Uncovering the correlation of regions, time and criminal incidents via representative learning

Yanfang Hou
Leiden University

Manlu He
Leiden University

ABSTRACT

As crime incidents that occur at certain locations and times might share some common characteristics, spatiotemporal patterns can be drawn from the crime data, which consists of three main units that describe the spatial, temporal and textual information of the crime. For uncovering the correlation of the units, we leverage the representative learning method via graph embedding. We map spatial, temporal, and textual units denoted by keywords into a joint vector space to encode the co-occurrence relationship and neighborhood relationship of the units and learn embeddings. After representing each unit as an embedding, we are able to make predictions for each unit based on similarity among vector representations. To evaluate the performance of the graph embedding, we compare it with Singular Value Decomposition (SVD) and TF-IDF weighting. The results on crime data collected by the Chicago Police Department show that the graph embedding works properly and is better than SVD, but performs slightly worse than TF-IDF.

KEYWORDS

crime data, representative learning, graph embeddings

ACM Reference Format:
Yanfang Hou and Manlu He. 2020. Uncovering the correlation of regions, time and criminal incidents via representative learning. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Urban crimes happen everywhere in the world and large volumes of past crime data are recorded by the police department. The analysis of crime data was time consuming as this work used to be done manually by reading criminal reports and making comparisons among cases but now the analysis can be done in a more automated way thanks to digital revolution [10].

Crime data mainly consists of three units: timestamp, location, and texts which provide primary descriptions of the crime. A simplified example of a crime record is given in Table 1. Criminal events are unevenly distributed within the city and it is possible to detect patterns based on past crime data with a multidimensional view, because empirical evidences show that the crime concentration in cities can be found in different dimensions regarding their contexts and features (e.g., time, location, texts) [5]. The Figure 1 and Figure 2 based on Chicago crime dataset [2] illustrate the spatial and temporal correlation where spatio-temporal hotspots can be detected and some spatio-temporal neighborhoods tend to have similar distributions or similar values. The goal of our project is to uncover the correlation of regions, time and criminal incidents using the graph embedding strategy which is discussed in more detail in section 3.

Table 1: An example of a crime record

Community Area	32
Date	3/28/2017 2:00:00 PM
Text	Deceptive practice, financial identity theft $300 and under, hospital building/grounds

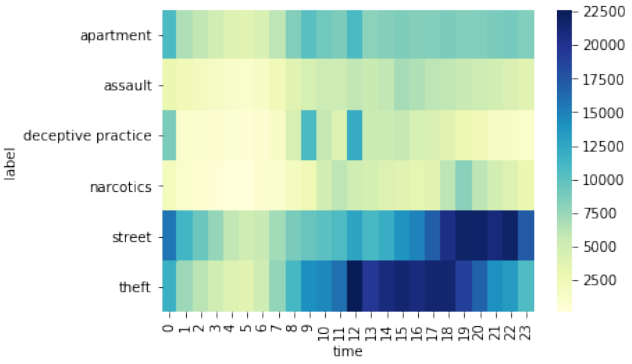

Figure 1: The occurrence frequency of six main candidate keywords which describe crime incidents in Chicago from 2015 to 2019.

Since urban crimes are threats to the daily life and they hinder the city's development, better understanding the correlation within crime data might be useful in crime prevention. With more crime patterns revealed, police department can allocate resources more efficiently in response to different regions and time, such as allocating more resources to areas which have higher crime rates at certain times; citizens can be provided with useful crime information to improve personal safety, such as being cautious when entering some neighborhoods at certain times.

Our model is able to make prediction for a unit given the other two units. Specifically: Given location and time, the output of the model will be all the candidate keywords of the crime along with

