Evaluation of crime topic models: topic coherence vs spatial crime concentration

Ritika Pandey*, George O. Mohler†,

Department of Computer and Information Science
Indiana University-Purdue University, Indianapolis, Indiana - 46202
ripande@iupui.edu, gmohler@iupui.edu

***Abstract*—Non-negative matrix factorization (NMF) topic modeling has recently been introduced for the categorization and analysis of crime report text. Topic modeling in this context allows for more nuanced categories of crime compared to official UCR categorizations. In this paper we suggest two metrics for the evaluation of crime topic models: coherence and spatial concentration. The importance of space comes into play through Weisburd's law of crime concentration, that states a large percentage of crime occurs in a small area of a city. We investigate the extent to which topic models that improve coherence lead to higher levels of crime concentration. Through analyzing a dataset of crime reports from Los Angeles, CA, we find that Latent Dirichlet Allocation (LDA) generates crime topics with both higher coherence and crime concentration. While NMF improves the coherence compared to UCR categorization, the spatial concentration is not as high. These findings have important implications for hotspot policing.**

***Keywords:* Crime topic modeling, LDA, NMF, Crime hotspot, Gini index**

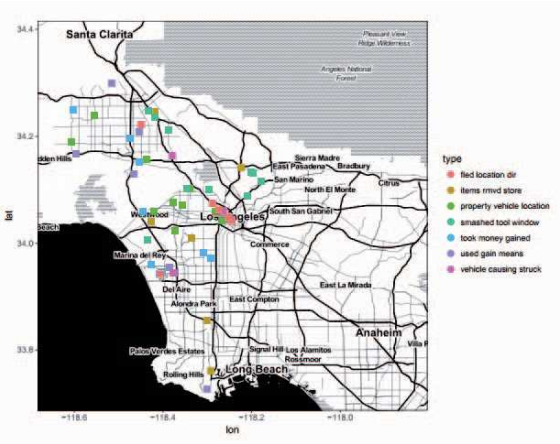

Fig. 1: Crime Hotspots for seven LDA topics.

I. Introduction

Kuang, Brantingham and Bertozzi [1] recently introduced *crime topic modeling*, the application of NMF topic modeling to short (several sentence) text descriptions accompanying crime incident reports. The idea behind crime topic modeling is that crime categories resulting from the FBI Uniform Crime Reporting (UCR) categorization system may lead to a loss of information and NMF topics exhibit a more nuanced model of the text. Under the UCR system crime incidents that reflect a complex mix of criminal behaviors are combined into one of only a few broad categories. For example in the following two crime text reports from Los Angeles, CA, both reports correspond to the same category (aggravated assault) despite the fact that the two suspects exhibit different motives and behaviors.

- *S APPROACHED V ON FOOT AND FOR NO APPARENT REASON STABBED VICT IN CHEST S FLED LOC IN UNK VEH UNK DIR*
- *VICT WAS WALKING OBSD SUSP GRAB A TEMPERED GLASS CANDLE HOLDER AND THROW IT AT HER HITTING HER ON THE ARM*

In [1], non-negative matrix factorization is combined with hierarchical clustering using cosine similarity to achieve a hierarchical topic model for crime incidents. While the resulting topics are qualitatively analyzed in [1], how to evaluate and choose the appropriate topic model for crime reports remains an open question. In this paper we make several contributions along this direction:

1) We motivate two quantitative metrics: topic coherence and spatial concentration.
2) We apply the metrics to evaluate the two most popular topic models, LDA and NMF, for crime topic modeling.
3) We show that by increasing the topic coherence of crime topic models, we may also be able to increase the spatial concentration of crime, which has important consequences for hotspot policing.

Topic coherence is a standard metric for the quantitative evaluation of topic models and has been shown to have good correlation with human evaluations [2]. In our methods section we describe in greater detail coherence, but informally topics have higher coherence when co-occurring words appear more frequently in the same topic. For example, gun and shot may co-occur together in one topic while knife and stab may occur in a different topic.

However, there is also a significant spatial aspect to crime that has important implications for policing. Crime is associated with the physical environment in which it occurs, along with the behavioral and situational conditions that ultimately link suspect to victim to place [3]. Weisburd's law of crime concentration states that a small proportion of the city, known as crime hotspots (see Figure 1), contains the majority of

978-1-5386-7848-0/18/$31.00 ©2018 IEEE

criminal events [4]. Place based interventions in crime hotspots are known to lead to crime reductions in those areas and allow police to focus limited resources on a small area of the city [5]. Our hypothesis is that crime topics that have greater coherence may also have higher levels of crime concentration, facilitating more effective policing interventions. Referencing the two assault reports above, a topic corresponding to the first report may necessitate a gang intervention task force whereas the second report may belong to a mental health topic. These two topics may individually be more concentrated in space compared to when combined.

The outline of this paper is as follows. In Section II we present a brief overview of LDA, NMF, topic coherence, and the gini coefficient for measuring crime concentration. In Section III we analyze a data set of crime reports in Los Angeles from 2009-2014. We show that both NMF and LDA improve upon the coherence of UCR categories, whereas LDA is also able to improve spatial concentration of crime. In Section IV we provide a conclusion and discuss several future directions for research in this area.

