Crime Rate Inference with Big Data

Hongjian Wang[†], Daniel Kifer[‡], Corina Graif[§], Zhenhui Li[†]
[†]College of Information Sciences and Technology
[‡]Department of Computer Science & Engineering
[§]Department of Sociology and Criminology
Pennsylvania State University, University Park, PA, USA
{hxw186, jessieli}@ist.psu.edu, dkifer@cse.psu.edu, corina.graif@psu.edu

ABSTRACT

Crime is one of the most important social problems in the country, affecting public safety, children development, and adult socioeconomic status. Understanding what factors cause higher crime is critical for policy makers in their efforts to reduce crime and increase citizens' life quality. We tackle a fundamental problem in our paper: crime rate inference at the neighborhood level. Traditional approaches have used demographics and geographical influences to estimate crime rates in a region. With the fast development of positioning technology and prevalence of mobile devices, a large amount of modern urban data have been collected and such big data can provide new perspectives for understanding crime. In this paper, we used large-scale Point-Of-Interest data and taxi flow data in the city of Chicago, IL in the USA. We observed significantly improved performance in crime rate inference compared to using traditional features. Such an improvement is consistent over multiple years. We also show that these new features are significant in the feature importance analysis.

1. INTRODUCTION

Understanding how to control crime is important because exposures to violence and crime have been unusually high in the U.S. for several decades and, while declining, they remain high [6, 16]. Over half a million children and youth aged 10-24 years were treated in 2012 in emergency departments for nonfatal physical assault injuries related to gun shots, cuts and stabbings, among others [17]. Understanding the neighborhood context of crime is particularly important because victimization and other forms of crime exposures have many severe consequences. Beyond the high medical bills and violent death, consequences include behavioral and mental health problems, aggression, substance abuse, post-traumatic stress disorder, and suicide, lower academic achievement, and engaging in further violence [22].

In this paper, we study the problem of crime rate inference of communities. We select Chicago as the target of study for the following reason. Chicago has more homicides and non-negligent manslaughter rates (15.2) per 100,000 residents than New York (4.0) and Los Angeles (6.5) according to the FBI crime statistics for 2013 and has experienced no decline in the past decade compared to the other two large cities, which have been on a slow declining slope [39].

KDD'16 August 13-17, 2016, San Francisco, CA, USA

DOI: http://dx.doi.org/10.1145/2939672.2939736

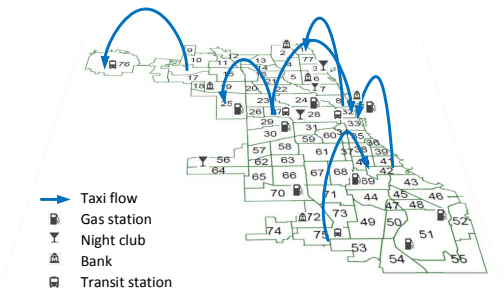


Figure 1: An illustration of various types of features we used in Chicago. The POI distribution across community areas reflects profiles of the region functionality. The taxi flow connects non-adjacent regions and act as "hyperlinks" on the space.

Traditionally, researchers have used demographic information (e.g., population poverty level, socioeconomic disadvantage, racial composition of population) to estimate the crime rate in a community [24]. However, such demographic information only contains partial information about the neighborhoods and does not dynamically reflect the changes in the community (e.g., official counts are collected by the U.S. Census Bureau every 10 years). Using only demographic information will result in a relative error of at least 30% for crime rate estimation in Chicago (refer to experiment section in the paper). Existing studies also use the geographical influence [4] to estimate the crime rate, i.e., the crime in the nearby communities can be propagated to the focal community. But this geographical influence is of little help in improving the crime inference on top of demographic feature, with at most 0.4% relative improvement in our experiments. This is probably because the nearby communities also share similar demographics, which limits the additional benefit of geographical influence.

Recently, big data reflecting city dynamics have become widely available [45], e.g., traffic flow, human mobility, social media, and crowd-generated Points-Of-Interest (POI). As shown in Figure 1, such newer types of big data could provide us new insights to understand some traditional socioeconomic urban problems, such as the crime rate inference problem we focus on in this paper. In particular, we propose to study two newer types of urban data: POI and taxi flow.

POI data. POI data provide venue information such as GPS coordinates, category, popularity, and reviews. These POIs mostly belong to categories such as food, shop, transit, education, etc. Recent

studies have shown that using such categorical information of POIs are useful to profile neighborhood functions [44]. Such neighborhood functions could further help us predict crime rate (e.g., communities with less education or entertainment facilities may have a higher rate of crime). Our experiments show that incorporating POI features significantly improve the crime rate inference. Adding POI features in addition to demographics features reduces the relative error by at least 5% in our experiments. This demonstrates that POI data provide additional information about the communities that is not covered by the demographics.

Taxi flow data. A huge amount of taxi flow data reflect how people commute in the city. In previous studies, when using geographical influence [4], people assume that a community is affected by the spatially nearby communities. However, communities are not only affected by spatially-close communities. Even if two communities are distant in geographical space, they could have a strong correlation if many people frequently travel between these two communities [23]. We hypothesize that taxi flows may be considered as "hyperlinks" in the city that connect the locations and we use such data to estimate crime rates. Taxis may be preferred to public transportation by offenders traveling to a crime location as they offer more privacy and more flexible pick-up and drop-off points. Even if taxis do not constitute the main transportation mode in committing crime, taxi flows may be a proxy for broader patterns of population routine activity and mobility, commuting flows, and other forms of social and economic exchanges between two communities over space. Such exchanges may increase the number of potential targets and opportunities for crime [13, 9] or contribute to inter-community diffusion of information about successful local strategies to control or prevent crime (e.g., successful features of neighborhood watch programs). Our experiments show very promising results – adding taxi flow data on top of all other features can further decrease the error by 5%.

We conduct extensive experiments including a systematic comparison between linear regression and negative binomial models, tests of different combinations of features, detailed discussions of how to construct features, analysis of the relative importance of features, and theoretical interpretations of the results from a social scientist (a co-author in the paper). The experiments are conducted on the crime data over multiple years. We demonstrate that using the big urban data shows significant improvements.

In summary, the contribution of this paper are: 1) We study an old but very important crime inference problem by utilizing new urban data: POIs and taxi flows. 2) We find that utilizing these new types of big urban data significantly improves the crime rate inference. 3) We conduct systematic experiments to compare different results and feature combinations. The significantly better performance could serve as a new baseline for future crime inference problems.

