Crime Rate Prediction with Region Risk and Movement Patterns

Shakila Khan Rumi  
*RMIT University*  
Melbourne, VIC  
shakilakhan.rumi@rmit.edu.au

Phillip Luong  
*RMIT University*  
Melbourne, VIC  
phillip.luong@rmit.edu.au

Flora D. Salim  
*RMIT University*  
Melbourne, VIC  
flora.salim@rmit.edu.au

Abstract—Location Based Social Network, Foursquare helps us to understand the human movement of a city. It provides data that characterises the volume of movements across regions and Places of Interests (POIs) to explore the crime dynamics of the city. To fully exploit human movement into crime analysis, we propose region risk factor which combines monthly aggregated crime and human movement of a region across different time intervals. Number of features are derived from this risk factor. Extensive experiments with real world data in multiple cities verify the effectiveness of the features.

I. INTRODUCTION

Safe and secure living place is one of the basic demands to every person in society. Therefore, it is important to find ways to control the crime rate. High crime rate hinders the economic development of a city. Understanding the root causes of what increase the likelihood of any particular crime event occurring at any time has great benefits for law enforcement to prevent crime events. According to the criminology theory, the surrounding environment including neighbourhood regions and movement of people play crucial role in crime event prediction. The widespread use of location based social networks such as *Foursquare* open the door of opportunities to analyse crime event occurrences in a timely manner. In this paper, we study the crime rate prediction with the help of urban mobility data.

Recently, there has been research exploring the link between crime events and urban dynamics using FourSquare data [1]. However, this study focused on a region’s check-in information to predict the crime events. There was no focus on hyperlinking human movement between two regions. In [2] the authors considered the movement between regions using taxi flow data for crime inference, however, did not consider the variation of movement in different periods of time. For example, the people who move to work place from home will move to the opposite direction in the afternoon. In this paper, we analyse the crime inference with the movement of people at different period of day. In Figure 1 we can observe the difference in the amount of people moving in morning and afternoon for New York City. The changes of human movements are highlighted with different colour. To fully exploit the human dynamics in crime inference problem, we further introduce region risk which associates crime and people movement in that region in a certain time interval, which is denoted as Region Risk. The hypothesis is that if the people are coming to a place from a high crime risk area, such movement implies a high crime risk in the arrival area too. We derive numerous features with assistance of this region risk. The significance of this features alongside other features is verified for three different cities including Chicago, Los Angeles and New York City across different time interval of a day. The contributions of this paper are summarized as follows:

- This is the first work that predicts crime rate based on the dynamic features that associate region risk and movement patterns between regions.
- Different features associated with region risk and the human mobility in different periods of time are crafted.
- The work verifies the effectiveness of different features in crime inference problem using correlation and regression analysis. Real-world crime data and FourSquare movement data are used for evaluation. The experimental results show that the region risk features are highly correlated with crime count of a region.

II. RELATED WORK

Many data mining research have been developed recently to verify the impact of human mobility into crime study. In [3], the authors extracted human behavior from mobile network activities and demographic features of people connected to the network over different regions and times. The study showed that the combination of mobile activity data and demographic data can predict crime event in a region with better accuracy.

Fig. 1: Few check-in movements in New York in different time intervals for "2018-03"
Ambient population is measured through FourSquare check-in data and is used to understand the long-term crime event occurrences [4], [5]. In [1] the authors proposed several dynamic features using FourSquare data to measure the social diversity of a region and predict short-term crime event occurrences. To understand crime event occurrences, it is important to explore the correlation between places. The mobility data represented by taxi flows and Points of Interest (POI) can lift the performance of crime rate inference [2]. Here, the authors’ hypothesis is that the social interaction between two places can be inferred through taxi trips and the crime rate propagate based on the connection between places. In [6], the authors proposed crime-specific dynamic features by analyzing individual risk factor of the users and extracted multiple features based on the risk analysis.

However, none of this work correlate the large scale human movement in different time period of day with crime in a region. Our work attempts to fill this gap.

III. DATASET DESCRIPTION

The datasets are collected for three different cities in USA including Chicago, Los Angeles and New York City. We collect different types of data including check-ins and crime events for each city. We segment each city into 400×400 grid.