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

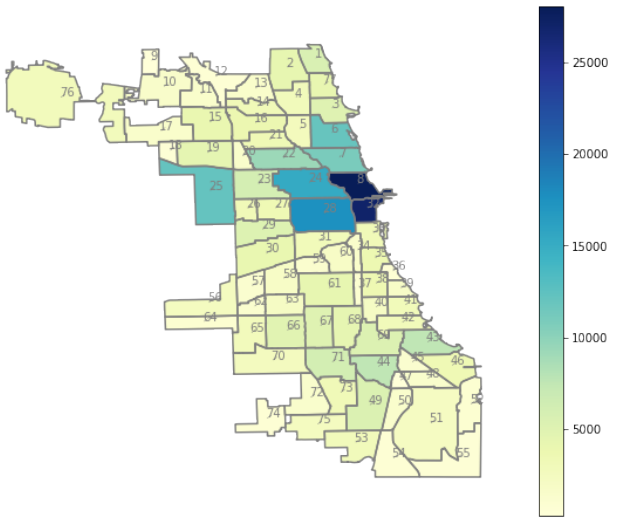

Figure 2: The distribution of the theft in Chicago from 2015 to 2019. The color bar denotes the count of incidents. Each community area is labeled by a number.

the probability; Given keywords and time, the model will predict all the possible locations and their probability; Given location and keywords, crime start time and their probability will be the output of the model.

The major challenges of our project come from two sides: (1) The basic three units of the crime data: timestamp, location, and texts, have different data types. Especially, the texts do not have a natural vector representation. However, in order to capture the correlation among these three units and preserve the internal relationship within each unit, we need to project different units into a common space for discovering spatio-temporal patterns of the crime data. (2) The probability of a crime occurrence is related to many factors [6]. Therefore, using timestamp, location and basic crime information might not be able to capture the complex correlations within huge amounts of past crime data. Given this, another challenge is to choose additional factors to help modeling the correlation.

To solve the first challenge, we adopt the graph embedding in our project. The strategy maps spatial, temporal, and textual units denoted by keywords into a joint vector space to encode the co-occurrence relationship and neighborhood relationship of the units and learn embeddings [11]. As for the second challenge, we leverage the location description and the description of the crime as contributing factors. By combining these two types of additional information with crime type as the text unit, the crime data used for analyzing are provided with some contextual information which better describes each crime incident.

The primary contributions of our project are as follows: (1) We leverage the graph embedding strategy to map the time, location, and text units of the crime data into same space and encode the correlations among the units for discovering crime patterns. (2) We compare the graph embedding with Singular Value Decomposition (SVD) embedding and TF-IDF embedding by experiments to investigate the performance of the graph embedding. Overall, the results of the experiment show that the graph embedding works properly and is able to outperform SVD, but produces slightly worse performance than tf-idf.

The rest of this paper is organized as follows. In Section 2, we discuss related work that relates to crime prediction and graph embedding. Section 3 describes the problem that needs to be addressed in this work and briefly explains the graph embedding methods. Section 4 describes the experimental setup, including the dataset, the baselines and metrics we use. In section 5, initial results of experiments on graph embedding, SVD and TF-IDF are provided to evaluate performance. Finally, we discuss the work and conclude the project.

2 RELATED WORK

Crime prediction. As mentioned in section 1, urban crimes have spatio-temporal patterns. For example, crimes are dense in some regions and sparse in others. Also, crime temporal pattern can be structured in various intervals like weeks, hours and months [12]. Given this complex property of the crime data, various methods are proposed to help detect the crimes patterns. Apart from location and temporal information, new types of data units are added as to help crime prediction. [3] incorporates the geo-tagged twitter information into a standard crime prediction model based on kernel density estimation (KDE). The method improves prediction performance for most of the crime types but has a major drawback that semantics of tweets are not taken into consideration. [9] learns vector representations for regions using the taxi flow data, which is used to supplement the geographical relations for crime rate forecast.[1] leverages twitter sentiment and weather as additional predictive factors for criminal incident prediction and achieves sound results, but it does not preserve the relationships among specific kinds of crime, twitter sentiment and weather. In our case, we add location description and the crime description as new factors to supply data.

Graph embedding. Many studies represent the data as embedding vectors which can capture the significant features of data. [8] proposes a network embedding model called the "Line", which is suitable for arbitrary types of information networks: directed or undirected, weighted or unweighted graphs. The model optimizes an objective which preserves both the first-order and second-order proximities. The first-order proximity refers to the observed links between nodes, while the second-order proximity means that nodes with shared neighbors being likely to be similar. Based on line embedding model, [11] studies the correlations between regions, periods and activities with massive geo-tagged social media data. The key idea is to map all spatial, temporal, and textual units into the same space and learns vector representations for different units such that the observed relations between variables can be preserved. The data structure of crime incidents is similar to GTSM data, since crime data also contains spatial, temporal and textual information. Therefore, in this project, we try to apply the graph embedding method presented in [11] to crime incidents data for discovering crime patterns.

