Visual crime pattern analysis

Germain Garcia-Zanabria*† and Luis Gustavo Nonato†
*Research and Innovation Center in Computer Science
Universidad Católica San Pablo, Arequipa, Peru
†Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo

***Abstract*—Studying and analyzing crime patterns in big cities is a challenging Spatio-temporal problem. The problem's difficulty is linked to different factors such as data modeling, unsophisticated hotspot detection techniques, Spatio-temporal patterns, and study delimitation. Previous works have mostly focused on the analysis of crimes with the intent of uncovering patterns associated to social factors, seasonality, and urban activities in whole districts, regions, and neighborhoods. Those tools can hardly allow micro-scale crime analysis closely related to crime opportunity, whose understanding is fundamental for planning preventive actions. Visualizing different patterns hidden in crime time series data is another issue in this context, mainly due to the number of patterns that can show up in the time series analysis. In this dissertation, we propose a set of approaches for interactive visual crime analysis. Relying on machine learning methods, statistical and mathematical mechanisms, and visualization, each proposed methodology focus on solving specific crime-related problems. These proposed tools to explore specific city locations turned out to be essential for domain experts to accomplish their analysis in a bottom-up fashion, revealing how urban features related to mobility, passerby behavior, and the presence of public infrastructures can influence the quantity and type of crimes. The effectiveness and usefulness of the proposed methodologies have been demonstrated with a comprehensive set of quantitative and qualitative analyses, as well as case studies performed by domain experts involving real data from different-sized cities. The experiments show the capability of our approaches in identifying different crime-related phenomena.** [1][2]

I. Introduction

The hardness of the crime ***spatio-temporal*** analysis problem is linked to the patterns' great variability among the different types of crimes and the large amount of data involved in such analysis. In recent years, it is undeniable that crimes have not only grown but also become more violent and modernized. In contrast, agencies in charge of law and order (*e.g.*, police and the criminal justice system) have not kept up with these trends. The gap between the dynamics of crime and violence and the state's ability to contain them within the law rule has widened. Therefore, introducing modern instruments for managing public order and crime containment is imperative to make public security policies more efficient in any big city.

Crime hotspot analysis has been one of the main resources employed by public security agencies to plan police patrolling and design preventive actions. Although sophisticated mechanisms have been proposed to detect hotspots, the search for a high prevalence of crimes ends up neglecting sites where certain types of crimes are frequent but not sufficiently intense to be considered statistically significant, which can be more harmful to the community than intensive crime waves that occur in a short period of time. Moreover, most techniques enable only rudimentary mechanisms to analyze an essential component of unlawful activities, the temporal evolution of crimes, and corresponding patterns.

There is another important aspect in the context of crime analysis, the ***spatial discretization*** of the urban areas under analysis. The spatial discretization directly impacts the computation and detection of hotspots. Moreover, according to environmental criminology, the concentration and persistence of crimes in certain locations are not random; that is, they occur due to prevalent characteristics present in those locations (demography, socioeconomic level, and unemployment). Therefore, characteristics' changes of particular locations impact crime activity over time, making Spatio-temporal hotspot analysis a fundamental task. Crime events are typically examined across various discretizations, such as states, cities, neighborhoods, and blocks. Nevertheless, urban crime activities happen on micro-places (*i.e.*, street segments and corners), which are thus a more meaningful representation of locations than arbitrarily-defined regions.

Understanding the ***dynamic of crime patterns*** over time is another important aspect of crime analysis. Space-time hotspot researchers sustain empirical evidence that locations of crime incidents tend to exhibit both spatial and temporal concentrations. Moreover, spatial and temporal crime concentration patterns are deciding factors for planning crime prevention measures. However, most hotspot-based analytic tools, mainly those used by security agencies, do not enable resources to identify and group hotspots according to their temporal behavior, hampering the identification of factors that can make crime viable or not over time.

