Crime Event Embedding with Unsupervised Feature Selection

Shixiang Zhu, Yao Xie

H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology
Atlanta Police Department

Introduction

One of the most important problems in crime analysis is that of crime series detection. Technically speaking, crime series is a subset of crime events committed by the same individual or group. Generally, criminals follow a **modus operandi (M.O.)** that characterizes their crime series. Finding crime series based on M.O. critically depends on learning **spatio-temporal** patterns, as well as informative **text narratives** for crime incidents, which is usually done by the human. However, this is not scalable to larger and ever-growing crime data set.

Figure: *An illustration of a crime series*

The main scope of the project is to develop an event embedding algorithm that can jointly capture time, location, and the complex free-text component of each event, using large-scale streaming police report data, both the structured (e.g., time, location) and unstructured (the so-called free-text).

Problem Setup

Challenges

- How to incorporate unstructured text narratives into spatio-temporal model?
- How to reduce the redundant information in text narratives (e.g. irrelevant words), and extract key features?

Problem setup

- A single event data point consists of a set of observed variables $\mathcal{X} = \{x_s, x_t, \} \cup \{x_i\}_{i \in \mathbb{Z}^+, i \le V}$, where $x_s \in \mathbb{R}^D$ (D is the dimension of the space), $x_t \in \mathbb{R}$ are temporal and spatial variables respectively, which are explicitly retained in the model (meaning that they will not be eliminated by the variable selection). And $\{x_i\}_{i \in \mathbb{Z}^+, i \le V}$ is a set of observed variables represent the tf-idf value of the keywords in the vocabulary that appeared at least once in the corpus where $x_i \in \mathbb{R}$, i indicate the index of the keyword and V is the total number of the keywords.
- Given an event point $(x_s, x_t, \mathbf{x})$, we define its embedding as $\mathbf{h} \in \{0,1\}^H$ with the dimension of H (in our later examples we set $H = 1,000$).
- The similarities between two embedding vector can be evaluated by their cosine distance $\mathbf{h} \cdot \mathbf{h}' / \|\mathbf{h}\| \cdot \|\mathbf{h}'\|$, where $\|\mathbf{h}\|$ denotes the ℓ_2 norm of vector $\mathbf{h}$.

Regularization Design

We introduce a ℓ_1-regularizer to the log-likelihood of RBM to mitigate the impact of noisy variables and achieve keywords selection, we impose an ℓ_1 penalty on the probability $1 - P(x_i < t|\mathbf{x})$, where

$$|1 - P(x_i < t|\mathbf{x})| = 1 - \int_{-\infty}^{t} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2\sigma^2}(x_i - b_i - \sigma \sum_{j=1}^{H} h_j w_{ij})^2} dx_i.$$

Gaussian-Bernoulli Restricted Boltzmann Machines

A Restricted Boltzmann Machine is a two layers neural networks and it can be viewed as a probabilistic graphical model. The weights of the network, represented as a matrix $\mathbf{w} = (w_{ij})$, visible bias $\mathbf{b} = (b_i)$ and hidden bias $\mathbf{c} = (c_j)$, which associate H hidden variables $\mathbf{h} = h_{1 \dots H}$ and V observed (visible) variables $\mathbf{x} = x_{1 \dots V}$. Here, for the convenience of demonstrating our application, we assume the real observed variables $\mathbf{x} \in \mathbb{R}^V$ take the Bag-of-Words as input, and binary hidden variables $\mathbf{h} \in \{0,1\}^H$ produce the embeddings. The models can be easily generalized to other types of variables with different activation probabilities. A probability is associated with configuration $(\mathbf{x}, \mathbf{h})$ as follows:

$$p(\mathbf{x}, \mathbf{h}) = \frac{1}{Z} e^{-E(\mathbf{x}, \mathbf{h})},$$

where the partition function Z is defined as $Z = \sum_{\mathbf{x}, \mathbf{h}} e^{-E(\mathbf{x}, \mathbf{h})}$, and the energy function $E(\mathbf{x}, \mathbf{h})$ is defined as

$$E(\mathbf{x}, \mathbf{h}) = -\sum_{i=1}^{V} \sum_{j=1}^{H} w_{ij} h_j \frac{x_i}{\sigma} - \sum_{i=1}^{V} \frac{x_i - b_i}{2\sigma^2} - \sum_{j=1}^{H} h_j c_j.$$

The RBMs' model parameters, $\theta = (\mathbf{w}, \mathbf{b}, \mathbf{c})$, can be learned by maximizing the log likelihood of marginal probability of a set of observed data,

$$\log \mathcal{L}(\theta | \{\mathbf{x}^{(k)}\}) = \sum_{k=1}^{K} \log p(\mathbf{x}^{(k)}),$$

where the marginal probability $p(\mathbf{x})$ can be derived as follows: $p(\mathbf{x}) = \sum_{\mathbf{h}} p(\mathbf{x}, \mathbf{h})$.