3 METHODS

In this section, we describe the embedding method that encodes all spatial, temporal and text units into a joint low-dimensional space. High quality embedding should be able to detect two important relationships: the co-occurrence and neighborhood relationships. The co-occurrence relationship exists between two units when they co-occur in the same crime incident record. The neighborhood relationship are used to represent spatio-temporal proximities. Specifically, near spatial or temporal units are considered to be correlated [11].

3.1 Problem Definition

We build a heterogeneous graph illustrated in Figure 3 to encode the relations among the three units. The graph is composed of the six bipartite graphs: region-word, time-word, word-word, region-time, region-region and time-time graphs, where the vertices of the three units are shared by the six bipartite graphs.

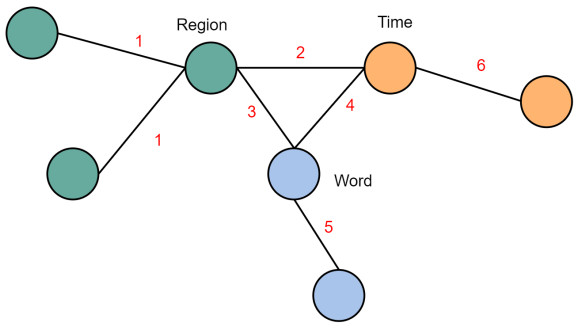

Figure 3: The heterogeneous graph to encode the relations between the three units. The edge 1 represents neighborhood relationships among regions; The edge 2, edge 3 and edge 4 denote co-occurrence relationships between region and time, region and word, time and word respectively; The edge 5 represents co-occurrence relationships between words; The edge 6 represents neighborhood relationships among timestamps.

Region-word graph. Region-word graph, denoted as $G_{rw} = (R \bigcup W, E_{rw})$, is a bipartite graph where R is a set of regions where W is a vocabulary of words and E_{rw} is the set of edges between regions and words. The weigh w_ij of edge between region r_i and word w_j is defined as the frequency of criminal incidents with w_j occurring in region r_i.

Time-word graph. Time-word graph, denoted as $G_{hw} = (H \bigcup W, E_{hw})$, is a bipartite graph where H is a set of timestamps and E_{hw} is the set of edges between time and words. The weigh w_ij of edge between h_i and w_j is defined as the frequency of criminal incidents with w_j happening at time slot h_i.

Word-word graph. Word-word graph, denoted as $G_{ww} = (W, E_{ww})$, captures the word co-occurrence relationships. E_{ww} is the set of edges between words. The weight w_ij of edge between node w_i and w_j is the frequency of the two words appearing in the same criminal incidents.

Region-time graph. Region-time graph, denoted as $G_{rh} = (R \bigcup H, E_{rh})$, is a bipartite graph where E_{hw} is the set of edges between regions and time. The weigh w_{ij} of edge between node r_i and h_j is defined as the number of criminal incidents that occur in region r_i at time slot h_j.

Region-region graph. Region-region graph, denoted as $G_{rr} = (R, E_{rr})$, captures the spatial neighborhood relationship. We use kernel densities to quantity this spatial proximity, given by

$$w(x,y) = \begin{cases} exp(-||x-y||^2/(2\sigma_r^2))/(2\pi\sigma_r^2) & if\ ||x-y|| \le \sigma_r \\ 0 & otherwise \end{cases} \tag{1}$$

where x and y are the centroids of the two regions, σ_r is a kernel bandwidth for spatial continuity. The two regions whose distance is within σ_r are considered to be neighboring. The weight w_{ij} of edge between r_i and r_j is their kernel density.

Time-time network. Time-time graph, denoted as $G_{rr} = (R, E_{rr})$, captures the temporal neighborhood relationship. We use kernel densities to quantity this spatial proximity, given by

$$w(x,y) = \begin{cases} exp(-||x-y||^2/(2\sigma_h^2))/(2\pi\sigma_h^2) & if\ ||x-y|| \le \sigma_h \\ 0 & otherwise \end{cases} \tag{2}$$

where x and y refer to the timestamps, σ_h is a kernel bandwidth for temporal continuity. The two timestamps whose distance is within σ_h are considered to be neighboring. The weight w_{ij} of edge between h_i and h_j is their kernel density.

