

ANÁLISE DE DADOS CRIMINAIS ATRAVÉS DA TEORIA DOS GRAFOS

2023005290 - Amon Lemes dos Santos
2023005010 - Hiara Faustino Gonçalves
2023004775 - João Guilherme Ferreira da Silva
2023006359 - Paulo Vitor Carvalho Rodrigues
2023006537 - Pedro Reimberg de Oliveira

SMAC03 - GRAFOS

Prof. Rafael Frinhani



Análise de Dados Criminais Através da Teoria dos Grafos

1 Introdução

A cidade de Los Angeles, situada no estado da Califórnia, é uma das mais populosas dos Estados Unidos da América. No entanto, apresenta desafios significativos relacionados à segurança pública devido aos altos índices de criminalidade. Mesmo que tenha ocorrido uma melhora significativa se comparada com os anos 80 e 90, a Los Angeles Police Department (LAPD) ainda enfrenta dificuldades para controlar os crimes cometidos na sua jurisdição.

Os Índices de criminalidade e violência tendem a crescer, principalmente em determinadas regiões, como por exemplo o Centro Los Angeles, uma das localidades com maior número de crimes, especialmente na parte sudoeste. O bairro tem altas taxas de crimes violentos, com aproximadamente 941 por 100000 habitantes ([AreaVibes, 2024](#)), incluindo assaltos e roubos, além de crimes contra a propriedade, como furtos e arrombamentos. Outras regiões notáveis incluem West Adams que tem um alto índice de criminalidade sendo um dos mais altos da cidade principalmente devido à atividade de gangues ([to Travels, 2024](#)) e Skid Row. Skid Row possui taxas criminais mais altas do que a média de Los Angeles e é particularmente conhecido por ser uma área com desafios sociais intensos, com altos índices de pobreza, desabrigados e usuários de drogas, o que contribui para a ocorrência de crimes. Além disso, os crimes envolvendo armas continuam sendo uma preocupação para a população, com o aumento no número de vítimas de tiros e crimes armados, mesmo com a apreensão de armas tendo aumentado. Esse problema se agrava devido a fatores como a alta densidade populacional, atividades comerciais intensas e desafios sociais.

Nesse contexto, o LAPD mantém um extenso banco de dados que abrange incidentes criminais registrados na cidade entre os anos de 2020 e 2024, com um total de quase 1 milhão de registros. Dessa forma, com a aproximação das Olimpíadas de Verão de 2028, que serão sediadas na cidade, é de grande interesse utilizar esses dados para extrair informações, padrões e insights, a fim de auxiliar na elaboração de estratégias de prevenção aos crimes, já que a cidade enfrenta desafios contínuos em manter a segurança pública.

Com isso, é importante destacar os locais da competição para ter uma visão da segurança dessas áreas. O Los Angeles Memorial Coliseum que sediará atletismo, no bairro de Exposition Park, ao sul de Los Angeles, passou por revitalizações juntamente com seus arredores, mas ainda sim continua apresentando altas taxas de criminalidade sendo quase quatro vezes mais frequente do que em outras regiões do país e é também próximo a bairros como West Adams ([AreaVibes, 2024](#)). LA Convention Center e a Crypto.com Arena que sediarão esportes como basquete, estão bem perto um do outro e também

estão em regiões movimentadas por ser um ponto comercial e cultural, localizadas no centro de Los Angeles, ambas têm recebido atenção, mas como apresentado anteriormente, algumas áreas do centro ainda enfrentam problemas de segurança como furtos e roubos principalmente por ter bairros ao redor como Skid Row.

O Sepulveda Basin e o Lake Balboa, localizados ao norte do Vale de San Fernando, no Sepulveda Basin, a segurança varia dependendo do horário, sendo uma das preocupações relatadas da região o aumento de desabrigados em algumas partes do parque. Já no Lake Balboa, a área central do bairro possui uma maior incidência de crimes e tem uma taxa média de crimes violentos e contra a propriedade acima da média da cidade ([Johnston, 2024](#)). Além disso, ambos ficam próximos ao bairro Van Nuys, que possui uma taxa de 28% de crimes maior que a média nacional ([AreaVibes, 2024](#)).

Venice Beach, localizada no bairro de Venice, também enfrenta problemas relacionados à segurança, as taxas de criminalidade são relativamente altas em comparação com outras áreas da cidade, especialmente para os crimes contra a propriedade e crimes violentos que chegam a ser 77% mais frequente do que na média do país ([AreaVibes, 2024](#)), o elevado número de turistas e a crise de pessoas em situação de rua na região contribuem para esta situação, tornando a região um ponto vulnerável durante as Olimpíadas ([LA Times Today Staff, 2021](#)).