This dissertation's core consists of presenting different methodologies that allow a visual Spatio-temporal crime pattern analysis of urban areas considering different characteristics (socio-economic, infrastructure, and social factors). To do this, we have to sort out different problems: hotspot definition and detection, space modeling, and identifying patterns related to the dynamics of crime patterns. Our approaches faced these problems from different fronts: (1) Given the crime events in an urban space, we propose two different methods to identify and present hotspots considering not only the intensity but also the frequency of crimes; (2) street-

[1]Ph.D. thesis at ICMC-Universidade de São Paulo.
[2]Email: germain.garcia@ucsp.edu.pe, gnonato@icmc.usp.br

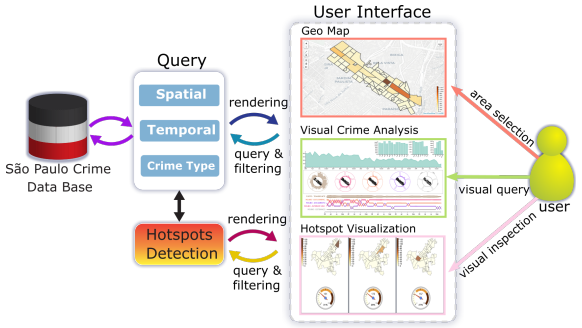


FIG. 1: Pipeline overview of the CrimAnalyzer System.

level domain discretization, switching from grid-based to a street-based spatial discretization; (3) Spatio-temporal crime patterns analysis, supported by visualization and machine learning mechanisms to extract and visually present different Spatio-temporal patterns. Considering these solutions, we have divided our proposals into three projects presented above.

Essentially, this document presents three proposed methodologies (CrimAnalyzer [1], Mirante [2], and CriPAV [3]) for interactive visual crime analysis developed in close collaboration with domain experts. Moreover, for each project, we have presented different qualitative and quantitative experiments with real data from large and mid-sized cities, which have been validated by domain experts.

II. CRIMANALYZER

CrimAnalyzer is a new visual analytic tool customized to support the analysis of criminal activities in urban areas with specific characteristics, that is, high criminality rates with great variability in the pattern of crimes, even in geographically close regions. CrimAnalyzer enables several linked views tailored to reveal patterns of crimes and their evolution over time, assisting domain experts in their decision-making process; providing guidelines not only for repressive but, above all, preventive actions, strengthening the planning and implementation of institutional actions, especially from the police.

In collaboration with a team of domain experts, we have designed visual analytic functionalities that allow users to select and analyze regions of interest in terms of their hotspots, crime patterns, and temporal dynamics. The modules and system architecture are illustrated in Fig. 1. Furthermore, CrimAnalyzer implements a methodology based on Non-Negative Matrix Factorization to identify hotspots based not only on the number of crimes but also on the rate they occur.

A. Hotspot Identification Model

As discussed, hotspot identification is one of the most important tasks for crime analysis. Here, hotspots have a more general connotation than in previous works, corresponding to sites where criminal activity is high but also to locations where the number of crimes is not large but frequent enough to deserve detailed analysis. For instance, in a given region, sites whose number of crimes is much larger than in any other sites are clearly important hotspots. However, the region can also contain a particular site where crimes are frequent but happening in a much smaller magnitude compared to the prominent ones. The region can also contain sites where crimes are not frequent at all but present spikes in particular time frames. We consider the three different phenomena as hotspots, seeking to identify: *(i)* sites where crimes are frequent and in large number, *(ii)* sites where crimes are frequent but do not in large number, and *(iii)* sites where crimes are not frequent, but happen in large numbers in particular time frames.

To get around the difficulties pointed above, we opted to an approach based on Non-Negative Matrix Factorization (NMF), which worked pretty well for us in identifying hotspots according to our needs. As far as we know, this is the first work to employ NMF as a mechanism to detect hotspots in the context of Crime Mapping.