3.2 Bipartite graph embedding

In this part, we adapt the LINE model for embedding bipartite graph [8]. Given a bipartite graph $G = (V_A \bigcup V_B, E$, we model the likelihood of generating node v_j in set V_B given node v_i in set V_A as

$$p(v_j|v_i) = exp(u_j'^T \cdot u_i) / \sum_{j' \in B} exp(u_{j'}'^T \cdot u_i) \tag{3}$$

where u_i and u_j are the embedding vectors of vertex v_i and v_j respectively.

On the other hand, the empirical distribution between node v_i and v_j is given by

$$\hat{p}(v_j|v_i) = w_{ij}/d_i \tag{4}$$

where d_i is the total out-degree of v_i, i.e. $d_i = \sum_{k' \in N(i)} w_{ik}$, where $N(i)$ is the set of out-neighbors of v_i.

Our objective is to make the joint distribution be close to the empirical distribution, which can be achieved by minimizing the following objective

$$O = \sum_{i \in A} d_i KL(\hat{p}(\cdot|v_i), p(\cdot|v_i)) \tag{5}$$

where KL is the KL divergence and we use d_i to weight the importance of the vertex i.

Two strategies are used to optimize the model. First, to calculate the conditional probability $P(\cdot|v_i)$ in equation 3, it requires the summation over the entire set of vertices, which is computationally expensive. We adopt negative sampling presented in [4] to address

this problem. Specifically, we randomly sample multiple negative edges from some noise distribution for each edge and change the objective equation 5 by replacing $p(\cdot|v_i)$ with the equation 6, where $\sigma(x) = 1/(1 + exp(-x))$ is the sigmoid function, $P_n(v)$ is the noise distribution and K is the number of negative samples. Second, as described in [8], the weights of edges with a high variance are more likely to result in gradient divergence problem. For this issue, we randomly sample an edge with the sampling probabilities proportional to the original edge weights.

$$log\sigma(u_j'^T \cdot u_i) + \sum_{i=1}^{K} E_{v_n \sim P_n(v)}[log\sigma(-u_n'^T \cdot u_i)] \tag{6}$$

3.3 Joint graph embedding

To learn embeddings of heterogeneous graph, we combine the above six bipartitite graphs by minimizing the following objective function:

$$O = O_{rw} + O_{hw} + O_{rh} + O_{ww} + O_{rr} + O_{hh} \tag{7}$$

Since the whole graph is heterogeneous, the edge weights between different types of nodes are incomparable to each other. Therefore, we alternatively sample from the six sets of edges and update the corresponding embeddings [7].

4 EXPERIMENT

In this section, we first describe the settings of experiments and then demonstrate the experimental results.

4.1 Experiment setup

4.1.1 Dataset. The dataset [2] is collected from Chicago Police Department from 2001 to present. In our experiments, we use 175,086 crime records from 2015 to 2019. Each crime record mainly includes the following attributes: case number, data, community area, primary type, secondary description and location description. In our experiment, we extract hour from "date" as timestamp, so there are 24 timestamps in total. The location is defined as the community area where criminal incidents occur. For text unit, each criminal incident is characterized by a combination of three factors: "primary type", "secondary description" and "location description".

4.1.2 Parameters setting. The major parameters settings of graph embedding are: (1) spatial kernel bandwidth $\sigma_r = 0.05$; (2) temporal kernel bandwidth $\sigma_h = 4$; (3) the latent embedding dimension $D = 200$; (4) number of negative edges $K = 5$; (5) number of iterations $T = 10000$. We use RMSprop optimizer to train the GE model at a learning rate of 0.001.

4.1.3 Comparative approches. We apply the following approaches to the dataset for comparison(more baselines in progress):

TF-IDF. TF-IDF model is a conventional method for vector representation. We first construct the three co-occurrence matrix: region-words, time-words and words-words. Then as [11] suggested, the tf-idf model is used to weight each unit in the vector by treating each row as document and each column as a word. Finally we can obtain the vector representations for each unit. The intuition of this method is that the two units that co-occur in many records but rarely appear in the same record as other units are considered to be more related.