The rest of this paper is organized as follows. We first review the related work in Section 2. The crime inference problem is formulated in Section 3. We discuss the inference model in Section 4 and feature extraction procedure in Section 5. The Section 6 presents the quantitative evaluation results on real data. Finally, we conclude the paper in Section 7.

2. RELATED WORK

In the criminology literature researchers have studied the relationship between crime and various features. Examples are historical crime records [28, 40], education [15], ethnicity [8], income level [32], unemployment [18], and spatial proximity [4]. In data mining, newer types of data are used in the study. For example, there are studies using twitter to predict crime [41, 20], and studies using cellphone data [38, 7] to evaluate crime and social theories at scale. Overall, the existing work on crime prediction can be categorized into three paradigms.

Time-centric paradigm. This line of work focuses on the temporal dimension of crime incidents. For example, in a study [28], the authors propose to use a self-exciting point process to model the crime and gain insights into the temporal trends in the rate of burglary. In another study [33], the authors investigate the temporal constraints on crime, and propose an offender travel and opportunity model. This paper validates the claim that a proportion of offending is driven by the availability of opportunities presented in the routine lives of offenders.

Place-centric paradigm. Most existing work adopt a place-centric paradigm, where the research question is to predict the location of crime incidents. The predicted crime location is usually referred by the term *hotspot*, which has various geographical size. There are plenty of studies on exploration of the crime hotspots. For example, in a study [37] the authors use criminal offense records to identify spatio-temporal patterns at multiple scales. They employ various quantitative tools from mathematics and physics and identify significant correlation in both space and time in the crime behavioral data. Short *et al.* [36] use a simple model to study the dynamics of crime hotspots and identify stable hotspots, where criminals are modeled as random walkers. Bogomolov *et al.* [7] use human behavioral data derived from mobile network and demographic sources, together with open crime data to predict crime hotspots. They compare various classifiers and find random forests have the best prediction performance. The paper [41] uses automatic semantic analysis to understand natural language Twitter posts from which the crime incidents are reported. Some other work [12, 14] employ kernel density estimation (KDE) to identify and analyze crime hotspots. Those studies form another form of crime prediction, which relies on the retrospective crime data to identify areas of high concentrations of crime. In [30], the authors extend the crime cluster analysis with a temporal dimension. They employ the space-time variants of KDE to simultaneously visualize geographical extent and duration of crime clusters.

Population-centric paradigm. In the last paradigm, research focuses on the criminal profiling at individual and community levels. At the individual level, [40] aim to automatically identify crimes committed by the same individual from a historical crime database. The proposed system, called *Series Finder*, is designed to find and classify modus operandi (M.O.) of criminals. At the community level, Buczak *et al.* [10] use fuzzy association rule mining to find crime patterns. The rules they found are consistent across all regions. The paper constructs association rules from population demographics in communities. In another paper [38], the authors use computational methods to validate various social theories at a large scale. They used mobile phone data in London, from which they mine the people dynamics as features to correlate with crime.

Our problem is different from the first two categories of work, mainly because our innovation lies in using newer type of data to enhance the commonly used traditional counterparts. More specifically, we use POI to enhance the demographics information and use taxi flow as hyperlinks to enhance the geographical proximity correlation. Although our problem does not consider the temporal dimension of crime in depth, it could be a promising supplement to better profile crime. Our problem does not predict the location of any particular crime incident. Therefore the methods proposed in place-centric methods are not applicable in our problem. However, the features we proposed may be incorporated in those crime prediction models.

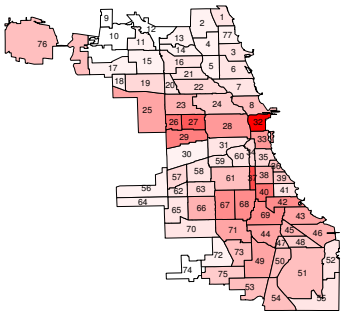


Figure 2: Crime rate of Chicago by community areas. The community area #32 is Chicago downtown, which has the highest crime rate.

Our problem falls into the third paradigm because we try to profile the crime rate for Chicago community areas. In our problem, the community areas are well-defined and stable geographical regions. The newly proposed POI features and taxi links providenew perspectives in profiling the crime rate across community areas.

3. OVERVIEW

The crime data collected in Chicago has detailed information about the time and location (i.e., latitude and longitude) of crime and the types of crime. In our problem, the term crime count refers to number of crime incidents in a region (i.e., community area) in a year. The *community area* is used as our geographical unit of study, since it is well-defined, historically recognized and stable over time [42]. In total, there are 77 community areas in Chicago. Crime rate is the crime count normalized by the population in a region. We use vector $\vec{y} = [y_1, y_2, \ldots, y_n]$ to denote the crime rates in regions. The crime rate inference problem is to estimate the crime rate in one region using the crime rate of other regions in the same year by considering the features of regions and correlations between regions.

The crime data of Chicago are obtained from the City of Chicago data portal [3]. Chicago is the city with most complete crime data that are made public online. The crime dataset contains the incident date, location (strict name and GPS coordinates), and primary type from year 2001 to 2015. In total there are $5{,}856{,}414$ recorded crime incidents over 15 years, which is an average $390{,}417$ crimes incidents per year. We visualize the crime normalized by population in Figure 2, from which we can see that the downtown area has the highest crime rate.

In this paper we study the crime rate inference problem. More specifically, we estimate the crime rate of some regions given the information of all the other regions. Without loss of generality, we assume there is one community area t with crime rate y_t missing, and we use the crime rate of all the other regions $\{y_i\}\backslash y_t$ to infer this missing value. Our problem is mathematically formalized as follows

$$\hat{y_t} = f(\{y_i\}\backslash y_t, X), \tag{1}$$

where X refers to observed extra information of all those community areas.

We consider two types of features X for inference:

- Nodal feature. Nodal features describe the characteristics of the focal region. Such features include demographic information and Point-of-Interest (POI) distribution. Demographics are frequently used in literature, but POI is a newer type of big data, which we find significantly improve the crime inference accuracy.
- Edge feature: (1) Geographical influence. Geographical influence considers the crime rate of the nearby locations. This feature has been extensively used in literature as well. To estimate the focal region, the crime rate of nearby regions are weighted according to spatial distances. (2) Hyperlink by taxi flow. Locations are connected through the frequent trips made by humans, which can be considered as the hyperlinks in space. This type of feature has never been studied in literature. We propose to use taxi trips to construct the social flow. Our hypothesis is that two regions that are more strongly connected through social flow will influence each other's crime rate.