A. POI and Check-in Data

The POI and check-in information is collected from FourSquare check-in. The dataset is provided as part of the Future Cities Challenge at Netmob 1. The check-in information describes an aggregated count of all movements from one venue to another separated by month and five time intervals including Morning, Midday, Afternoon, Night and Overnight. In the collected data, we focus on the three cities mentioned above for the year of 2018. The aggregated number of venues and different venue movements are summarised in Table I.

B. Crime Data

We collect crime event records that lie in 2018 for Chicago, Los Angeles and New York from Open Data Portal of the respective city councils 2,3,4. Each dataset consists of the longitude, latitude, and the time and date of crime event occurrences. The total number of crime event occurrences are 263,515, 226,498 and 452,958 for Chicago, Los Angeles and New York respectively.

IV. FEATURE DESCRIPTION

This section presents the detailed description of the features. We consider each city as a directed graph, $G = (V, E)$. Each grid, $v \in V$, represents a node in the graph. $E$ represents the set of edges between two nodes, which are weighted and directed. If there is a directed edge between two nodes, $i$ and $j$ (denoted as $(i, j)$), in a time interval, $t$, there is at least one check-in from $i$ to $j$ over $t$. The cost (or weight) of each node is the aggregated number of check-ins in a month. For each node $v \in V$ in a time interval, $t$, the following nodal and edge features are calculated.

A. Nodal Features

Nodal features describe the characteristics of the focal grid only.

1) Historical Features: To retain the historical knowledge about crime event occurrence, we calculate the following feature:

**Crime Event History:** We measure the number of crime events in node, $v$, during interval, $t$, in the past month. This is represented as:

$$NH(v, t) = \sum_{j \in m} Cr_j(v, t).$$

Here, $Cr_j(v, t)$ denotes the number of crime events that happened on $j$-th day in node, $v$, during time interval $t$. The variable, $m$, represents the day in the past month.

2) POI Features: The regional information of a node is described using the following features.

**POI Density:** For each node, $v \in V$ the POI density is calculated as follows:

$$NP(v) = \frac{N(v)}{N(V)}.$$ (2)

The total POI of the city is represented as $N(V)$. $N(v)$ denotes the number of POI in focal node, $v$.

**Venue Category Distribution:** Each type of venue has different impact on crime. Hence, it is important to extract the distribution of venue types in node, $v$. It is calculated as follows:

$$ND_i(v) = \frac{N_i(v)}{N(v)}.$$ (3)

Here, $N_i(v)$ represents the number of $i$-th category venue in node, $v$.

**Venue Diversity:** Shannon’s entropy [7] measurement is applied to determine the diversity of venue types, $P$, in node, $v$. Thus, Venue Diversity is modelled as:

$$NE(v) = -\sum_{i \in P} \frac{N_i(v)}{N(v)} \log_2 \left( \frac{N_i(v)}{N(v)} \right).$$ (4)

3) Movement Features: The human dynamics of region in time interval is represented using the following features:
TABLE II: Feature Correlation Analysis for the New York and Chicago

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Chicago</th>
<th>New York</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Morning</td>
<td>Midday</td>
</tr>
<tr>
<td>Check-in Entropy</td>
<td>0.365</td>
<td>0.277</td>
</tr>
<tr>
<td>Risk Count</td>
<td>0.210</td>
<td>0.088</td>
</tr>
<tr>
<td>Risk Ratio</td>
<td>0.113</td>
<td>-0.010</td>
</tr>
<tr>
<td>Risk</td>
<td>0.154</td>
<td>0.095</td>
</tr>
<tr>
<td>Number of Crimes</td>
<td>0.001</td>
<td>0.039</td>
</tr>
<tr>
<td>Historical Density</td>
<td>0.637</td>
<td>0.867</td>
</tr>
<tr>
<td>Food</td>
<td>0.175</td>
<td>0.116</td>
</tr>
<tr>
<td>Total POI</td>
<td>0.517</td>
<td>0.651</td>
</tr>
<tr>
<td>Venue Diversity</td>
<td>0.318</td>
<td>0.287</td>
</tr>
</tbody>
</table>

**Incoming Movement:** The density of incoming movement into node, \( v \), in time interval, \( t \), is measured here. If a check-in is performed from other nodes to node, \( v \), then it is considered as incoming movement. This we obtain the density of incoming movement as,

\[
NI(v, t) = \frac{C_i(v, t)}{C(v, t)},
\]

where \( C(v, t) \) and \( C_i(v, t) \) denote total check-in and incoming check-ins respectively, performed at node, \( v \), during time interval, \( t \), in each month.