Outras áreas como a Universidade da Califórnia em Westwood, é considerada mais segura e com baixas taxas de criminalidade em relação às que foram mencionadas e está 36% abaixo da média dos Estados Unidos ([AreaVibes, 2024](#)). Dessa maneira, é notável a proximidade de áreas que receberão turistas e atletas com locais mais perigosos.

A proposta deste projeto é realizar uma análise eficiente do grande volume de dados criminais e organizar essas informações em um grafo, onde os nós serão representados pelos dados dos crimes, como local, características da vítima, horário, modus operandi e características da arma, e as arestas serão ponderadas de acordo com a relação de similaridade entre os nós levando em consideração os dados de cada um por exemplo, quanto mais próximo um local do outro maior o peso, quanto maior a proximidade temporal maior o peso e assim por diante comparando cada uma das características. Com isso, é possível fazer uma clusterização no grafo e será possível ter informações detalhadas de padrões temporais e geográficos para obter conclusões relevantes.

Considerando os desafios de segurança nas áreas que irão sediar as Olimpíadas, é de suma importância realizar análises que possam prevenir possíveis crimes, e a aplicação da teoria de grafos pode ser uma ferramenta muito útil para a modelagem, análise e manipulação dos dados contidos no gigante da-

taset fornecido pela LAPD que poderá auxiliar as autoridades públicas através da identificação de padrões, com a finalidade de garantir a segurança da cidade de Los Angeles durante as Olimpíadas que exigem planejamento estratégico cuidadoso. Será possível auxiliar na tomada de decisões corretas e otimizar os recursos policiais, principalmente durante eventos dessa magnitude deixando um legado de melhoria na gestão de segurança da cidade.

2 Referencial Teórico

O tema de análise de dados criminais é extensamente estudado, acumulando soluções que utilizam diversas abordagens, como o artigo sobre redes neurais (Roshankar & Keyvanpour, 2023) em que o modelo desenvolvido busca analisar dados criminais de 5 anos da cidade de Chicago utilizando coordenadas espaciais e temporais para a predição de crimes, sendo de grande importância para nossa análise já que é bem semelhante ao que procuramos fazer que é otimizar os recursos policiais e prever os crimes.

Além disso, utilizamos o artigo sobre machine learning de Das & Das (2022) em nossos estudos no qual os autores utilizam uma metodologia que faz uso de um classificador de conjunto baseado em grafos para prever crimes. O estudo cria conjuntos independentes de características dos relatórios criminais e os utiliza para construir árvores de decisão e juntamente com o classificador de conjunto o sobreajuste é reduzido. Essa pesquisa aprimora a predição criminal ao utilizar a correlação entre diferentes tipos de infração e reforça o uso de técnicas utilizando grafos para esse tipo de estratégia que busca uma maneira eficaz de resolver esse problema.

O artigo que aborda análise espacial (Faria et al., 2019) faz uso de teoria dos grafos na modelagem de distribuição de crimes violentos na cidade de Belo Horizonte para encontrar padrões úteis aos setores responsáveis pela segurança pública. A pesquisa utiliza algoritmos como o de Dijkstra, para definir rotas seguras e otimizar a alocação de recursos policiais a partir de mapas de densidade criminal.

Muitas abordagens têm como objetivo identificar pontos de interesse, levando em conta o espaço geográfico, enquanto outros têm uma análise espaço-temporal, levando em consideração o fator tempo na análise dos crimes. O CrimAnalyzer de Zanabria et al. (2021) apresenta uma ferramenta de análise visual, que auxilia o estudo de atividades criminosas na cidade de São Paulo, que tem altas taxas de criminalidade e variabilidade criminal. O CrimAnalyzer proporciona ao usuário a exploração de padrões criminais e sua evolução em relação ao tempo, auxiliando, assim, em tomadas de decisão das autoridades.

Através desta revisão da literatura sobre a análise criminal, principalmente através de grafos, foi possível entender as possibilidades e limitações das abordagens. Desse modo, o estudo que tem a estrutura de dados mais similar ao dataset de crimes da LAPD é a do CrimAnalyzer, que foca em identificar zonas de interesse, também chamadas de hotspots. Essa abordagem é a mais adequada e que melhor se encaixa no contexto deste projeto.

3 Desenvolvimento

Para que a análise de padrões de crimes na cidade de Los Angeles fosse possível, utilizamos um dataset disponível no portal de dados abertos da cidade. Dessa forma, foi possível obter detalhes sobre as atividades criminais. O conjunto de dados contém informações sobre os crimes, onde cada linha representa uma ocorrência.

