Estimando a Vulnerabilidade à Exposição de Usuários em Dados de Mobilidade

Lucas G. S. Félix[1], [2], **Nadjib Achir** [2], **Anne Josiane Kouam** [3], **Aline Carneiro Viana** [2], **Jussara Marques Almeida** [1]

[1]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)

[2]INRIA-Saclay - France

[3]TU-Berlin - Germany

{lucas.da-silva-felix, nadjib, aline.viana}@inria.fr

kouam.djuigne@tu-berlin.de

jussara@dcc.ufmg.br

Abstract. *Mobility is a fundamental aspect of human life, and mobility data offers valuable insights into user behavior. Yet, this data also exposes users to privacy risks given the singularity in their trajectories. To address these concerns, techniques to quantify users' vulnerability have been proposed. However, these techniques often focus on the users spatial-temporal vulnerability, and leave aside the behavioral vulnerability of the users. Thus, we introduce the Hypercube, a novel interpretable metric for quantifying user vulnerability. This metric leverages features extracted from trajectories, modeling user behavior in a multidimensional space, to provide a comprehensive and intuitive assessment. Results shows that the hypercube effectively captures vulnerability within the behavioral space, identifying the singularity in such behavior.*

Resumo. *A mobilidade é um aspecto fundamental da vida humana, e os dados de mobilidade oferecem insights sobre o comportamento dos usuários. Contudo, esses dados também expõem os usuários a riscos de privacidade, dado a singularidade em suas trajetórias. Na literatura, técnicas foram desenvolvidas para quantificar a vulnerabilidade dos usuários. Porém, essas técnicas, no geral, focam na singularidade espaço-temporal dos usuários, deixando de lado a singularidade do comportamento dos usuários. Neste contexto, esse artigo introduz o hipercubo, uma nova métrica interpretável para quantificar a vulnerabilidade dos usuários. Para tal, o hipercubo explora métricas extraídas de trajetórias modelando o comportamento dos usuários em um espaço multidimensional. Nossos resultados mostram que o hipercubo captura a vulnerabilidade dentro do espaço das métricas comportamentais, identificando a singularidade nesse comportamento.*

1. Introdução

A mobilidade urbana é um aspecto fundamental da vida humana que influencia diversas dimensões do bem-estar individual e social [Recchi and Tittel 2023]. A mobilidade de uma pessoa funciona como uma assinatura do seu comportamento e, por meio de dados de mobilidade extraídos de telefones celulares, GPS e redes sociais é possível modelar

esse comportamento. Ao identificar os padrões de mobilidade de um indivíduo ou grupo, é possível obter *insights* sobre vários aspectos de suas vidas diárias, como os lugares que visitam, onde vivem e trabalham. Esse entendimento, quando exposto, também pode tornar os usuários vulneráveis quanto a possíveis violações de privacidade.

Na literatura, o conceito de unicidade (em inglês, *unicity*) é utilizado para representar a singularidade nos padrões de deslocamento de usuários. Este fenômeno pode ser facilmente observado em conjuntos de dados de mobilidade [Zang and Bolot 2011, Farzanehfar et al. 2021], expondo a vulnerabilidade dos usuários. Corroborando essa visão, um estudo de [De Montjoye et al. 2013] demonstrou que apenas quatro pontos espaço-temporais em um conjunto de dados de mobilidade celular são suficientes para identificar unicamente 95% dos indivíduos em uma população de 1,5 milhão de usuários.

Contudo, medir a vulnerabilidade de um usuário não é uma tarefa fácil, uma vez que seus padrões de mobilidade podem ser analisados sob diferentes perspectivas, cada uma com o potencial de revelar aspectos singulares distintos que podem colocar em risco (ou tornar vulnerável) a privacidade do usuário. Uma das técnicas mais amplamente usadas para estimar a vulnerabilidade em padrões de mobilidade foi proposta em [De Montjoye et al. 2013]. Os autores propuseram uma métrica chamada *uniqueness*, que mede o número de sequências *únicas* de lugares visitados por um usuário, ordenadas temporalmente, com base em suas trajetórias. Quanto mais singular forem os movimentos de um usuário, maior será sua propensão à exposição. Apesar de amplamente usada (e.g., [Pyrgelis et al. 2017]), a *uniqueness* tem alto custo computacional (para computar todas as possíveis combinações de pontos em uma trajetória) e oferece uma perspectiva única da vulnerabilidade do usuário, baseada apenas nas sequências de lugares visitados.