In the following sections, we first discuss the inference models based on these three types of features in Section 4 and then discuss how to construct these features using the real-world data in Section 5.

4. INFERENCE MODEL

4.1 Linear Regression

The most straightforward prediction model is the linear regression. This model assumes the error terms follow a Gaussian distribution $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

Equation 2 gives the linear regression formulation of our problem.

$$\vec{y} = \vec{\alpha}^T \vec{x} + \beta^f W^f \vec{y} + \beta^g W^g \vec{y} + \vec{\epsilon}, \tag{2}$$

where $\vec{x}$ represents the nodal features, including demographics and POI distribution, W^f is the flow matrix of taxi flow, and W^g is the spatial matrix representing the geographical adjacency. On the right-hand side, ϵ is the only stochastic variable, and all other terms are fixed observation values. Therefore, we incorporate all the fixed observations into one term X, and we get the standard regression problem

$$E(y) = Xw + \epsilon.$$

4.2 Negative Binomial Regression

In our problem, we aim to infer the crime rate, which is guaranteed to be a non-negative integer. However, linear regression does not ensure this property. *Poisson regression* is another form of regression, more appropriate for count data than linear regression [19][27]. With shortened notation X, the Poisson regression model has the exponential function as link function

$$E(y) = e^{Xw}. \tag{3}$$

This comes from the assumption that y follows Poisson distribution with mean λ. Additionally, the mean λ is determined by observed independent variables X, with the link function $\lambda = e^{Xw}$. Adding all together, the joint probability of y is

$$P(y|w) = \frac{e^{-e^{Xw}}(e^{Xw})^y}{y!}. \tag{4}$$

However, Poisson regression enforces the mean and variance of dependent variable y to be equal. This restriction leads to the "over-dispersion" issue for some real problems, that is the presence of larger variability in data set than the statistical model expected. To address this, we use the Poisson-Gamma mixture model, which is

also known as *negative binomial regression*. Negative binomial regression has been used in similar work [31].

Given that the crime rate y follows Poisson distribution with mean λ, in order to allow for larger variance, λ itself is a random variable having a Gamma distribution with shape $k = r$ and scale $\theta = \frac{1-p}{p}$. The probability function of y becomes

$$P(y|r,p) = \int_0^\infty P_{Poisson}(y|\lambda) \cdot P_{Gamma}(\lambda|r,p)d\lambda$$
$$= \frac{\Gamma(r+y)}{y!\Gamma(r)}p^k(1-p)^y \quad (5)$$

This is exactly the probability density function of negative binomial distribution.

In negative binomial regression, the link function is

$$E(y) = e^{Xw+\epsilon}. \quad (6)$$

The error term e^ϵ is the mixture prior, and we assume it follows Gamma distribution with shape parameter $k = \frac{1}{\theta}$, so that it has mean $E(e^\epsilon) = k\theta = 1$ and variance $Var(e^\epsilon) = k\theta^2 = \theta$. This setting ensures the $E(y) = e^{Xw} \cdot e^\epsilon = e^{Xw}$.

5. FEATURE EXTRACTION

In this section, we will discuss the details of features used in our method. The two types of new features we use are extracted from Point-Of-Interest data and taxi flow data. Below we describe the datasets used to construct features and the characteristics of these features.

5.1 Nodal Feature: Demographics

Socioeconomic and demographic features of neighborhoods have been widely used to predict crime [7, 25, 43, 34]. Previous studies have shown that crime rate correlates with certain demographics. For example, [26, 24] suggests that population diversity leads to less crime in certain neighborhoods. In our study, we include demographic information from the US Census Bureau's Decennial Census [2]. Using 2010 census information would overlap with the time in which crime is measured. Instead, we use year 2000 demographic data because we are interested in predictors that precede temporally the period in which crime rates are evaluated. The demographics include the following features:

total population, population density, poverty, disadvantage index, residential stability, ethnic diversity, race distribution.

The poverty index measures the proportion of community area residents with income below the poverty level. The disadvantage index is a composite scale based on prior work [35], a function of poverty, unemployment rate, proportions of families with public assistance income, and proportion of female headed households. The residential stability measures home ownership and proportion of residents who lived in the neighborhood for more than one year. Racial and ethnic diversity is an index of heterogeneity [24] based on six population groups, including: Hispanics, non-Hispanic Blacks, Whites, Asians, Pacific Islanders and others.

Figure 3 visualizes the crime rate and demographics features in Chicago by community areas. Comparing with Figure 2, it is clear that the crime rate and poverty index and disadvantage index are consistent, the ethnic diversity shows an inverse correlation, and the total population has little correlation with crime.

Table 1 shows the Pearson correlation coefficient between various demographics features and the crime rate at community area level. The corresponding p-value is also calculated and shown in the table to indicate the significance of the correlation coefficient. There are in total 77 community areas in Chicago. Table 1 shows such correlation with several most correlated features. We can see that the poverty index and disadvantage index positively and strongly correlate with crime, while the ethnic diversity negatively correlates with crime. Other features such as total population, population density, and residential stability have weaker correlations. One counter-intuitive observation is that the total population has a weak and negative correlation with crime. The reason is that we use crime rate in each community area, which is already normalized by the population, and therefore the total population and population density have less impact.

Table 1: Pearson correlation between demographic features and crime rate (* indicates significant correlations with p-value less than 5%).

Feature	Correlation	p-value
Total Population	-0.1269	0.2716
Population Density	-0.1972	0.0855
Poverty Index	**0.5573***	1.403e-07
Disadvantage Index	**0.5959***	1.082e-08
Residential Stability	-0.0453	0.6965
Ethnic Diversity	**-0.5545***	1.678e-07
Percentage of Black	**0.6696***	2.779e-11
Percentage of Hispanic	**-0.3820***	0.0006

5.2 Nodal Feature: Point-of-Interest (POI)

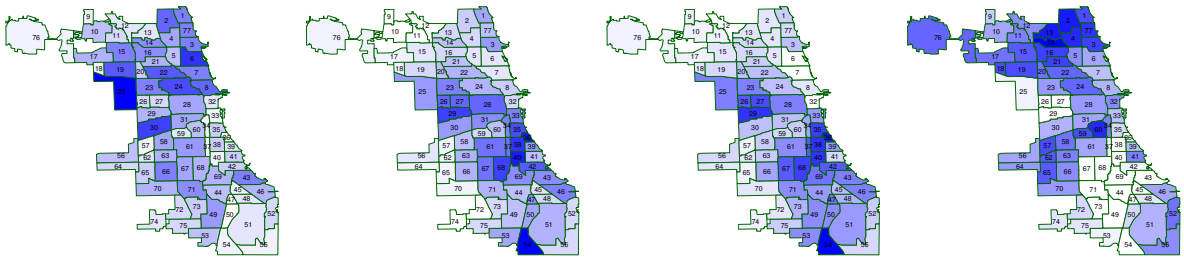
While demographics are traditional census data, POI is a type of modern data that provide fine-grained information about locations. We collect POI from FourSquare [1]. POI data from FourSquare provide the venue information including venue name, category, number of check-ins, and number of unique visitors. We mainly use the major category information because categories can characterize the neighborhood functions. There are 10 major categories defined by FourSquare:

food, residence, travel, arts & entertainment, outdoors & recreation, college & education, nightlife, professional, shops, and event.