B. Non-Negative Matrix Factorization

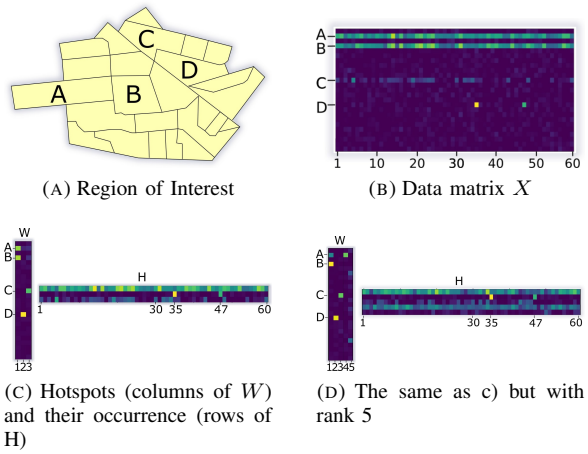
An $m \times n$ matrix X is said *non-negative* if all entries in X are greater or equal to zero ($X \geq 0$). The goal of NMF is to decompose X as a product $W \cdot H$, where W and H are non-negative matrices with dimensions $m \times k$ and $k \times n$, respectively (the roles of m, n, and k will be clear in Subsec. II-C). In mathematical terms, the problem can be stated as follows, $\arg\min_{W,H} \|X - WH\|^2$ subject to, $W, H \geq 0$.

A solution for the minimization problem provides a set of basis vector w_i, corresponding to the columns of W, and a set of coefficients h_j, corresponding to the columns of H, such that each column x_j of X is written as the linear combination $x_j = \sum_i h_{ij} w_i$, (or $x_j = W h_j$). In other words, for each line in X we have a corresponding column in H whose entries are coefficients associated to the columns (basis vectors) of W.

There are two important aspects in an NMF decomposition that will be largely exploited in the context of hotspot detection, namely, low rank approximation and sparsity. Low rank approximation accounts for the fact that the basis matrix W usually has much lower rank than the original matrix X, meaning that (the columns of) X is represented using just a few basis vectors. As detailed in the next subsection, we rely on low rank approximation to define the number of hotspots, that is, by setting the rank of W, we also set the number of hotspots. Sparsity means the basis and coefficient matrices contain many entries equal (or close) to zero, which naturally enforces only relevant information from X to be kept in W and H. This fact is important to identify particular sites within a hotspot and the time slices where each hotspot shows up.

C. Identifying Hotspots with NMF

We rely on NMF to identify hotspots, their rate of occurrence and "intensity". The matrix X to be decomposed as the product $W \cdot H$ comprises crime information in a particular region of interest. Specifically, each row in X corresponds to a site of the region and each column to a time slice. In order to facilitate discussion, we present the proposed approach using a synthetic example. Fig. 2 shows a region made up of 25



(A) Region of Interest (B) Data matrix X

(C) Hotspots (columns of W) and their occurrence (rows of H) (D) The same as c) but with rank 5

FIG. 2: (A) Region of interest. (B) Data matrix containing crime information from the regions in a). Rows correspond to sites while columns are time slices. The darker the color, the closer to zero the number of crimes is. (C) Rank 3 NMF decomposition from X. (D) Rank 5 NMF from X.

sites and we generated synthetic crime data in 60 time slices, representing months over five years. For sites denoted as A and B, we draw 60 samples from a normal distribution with mean 8 and variance 4, ensuring that A and B are correlated, that is, when the number of crimes in A is large the same happens with B (the number of crimes in B is generated by perturbing the values of A using a uniform random distribution with values between -3 and 3). This construction is simulating two regions with high prevalence of crimes over time. Crimes in the site denoted as C in Fig. 2 follows a normal distribution with mean 1 and variance 4, corresponding to a location where crimes are not large in number, but happening quite frequently. Finally, for site D we draw 60 samples from a normal distribution with mean 0 and variance 0.25, except for time slices 35 and 47, where we set the number of crimes equal to 15 and 10 respectively, simulating a site where crimes are no frequent, but happen in large numbers in particular time slices. For all the other sites we associated 60 samples drawn from a normal distribution with mean 0 and variance 0.25. Values for all sites are rounded to the closest integer and negative values set to zero. Fig. 2b illustrates the matrix X built from the synthetic data described above. Notice that the simulated crime dynamics is clearly seen in X.