Entretanto, a mobilidade de um usuário, do ponto de vista do seus padrões de comportamento, é fundamentalmente multifacetada e, portanto, deve ser analisada sob múltiplas perspectivas. De fato, estudos anteriores modelaram a mobilidade humana utilizando diferentes métricas de mobilidade, como raio de giro (do inglês *radius of gyration*), diversidade, regularidade, estacionaridade, entre outras [Esper et al. 2024, Gonzalez et al. 2008]. Essas métricas capturam diferentes perspectivas da mobilidade e podem revelar padrões únicos de usuários que não são explicitamente capturados pela métrica de *uniqueness*. Por exemplo, o raio de giro mede a mobilidade característica de um usuário, enquanto a diversidade avalia o quão variadas são as subsequências de lugares visitados. Ao considerar múltiplas métricas conjuntamente, é possível representar o comportamento de um usuário em um espaço multidimensional. A análise de tal representação pode revelar padrões de mobilidade bem distintos que, por sua vez, podem expor um usuário à medida em que ele se destaca dos demais.

De fato, indo além da perspectiva única oferecida pela métrica de *uniqueness*, outros trabalhos exploraram conjuntos de métricas de mobilidade para avaliar a vulnerabilidade do usuário considerando conjuntamente diferentes perspectivas [Pellungrini et al. 2017, Naretto et al. 2020]. Porém estes trabalhos vinculam a definição de vulnerabilidade a modelos de ataque específicos (e.g., ataque de reidentificação), e tratam a tarefa de estimar a vulnerabilidade de um usuário como uma tarefa de classificação supervisionada baseada em aprendizado de máquina, a qual exige dados rotulados [Pellungrini et al. 2017, Naretto et al. 2020]. Assim, esses métodos são inerentemente limitados, pois dependem do conhecimento adversário, e não se adequam

a tipos de ataques imprevistos. Como resultado, a métrica de *uniqueness*, que independe de modelos específicos, permanece a mais utilizada na literatura.

Neste contexto, nossa pesquisa investiga como estimar a vulnerabilidade de um usuário em dados de mobilidade, considerando múltiplas perspectivas capturadas por diferentes métricas de mobilidade, mas sem a vinculação com modelos adversários específicos. Buscamos explorar como uma *assinatura do comportamento* em dados de mobilidade pode representar *uma nova dimensão da exposição da privacidade, amplamente negligenciada*. Por exemplo, considere um usuário com o seguinte padrão de mobilidade: (i) Nos dias de semana, desloca entre casa ao trabalho e, (ii) Aos fins de semana, fica em casa. Mesmo que esse usuário resida em um prédio densamente populado e compartilhe um local de trabalho com diversos colegas, o que torna sua sequência de locais visitados não única, seu padrão de mobilidade altamente previsível ainda poderia expô-lo. O conhecimento dessa *assinatura de comportamento*, que distingue o usuário dos demais que visitam os mesmos locais, poderia comprometer a sua privacidade.

Especificamente, nós propomos uma nova técnica para estimar a vulnerabilidade de usuários em conjuntos de dados de mobilidade chamada *hipercubo*. O hipercubo atua sobre um espaço genérico de $|M|$ dimensões, definido a partir de um conjunto de M métricas de mobilidade urbana. Este espaço visa capturar a assinatura de comportamento nos padrões de mobilidade dos usuários, representados pelo conjunto de M métricas. Especificamente, a assinatura de um usuário é representada por uma região (um *hipercubo*) neste espaço multi-dimensional. O hipercubo de um usuário é centrado no ponto definido pelos valores das M métricas computados sobre a sua trajetória, e tem o tamanho definido por uma variação máxima v em todas as dimensões. Quanto mais vizinhos um usuário tiver dentro do seu hipercubo, mais usuários têm comportamentos similares ao seu, logo menos vulnerável ele estará. Ressaltamos que o hipercubo apresenta uma alta interpretabilidade pois permite identificar as métricas que mais contribuem para que um usuário esteja mais longe de seus vizinhos (logo mais vulnerável). Além disto, comparado com a literatura, nossa técnica apresenta um baixo custo computacional, pois não explora combinações espaço-temporais (como a *uniqueness*), mas sim, um conjunto definido de métricas. Ela também não exige dados de treinamento (rotulado) nem está atrelada a modelos específicos de ataques (como [Pellungrini et al. 2017, Naretto et al. 2020]).

Nós avaliamos nossa proposta usando dois conjuntos de dados anonimizados de redes celulares com diferentes propriedades espaço-temporais e populacionais, coletados por operadores de telecomunicações nas áreas metropolitanas de Shanghai e Shenzhen, na China. Comparamos os resultados alcançados pelo hipercubo com a métrica mais usada na literatura, a *uniqueness*. Nossa análise revela usuários que, apesar de não serem únicos em termos das sequências de lugares visitados, exibem padrões comportamentais de mobilidade bastante distintos dos outros e, como tal, podem ser facilmente expostos. Estes resultados mostram como o hipercubo contribui para o estado da arte ao oferecer uma nova análise da exposição de um usuário em dados de mobilidade.