In total, we have crawled $112{,}000$ POIs from FourSquare for Chicago. Most of these POIs are in the downtown area of Chicago. For the purpose of visualization, we normalize the POIs count per category by the total POI count in a neighborhood and plot two selected categories, i.e. nightlife and professional, in Figure 4. The darker colored neighborhoods in Figure 4 are the ones with a higher proportion of residence POIs.

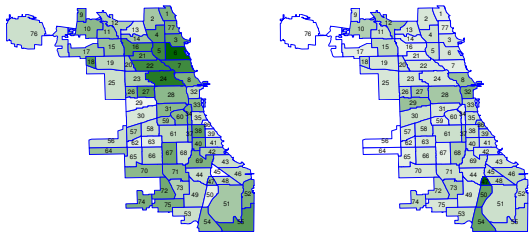
Table 2: Pearson correlation between POI category and crime rate (* indicates significant correlations with p-value less than 5%).

POI category	Correlation	p-value
Food	-0.1543	0.1803
Residence	-0.0610	0.5984
Travel	-0.0017	0.9883
Arts & Entertainment	-0.0049	0.9661
Outdoors & Recreation	0.0668	0.5637
College & Education	-0.0078	0.9473
Nightlife	-0.1553	0.1775
Professional	**0.3221***	0.0043
Shops	-0.1676	0.1450
Event	0.2196	0.0549



(a) Total population (b) Poverty index (c) Disadvantage index (d) Ethnic diversity

Figure 3: (a)-(d) Demographics in Chicago by community areas. Darker colors indicate higher values. Each demographic feature is normalized into $[0, 1]$.



(a) Nightlife (b) Professional

Figure 4: POI ratio per neighborhood. The saturation of color is proportional to the ratio value. The "professional" category distribution is more consistent with the crime distribution, and therefore it is the most correlated with crime. Meanwhile, the "nightlife" category is negatively correlated with Chicago crime. The POI ratios are independently normalized for different POI categories.

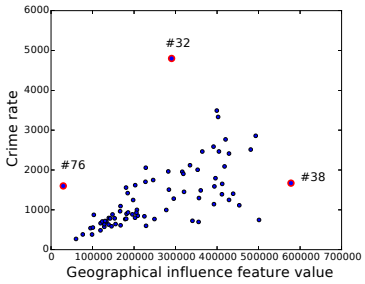


Figure 5: The correlation between geographical influence feature and crime rate. In the plot we marked out three outliers and their corresponding community area ID.

In Table 2 we show the Pearson correlation between POI category and crime rate. The category "professional" is most significantly correlated with the crime rate. Under the professional POI category, there are some venues with a large population concentration, such as transportation center, convention center, community center, and coworking space. In those venues, the population volume is high and residential stability is low, therefore the professional POI counts positively correlates with crime rate. One counter-intuitive observation is that "nightlife" category is not positively correlated with crime (-0.1553). This can be seen in Figure 4(a). The majority of nightlife venues in Chicago are located in the northern area, while most crime incidents occur in the downtown area.

5.3 Edge: Geographical Influence

Together with the US census demographics data, we also collected the boundary shape files of Chicago, which are used to calculate the geographical influence feature. Previous studies have also shown that the crime rate at one location is highly correlated with nearby locations [21, 11]. Such geographical influence is also frequently used in the literature [5, 29]. It is calculated as:

$$\vec{F^g} = W^g \cdot \vec{Y}, \tag{7}$$

where W^g is the spatial weight matrix. If region i and j are not geospatially adjacent, $w^g_{ij} = 0$; otherwise, $w^g_{ij} \propto distance(i, j)^{-1}$.

In Figure 5, we plot crime rate with respect to geographical influence calculated in Eq. 7. We observe an obvious positive correlation, which means if nearby neighborhoods have a high crime rate, the focal neighborhood is more likely to have a high crime rate. We also do observe a few outliers in Figure 5. These neighborhoods show different crime rate in their nearby neighborhoods compared to their own. For example, as we can also see in Figure 2, community area #38 locates in an area where the the neighbors have high crime rates but its crime rate is relatively low; in contrast, neighborhood #32 has a high crime rate even though its neighbors have relatively low crime. The community area #76 home of the O'Hare International Airport is far from most of other community areas, however its own crime rate is relative high.

5.4 Edge: Hyperlinks by Taxi Flow

In our Chicago taxi dataset, there are 1,048,576 taxi trips in total from October to December in 2013. For each trip, the following information are available: pickup/dropoff time, pickup/dropoff location, operation time, and total amount paid. We requested the taxi trip records from Chicago under the Illinois Freedom of Information Act. Figure 6 shows a visualization of the major flows at community level.

One of our hypotheses is that the social interaction among two community areas propagates crime from one region to another. The Chicago taxi data captures the social interactions among various community areas. To calculate this, we first map all taxi trips to community areas to get the taxi flow w_{ij} $\forall i, j \in \{1, 2, \cdots n\}$. Then the taxi flow lag is constructed by the product of social flow

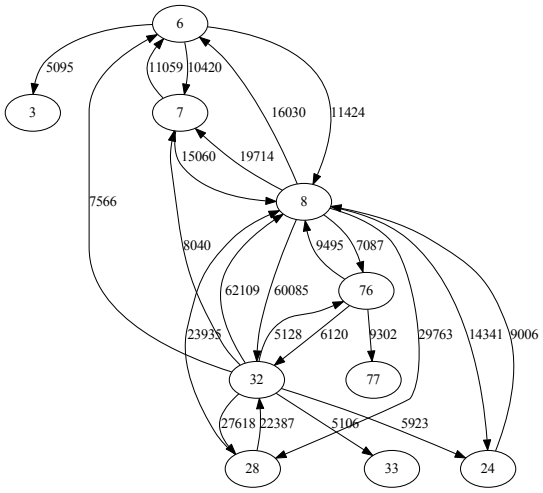


Figure 6: Major taxi flows between neighborhoods. The label on the edge shows the count of taxi trips commuting between two community areas from October to December months in 2013. We set a threshold (more than 5,000 trips) on the flow and only plot high volume flows. The label on a node is the ID of its corresponding community area. We can see that there are several hub community areas, such as #6, #8, #32, which are all in the downtown areas.

and the crime rate of neighboring regions as follows

$$\vec{F^t} = W^t \cdot \vec{Y}. \tag{8}$$

The taxi flow W^t is a matrix with entry w_{ij} denoting the taxi flow from i to j. Note that $\forall i$, $w^s_{ii} = 0$ in matrix W^t, because we have to exclude the crime in the focal area from its own predictor. The semantic of this taxi flow feature is how much crime in the focal area is contributed by its neighboring areas through social interaction.