SVD. Singular value decomposition is a common approach for dimension reduction. We apply SVD to decompose the co-occurrence matrix between each pair of location, time and words. Then we represent each unit by a k-dimensional vector by averaging the vectors for each unit. In our experiment, the number of dimensions is set to be 24.

4.1.4 Metrics. We evaluate the model performance by using the Mean reciprocal rank (MRR) and Mean Average Precision (MAP). Consider keyword prediction as an example. We first mix the true keywords with the randomly chosen negative keywords. Then we compute the average cosine similarity of each candidate keyword to the observed location and timestamp, and rank them in the descending order of the similarity. Finally we use the following metrics to measure the model performance. In our setting, we have 20 candidate examples for keywords prediction for each test record, 10 candidate examples for location and time prediction.

MRR. The metric is given by equation 8, where $rank_i$ refers to the position of the first correct answer for the ith test record. A better model is able to rank the ground truth location to top positions, thus obtaining the high MRR value.

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \tag{8}$$

MAP. This metric is defined in equation 9, where Q is the number of test records. For each test record, the average precision is the mean of the precision scores after each correct answer is retrieved. Then the MAP is the mean of the average precision scores for all test questions.

$$MAP = \frac{1}{|Q|} \sum_{i=1}^{|Q|} AveP(q) \tag{9}$$

4.2 Experimental results

Table 2 shows the experimental results. Overall, the graph embedding (GE) model performs better than SVD but slightly worse than tf-idf. The performance of the three models for the three kinds of prediction tasks are discussed as follows.

For keywords prediction, both of GE and tf-idf achieve better performance than SVD. There are mainly three reasons. First, some types of crimes occur much more frequently than others. Table 3 shows that theft, battery and criminal damage account for 50% of criminal incidents, while many crime types have a much lower frequency than them. Due to the imbalance of keywords distribution, the method based on frequency like tf-idf can reach a high accuracy. The ability of the tf-idf model to predict minority keywords requires further research. Second, compared to the tf-idf model with raw co-occurrence information, SVD inevitably loses some co-occurrence information when projecting data into a low-dimension space, resulting in lower accuracy. Third, the GE model strides a relatively good balance between dimension reduction and relationship preservation compared to SVD. This mainly because SVD utilizes co-occurrence information between pairs of units, while

GE incorporates the co-occurrence and neighborhood information of all units into the model.

Compared with keywords prediction, the predictive accuracy for location of all methods has dropped significantly and tf-idf performs slightly better than others. First, the result of tf-idf model is expected since the region difference for the given timestamp and keywords is not as large as the keywords difference in certain region at a given timestamp. Second, the GE model fails to improve the results, which indicates that the vector representations learned from GE model are not able to explain the complicated relationship between crime, location and time. We guess there are two main reasons: (1) Due to the imbalance of criminal data, the GE model would repeatedly learn embeddings from some edges with high frequency. Consequently, a large number of edges would not be sufficiently learned because of lower frequency. (2) There is a large difference in graph size among the six bipartite graphs, as shown in table 4. This will make it difficult to choose an appropriate number of iterations. Specifically, small graphs need fewer iterations to converge and multiple iterations might lead to overfitting. Conversely, learning from large graphs requires a large number of iterations and fewer iterations will result in edge undersampling. The above two points might make it challenging for GE model to detect all correlations among the three units.

Time prediction performs worst among the three subtasks. Apart from the reasons mentioned in location prediction part, complicated temporal cyclic effect will also increase the difficulty of prediction. Criminal incidents may periodically happen at different time intervals, like hourly, weekly or monthly. Currently our GE model can only detect temporal pattern with a period of 1 hour, which might be insufficient to predict occurring time of criminal incidents accurately.

Table 2: Experimental results on crime dataset. For location and time prediction, there is only one correct answer for each test record, so the MRR value is equal to MAP value.

	Text		Region		Time	
Method	MRR	MAP	MRR	MAP	MRR	MAP
TF-IDF	0.865	0.703	0.446	0.446	0.358	0.358
SVD	0.550	0.420	0.390	0.390	0.325	0.325
Graph Embedding	0.887	0.713	0.405	0.405	0.307	0.307

5 DISCUSSION

In this work, we study the problem of using the criminal incidents dataset for three prediction tasks: crime features prediction, crime location prediction and crime time prediction. We build a graph to encode the co-occurrence and neighborhood relationships among the three units and then learn the embedding vector to preserve such correlations. Additionally, we compare its performance with tf-idf and SVD model. The experiment shows that the overall performance of graph embedding is better than SVD but slightly worse than tf-idf. The imbalance of crime data and the big difference in in size of bipartite graphs might make it challenging for GE model to sufficiently learn correlations between unis.