Esta coleção de dados é composta por 28 colunas, sendo cada uma delas um atributo. Antes de iniciar a modelagem, foi necessário pré-processar os dados, para garantir maior qualidade e informações consistentes. Dessa maneira, várias etapas foram necessárias, como limpeza de valores nulos, transformação dos tipos dos atributos e seleção dos atributos relevantes ao projeto, resultando nas seguintes colunas:

- **Data da ocorrência:** MM/DD/AAAA;
- **Tempo da ocorrência:** Horário no formato militar de 24 horas;
- **Área:** O LAPD tem 21 delegacias de polícia comunitárias chamadas de Áreas Geográficas dentro do departamento. Essas Áreas Geográficas são numeradas;
- **Nome da Área:** As 21 Áreas Geográficas ou Divisões de Patrulha recebem uma designação que faz referência a um marco ou à comunidade pela qual são responsáveis;
- **Subárea:** Um código de quatro dígitos que representa uma subárea dentro de uma Área Geográfica;
- **Código penal do crime:** Indica o crime cometido;
- **Descrição do código penal:** Define o Código Penal fornecido; e
- **Códigos MO (Modus Operandi):** Atividades associadas ao suspeito na prática do crime.
- **Atributos da vítima:** Deixamos também todos os atributos da vítima dos crimes, para ser possível obter uma análise mais certeira.

Após a limpeza e transformação dos dados, filtrou-se os delitos. Foram selecionados e separados os crimes de acordo com o sua categoria. Sendo elas: Homicídio, Crime Sexual, Roubo, Agressões leves, Agressões graves, Vandalismo, Sequestro e Golpes. Dessa maneira, conseguimos uma melhor organização e visualização do problema, filtrando os crimes mais relevantes e concentrando em categorias mais relevantes para o objetivo do projeto. Assim, é possível realizar a modelagem utilizando as colunas mencionadas para identificar padrões e características dos incidentes criminais mais sérios e preveníveis com patrulhamento policial.

Para proporcionar uma visão mais clara sobre como os crimes se relacionam, utilizamos as ocorrências criminais como vértices. Para as arestas, foi calculado um peso final considerando múltiplos fatores relacionados às características dos incidentes, ponderando-os de acordo com sua relevância para o estudo. Esta fórmula foi definida como:

```

1 peso_final = peso_distancia * 0.25 +
  peso_horario * 0.1 + peso_crime * 0.25
  + peso_mocodes * 0.1 + peso_vitima *
  0.1 + peso_arma * 0.15 + peso_crm_cds *
  0.05

```

A seguir são apresentados detalhadamente os componentes utilizados no cálculo.

1. **Peso Distância (peso_distancia):** Avalia a proximidade geográfica entre as ocorrências com base na distância esférica entre as coordenadas. A fórmula considera a razão entre a distância calculada e o limite máximo definido de 250 metros(DISTANCIA_OCORRENCIAS):

$$\text{peso_distancia} = 1 - \frac{\text{distância_m}}{\text{DISTANCIA_OCORRENCIAS}}$$

2. **Peso Horário (peso_horario):** Mede a diferença temporal entre as ocorrências, ajustando-se com uma função exponencial que favorece horários mais próximos:

$$\text{peso_horario} = \exp\left(-\alpha \cdot \frac{\Delta t}{3600}\right)$$

Onde α é um coeficiente ajustável, e Δt é a diferença de tempo em segundos.

3. **Peso Crime (peso_crime):** Avalia a semelhança entre as categorias criminais envolvidas. Utiliza uma função de comparação específica que retorna valores normalizados.
4. **Peso M.O. (Modus Operandi) (peso_mocodes):** Representa a similaridade entre os métodos de execução dos crimes, baseada nos códigos de Modus Operandi. atribuídos às ocorrências.
5. **Peso Vítima (peso_vitima):** Compara perfis das vítimas (como idade e gênero), utilizando uma métrica personalizada:
$$\text{peso_perfil} = \text{peso_idade} \times 0.40 + \text{peso_sexo} \times 0.30 + \text{peso_descendencia} \times 0.30$$
6. **Peso Arma (peso_arma):** Mede a relação entre os tipos de armas utilizadas nos crimes, similar com o método de separação por categorias utilizadas no cálculo de peso dos crimes.
7. **Peso Crimes Secundários (peso_crm_cds):** Analisa a correspondência entre crimes secundários.

Para visualizar a solução, foram feitas simulações de grafos com amostras do dataset, a fim de validar as expectativas da modelagem. Usando o programa Gephi, foi possível visualizar o grafo criado, com os nós posicionados espacialmente de acordo com as coordenadas das ocorrências. Foram feitos testes com 100, 500, 750 e 2000 vértices e limites de distância para a criação de arestas de 500, 750 e 1000 metros.