Given an $m \times n$ matrix $X \geq 0$, an NMF decomposition of X results in matrices $W \geq 0$ and $H \geq 0$. In practice, the rank of W is significantly less than both m and n, i.e., $k = (W) \ll m, n$. In our context, the columns of W correspond to hotspots while entries in the rows of H indicate the "intensity" of the hotspot in each time slice. Fig. 2c illustrates matrices W and H obtained from matrix X in Fig. 2b using a NMF decomposition with rank $k = 3$. Notice that the entries in the first (left most) column of W have values close to zero almost everywhere, except in the entries corresponding to the sites A and B. Therefore, the hotspot derived from the first column of W highlights sites A and B as the relevant ones. The high prevalence of crimes in those regions can be seen from the first (top) row of matrix H, which has most of its entries with non-zero values. The second column of W is mostly null, except in the entry corresponding to site D, where crimes are not frequent but happen with high intensity in particular time slices. Notice that the second row of H has two entries different from zero, corresponding exactly to the time slices 35 and 47, when the site D faces a large number of crimes. Finally, the last column of W gives rise to a hotspot that highlights site C, where crimes are frequent but in smaller magnitude when compared to A and B. The incidence and intensity of crimes in C are clearly seen in the third (bottom) row of H.

One can argue that the results presented in Fig. 2c worked so well because we wisely set the rank of W equal $k = 3$ and that in practice, it is difficult to find a proper value for the rank. To answer this question, we have presented many experiments in the manuscript; for instance, Fig. 2d shows an experiment with rank $k = 5$. With the experiments, we have noticed that increasing k tends to split meaningful hotspots while creating some noisy, not-so-important ones, which can easily be identified from almost zero rows in H.

D. Experimental Results Summary

CrimAnalyzer provides visual resources supported by mathematical and computational machinery and validated by domain experts. To show its effectiveness and usefulness, we have reported qualitative and quantitative comparisons as well as case studies run by domain experts involving real data. We have validated the outperforming of our new methodology to identify crime hotspots over statistical methods by quantitative comparisons in 300 regions over $30,815$ census blocks of São Paulo (the largest city of South America). We also reported three case studies revealing interesting phenomena in the crime dynamic. Finally, we have presented experts' evaluation reporting about their experience and methodology feedback.

III. MIRANTE

An important aspect of crime mapping is the spatial distribution. Most techniques rely on regular grids with crime data aggregated on grid cells, each possibly covering hundreds of square meters. However, crime events rarely concentrate on regions larger than street segments or corners since those places attract distracted and vulnerable people who carry money and valuables. Therefore, relying on spatial discretizations such as the regular grids renders fine-grained crime analysis a quite challenging task since the definition of a proper grid resolution and the identification of urban factors impacting the crime opportunity is not so straightforward when crimes are aggregated in a cell containing several street blocks. Even when a small grid resolution is used, the alignment of the grid cell, streets' segments, and other urban structures are not easy to do, hampering the detailed analysis of crime patterns and their possible causes. In addition, the grid representation also limits the analysis of the temporal behavior of crimes.

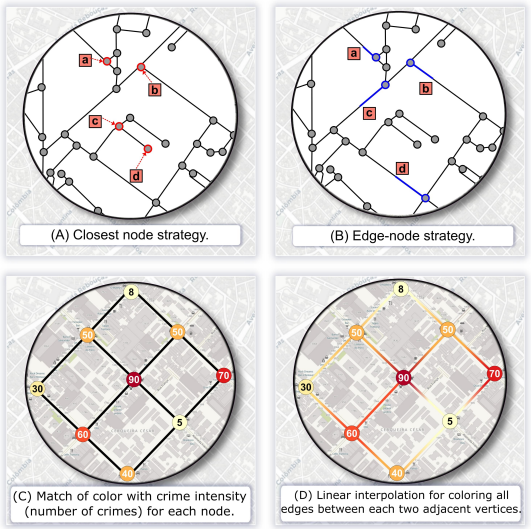


FIG. 3: (A, B) Two ways to build a crime-based street-network by closest node based on: (A) Euclidean distance, and (B) edge-node strategy. (C, D) Street-level heatmap construction in two steps: Color assignation and linear interpolation.

For instance, suppose that a type of crime occurs consistently on a street corner during a period and, after a while, moves to a nearby corner. In a grid representation, such a temporal behavior can hardly be caught if both corners lie on the same grid cell.