O artigo está organizado da seguinte forma: a Seção 2 apresenta a contextualização e os trabalhos relacionados. A técnica de hipercubo é descrita na Seção 3. Os conjuntos de dados e a avaliação da técnica proposta, em comparação com o estado da arte, estão nas Seções 4 e 5. Por fim, a Seção 6 traz as conclusões.

2. Contextualização

O uso de dados de mobilidade urbana tem o potencial de trazer grandes benefícios para a rotina diária das pessoas, apoiando o planejamento urbano, a avaliação da ocupação do espaço e o projeto de serviços de recomendação [Sojahrood et al. 2023, Blondel et al. 2015, Cui et al. 2016]. Neste trabalho, focamos em dados de trajetórias, onde a trajetória de um usuário é constituída por uma sequência temporalmente ordenada de locais visitados por ele.

Vale ressaltar que, os dados de mobilidade também podem revelar hábitos e interesses dos indivíduos, como estilos de vida, visitas a lugares importantes ou sensíveis (por exemplo, casa, trabalho, hospitais), filiações políticas ou religiosas. A possibilidade de exposição dessas informações sensíveis levanta preocupações significativas sobre a privacidade dos indivíduos, representando uma grande barreira para compartilhamento de conjuntos de dados de mobilidade [Pyrgelis et al. 2020].

De fato, já foi mostrado que 50% das trajetórias extraídas de um conjunto de dados de telefonia móvel de 25 milhões de usuários eram únicas ao considerar os três lugares mais frequentemente visitados [Zang and Bolot 2011]. Por outro lado, em um conjunto de dados de 60 milhões de usuários, o risco de exposição individual estimado chegou a 93% usando 4 pontos espaço-temporais diferentes [Farzanehfar et al. 2021]. Na prática, estes resultados sugerem que um conhecimento mínimo sobre a mobilidade de um indivíduo é suficiente para expor dados sensíveis de um usuário em dados de mobilidade. Esta vulnerabilidade à exposição se deve à *singularidade ou unicidade intrínseca da mobilidade humana*, geralmente influenciada por diversos fatores de natureza geográfica, temporal e social [Recchi and Tittel 2023].

Vulnerabilidade em mobilidade é um conceito que não é bem estabelecido, podendo admitir múltiplas definições. Neste trabalho, vinculamos a definição de vulnerabilidade ao conceito de singularidade de padrões (como em [De Montjoye et al. 2013]) e definimos: *um usuário é considerado vulnerável se ele tiver um comportamento distinto (i.e., único) dos outros usuários ao seu redor, de forma que possa ser facilmente destacado do grupo. Isto é, neste trabalho, vulnerabilidade refere-se ao grau com que os padrões de mobilidade individuais, extraídos das trajetórias dos usuários, podem ajudar a distinguir um indivíduo do resto.* A seguir discutimos esforços anteriores de avaliar a vulnerabilidade individual em dados de mobilidade.

Abordagens prévias: No trabalho [De Montjoye et al. 2013], os autores introduziram a métrica *uniqueness*, que estima a vulnerabilidade à exposição de um usuário a partir do cálculo de todas as possíveis subsequências ordenadas de locais visitados extraídas da sua trajetória. Ela é calculada em janelas de tempo $|t_w|$: uma trajetória é considerada única se a subsequência associada a qualquer janela for única. *Uniqueness* é dada pela frequência com que uma subsequência ordenada de lugares aparece; valores mais baixos indicam maior exposição. Usuários com *uniqueness* igual a 1 são os mais vulneráveis. Neste sentido, a noção de vulnerabilidade por trás do conceito de *uniqueness* é o mesmo adotado neste trabalho. Porém, medir todas as subsequências existentes é computacionalmente caro. Estratégias para reduzir custos consideram apenas uma parte das subsequências possíveis [Pyrgelis et al. 2017]. Como definido, a *uniqueness* e suas adaptações estimam a vulnerabilidade do usuário à exposição a partir dos dados brutos de trajetória, os quais mapeiam a mobilidade do usuário em um espaço bidimensional de espaço versus tempo.