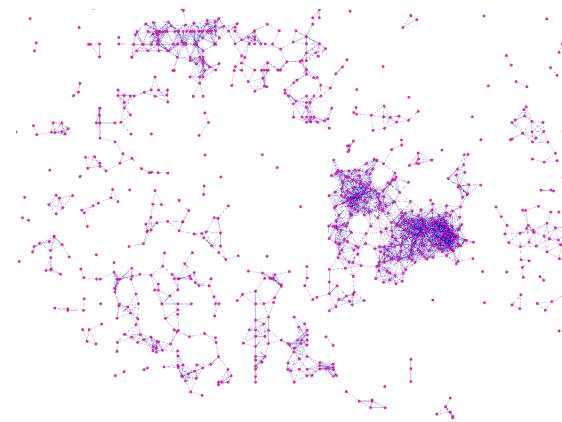


Figura 1: Parte do grafo resultante de simulação com 2000 ocorrências e 500m de limite de distância.

Esses testes iniciais mostraram que, com o aumento da quantidade de dados, um limite de distância menor é mais adequado para o objetivo do trabalho, pois mantém as arestas mais localizadas e, ao considerar que em zonas de interesse haverão muitos vértices vizinhos, arestas com distâncias maiores não adicionam informação relevante ao grafo. É importante ressaltar que, dependendo do funcionamento da solução, será necessário mudar o limite de distância para que o sistema funcione corretamente.

O método de solução proposto é a aplicação de um algoritmo de clusterização no grafo modelado. Este algoritmo serve para identificar as regiões com maior concentração de ocorrências de crimes. Cada cluster formado representa uma área prioritária para o patrulhamento.

Após a clusterização, cada cluster é percorrido como se fosse um grafo a parte, usando um algoritmo de Busca em Largura para passar por cada ocorrência, ou seja, vértice, e coletar informações de horário e tipo de crime. Os dados coletados permitem estabelecer dados gerais sobre cada zona de interesse, como os horários com mais ocorrências, probabilidade de cada tipo de crime, áreas e subáreas geográficas do local e modus operandi dos infratores.

Com os os clusters formados e seus dados obtidos, é aplicada uma interface que calcula com base numa entrada de horário e tamanho de contingente, qual a melhor distribuição de agentes de segurança entre os locais para determinado período do dia, além de informações que ajudem a preparar a força policial para o patrulhamento.

Para definir uma similaridade entre duas ocorrências é utilizado um calculo que avalia localização, data, tipo de crime, caracteristica das vitimas e a arma utilizada. A localização, por ser uma grandeza numérica, é calculada pela distância entre as duas localizações. As demais características são avaliadas por tipagem e comparação. Com o tipo de crime por exemplo, foram agrupados crimes considerados similares pelo grupo, levando em consideração a gravidade do crime. Com base nesses agrupamentos, caso o crime das duas ocorrências estiver no mesmo grupo de crimes é somado um valor ao peso final.

De tal maneira, ocorrencias (vertices) que forem ligadas por arestas com um alto valor em seu peso significa que as ocorrências são similares. Permitindo assim, refletir sobre quais tipos de crimes ocorrem na cidade de Los Angeles.

Referências

AreaVibes (2024). Areavibes - crime in los angeles neighborhoods. Acesso em: 17 nov. 2024.

Das, A. K. & Das, P. (2022). Graph based ensemble classification for crime report prediction. *Applied Soft Computing*, 125, 109215.

Faria, A., Alves, D., & Barroso, L. (2019). *APLICAÇÃO DA TEORIA DE GRAFOS E ANÁLISE ESPACIAL PARA SOLUÇÃO DE PROBLEMAS GEOGRÁFICOS: UM ESTUDO DA CRIMINALIDADE VIOLENTA NO HIPERCENTRO DE BELO*, (pp. 65–79).

Johnston, M. (2024). Is lake balboa a safe place to live? Acesso em: 17 nov. 2024.

LA Times Today Staff (2021). How will the city fix venice beach's homeless crisis? Acesso em: 17 nov. 2024.

Niche (2024). Downtown los angeles neighborhood. Accessed: Nov. 17, 2024.

Roshankar, R. & Keyvanpour, M. R. (2023). Spatio-temporal graph neural networks for accurate crime prediction. (pp. 168–173).

to Travels, D. (2024). Safety in los angeles: Complete guide. Acesso em: 17 nov. 2024.

Zanabria, G. G., Silveira, J. A., Poco, J., Paiva, A., Nery, M. B., Silva, C. T., Adorno, S., & Nonato, L. G. (2021). Crimanalyzer: understanding crime patterns in são paulo. *IEEE Transactions on Visualization and Computer Graphics*.