Mirante is a scalable and versatile visualization tool designed in close collaboration with domain experts that has been tailored to explore crime data in a street-level of detail. Considering street corners as nodes and street segments as edges, Mirante assumes the city street network as the spatial discretization. Crime data is spatially aggregated on street corners using an edge-node strategy rather than Euclidean distance, which avoids several issues presented in grid cell aggregation. The result is a graph-based heatmap when a region of interest is selected. Fig. 3 (C and D) show the street level heatmap construction (aggregation and interpolation). The heatmap is updated according to interactive users' filters applied to the data. Mirante provides several interactive resources to explore the spatial distribution of crimes and their dynamics over time, making it possible to identify temporal patterns such as the shift of crime hotspots among nearby locations. Interactive filters allow users to focus their analysis on particular hours of the day, days of the week, and months of the year, making it easy to scrutinize crimes seasonality. Using different selection mechanisms, users can interactively select regions of interest on various scales, enabling the Spatio-temporal analysis of large regions and specific city locations, a trait not available in most crime analysis tools. Simplicity and ease of use are other characteristics that render Mirante an interesting alternative in crime mapping.

A. Building the spatial representation

To build the graph corresponding to the spatial discretization, we use the OpenStreetMap API, which allows for generating a street-graph containing roads and intersections for entire cities. It is possible to define the type of map to use, *e.g.*, pedestrian, bike, and car drive roads. We opt to use the pedestrian map, as it comprises drive roads and pedestrian walkways.

The number of nodes and edges derived from the map varies considerably depending on the city. However, a large number of vertices do not correspond to street intersections. To remove non-intersection vertices and all the points along a single street segment, we run a procedure that topologically simplifies the graph.

B. Assigning data to the nodes of the spatial graph representation

Let $L_{crime} = \{c_0, c_1, ..., c_n\}$ be a list of n crime records, where each c_i contains information such as record id (unique identifier), location (latitude, longitude), crime type, date, number of people involved, among other information. Let $G = (V, E)$ be the graph corresponding to the city's spatial representation. Each vertex has a unique geo-referenced coordinate (identifier, latitude, longitude), and each edge represents a segment joining two intersections.

In our context, each crime record c_i must be assigned to a vertex of the graph G. The easiest solution would be to assign each c_i to its nearest vertex using the Euclidean distance. However, using Euclidean distance is not appropriate because it does not consider the topology of the spatial representation. We illustrated this issue in Fig. 3 (A, B). Notice that using Euclidean distance the crime records "a" and "b" are properly assigned vertices, however, records "c" and "d" are not, since it is clear that the corresponding crimes took place on the street segments closer to them, so they should be assigned to one of the vertices defining the segments. Fig. 3 (B) shows the correct procedure, which we call *edge-node strategy*, where first, the nearest edge (e_{near}) is found and then the closest vertex.

The crime-vertex assignment starts by traversing the list of crime records L_{crime} to compute their nearest edge e_{near} in the graph G. Different strategies can be used to efficiently perform this step, *e.g.*, using a spatial data structure as Quad-tree or Ball-tree. In our case, we use an R-tree implemented in the OSMnx library ($e_{near} = G.get\_nearest\_edge(c_i)$). Once the nearest edge e_{near} and end-nodes ($(v_1, v_2) = G.get\_vertices(e_{near})$) is found for each record, it is assigned to the closest edge node. For that, we compute the distance to both nodes ($d_{\{1,2\}} = greatCircleDistance(c_i, v_{\{1,2\}})$) and crime record in c_i are stored into the list $crimes$ associated to each node, that is, $v_1.crimes.append(c_i)$ if $d_1 < d_2$ or $v_2.crimes.append(c_i)$ otherwise. List per-vertex is used to temporally aggregate crime records (hourly aggregation in our case), giving rise to a time series associated to each vertex.