The correlation between taxi flow and crime rate is shown in Figure 7. From the scatter plot, we can see that overall the crime rate is positively correlated with the taxi flow. There are two outliers clearly shown in Figure 7. The community area #32 is the downtown Loop, which has the highest crime rate and is hard to predict by taxi flow. Another anomalous community area (#47) has relatively low crime rate by itself. However, this area has a lot of in flows from high-crime communities.

6. EXPERIMENTS

6.1 Settings

We adopt leave-one-out evaluation to estimate the crime rate of one geographic region given all the information of all the other regions. When we construct the spatial/social lag variable for the training data, the effect of testing region is completely removed. For example, if region y_t is the testing region, the remaining $\{y_i\}\backslash y_t$ become the training set. For any y_j in the training set, its geographical influence feature and taxi flow feature are constructed from $\{y_i\}\backslash\{y_t, y_j\}$.

In the evaluation, we estimate the crime rate for testing community areas. The accuracy of estimation is evaluated by mean absolute error (MAE) and mean relative error (MRE).

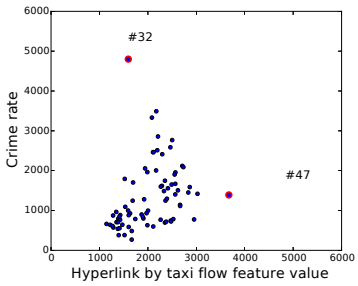


Figure 7: Correlation between taxi flow feature and crime rate. In the plot, we marked out two outliers and their corresponding community area ID.

$$MAE = \frac{\sum_i^n |y_i - \hat{y}_i|}{n} \tag{9}$$

$$MRE = \frac{\sum_i^n |y_i - \hat{y}_i|}{\sum_i^n y_i} \tag{10}$$

6.2 Performance Study

We evaluate the estimation accuracy under various feature combinations. The leave-one-out evaluation results are shown in Table 3. We run both the linear regression and the negative binomial regression on five consecutive years, 2010 – 2014. Both MAE and MRE are shown in the table. We have four types of features: demographics, POI, geographical influence and taxi flow. We test the various settings of feature combinations.

6.2.1 *Negative Binomial Regression vs. Linear Regression*

In Table 3, we can see that in different years and under most settings, the negative binomial regression significantly outperforms the linear regression (with only a few exceptions when using only demographic feature). When using all the features, NB is significantly better than LR with at least 6% improvement in relative error. One reason is that negative binomial regression is a count prediction model, which guarantees the prediction variable is non-negative . Another reason is that it is difficult to get very precise estimates of crime rate, and the negative binomial regression allows a large variance in the estimated crime rate. Therefore negative binomial is more appropriate for crime rate estimation than linear regression.

In the following discussions, we only refer to the performance of the negative binomial regression.

6.2.2 *POI Feature*

Adding POI features always improves the accuracy (see NB for column 2 vs. column1, column 6 vs. column 5, column 8 vs. column 7). The POI distribution reflects the functionality of a region. The most correlated POI major category is "professional", under which there are a lot of venues like transportation center and conventional center. These are locations with more dynamic movements of people. Such location information is not reflected in any of other features. POI thus provides unique information and it shows that using big data can benefit us in advancing the study of traditional crime inference problems.

Another issue that is worth discussing is whether POI is a surrogate of population features from demographics. That is, a region with POIs is a region with a higher population. However, as we

Table 3: Performance evaluation. Various feature combinations are shown in each column. The linear regression model and negative binomial results are compared by year group.

			Settings							
Column ID			1	2	3	4	5	6	7	8
Features[1]		**D**emo	✓	✓	✓	✓	✓	✓	✓	✓
		Geo					✓	✓	✓	✓
		POI		✓		✓		✓		✓
		Taxi			✓	✓			✓	✓
Year	Model[2]	Error								
2010	LR	MAE	394.41	416.98	408.09	406.93	394.78	432.45	402.25	416.41
		MRE	0.294	0.311	0.304	0.304	0.295	0.323	0.300	0.310
	NB	MAE	391.53	333.14	395.64	323.47	389.55	350.06	387.43	**320.75**
		MRE	0.292	0.249	0.295	0.241	0.290	0.261	0.289	**0.239**
2011	LR	MAE	380.22	409.30	396.97	401.11	379.61	422.94	389.39	408.91
		MRE	0.295	0.318	0.309	0.312	0.295	0.328	0.302	0.320
	NB	MAE	381.11	332.62	388.81	328.94	378.84	345.24	381.33	**335.97**
		MRE	0.296	0.259	0.302	0.256	0.294	0.268	0.296	**0.253**
2012	LR	MAE	378.91	412.95	401.54	412.20	376.53	423.88	399.25	419.93
		MRE	0.306	0.334	0.325	0.333	0.304	0.343	0.322	0.339
	NB	MAE	386.31	337.24	389.58	331.41	384.23	352.22	381.67	**345.49**
		MRE	0.312	0.273	0.315	0.268	0.310	0.284	0.308	**0.279**
2013	LR	MAE	367.89	420.81	390.75	402.75	369.24	433.48	388.92	412.31
		MRE	0.324	0.370	0.344	0.354	0.325	0.381	0.342	0.362
	NB	MAE	376.08	333.92	373.08	312.63	377.57	350.33	368.49	**319.86**
		MRE	0.331	0.294	0.328	0.275	0.332	0.308	0.324	**0.281**
2014	LR	MAE	331.28	375.53	349.00	350.31	329.93	386.90	345.79	361.28
		MRE	0.326	0.369	0.343	0.345	0.324	0.380	0.340	0.355
	NB	MAE	340.73	293.52	339.17	274.45	336.09	308.18	326.07	**273.27**
		MRE	0.335	0.289	0.334	0.270	0.331	0.303	0.321	**0.269**

[1] D – demographic features, G – geographical influence, P – POI features, T – taxi flow feature.
[2] LR – Linear Regression, NB – Negative Binomial Regression.

see from Table 3, adding POI in addition to demographics always outperforms the features without POI. This is because population from demographics reflects the number of residents in that region, but POI reflects dynamics of population (e.g., people go to venues for food, entertainment, or travel). Therefore, the dynamic population in POI further complements the residential population in demographics.