II. Methods

Latent Dirichlet Allocation (LDA) is a Bayesian graphical model for text document collections represented by bag-of-words [2][6]. LDA is given by a generative probabilistic model, where each word in a document is generated by sampling a topic from a multinomial distribution with Dirichlet prior and then sampling a word from a separate multinomial with parameters determined by the topic.

Non-negative matrix factorization (NMF) is a widely used tool for the analysis of high-dimensional data as it automatically extracts sparse and meaningful features from a set of nonnegative data vectors [7]. NMF uncovers major hidden themes by factoring the term-document matrix of a corpus into the product of two non-negative matrices, one of them representing the relationship between words and topics and the other one representing the relationship between topics and documents in the latent score topic space [8].

Coherence is a quantitative measure of the similarity of words in a topic. In particular, given a set V of topic words in a corpus (we will use the top 10 most frequent words in each topic), coherence is computed as a sum of similarity scores over all pairs of words in V. While different similarity scores may be used, we consider the intrinsic measure UMass [9] to calculate the coherence. The Umass similarity score measures the extent to which words tend to co-occur in topics together:

$$score(w_i, w_j) = \log \frac{D(w_i, w_j) + \epsilon}{D(w_j)}$$

where $D(w_i, w_j)$ is the number of documents containing both words w_i and w_j and $D(w_j)$ is the number of documents containing word w_j.

Gini index in the context of crime, is a measure of the extent to which a large percentage of crime falls within a small area percentage of a city [10][11]. Consider a city divided into grid cells where the amount of crime falling in each cell is calculated over an observation period. The Lorenz curve is computed by rank ordering the cells by count and then plotting the cumulative percentage of crime against the cumulative percentage of land area. The Gini index, G, ranges from 0 to 1 and is the ratio of the line of equality (representing equal hotspots across the city) and the area under the Lorenz curve. In particular, $G = 0$ corresponds to equal distribution of crime at all grid cells and $G = 1$ corresponds to maximal concentration at a single hotspot. Since, the number of crimes may be less than the number of places, we measure the crime concentration using an adjusted gini coefficient G', defined as the area between the Lorenz curve and line of maximal equality [10],

$$G' = \max(\frac{1}{c}, \frac{1}{n})(2 * \sum_{i=1}^{n} iy_i - n - 1) - \max(\frac{n}{c}, 1) + 1$$

where c is the total number of crimes, n is the total number of places, y_i is the proportion of crimes occurring in place i and, i is the rank order of the place when places are ordered by the number of crimes y_i.

TABLE I Coherence vs. spatial concentration 2009-2014

type	coherence	std. error	gini	std. error
category	-0.817	0.0068	0.3308	0.0021
lda	**-0.287**	0.031	**0.360**	0.007
nmf	-0.300	0.012	0.308	0.002

III. Results

We analyze a data set of crime incidents in Los Angeles that spans the years 2009 to 2014. Each incident is accompanied by a date, latitude, longitude, and text description of the incident that is a short paragraph. For measuring the Gini index, we divide Los Angeles into a grid of size 100x100 and measure the number of crimes of each topic falling in each grid cell.

As part of text preprocessing we remove stop words [12]. We extend the stop-words list from the python NLTK package with common words such as victim, suspect and unknown. We discard all the stop-words and any word whose length is less than 3 characters. We then process the document term matrix using Term Frequency Inverse Document Frequency (TFIDF) weighting factors [13] to emphasize words that occur frequently, but penalizing words that occur in a large percentage of documents (for example stop words not found in our annotated list).

For each year, we sample a balanced data set of 35,000 events, 5000 events from each of seven UCR categories: vandalism, theft, burglary theft from vehicle, burglary, robbery, aggravated assault, and other. We then estimate LDA and NMF using $k = 7$ topics each for a fair comparison to the UCR categories.

In Table I we present the average coherence across years along with the average Gini coefficient. We use a weighted average where the average is weighted by the number of events in each category to take into account the fact that some topics may have more or less than 5000 events. Here we see that LDA has both the highest coherence of topics and highest gini coefficient. In Table II we display the most frequent words of

TABLE II Category 2014

coh.	gini	frequent words
-0.810	0.320	used location fled vehicle info face verbal without punched became
-0.892	0.296	vehicle fled location window used causing door damage side smashed
-0.883	0.433	property location fled removed took store entered items without paying
-0.672	0.282	vehicle property location fled removed window entered took smashed door
-0.720	0.364	prop. fled approach location took vehicle demand money removed punch
-0.894	0.257	location property fled door entered removed window open rear entry
-0.825	0.353	vehicle fled location struck head hit verbal knife causing argument

TABLE III LDA 2014

coh.	gini	frequent words
0.000	0.467	items rmvd store phone paying business exited selected cell concealed
-0.122	0.146	used gain means smash open remove merchandise permission card ifo
0.000	0.180	took money gained secured returned parked demanded missing stated gave
-0.231	0.377	vehicle causing struck punched approached face head property verbal times
-0.462	0.390	property vehicle location removed entered door window rear entry ransacked
-0.185	0.346	fled location dir direction resid hit entered approached foot open
-0.254	0.163	smashed tool window open residence pushed res pry produced glass

each UCR crime category in 2014 and in Table III we display the same table for LDA topics in 2014. For example, the burglary theft from vehicle (BTFV) category has a coherence value of -.672 and a gini index of .282. The closest topic of LDA to BTFV is topic 5, however this topic has higher coherence of -.462 and a higher gini index of .39. There are several topics where LDA has a lower gini index, for example in the case of theft. These topics with lower gini index have lower number of events, resulting in more zero count cells, and the adjusted gini index is lower in these cases. However, in the weighted average across topics LDA has a higher over all gini index.