C. Experimental Result Summary

We have employed real data set of two different-sized cities to show the effectiveness of Mirante in identifying crime patterns, making it easier to establish relations between crimes and other factors, such as urban infrastructure. Moreover, the

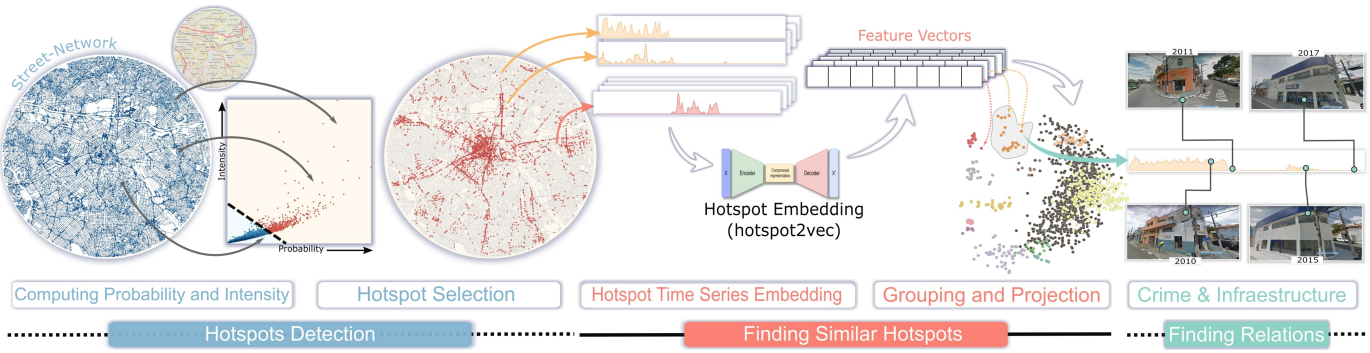


FIG. 4: The proposed street-level crime visualization methodology, *CriPAV*, comprises three main steps. Hotspot Identification: identifying hotspots based on crime intensity and crime probability. Finding Similar Hotspots: hotspot time series embedding (Hotspot2Vec), clustering, and projection into a visual space. Cripav system: finding the relation between urban infrastructure and crimes.

case studies were addressed in colaboración of domain experts who gave us positive feedback; more details can be found in https://youtu.be/SeFNScNMgQY.

IV. CRIPAV

The problem of hotspot detection methods described in Sec. II derives from the fact that there is no consensus about the definition of a hotspot, as such definition strongly depends, among other factors, on the spatial discretizations that support hotspot computation. The most common spatial discretization is a regular grid with cell granularity varying according to the scale of the analysis, which can range from dozen of meters to large areas covering entire neighborhoods. However, crimes are mostly concentrated in 'micro' places that are relatively stable over time. Therefore, fine-grained crime analysis demands a level of discretization that should reach the scale of streets, which is difficult to be obtained with regular grids, as the density and arrangement of streets may vary considerably across the city.

Another important aspect related to hotspot analysis is related to the reasons that lead to the appearance of a hotspot in a given location. Crime events are related to each location's features; consequently, changes in the characteristics of particular locations impact crime activity over time, making the temporal analysis of hotspots a fundamental task.

In collaboration with domain experts and supported by mathematical and machine learning tools, we have designed a visual analytic methodology to scrutinize crime activities in a street-level of detail. Specifically, we use the same modeling of Mirante, which avoids the issue of finding a proper level of refinement to accomplish the analysis. The proposed methodology relies on mathematical theories to identify hotspot based not only on the intensity of crimes but also on the probability. Moreover, we rely on a deep learning model to embed crime time series in a high-dimensional space so as to make possible the identification of hotspot groups with similar temporal behavior, a task difficult to be performed with conventional hotspot analysis tools. The proposed methodology, illustrated in Fig. 4, has been assembled in a visualization system called *CriPAV*, which, besides enabling a more general characterization of hotspots, provides visual resources to identify hotspot locations with similar temporal behavior. The identification of hotspot locations with similar temporal crime behavior helps the understanding of how changes in urban infrastructure impact criminal activity over time.

A. Hotspots Detection

As illustrated on the left of Fig. 4 (Hotspot Detection), hotspot identification is a primary component of CriPAV. Hotspots are visually defined from a 'Probability × Intensity' scatter plot, where each dot corresponds to an anchor point. The intensity axis of each anchor point is the temporally aggregated number of crime events in the anchor point divided by the maximum number of crime events among all anchor points. The computation of how likely crimes are in each anchor point is more intricate, and it will be detailed below.