6.2.3 Taxi Flow

The taxi flow is shown to improve the inference accuracy (see NB for column 3 vs. column 1, column 7 vs. column 5, column 8 vs. column 6). This validates our hypothesis that crimes do not only correlate with nearby regions but also correlate through hyperlinks on the space (i.e., the taxi flow).

Comparing column 7 (D+G+T) with column 5 (D+G), we find that the improvement by taxi flow is not obvious. However, comparing column 8 (D+G+P+T) with column 6 (D+G+P), we observe a much significant accuracy boost. The reason could be that the taxi flow further complements the POI data. When POI information is missing from the predictor, the city dynamics captured by taxi flow are weakened as well.

6.3 Feature Construction

There are different ways to use the POI and taxi datasets. In this section, we share our insights into the more effective ways in constructing the features.

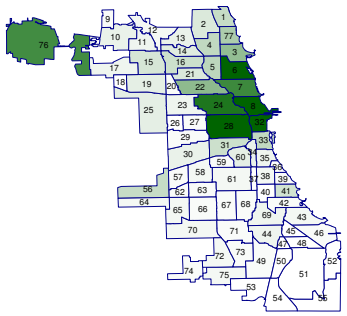


Figure 8: Absolute POI count distribution. In our crawled POI dataset, most community areas have less than 100 venues. Meanwhile, the downtown area there are over $10,000$ venues for one community area, e.g. #8, #32.

6.3.1 POI Normalization

The straightforward definition of POI distribution is calculated by normalizing the POI count in each category by the total POI counts. However, the POIs in Chicago are not evenly distributed. As shown in Figure 8, most POIs are in the downtown area and some areas only have a few POIs. If normalized by the total number of POIs in a neighborhood, two neighborhoods may show similar distributions but they are quite different. For example, a downtown

neighborhood and a distant neighborhood may both have a high ratio of the food category but the downtown neighborhood has many more POIs in total and is more dynamic in population constitution. Therefore, using the raw count instead of normalized distribution is more effective. This is also demonstrated in estimation accuracy as shown in Table 4.

Table 4: Using POI count instead of POI percentage improve the estimation accuracy. Estimation for crime in 2014 with all other features.

Scheme	NB	
	MAE	MRE
POI count	273.27	0.269
POI percentage	283.16	0.278

6.3.2 *Taxi Flow Normalization*

The taxi flow represents the interactions among community areas. There are several different approaches to incorporate the taxi flow into the model. First, we can use the raw taxi count as a weight on crime from other neighborhoods. One issue with the raw count is the concentration of taxi trips distribution in the downtown area. Consider the following example. In the downtown area, the average taxi flow count is 1000 between any pair of community areas, while the average of suburbs is 100. When we propagate crime by raw taxi count, the same amount of crime in downtown is propagated with a 10 times higher coefficient than that of suburb.

To address this issue, we can normalize the taxi flow, and there are two different approaches to normalize. 1) We can normalize the taxi flow by the total incoming traffic of the destination community area, and the semantics of this normalization is splitting the crime in the destination to all its neighbors. 2) Alternatively, we can normalize the taxi flow by the outgoing total trips in the source community area. This normalization assumes the crime in each source community is spread out by the flow. The two normalization methods are shown in Figure 9.

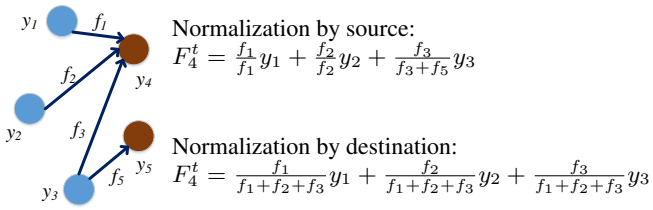


Figure 9: Two different normalization schemes.

Table 5: Various approaches to construct taxi flow feature. Estimation for crime in 2013 with all other features.

Settings	NB	
	MAE	MRE
Taxi flow count	368.71	0.324
Taxi flow normalized by source	349.38	0.307
Taxi flow normalized by destination	319.86	0.281

In Table 5 we compare the different approaches to handle the taxi flow. Using raw taxi flow count is clearly not a good option, due to the unbalanced data distribution. We also observe that normalizing taxi flow by destination is better than normalization by source. The reason could be explained by the example given in Figure 9. Suppose the focal region is a transportation hub, which has a lot of isolated regions connected to it. If we normalize the crime by source region, then the taxi flow feature of focal region is overestimated, since the coefficients of its neighbors do not sum to one.

6.4 Feature Importance

In this subsection, we study the importance of features through significance tests and coefficient changes over the years.

6.4.1 *Significance Test*

From previous results, we see that combining POI features and taxi flow will help improve the estimation accuracy. Now we try to measure the significance of this accuracy boost by permutation tests. If a feature correlates with crime, when we randomly permute the values of this feature among neighborhoods, we will expect a higher error in crime estimation. So in each round of permutation, we can get an error in estimation. We compare the error with the original feature to the error distribution obtained from permutations. We conduct 1,000 rounds of permutations to approximately estimate the error distribution. The position of the original error in this distribution indicates the significance of this feature. For example, if the original error is smaller than 99% of the errors from the permutations, the p-value is 0.99.

Table 6: Estimated p-value for each feature. The p-value is defined as the possibility that a smaller error measure is observed under the null hypothesis.

Settings: D+S+P+T	LR		NB	
	MAE	MRE	MAE	MRE
	412.31	0.363	319.86	0.281
Feature	p-value			
D (demographics)	0.000	0.000	0.000	0.000
G (geographic inf.)	0.640	0.664	0.602	0.565
P (POI distribution)	0.025	0.025	0.001	0.001
T (taxi flow)	0.000	0.000	0.000	0.000

In Table 6, the p-values of different features are given. The demographics feature is the most significant with estimated p-value equals to 0.00. In all the $1,000$ random permutations of demographic feature, we never observe an error lower than the original error. The proposed POI distribution and taxi flow are significant as well, with a p-value of 0.5% and 1.3% for the negative binomial model. One interesting observation is that the geographical influence is not significant at all. One possible reason is that the demographics features capture the similarity of geographical neighbors, and therefore are surrogates of geographical influence.