Por sua vez, outros trabalhos propuseram avaliar a exposição do usuário extraindo um conjunto de métricas a partir das visitas espaço-temporais [Pellungrini et al. 2017, Naretto et al. 2020], capturando assim uma perspectiva multidimensional da mobilidade. Em comum, esses estudos mapearam o problema em uma tarefa de classificação baseada em aprendizado de máquina supervisionado, a qual exige dados rotulados para o treinamento do modelo. A rotulação de dados quanto à vulnerabilidade é uma tarefa custosa. Para lidar com isto, os autores definiram a vulnerabilidade de um usuário como a probabilidade dele ser re-identificado por um modelo de ataque. Especificamente, em [Pellungrini et al. 2017], os autores mediram a vulnerabilidade de um usuário em relação a 9 tipos de ataque. Eles treinaram um modelo de aprendizado de máquina baseado em árvores, usando um conjunto de métricas de mobilidade como atributos de entrada, para prever a probabilidade de re-identificação por um desses modelos. Quanto maior a probabilidade, maior a vulnerabilidade do usuário. Já em [Naretto et al. 2020], os autores propõem *EXPERT*, um arcabouço que explora métricas de mobilidade para classificar um usuário em vulnerabilidade alta ou baixa com relação a um ataque de sequência de localizações. Em comum, estas estratégias vinculam o conceito de vulnerabilidade a modelos adversários específicos, que, por sua vez, dependem do cenário ou aplicação [Gadotti et al. 2024, Wagner and Eckhoff 2018]. Tal dependência pode oferecer uma perspectiva distorcida de proteção do usuário, pois ataques no mundo real podem ser diferentes daqueles modelados. Além disso, assim como o *uniqueness*, estas abordagens têm alto custo computacional.

Em comparação, a estratégia aqui proposta tem baixo custo computacional, alta interpretabilidade, além de independência de modelos de ataques específicos e de dados rotulados. Mais ainda, a proposta é genérica e, ao considerar simultaneamente múltiplas métricas derivadas da trajetória do usuário, captura conjuntamente múltiplas perspectivas dos padrões de mobilidade, Por fim, a proposta pode ser adaptada facilmente aos dados disponíveis, a partir da introdução de diferentes conjuntos de métricas deles derivados.

3. Hipercubo de vulnerabilidade

Nesta seção nós apresentamos a nossa técnica, chamada *hipercubo*, para estimar a vulnerabilidade de usuários, considerando seus padrões de mobilidade e o de outros usuários ao seu redor. Como em trabalhos anteriores, a técnica opera sobre dados de mobilidade definidos por um conjunto de usuários $U = \{1, 2, \ldots, i, \ldots, N\}$, no qual cada usuário i é associado a uma trajetória Φ_i. Cada trajetória Φ_i, por sua vez, é composta por tuplas (i, t, l), onde i se refere ao identificador do usuário, t ao instante de observação e l à localização (par latitude e longitude) do usuário i em t.

A técnica é instanciada a partir da seleção de um conjunto M de métricas, $M = \{m_1, m_2, \ldots, m_r, \ldots, m_{|M|}\}$, que podem ser computadas nos dados de entrada e que capturam diferentes perspectivas do comportamento de mobilidade dos usuários expressas nos dados. O conjunto de métricas é usado para definir um espaço de $|M|$ dimensões, em que cada métrica m_r corresponde a uma dimensão. Assim, cada usuário $i \in U$ é representado por um vetor $(m_1^i, m_2^i, \ldots, m_{|M|}^i) \in \mathbb{R}^{|M|}$, onde cada componente do vetor corresponde ao valor da métrica m_r^i daquele usuário, computado sobre sua trajetória. Neste cenário, *nossa hipótese é que usuários que possuem um comportamento distinto estarão isolados no espaço de métricas definido, podendo assim ser mais facilmente distinguidos dos demais usuários em U e logo estarão mais vulneráveis à exposição.*

Para estimar a vulnerabilidade de um usuário alvo i (e assim identificar os mais vulneráveis), nós propomos uma técnica que mensura a quantidade de *vizinhos* a i no espaço de métricas, considerando uma região de similaridade neste espaço definida por um parâmetro $v \in \mathbb{R}_+$. O termo *vizinho* é usado para referenciar usuários com um comportamento muito similar (com relação às métricas M) ao usuário i. A região de similaridade de i é definida como um *hipercubo* centrado no ponto $(m^i_1, m^i_2, \ldots, m^i_{|M|})$ (i.e., localização de i no espaço de métricas) e cujo lado na dimensão r é dado por $m^i_r \pm v \times m^i_r$. Ou seja, v pode ser interpretado como um limite na variação do comportamento de um usuário (capturado por cada métrica em M) e representa a diferença relativa máxima permitida para que um outro usuário j seja considerado similar a i com relação a qualquer métrica considerada.

Nosso objetivo é identificar para cada usuário i o número de $vizinhos$ $N_i \in \mathbb{Z}_+$, ou seja, usuários cuja representação no espaço de métricas esteja localizada dentro do hipercubo de i. Tais usuários têm um comportamento muito similar (controlado por uma diferença relativa máxima v) a i no que tange as métricas em M. Assim, quanto menor N_i, mais vulnerável o usuário i está. A Figura 1 ilustra em um espaço de $|M|$ = 3 (X, Y, e Z) como a vizinhança é mensurada no hipercubo para um usuário alvo i. Em amarelo temos os usuários que estão dentro do hipercubo de i, em verde caso contrário. Perceba que o hipercubo é construído ao redor de um único parâmetro v, o qual é simples e interpretativo, visto que pode ser visto como a variação aceita no comportamento para que outros usuários façam parte da vizinhança do usuário.