**Outgoing Movement:** If a check-in is performed from node, \( v \), to other nodes then it is considered as outgoing movement. The density of outgoing movement is represented as,

\[
NC(v, t) = \frac{|C_o(v, t)|}{|C(v, t)|},
\]

where \( C_o(v, t) \) represents outgoing check-ins performed in node, \( v \), during time interval, \( t \), in each month.

**Stationary Movement:** When the origin and destination of a check-in is the same node, it is denoted as stationary movement. The density of stationary movement for \( C_s(v, t) \) stationary check-ins in node, \( v \), in time interval, \( t \), is as follows:

\[
NS(v, t) = \frac{|C_s(v, t)|}{|C(v, t)|}.
\]

**Diversity of Movement:** The heterogeneity of movement type is measured here:

\[
NM(v, t) = \sum_{i \in M} \frac{|C_i(v, t)|}{|C(v, t)|} \log_2 \left( \frac{|C_i(v, t)|}{|C(v, t)|} \right).
\]

The set, \( M \), consists of three movement types: incoming, outgoing and stationary.

**B. Edge Features**

Edge features determine how crime rate of a region is influenced by its connected region. Two types of feature have been crafted. One is based on the crime rate in the neighbourhood region and another one is based on the risk analysis of a region based on movement data.

1) **Neighbourhood Crime:** The number of crime events for each adjacent node of focal node, \( v \), is computed in each time interval. It reflects the situation of the surroundings.

2) **Region Risk:** To compute region risk, we analyse the risk associated with each node to support the following intuition. If the incoming check-ins of a node in a time interval are from high risk area, it imposes high risk of crime event occurrence in that node in that time interval. The region risk of node, \( v \), for time interval, \( t \), is as follows:

\[
RR(v, t) = \frac{|Cr(v, t)|}{|C(v, t)|},
\]

where \( Cr(v, t) \) denotes the crime events that happened in node, \( v \), in time interval, \( t \). Several edge features are crafted which consider this region risk, \( RR(v, t) \).

**Risk Distribution:** The risk distribution consists of the mean and median of the region risk associates with the regions \( r \in R(v, t) \) from where people are moving to focal node, \( v \), in time interval \( t \).

**Risk Count:** Risk count in node, \( v \), in time interval, \( t \), determines the number of regions with high risk than average from where incoming movement occur in node, \( v \), in that time interval. The risk count is denoted by,

\[
RC(v, t) = |\{ r : r \in R(v, t) \text{ and } RR(r) > \frac{1}{|R|} \sum_{n \in V} RR(n, t) \}|.
\]

Here, \( R(v, t) \) denotes the regions which are origin of check ins to node \( v \), and \( |R| \) is the total number of regions.

**Risk Ratio:** Risk count determines the absolute number of regions with high risk. We normalise this feature based on total regions with incoming movement. The risk ratio is modelled using,

\[
RT(v, t) = \frac{RC(v, t)}{|R(v, t)|}.
\]

**Region Risk:** This feature represents absolute value of region risk that is associated to the focal node, \( v \), in time interval, \( t \), \( RR(v, t) \).

**C. Feature Correlation Analysis**

We conduct Pearson correlation analysis to see how the proposed features are individually correlated with the monthly crime count of a region in a certain time interval of day. The correlation value between the features and crime count is illustrated in Table II for Chicago and New York City across different periods of a day. Here, we note only the features which have high correlation with crime count due to space limit. We observe that crime count is highly correlated with
Crime Event History in both cities in many intervals. According to the Near Repeat theory [8], crime tends to happen in the vicinity of past crime. The features derived from region risk analysis also have good correlation with crime count for both cities especially overnight. Such positive correlation proves the importance of such features in crime count. Surprisingly, many POI densities are negatively correlated with crime count. Although venue diversity has good correlation with crime which verifies that mixed land use has good impact on crime.

V. MODEL DESCRIPTION

The main purpose of this paper is to show the effectiveness of the features derived from human movement in crime inference. To serve this purpose, we apply Linear Regression (LR) to count the features of a node in a time interval. We adopt LR as inference model because it is a simple and most straightforward regression technique.