Probability of Crimes The probability of crimes take place in each particular anchor point is derived from the stationary state of a stochastic matrix built from the temporal crime data.

In our context, the probability of crimes in each anchor point is given by the stationary vector of a stochastic matrix built from the crime time series in each anchor point. The construction of such stationary matrix is detailed in the following.

Computing the Stochastic Matrix

Given a *street network* with a set of n anchor points $V = \{\tau_1, \tau_2, \ldots, \tau_n\}$ associated to m time units $T = \{t_1, t_2, \ldots, t_m\}$ describing crime events aggregated into m time instants. We can define a function $f : V \times T \to R$ that associates the number of crime events $f(\tau_i, t_j)$ in the anchor point τ_i in the time slice t_j. We denote by D the $n \times m$ matrix where each entry D_{ij} corresponds to $f(\tau_i, t_j)$. From $f(\tau_i, t_j)$ we define an occurrence matrix $\widehat{D}$ where $\widehat{D}_{ij} = 1$ if $f(\tau_i, t_j) > 0$ and $\widehat{D}_{ij} = 0$ if $f(\tau_i, t_j) = 0$. $\widehat{D}$ is a binary matrix where each entry $\widehat{D}_{ij}$ indicates whether crimes took place in the anchor point τ_i in the time slice t_j.

From $\widehat{D}$ we define the $n \times n$ co-occurrence matrix $\widehat{P}$: $\widehat{P} = \widehat{D} \cdot \widehat{D}^T$. Each entry $\hat{P}_{ij}$ of $\widehat{P}$ corresponds to the number of times that the anchor points τ_i and τ_j faced crime events in the same time slice, that is, the number of times that crimes took place simultaneously in τ_i and τ_j. A large value of $\hat{P}_{ij}$ indicates that τ_i and τ_j present similar crime activity over time. Dividing each row of $\widehat{P}$ by the sum of its values, we end up with a stochastic matrix P, that is, $P_{ij} = \hat{P}_{ij} / \sum_{k=1}^{n} \hat{P}_{ik}$. The entry P_{ij} corresponds to the probability $Pr(\tau_i, \tau_j)$ of a crimes take place simultaneously in τ_i and τ_j.

The reasoning behind the construction of the stochastic matrix P is that certain crime types are seasonal, occurring concurrently in different city locations depending on the day of the week, the fortnight of the month, and the month of the year. Matrix P, as defined above, captures such a seasonality, being able to point as likely an anchor point where crime activity is not intense, but occur concurrently with other anchor points.

Given the stochastic matrix P the *probability* of crime occurrence in each anchor point is given by the stationary vector π of P, that is, the probability of a crime event occur in τ_i is the value in the i-th entry in π.

Selecting Hotspots The *probability* and *intensity* values summarizing crime activities in each anchor point enable the use of a *Probability vs. Intensity* scatter plot to visually identify anchor points based on their intensity, probability, or both.

In order to filter out relevant anchor points (*i.e.*, high probability and/or high intensity), we use a function $g = [0,1] \times [0,1] \rightarrow [0,1]$ that assigns a value to each anchor point, as for example $g(probability, intensity) = ((1-\alpha) * probability + \alpha * intensity)$. The value of α is the weight one wants to give to intensity and probability when filtering the hotspots. The scatter plot in Fig. 4-Hotspot Detection shows the selection with $\alpha = 0.5$.

We have implemented the linear hotspot selection mechanism as an alternative to the interactive brush-based interactive tool. Domains experts deemed the linear approach easier to use than a brush-based one.

B. Finding Similar Hotspots

Another essential task that our methodology must accomplish is identifying hotspots with similar temporal behavior (see Fig. 4-Finding Similar Hotspots). Finding the temporally similar hotspot means searching for a similar time series, which is a difficult problem. Methods such as Discrete-Time Wrapping can be used to this end but with the price a high computational cost and instability to noise. Instead, we opt for a deep learning embedding technique we called *Hotspot2Vec*.