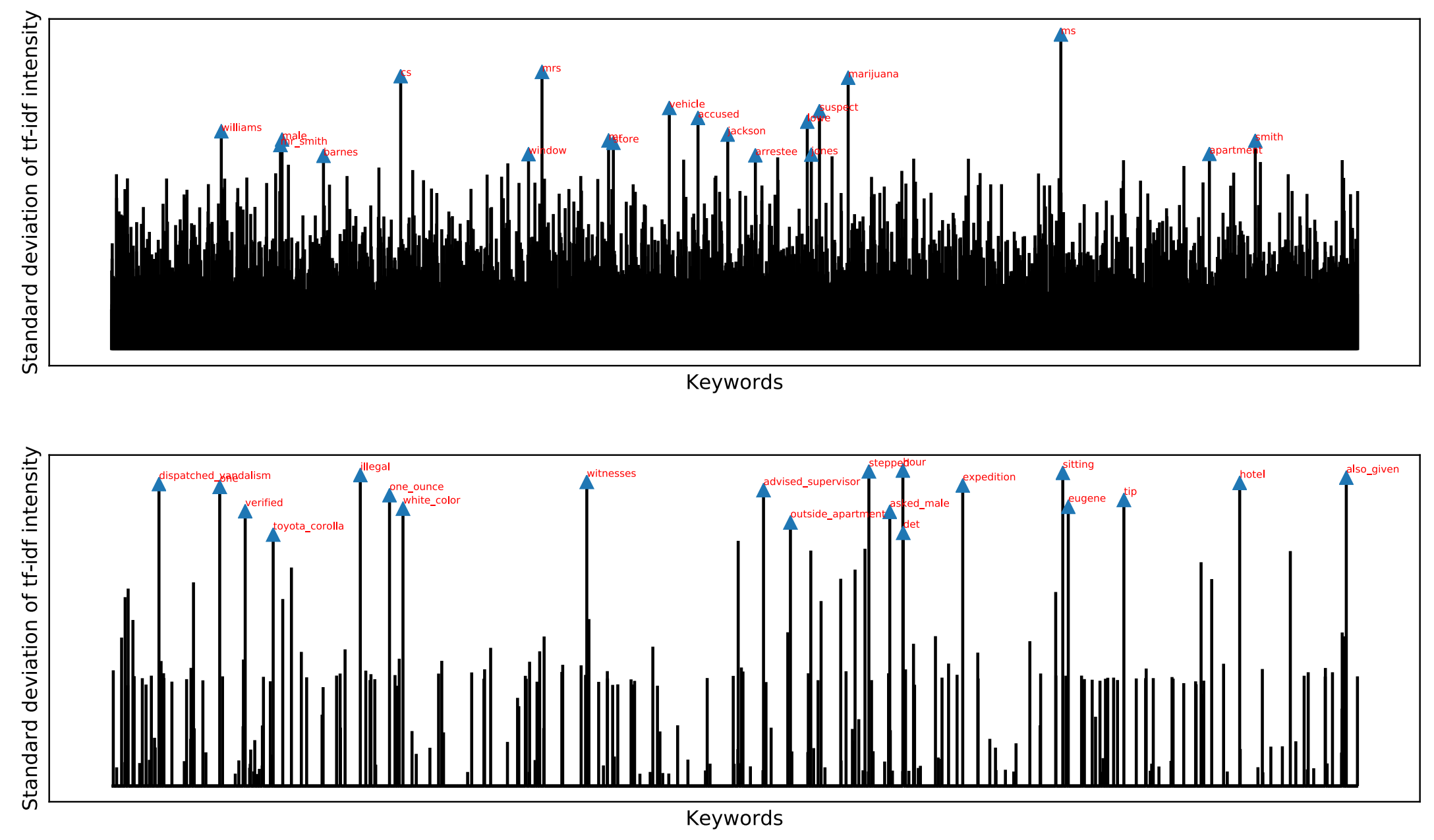

Figure: *The x-axis is the 7,038 keywords, and the y-axis is the standard deviations of each keyword. The standard deviations can be viewed as a kind of indicator for discriminating crime events. Top: Intensities for the words in raw corpus; Bottom: Intensities for the selected words by regularized RBM.*