In this study, we only use the region and time interval where check-in movement exists to train the model. For example, in a month, \( m \), for a region, \( v \), during time interval, \( t \), if there is any type of movement, it generates the training and test data based on the features described in previous section. Finally, LR model is trained with different feature settings to know the effectiveness of the feature. The performance metric compares which feature sets are significant for the crime inference.

VI. EXPERIMENT

A. Settings

The dataset used in this experiment is introduced in Section 3. Each day is segmented in five intervals and for each interval the are aggregated in monthly level. To prevent extreme sparsity situations, only check-in movement data with 10 or more unique movements in a month. The aim of the experiment is to examine the effectiveness of the proposed features in crime inference model for a month in a certain period of day. We partition the data about 75% as training set and the rest 25% as test set for all three cities. Particularly, the data lies between January 2018 and September 2018 (inclusive) is used as training data and the data between October 2018 to December 2018 (inclusive) as test data. If new regions are found with check-in movement greater than, or equal to 10 in test data for a time interval, the risk value for that region is set 0.

Two performance metrics, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) have been applied to verify the effectiveness of the crime inference model.

B. Performance Study

We evaluate the performance of the proposed features for monthly crime count. We build linear regression based model and apply ‘leave-one-out’ to measure the performance of a set of features. First, we train a model with all the features. Next, we train another model without a set of features to identify the effectiveness of that set of features. If the MAE and RMSE increase for second model compare to the first one, that set of features is considered important in monthly crime count.

C. Feature Importance

We measure the importance of each group of feature using regression method. If the MAE and RMSE value is higher without a set of features, it verifies the importance of that feature set. The importance of each feature set is illustrated in Figure III for New York City. We observe that historical features are the dominating set of features among all of the features. The edge features based on Region Risk analysis also show the effectiveness across different time interval. The same analysis has been done for the other two cities.

<table>
<thead>
<tr>
<th>Time</th>
<th>Error 1</th>
<th>Error 2</th>
<th>Error 3</th>
<th>Error 4</th>
<th>Error 5</th>
<th>Error 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morning</td>
<td>MAE</td>
<td>7.76</td>
<td>12.25</td>
<td>7.69</td>
<td>7.78</td>
<td>7.86</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>11.1</td>
<td>17.01</td>
<td>11.14</td>
<td>10.53</td>
<td>11.32</td>
</tr>
<tr>
<td>Midday</td>
<td>MAE</td>
<td>12.38</td>
<td>23.08</td>
<td>12.2</td>
<td>12.24</td>
<td>12.35</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>16.36</td>
<td>32.28</td>
<td>16.28</td>
<td>15.22</td>
<td>16.36</td>
</tr>
<tr>
<td>A’noon</td>
<td>MAE</td>
<td>21.29</td>
<td>23.45</td>
<td>21.3</td>
<td>21.88</td>
<td>21.29</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>22.55</td>
<td>42.61</td>
<td>42.61</td>
<td>44.11</td>
<td>42.54</td>
</tr>
<tr>
<td>Night</td>
<td>MAE</td>
<td>12.62</td>
<td>23.71</td>
<td>12.44</td>
<td>13.47</td>
<td>12.7</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>16.56</td>
<td>31.39</td>
<td>16.51</td>
<td>17.82</td>
<td>16.66</td>
</tr>
<tr>
<td>O’night</td>
<td>MAE</td>
<td>13.39</td>
<td>17.34</td>
<td>14.06</td>
<td>14.78</td>
<td>13.35</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>17.5</td>
<td>21.3</td>
<td>18.09</td>
<td>18.55</td>
<td>17.46</td>
</tr>
</tbody>
</table>

*List of experiments:
1. All Features Present, 2. Historical Features Omitted
3. Nodal Check-In Omitted, 4. Geographic (POI) Features Omitted,
5. Neighbourhood Features Omitted, 6. Region Risk Features Omitted

VII. CONCLUSION

This work provides new perspective to understand crime dynamics with the help of human mobility. It captures the relationship between the monthly aggregated crime data and the movement of people in a region across different time period of a day. The experiments verify that a group of people from high crime risk area increase the crime risk of their destination.

REFERENCES