In Figure 2 we display coherence and the gini coefficient over time to assess the stability of these results. Here we see that LDA consistently has a higher gini index over time. For some years NMF has a higher coherence, though both NMF and LDA have consistently higher coherence than the UCR crime categories.

IV. Conclusion

We suggested two performance metrics for crime topic models: topic coherence and the gini coefficient for measuring spatial concentration. We showed that the choice of topic model has important implications for detecting crime hotspots. In particular, it is possible to achieve more coherent topics that simultaneously concentrate to a higher degree in space, allowing for more targeted police interventions given limited resources. For the data set analyzed in Los Angeles, our results show that LDA has the highest coherence and gini coefficient compared to NMF and UCR crime categories.

Future research may focus on the joint optimization of coherence and spatial concentration. LDA and NMF in this paper were provided with no spatial information. Methods may be developed that can improve both coherence and concentration jointly using supervised learning. Additionally, such methods may be extended to spatio-temporal models where topics and spatial hotspots evolve over time [14].

V. Acknowledgements

This work was supported in part by NSF grants S&CC-1737585, SES-1343123, ATD-1737996. G.M. is a co-founder of PredPol, a company offering predictive policing services to law enforcement agencies and serves on the board of directors.

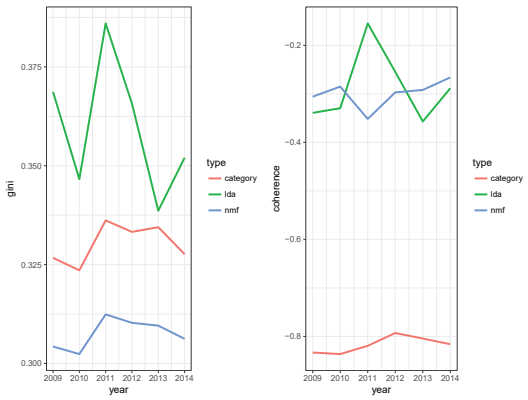

Fig. 2: Stability over time of coherence vs generalized gini coefficient over time.

References

[1] D. Kuang, P. J. Brantingham, and A. L. Bertozzi, "Crime topic modeling," *Crime Science*, vol. 6, no. 1, p. 12, 2017.

[2] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, (Stroudsburg, PA, USA), pp. 100–108, Association for Computational Linguistics, 2010.

[3] P. L. Brantingham and P. J. Brantingham, "Nodes, paths and edges: Considerations on the complexity of crime and the physical environment," *Journal of Environmental Psychology*, vol. 13, no. 1, pp. 3 – 28, 1993.

[4] D. Weisburd, "The law of crime concentration and the criminology of place*," *Criminology*, vol. 53, no. 2, pp. 133–157.

[5] A. A. Braga, A. V. Papachristos, and D. M. Hureau, "The effects of hot spots policing on crime: An updated systematic review and meta-analysis," *Justice quarterly*, vol. 31, no. 4, pp. 633–663, 2014.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[7] N. Gillis, "The why and how of nonnegative matrix factorization," *Regularization, Optimization, Kernels, and Support Vector Machines*, vol. 12, no. 257, 2014.

[8] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, pp. 267–273, 01 2003.

[9] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, (Stroudsburg, PA, USA), pp. 262–272, Association for Computational Linguistics, 2011.

[10] W. Bernasco and W. Steenbeek, "More places than crimes: Implications for evaluating the law of crime concentration at place," *Journal of Quantitative Criminology*, vol. 33, pp. 451–467, Sep 2017.

[11] J. E. Eck, Y. Lee, O. SooHyun, and N. Martinez, "Compared to what? estimating the relative concentration of crime at places using systematic and other reviews," *Crime Science*, vol. 6, no. 1, p. 8, 2017.

[12] M. Rajman and R. Besançon, "Text mining - knowledge extraction from unstructured textual data," in *Advances in Data Science and Classification* (A. Rizzi, M. Vichi, and H.-H. Bock, eds.), (Berlin, Heidelberg), pp. 473–480, Springer Berlin Heidelberg, 1998.

[13] J. Ramos *et al.*, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, vol. 242, pp. 133–142, 2003.

[14] G. Mohler and P. J. Brantingham, "Privacy preserving, crowd sourced crime hawkes processes," in *Social Sensing (SocialSens), 2018 International Workshop on*, pp. 14–19, IEEE, 2018.