Motivation for Text Embeddings

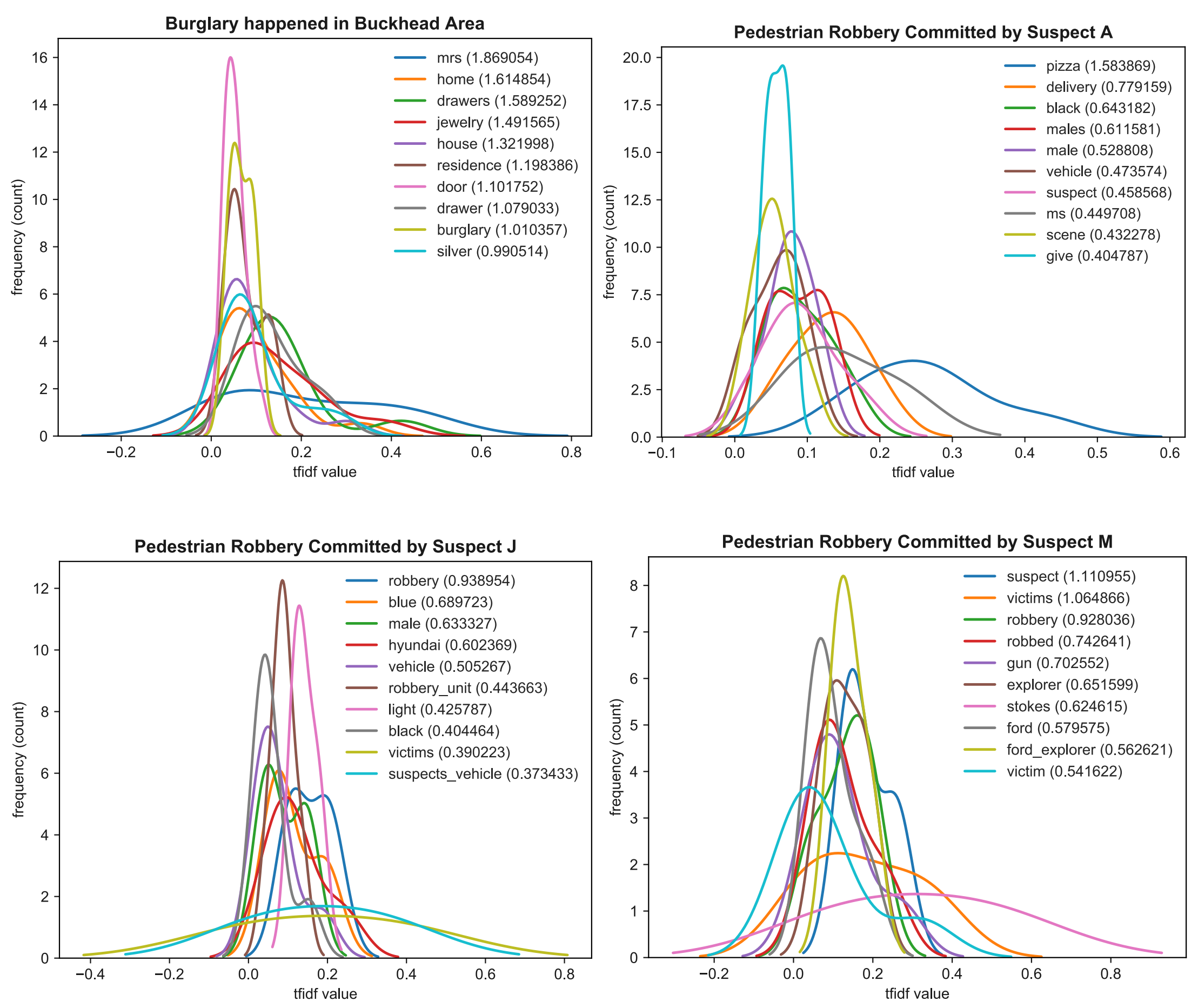

Figure: *Histograms of top 10 high-frequency keywords for 4 crime series. The co-occurrence of high-frequency keywords of different crime series have very distinctive pattern. The specific co-occurrence of keywords reveals key clues of the pattern of a series of crime events, which are able to uniquely identify a crime series.*

Unsupervised Learning Text Embeddings

We propose a *regularized Restricted Boltzmann Machine (RBM)* with built-in feature selection by imposing a ℓ_1-regularizer to the original log likelihood of RBM.

$$\max_{\mathbf{w},\mathbf{b},\mathbf{c}} \left\{ \log \ \mathcal{L}(\theta | (x_s, x_t, \mathbf{x})) - \lambda \sum_{i \le V} |1 - P(x_i < t|\mathbf{x})| \right\}$$