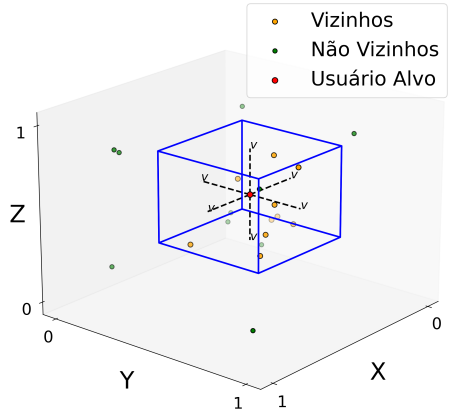


Figura 1. Hipercubo de variação do usuário alvo i.

Para encontrar N_i, primeiramente, nós identificamos os vizinhos de i considerando cada dimensão separadamente. Para tal, para cada métrica m^r e cada outro usuário $j \in U$, definimos a variável binária $X^r_{i,j}$ para indicar se j tem um comportamento muito similar a i com relação à métrica m^r, ou seja:

$$X^r_{i,j} = \begin{cases} 1, \ se \ m^r_i - v \times m^r_i \leq m^r_j \leq m^r_i + v \times m^r_i \ \forall \ j \ \in \ U \\ 0, \ \text{caso contrário} \end{cases} \tag{1}$$

Em seguida, nós identificamos os usuários que são vizinhos a i considerando *todas as dimensões* conjuntamente, definindo a variável $A_{i,j}$ como:

$$A_{i,j} = \begin{cases} 1, \ se \ \sum_{r=0}^{|M|} X_{i,j}^{r} = |M| \ \forall \ j \in \ U \\ 0, \ \text{caso contrário} \end{cases} \tag{2}$$

O número de vizinhos de i, N_i é definido como $N_i = \sum_{j=0}^{|U|} A_{i,j} - 1$. Subtraímos 1 do somatório para não incluir o próprio i na contagem.

Considerando uma estimativa binária de vulnerabilidade, como em [De Montjoye et al. 2013], usuários com N_i=0 seriam considerados vulneráveis, enquanto usuários com pelo menos um vizinho já poderiam ser considerados protegido, dado que seria difícil distinguir seus comportamentos dos vizinhos mais próximos. Porém, a técnica proposta pode ser utilizada também para prover estimativas contínuas de vulnerabilidade, as quais podem auxiliar a direcionar melhor técnicas de proteção, permitindo assim uma maximização de proteção e utilidade dos conjuntos de dados. Neste cenário, propomos estimar a vulnerabilidade do usuario i, V_i como inversamente proporcional ao seu número de vizinhos, isto é:

$$V_i = \frac{1}{N_i + 1} \tag{3}$$

Usuários totalmente isolados (i.e., $N_i = 0$), terão uma vulnerabilidade igual a 1. A vulnerabilidade reduz à medida que o usuário possui mais vizinhos.

Até então, nós não definimos o conjunto de métricas M. De fato, a técnica de hipercubo é genérica e pode ser aplicada para qualquer conjunto M. Especificamente neste trabalho, nós consideramos $|M| = 13$ métricas amplamente utilizadas na literatura para caracterizar os padrões de mobilidade de usuários de acordo com diferentes perspectivas [Esper et al. 2024]. Elas são categorizadas em três grupos:

- **Métricas espaciais:** `Raio de Giro (RG)`, `2-RG` (RG dos dois locais mais visitados), `Distância Máxima` (Maior deslocamento histórico), `Média e desvio padrão do tamanho dos saltos` (i.e. distância entre locais visitados sucessivamente), `número de visitas e número de locais distintos`, as quais são dependentes apenas de informações geográficas.
- **Métricas temporais:** `Média e desvio padrão do tempo de espera` (tempo médio entre registros); as quais dependem apenas do tempo.
- **Métricas estruturais:** `diversidade, regularidade, estacionariedade e entropia`, as quais são dependente de informações geográficas ordenadas no tempo

Note que essas métricas capturam rotinas espaço-temporais, preferências de mobilidade, e incerteza no comportamento humano (vide definição em [Esper et al. 2024]), assim, diferentes combinações de métricas podem capturar juntas usuários que estão isolados.

Por fim, ressaltamos que uma das vantagens do hipercubo de vulnerabilidade é sua alta interpretabilidade. Note que é possível recuperar individualmente a diferença relativa em cada métrica do usuário i para seu vizinho mais próximo. A análise destas diferenças permite entender que padrão de comportamento (capturado por uma ou mais métricas) está contribuindo mais para a vulnerabilidade de i.