6.4.2 *Coefficient Study*

In our regression model, the coefficient also indicates the importance of features. We normalize the values of all features to the range $[0, 1]$, so that coefficients are comparable. The top-6 features with the most significant coefficients are shown in Table 7. The top 3 rows in Table 7 are features with positive coefficients, which implies the positive correlation with Crime. The three features with negative coefficients are negatively correlated with crime.

By comparing the coefficients over different years, we observe that the coefficients are relatively stable with respect to time. The most important feature is always POI professional category, which represents many populated public areas. The other two important demographic features are disadvantage index and percentage black.

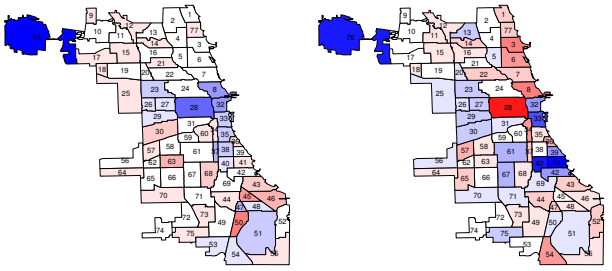
We also find three POI categories are among the top negatively correlated features. They are the residence, shop, and education categories. The reason is that at those places the population is relatively stable, which provides less opportunity for crime.

Table 7: The coefficients of the top-6 features over different years. There are 21 different features in total. Due to limited space, we only show the top 3 features with the highest positive/negative coefficients respectively.

Feature	Year				
	2010	2011	2012	2013	2014
POI professional	1.414	1.733	1.905	2.206	1.874
pct black	1.376	1.370	1.301	1.296	1.252
disadvantage index	1.237	1.055	1.270	1.700	1.462
POI education	-1.171	-1.265	-1.735	-2.041	-1.871
POI shops	-2.671	-2.747	-2.687	-2.549	-2.834
POI residence	-3.059	-2.719	-2.424	-2.151	-2.459

6.5 Improvements on Different Regions

The POI distributions are different from region to region. It is interesting to find out whether POI distribution is consistently positive in making the crime estimation better. We calculate the difference in estimation error (MAE) between two settings: 1) using demographics, geographical influence, taxi flow; and 2) using all these three features plus POI distribution. The similar measurement is calculated for the taxi flow feature. The results are shown in Figure 10. A positive difference (blue area) indicates that adding the new feature will help reduce the estimation error, while a negative difference (red area) indicates that the new feature adds more noise to the data.



(a) POI (b) Taxi flow

Figure 10: Performance improvement per region by using POI or taxi flow features on 2014 crime. The difference of MAEs in estimating crime with/without POI feature is shown on the left, and the same measure of taxi flow is shown on the right. The color blue means the MAE is reduced by adding corresponding feature (i.e., better performance), while the red means the MAE is increased (i.e., worse performance). The color saturation indicates the value of difference.

It is interesting to find out that in the downtown area, i.e. community area #8, #32, #28, and #33, POI significantly improves the estimation accuracy. The reason is two fold. 1) The demographics information from census is mostly about the residing population in the focal area. However, in the downtown area there are a lot of floating population groups conducting various social activities, and this is not reflected by the census demographics. The POI information, on the other hand, reflects the functionality of a region, and plays a complementary role of demographic information. 2) In the downtown area, there are much more POIs than any other places, which provides more complete information about the community profile.

As for the taxi flow feature, it helps the most in those suburb area, because the taxi flow reflects the social interaction in those areas. In the downtown, the taxi flow feature incurs a relatively large estimation error. The reason is that the taxi flow distribution in Chicago is extremely skewed. Roughly 61% of the Chicago taxi trips have a destination in the downtown area, which may result in the model over-propagating crime estimates from all of Chicago into the downtown area.

7. CONCLUSION

In the social science literature, the demographics and geographical neighbors are known to exhibit strong correlations with crime. In this paper we solve the problem of crime rate inference with new features. More specifically, we propose to use POI features to assist the demographic features, and to use taxi flow as hyperlinks to supplement the geographical neighbors. The intuition behind the POI feature is that the POI distribution across community areas reflects profiles of the region functionality. The intuition behind the hyperlinks is that the taxi flow models the social interaction among nonadjacent regions, which potentially propagate crime or resources and information used in crime control. We adopt the negative binomial regression modal over the linear regression model, mainly because the count based regression models and guarantees positive prediction, while the linear regression may give negative crime rate as prediction. Both POI and taxi flow features from a publicly accessible dataset in Chicago are evaluated to be helpful. In the best scenario, the POI distribution and taxi flow reduces the prediction error by 17.6%.

8. ACKNOWLEDGEMENTS

The work was supported in part by NSF award #1544455, #1054389, and funding from NICHD R24-HD044943. The views and conclusions contained in this paper are those of the authors and should not be interpreted as reprinting any funding agencies.

9. REFERENCES

[1] Foursquare venues service. https://developer.foursquare.com/overview/venues.html.

[2] United states census bureau. http://www.census.gov.

[3] City of chicago data portal. https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2, 2015.

[4] ANSELIN, L. Under the hood: issues in the specification and interpretation of spatial regression models. *Agricultural economics 27*, 3 (2002), 247–267.

[5] ANSELIN, L., COHEN, J., COOK, D., GORR, W., AND TITA, G. Spatial analyses of crime. *Criminal justice 4*, 2 (2000), 213–262.

[6] BAUM, K. *Juvenile victimization and offending, 1993-2003*. US Department of Justice, Office of Justice Programs, Bureau of Justice Statistics, 2005.

[7] BOGOMOLOV, A., LEPRI, B., STAIANO, J., OLIVER, N., PIANESI, F., AND PENTLAND, A. Once upon a crime: towards crime prediction from demographics and mobile data. In *Proceedings of the 16th international conference on multimodal interaction* (2014), ACM, pp. 427–434.

[8] BRAITHWAITE, J. *Crime, shame and reintegration*. Cambridge University Press, 1989.

[9] BRANTINGHAM, P., AND BRANTINGHAM, P. Criminality of place. *European journal on criminal policy and research 3*, 3 (1995), 5–26.

[10] BUCZAK, A. L., AND GIFFORD, C. M. Fuzzy association rule mining for community crime pattern discovery. In *ACM SIGKDD Workshop on Intelligence and Security Informatics* (2010), ACM, p. 2.

