CONSTRUCTION"
21	0.914*"Advocate-Personal Exemptions" + 0.028*"OEM Disabled Vehicle" + 0.001*"GENERAL CONSTRUCTION"
22	1.000*"Derelict Vehicles" + 0.000*"Mosquitoes" + 0.000*"UNSANITARY CONDITION"
23	0.999*"Air Quality" + 0.000*"PLUMBING" + 0.000*"UNSANITARY CONDITION"
24	0.970*"City Vehicle Placard Complaint" + 0.000*"HEATING" + 0.000*"Snow"
25	0.977*"Investigations and Discipline (IAD)" + 0.002*"PLUMBING" + 0.000*"GENERAL CONSTRUCTION"
26	0.998*"Special Enforcement" + 0.000*"PLUMBING" + 0.000*"HEATING"
27	0.992*"UNSANITARY CONDITION" + 0.006*"Elder Abuse" + 0.002*"PLUMBING"
28	**0.650*"Standpipe - Mechanical" + 0.112*"OEM Disabled Vehicle" + 0.005*"Snow"**
29	0.545*"Transportation Provider Complaint" + 0.142*"OEM Disabled Vehicle" + 0.020*"HEATING"
30	0.996*"Asbestos" + 0.000*"PLUMBING" + 0.000*"Non-Emergency Police Matter"
31	0.996*"Electronics Waste" + 0.002*"Miscellaneous Categories" + 0.000*"GENERAL CONSTRUCTION"
32	0.999*"Root/Sewer/Sidewalk Condition" + 0.000*"Electronics Waste" + 0.000*"UNSANITARY CONDITION"
33	0.999*"Food Establishment" + 0.000*"Advocate-UBT" + 0.000*"FLOORING/STAIRS"
34	0.980*"Mosquitoes" + 0.005*"Interior Demo" + 0.002*"Building Use"