Conjunto de dados	# Usuários	# Locais	# Dias	Média de locais por usuário	Média de locais por usuário por dia	Média de locais por hora por dia	Média de locais dif. por usuário	Média de locais dif. por usuário por dia
Shanghai	58500	6322	10	180,37	18,03	1	12,48	3,38
Shenzhen	41159	814	15	158,33	10,74	1,26	7,81	3,20

Tabela 1. Estatísticas dos dados após pré-processamento.

4. Estudo de caso - Conjuntos de Dados

Avaliamos a técnica de hipercubo em dois conjuntos de dados anonimizados de telefonia móvel (*Call Record Details* - CDR), denominados *Shanghai* e *Shenzhen*. Eles foram escolhidos por sua relevância na literatura e características distintas. *Shanghai* contém trajetórias individuais com granularidade temporal de cerca de uma hora, enquanto *Shenzhen* registra posições apenas durante eventos, como chamadas ou mensagens, resultando em maior esparsidade. A Tabela 1 mostra um sumário dos dois conjuntos de dados.

Diferentemente de *Shanghai*, *Shenzhen* não informa a data associada a cada evento, apenas a hora. Assim, eventos no mesmo dia são identificados com base no horário do último evento (como feito em [Esper et al. 2024]). *Shanghai* abrange 10 dias de dados, com a mesma quantidade de usuários ao longo dos dias. Em contraste, *Shenzhen* contém trajetórias de até 949 dias, mas avaliamos apenas os primeiros 15 dias, devido à grande redução de usuários após esse período. Ambos os conjuntos passaram por pré-processamento para mitigar vieses causados pela esparsidade. Uma tesselação de 200×200 m baseada no *OpenStreetMap* foi aplicada, e usuários inativos (menos de 10 dias e 100 registros) foram removidos. Para preencher lacunas temporais, empregamos uma estratégia de completude de dados que preenche lacunas em torno dos locais de casa e trabalho [Esper et al. 2024].

Cidade	Raio de Giro (km)	2-Raio de Giro (km)	Distância Máxima (km)	# de Visitas	Média de Saltos (km)	DP Saltos (km)	# Locais Distintos
Shanghai	$3{,}52 \pm 0{,}031$	$1{,}896 \pm 0{,}026$	$17{,}309 \pm 0{,}126$	$180{,}37 \pm 0{,}058$	$1{,}131 \pm 0{,}01$	$3{,}184 \pm 0{,}026$	$12{,}484 \pm 0{,}061$
Shenzhen	$1{,}557 \pm 0{,}025$	$0{,}998 \pm 0{,}021$	$4{,}906 \pm 0{,}071$	$157{,}971 \pm 0{,}423$	$0{,}948 \pm 0{,}014$	$1{,}386 \pm 0{,}022$	$7{,}816 \pm 0{,}067$

(a) Média e IC (95%) das métricas Espaciais,

Cidade	Média de Tempo (H)	DP Tempo (H)
Shanghai	$1{,}332 \pm 0{,}00$	$0{,}63 \pm 0{,}001$
Shenzhen	$2{,}301 \pm 0{,}005$	$4{,}071 \pm 0{,}013$

(b) Média e IC (95%) das métricas Temporais,

Cidade	Entropia	Estacionaridade	Regularidade	Diversidade
Shanghai	$1{,}036 \pm 0{,}004$	$0{,}725 \pm 0{,}001$	$0{,}93 \pm 0{,}0$	$0{,}852 \pm 0{,}002$
Shenzhen	$1{,}083 \pm 0{,}005$	$0{,}543 \pm 0{,}002$	$0{,}95 \pm 0{,}0$	$0{,}843 \pm 0{,}002$

(c) Média e IC (95%) das métricas Estruturais,

Tabela 2. Média e IC (95%) das métricas por grupo

As Tabelas 2a, 2b e 2c, mostram a média e o Intervalo de Confiança (IC) de 95% para várias das métricas utilizadas divididas em três grupos: Métricas espacias, temporais e estruturais, para os dois conjuntos de dados (as demais métricas foram omitidas por restrições de espaço). Essas métricas caracterizam o comportamento dos usuários em cada cenário. Por exemplo, em Shanghai, que é uma cidade maior, os usuários tendem a se deslocar e visitar mais locais distintos, como é possível observar pelas métricas espacias (`Raio de Giro`, `2-Raio de Giro`, `Distância Máxima`), possuindo