[11] BURNELL, J. D. Crime and racial composition in contiguous communities as negative externalities: prejudiced household's evaluation of crime rate and segregation nearby reduces housing values and tax revenues. *American Journal of Economics and Sociology 47*, 2 (1988), 177–193.

[12] CHAINEY, S., TOMPSON, L., AND UHLIG, S. The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal 21*, 1 (2008), 4–28.

[13] COHEN, L. E., AND FELSON, M. Social change and crime rate trends: A routine activity approach. *American sociological review* (1979), 588–608.

[14] ECK, J., CHAINEY, S., CAMERON, J., AND WILSON, R. Mapping crime: Understanding hotspots.

[15] EHRLICH, I. On the relation between education and crime. In *Education, income, and human behavior*. NBER, 1975, pp. 313–338.

[16] FINKELHOR, D. *Childhood victimization: violence, crime, and abuse in the lives of young people: violence, crime, and abuse in the lives of young people*. Oxford University Press, USA, 2008.

[17] FOR DISEASE CONTROL, N. C., AND (CDC), P. Leading causes of nonfatal injury, united states 2001 - 2013. *Injury Prevention and Control: data and statistics* (2015).

[18] FREEMAN, R. B. The economics of crime. *Handbook of labor economics 3* (1999), 3529–3571.

[19] GARDNER, W., MULVEY, E. P., AND SHAW, E. C. Regression analyses of counts and rates: Poisson, overdispersed poisson, and negative binomial models. *Psychological bulletin 118*, 3 (1995), 392.

[20] GERBER, M. S. Predicting crime using twitter and kernel density estimation. *Decision Support Systems 61* (2014).

[21] GORMAN, D. M., SPEER, P. W., GRUENEWALD, P. J., AND LABOUVIE, E. W. Spatial dynamics of alcohol availability, neighborhood structure and violent crime. *Journal of studies on alcohol 62*, 5 (2001), 628–636.

[22] GRAIF, C. Toward a geographically extended perspective of neighborhood effects on children's victimization. *American Society of Criminology Annual Meeting* (2015).

[23] GRAIF, C., GLADFELTER, A. S., AND MATTHEWS, S. A. Urban poverty and neighborhood effects on crime: Incorporating spatial and network perspectives. *Sociology Compass 8*, 9 (2014), 1140–1155.

[24] GRAIF, C., AND SAMPSON, R. J. Spatial heterogeneity in the effects of immigration and diversity on neighborhood homicide rates. *Homicide Studies* (2009).

[25] HSIEH, C.-C., AND PUGH, M. D. Poverty, income inequality, and violent crime: a meta-analysis of recent aggregate data studies. *Criminal Justice Review 18*, 2 (1993), 182–202.

[26] JACOBS, J. *The death and life of great American cities*. Vintage, 1961.

[27] LAMBERT, D. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics 34*, 1 (1992), 1–14.

[28] MOHLER, G. O., SHORT, M. B., BRANTINGHAM, P. J., SCHOENBERG, F. P., AND TITA, G. E. Self-exciting point process modeling of crime. *Journal of the American Statistical Association* (2012).

[29] MORENOFF, J. D., AND SAMPSON, R. J. Violent crime and the spatial dynamics of neighborhood transition: Chicago, 1970–1990. *Social forces 76*, 1 (1997), 31–64.

[30] NAKAYA, T., AND YANO, K. Visualising crime clusters in a space-time cube: An exploratory data-analysis approach using space-time kernel density estimation and scan statistics. *Transactions in GIS 14*, 3 (2010), 223–239.

[31] OSGOOD, D. W. Poisson-based regression analysis of aggregate crime rates. *Journal of quantitative criminology 16*, 1 (2000), 21–43.

[32] PATTERSON, E. B. Poverty, income inequality, and community crime rates. *Criminology 29*, 4 (1991), 755–776.

[33] RATCLIFFE, J. H. A temporal constraint theory to explain opportunity-based spatial offending patterns. *Journal of Research in Crime and Delinquency 43*, 3 (2006), 261–291.

[34] SAHBAZ, O., AND HILLIER, B. The story of the crime: functional, temporal and spatial tendencies in street robbery. In *Proc of 6th International Space Syntax Symposium, Istanbul* (2007), pp. 4–14.

[35] SAMPSON, R. J., RAUDENBUSH, S. W., AND EARLS, F. Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science 277*, 5328 (1997), 918–924.

[36] SHORT, M. B., D'ORSOGNA, M. R., PASOUR, V. B., TITA, G. E., BRANTINGHAM, P. J., BERTOZZI, A. L., AND CHAYES, L. B. A statistical model of criminal behavior. *Mathematical Models and Methods in Applied Sciences 18*, supp01 (2008), 1249–1267.

[37] TOOLE, J. L., EAGLE, N., AND PLOTKIN, J. B. Spatiotemporal correlations in criminal offense records. *ACM Transactions on Intelligent Systems and Technology (TIST) 2*, 4 (2011), 38.

[38] TRAUNMUELLER, M., QUATTRONE, G., AND CAPRA, L. Mining mobile phone data to investigate urban crime theories at scale. In *Social Informatics*. Springer, 2014, pp. 396–411.

[39] TRIBUNE, C. A tale of 3 cities: La and nyc outpace chicago in curbing violence, 2015.

[40] WANG, T., RUDIN, C., WAGNER, D., AND SEVIERI, R. Learning to detect patterns of crime. In *Machine Learning and Knowledge Discovery in Databases*. Springer, 2013.

[41] WANG, X., GERBER, M. S., AND BROWN, D. E. Automatic crime prediction using events extracted from twitter posts. In *Social Computing, Behavioral-Cultural Modeling and Prediction*. Springer, 2012, pp. 231–238.

[42] WIKIPEDIA. Community areas in chicago — wikipedia, the free encyclopedia, 2015.

[43] WOLFE, M. K., AND MENNIS, J. Does vegetation encourage or suppress urban crime? evidence from philadelphia, pa. *Landscape and Urban Planning 108*, 2 (2012), 112–122.

[44] YUAN, J., ZHENG, Y., AND XIE, X. Discovering regions of different functions in a city using human mobility and pois. In *ACM SIGKDD* (2012), ACM, pp. 186–194.

[45] ZHENG, Y., CAPRA, L., WOLFSON, O., AND YANG, H. Urban computing: concepts, methodologies, and applications. *ACM TIST 5*, 3 (2014), 38.