By introducing this penalty term, the gradients ∇w_{ij} and ∇b_i in RBM can be rewritten as follows:

$$\begin{aligned} \nabla w_{ij} &= \langle x_i h_j \rangle_{p(\mathbf{h}|\mathbf{x})} - \langle x_i h_j \rangle_{p(\mathbf{x},\mathbf{h})} \\ &\quad - \lambda \frac{h_j}{\sqrt{2\pi}} \sum_{i \le V} \exp(-\frac{1}{2\sigma^2}(t - b_i - \sigma \sum_{j=1}^{H} h_j w_{ij})), \\ \nabla b_i &= x_i - \langle x_i \rangle_{p(\mathbf{x})} \\ &\quad - \lambda \frac{1}{\sigma\sqrt{2\pi}} \sum_{i \le V} \exp(-\frac{1}{2\sigma^2}(t - b_i - \sigma \sum_{j=1}^{H} h_j w_{ij})). \end{aligned}$$

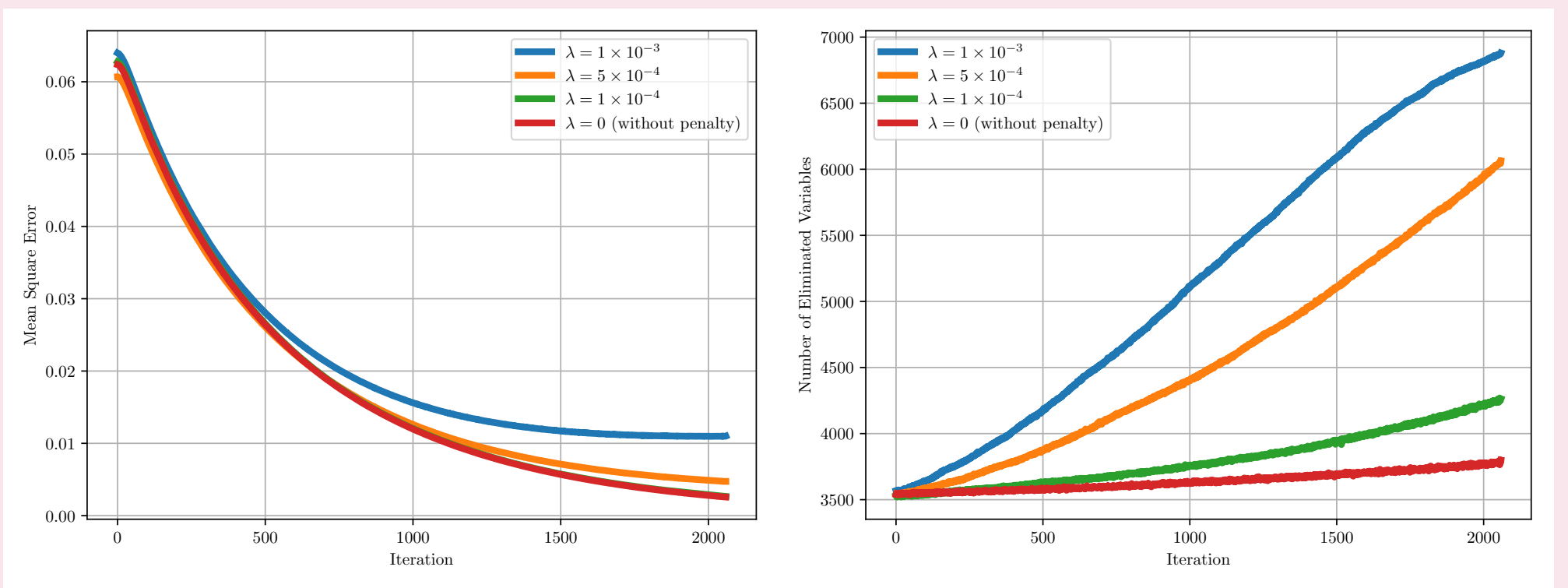

Figure: *Fitted RBM with and without designed penalty term over 2,056 crime events (7,038 words). Left: training errors over iterations; Right: numbers of eliminated variables over iterations ($\rho = 10^{-2}$)*