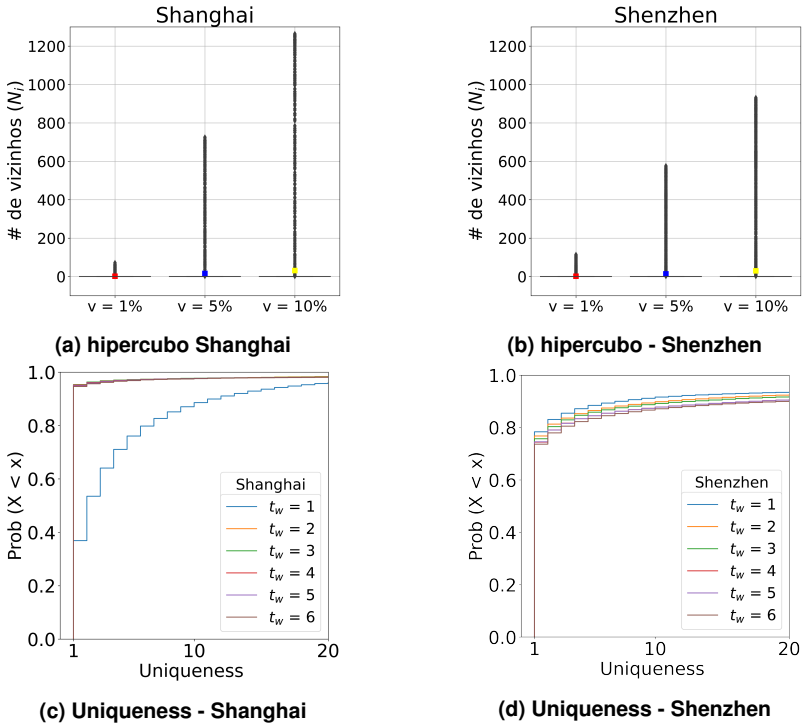


Figura 2. As Figuras (a) e (b) mostram a distribuição e média (pontos coloridos) da quantidade de vizinhos em diferentes valores de variação de comportamento v. As Figuras (c) e (d) *Uniqueness* em diferentes janelas de tempo de agregação (t_w)

uma menor regularidade. Porém, os usuários de Shanghai são mais estacionários, o que é parcialmente justificado por haver mais registros por usuário, o que permite capturar mais o tempo desses usuários em locais de longa estadia (e.g., casa e trabalho). Neste cenário, a coleta tem também um impacto na diversidade dos usuários, que apesar de visitarem mais locais diferentes que os usuários de Shenzhen, são mais diversos em períodos espacialmente mais distantes, enquanto os usuários de Shenzhen tendem a realizar mais deslocamentos entre cada registro coletado. Isso se reflete também na `Entropia` observada, que captura a previsibilidade no movimento dos usuários. Estas diferenças entre os dois conjuntos de dados permitem avaliar a técnica proposta em cenários diversos.

5. Resultados

5.1. Hipercubo

As Figuras 2a, 2b mostram a distribuição da quantidade de vizinhos dos usuários em Shanghai e Shenzhen utilizando a técnica do hipercubo em diferentes valores de variação do comportamento v. Vale destacar que, quanto menor a quantidade de vizinhos, mais vulnerável é o usuário. Chamamos o conjunto de usuários com quantidade de vizinhos igual a 0 de $Hiper^*$. Neste cenário, note que consideramos variações no comportamento de 1%, o que significa uma pequena variação, 5%, a qual é consideramos uma variação média, e 10% a qual considera vizinhos que estão mais distantes em seu comportamento. Note que ao aumentar o valor de v há uma redução na porcentagem de usuários com o número de vizinhos igual a 0, o que é é justificado pelo fato da variação igual a 1% ser muito restritiva, o que muda ao realizar o relaxamento dessa variável.

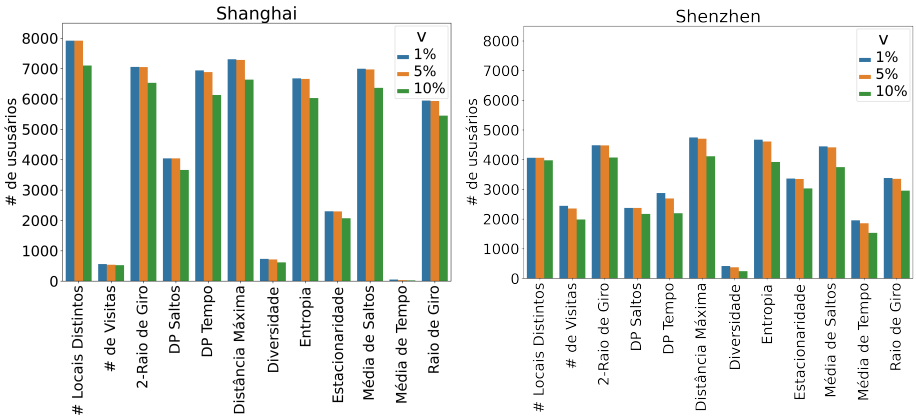
As Figuras 3, 4, mostram para o conjunto de usuários vulneravéis ($Hiper^*$), a variação média da métrica mais distante de cada usuário (Fig.3), e complementarmente, a quantidade de vezes que cada métrica foi considerada a como a mais distante (Fig.4). Neste cenário, quanto maior a quantidade de vezes que uma métrica foi considerada a mais distante, mais importante ela é para caracterizar a vulnerabilidade observada no conjunto de dados. Ao analisarmos as Figuras, o que notamos é que em Shanghai e Shenzhen grande parte da população (mais de 80%) são capturados em diferentes valores de v. Isso mostra que no geral os usuários possuem uma variação do comportamento maior que 10% para os vizinhos mais próximos, o que indica uma alta unicidade comportamental. É notável que ao aumentar o valor de v a variação nas métricas aumenta, visto que usuários capturados agora apresentam uma distância maior para seu vizinho mais próximo. Neste cenário, as métricas apresentam uma variação média entorno de 16%, em Shanghai e 17% em Shenzhen, sendo que essa diferença pode ser atribuída ao diferente método de coleta realizado em cada um dos conjunto de dados. Shenzhen, a qual coleta dados por eventos, acentua a diferença entre os usuários, visto que sua amostragem varia de acordo com a utilização de serviços móveis por parte do usuário. Note que a Regularidade apresenta importância nula em ambos cenários, o que é justificado pela sua baixíssima variação, como observado na Tabela 2c.