Table 3: The occurrence frequency of the five frequent crime types and the five least frequent crime types

Primary type	Proportion(%)
theft	0.2342
battery	0.1881
criminal damage	0.1085
assault	0.0733
deceptive practice	0.0660
kidnapping	0.0007
intimidation	0.0005
concealed carry license violation	0.0004
obscenity	0.0002
public indecency	4.5119e-05

Table 4: Number of edges for the six bipartite graphs

Graph type	Number of edges
Region-word	43374
Region-time	3696
Time-word	17082
Word-word	22772
Region-region	558
Time-time	192

However,compared with tf-idf model, graph embedding is capable of encoding data in a lower space with just minor decrease of performance. This can help to save a lot of space, which will become very advantageous in large dataset. Meanwhile, these low-dimensional feature vectors are much easier to be applied in other methods, such as regression and neural networks. Besides, the graph embedding method can be straightforwardly extended to incorporate other influential factors.

For future work, we may try the following strategies to improve the model:

- In our experiment, we combine primary description, secondary description and location description as the keywords to characterize each criminal incident. Thus the bipartite graphs involving keywords usually have more nodes and edges, which cause the imbalance in graph size between different graphs. One way to address this issue is to regard the three descriptions as individual units. In this case we need to construct more bipartite graphs, but this will narrow the gap of graph size for different bipartite graphs.
- Sometimes multiple units have overlapping features, so merging those multiple minority units would be helpful to reduce imbalance of crime data. However, this task requires professional knowledge of criminal field.
- In order to let the GE model have a sufficient learning, we might slightly adjust the edge sampling strategy. For example, we might decrease the sampling probability of edges with high frequency and increase the sampling probability of edges with low frequency.
- Parameters settings have an important effect on GE model performance. In the future, we would like to try different

parameter settings to investigate their influence on model performance.

REFERENCES

[1] Xinyu Chen, Youngwoon Cho, and Suk Jang. 2015. Crime prediction using Twitter sentiment and weather. *2015 Systems and Information Engineering Design Symposium, SIEDS 2015* (06 2015), 63–68. https://doi.org/10.1109/SIEDS.2015.7117012

[2] Chicago Police Department. 2020. *Crimes - 2001 to present.* https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

[3] Matthew S Gerber. 2014. Predicting crime using Twitter and kernel density estimation. *Decision Support Systems* 61 (2014), 115–125.

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[5] Marcos Oliveira, Carmelo J. A. Bastos-Filho, and Ronaldo Menezes. 2017. The scaling of crime concentration in cities. *PLoS ONE* 12 (2017).

[6] Shakila Rumi, Ke Deng, and Flora Salim. 2018. Crime event prediction with dynamic features. *EPJ Data Science* 7 (12 2018). https://doi.org/10.1140/epjds/s13688-018-0171-7

[7] Jian Tang, Meng Qu, and Qiaozhu Mei. 2015. Pte: Predictive text embedding through large-scale heterogeneous text networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1165–1174.

[8] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *Proceedings of the 24th international conference on world wide web*. 1067–1077.

[9] Hongjian Wang and Zhenhui Li. 2017. Region representation learning via mobility flow. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 237–246.

[10] Tong Wang, Cynthia Rudin, Daniel Wagner, and Rich Sevieri. 2013. Learning to Detect Patterns of Crime, Vol. 8190. https://doi.org/10.1007/978-3-642-40994-3_33

[11] Chao Zhang, Keyang Zhang, Quan Yuan, Haoruo Peng, Yu Zheng, Tim Hanratty, Shaowen Wang, and Jiawei Han. 2017. Regions, periods, activities: Uncovering urban dynamics via cross-modal representation learning. In *Proceedings of the 26th International Conference on World Wide Web*. 361–370.

[12] Xiangyu Zhao and Jiliang Tang. 2018. Crime in urban areas: A data mining perspective. *ACM SIGKDD Explorations Newsletter* 20, 1 (2018), 1–12.