Hotspot2Vec. We use an autoencoder to map each time-series $TS = \{ts_1, ts_2, \ldots, ts_m\}$ to a feature space. The autoencoder model is trained with a set of time-series $\widehat{TS} = \{\widehat{ts}_1, \widehat{ts}_2, \ldots, \widehat{ts}_m\}$, where $\widehat{ts}_i = 1$ if $ts_i > 0$ and $\widehat{ts}_i = 0$ otherwise, for all $i \in \{1, \ldots, m\}$. The idea is to train the deep learning model to capture the temporal behavior of crimes, without taking into account the intensity of crimes. Therefore, anchor points with crimes happening at the same time interval will be considered similar, no matter the intensity of crimes in each location.

Autoencoder is a well-known neural network model in which the input and output are the same. The middle layer of the network has a bottleneck that creates a compressed representation, aiming to reduce the data's dimensionality.

Grouping Similar Hotspots. The encoder's output is used as feature vectors, and a clustering algorithm is applied to group hotspots based on their proximity in the feature space. We choose a hierarchical variant of BDSCAN called HDBDSCAN because it can automatically find the number of clusters (as DBSCAN) without tuning several parameters, relieving users of this task, which is essential for domain experts with little training in machine learning.

Projection. Empirical tests showed that reducing the dimensionality of time series preserve good properties in terms of capturing their similarity. To visualize the resulting embedding, we relied on a modified version of the LAMP projection technique, which maps the embedded time series to a 2D visual space. LAMP is a computationally efficient projection method that can be tuned to preserve labeled clusters during the mapping. Fig. 4 (Finding Similar Hotspots) shows an example of the HDBSCAN clusterization and LAMP projection.

C. Experimental Result Summary

We have validated the proposed approach from different fronts. We have validated the hotspot detection and Finding similar hotspots (Hotspot2Vec) techniques in a region with more than $14,000$ street intersections. Moreover, in addition to some qualitative comparisons, we have presented three case studies considering São Paulo city, with about $1,650,000$ crime incidents. Each case study addresses different analytical tasks. Finally, we have presented a user study considering experts with a large experience in crime analysis; the quantitative metrics and comments highlighted CriPAV as a helpful, useful, friendly, and innovative tool.

V. Publications, Scientific Dissemination, Awards and Honors

Publication (and Qualis)

[1] (A1), [3] (A1), [2] (A2).

Award and Honors

- Best paper in *Graphics and Visualization at* SIBGRAPI 2020.
- Invited IEEE TVCG paper presented at IEEE Vis 2019.
- Invited IEEE TVCG paper presented at IEEE Vis 2021.
- Invited project presented at NetCrime—The structure and mobility of crime at NetSci 2019.

Scientific Dissemination

- **CrimAnalyzer:** *Journal USP* Link, *Journal USP* Link, *Gazeta do Povo* Link, *Saense* Link, *Portal USP São Carlos* Link.
- **MIRANTE:** *Revista Galileu - Globo* Link, *ICMC portal* Link, *CEMEAI portal Pesquisa* Link, *São Carlos Agora* Link, *Journal USP* Link.
- **CriPAV:** *Agencia FAPESP* Link, *Journal IMPA* Link, *G1 - O portal Globo* Link, *Governo de São Paulo* Link.

REFERENCES

[1] G. Garcıa-Zanabria, J. Silveira, J. Poco, A. Paiva, M. B. Nery, C. T. Silva, S. Adorno, and L. G. Nonato, "Crimanalyzer: Understanding crime patterns in são paulo," *IEEE transactions on visualization and computer graphics*, vol. 27, no. 4, pp. 2313–2328, 2019.

[2] G. Garcia-Zanabria, E. Gomez-Nieto, J. Silveira, J. Poco, M. Nery, S. Adorno, and L. G. Nonato, "Mirante: A visualization tool for analyzing urban crimes," in *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2020, pp. 148–155.

[3] G. Garcia-Zanabria, M. M. Raimundo, J. Poco, M. B. Nery, C. T. Silva, S. Adorno, and L. G. Nonato, "Cripav: Street-level crime patterns analysis and visualization," *IEEE Transactions on Visualization Computer Graphics*, no. 1, pp. 1–14, 2020.