Results

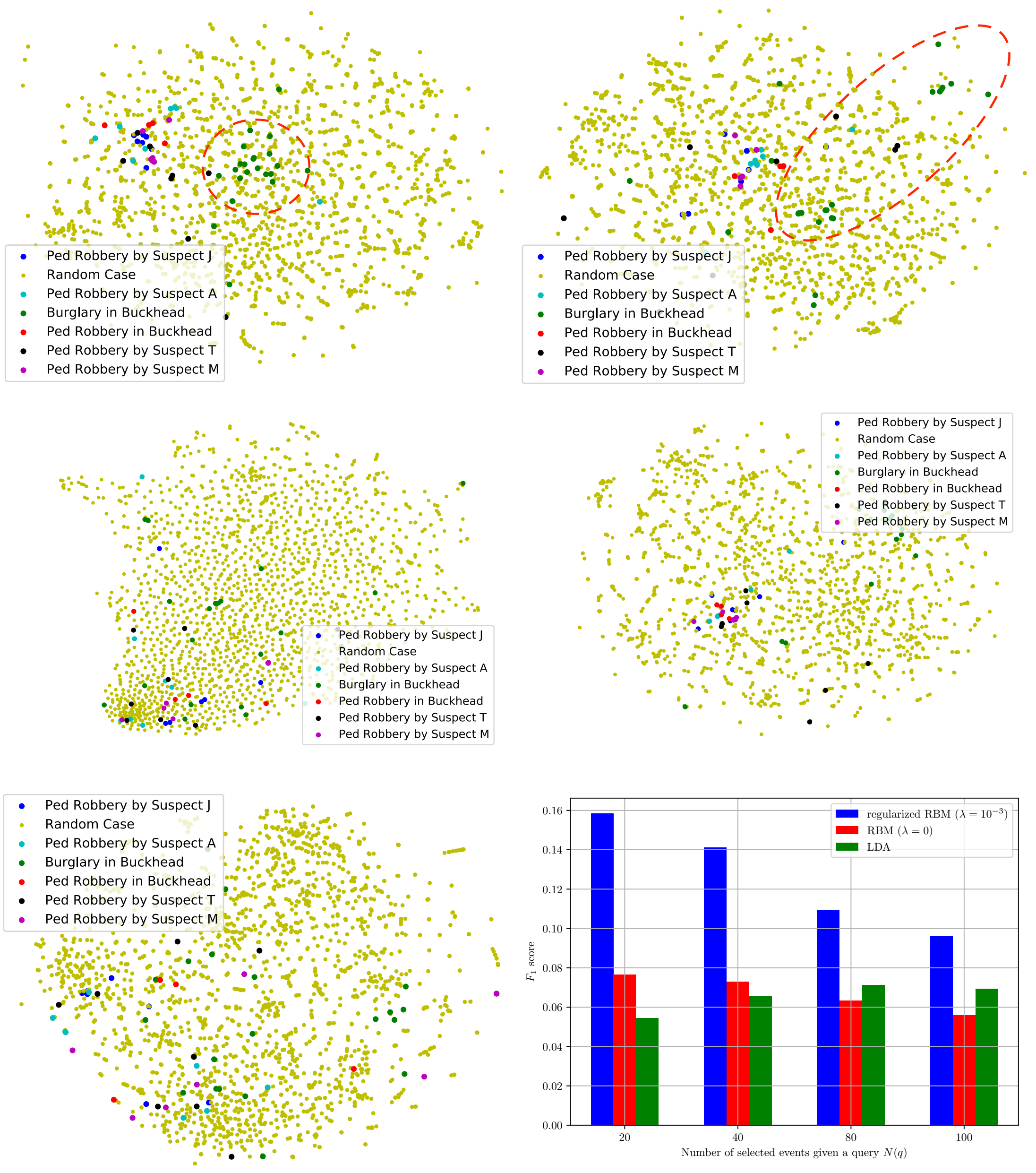

Figure: *Embeddings of 2,056 crime events projected into a 2D space by t-SNE. Top left: the RBM with regularization; Top right: vanilla RBM; Middel left: the autoencoder; Middel right: SVD; Bottom left: LDA; Bottom right: F_1 scores of above methods.*

Conclusion

We have presented a novel approach for learning embeddings for crime events with unsupervised feature selection. By imposing a well-designed ℓ_1 penalty on observed variables' activation probabilities that leads to simple gradient descent based algorithm, our regularized RBMs are able to produce high-quality embeddings as well as eliminate irrelevant and noisy features in observed variables. Additionally, regularized RBMs can select key features without supervision. The selected features are not only highly sparse but also interpretable to human. The techniques introduced in this paper can be also used for learning some other high-dimensional dataset with complex interdependencies between their features.

References

- **Code for Text Embeddings**: *https://github.com/meowoodie/Regularized-RBM*
- **Code for Correlation Estimation**: *https://github.com/meowoodie/Spatio-Temporal-Textual-Correlation-Detection*
- **Related publication**: S. Zhu and Y. Xie. "Crime Linkage Detection by Spatial-Temporal Text Point Processes"