Agora, observando as diferenças entre as cidades, notamos que em Shanghai as métricas de `Número de Locais Distintos`, `Distância Máxima`, e `2-Raio de Giro` são as aquelas que estão mais distantes do vizinho mais próximo, enquanto em Shenzhen, percebemos um padrão diferente, tendo métricas como `Distância Máxima`, `Entropia`, e `2-Raio de Giro` como aquelas que possuem mais vezes a maior variação.

O `Número de Locais Distintos`, o qual visa capturar o quanto o usuário explora novos locais é a variável mais afastada do vizinho mais próximo em Shanghai. Isso pode ser justificado pelas diferentes características apresentadas por usuários dentro da mesma cidade. Enquanto alguns são mais rotineiros, outros preferem explorar mais essas cidades. Essa variação por sua vez impacta em outras métricas como `Distância Máxima` e `2-Raio de Giro`, as quais justificadas também pelos diferentes padrões de deslocamento espacial dos usuários. Neste cenário os vizinhos mais próximos, apesar de ter um comportamento no geral semelhante, tendem a visitar locais mais afastados, variando do padrão de visita casa-trabalho, consequentemente impactando na métrica de `2-Raio de Giro`.

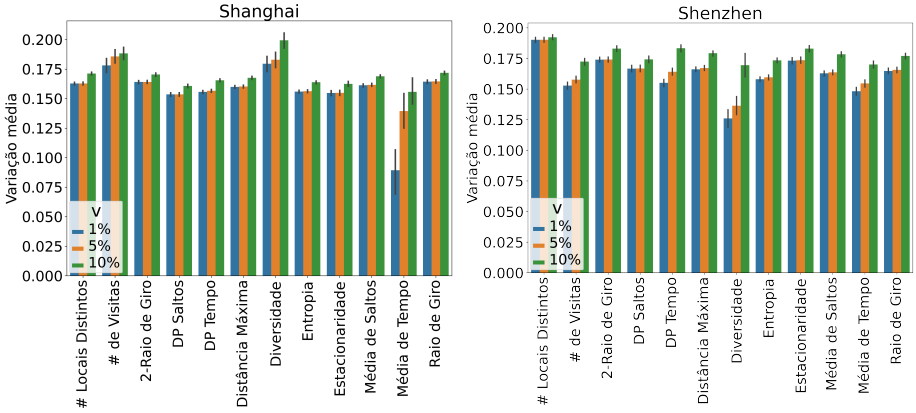
Já em Shenzhen, o que observamos é que a `Distância Máxima`, é a dimensão que, no geral, está mais distante, juntamente com a `Entropia` e `2-Raio de Giro`. A primeira métrica, que captura *outliers* no deslocamento espacial dos usuários, é altamente influenciada pelo comportamento espacial do usuário, sendo justificada pelo fato que por mais que nas atividades cotidianas o vizinho mais próximo possua um comportamento semelhante, em períodos que fogem ao padrão (e.g. finais de semana, feriados) os usuários possuem visitas únicas a locais mais distantes, a qual distancia o comportamento desses usuários. Por sua vez, a `Entropia` e o `2-Raio de Giro` são influenciados pela coleta de dados realizado em Shenzhen. Neste cenário, a `Entropia` é altamente dependente da regularidade em locais visitados, enquanto o `2-Raio de Giro` é dependente da regularidade da re-visitas em locais como casa e trabalho. Neste cenário, em casos

de coletas muitas variações nas coletas, tais locais podem ser erroneamente identificados impactando assim no comportamento modelado.



(a) Métricas com maior variação - Shanghai **(b) Métricas com maior variação - Shenzhen**

Figura 3. # de usuários para cada métrica mais distante do vizinho



(a) Variação das métricas em Shanghai **(b) Variação das métricas em Shenzhen**

Figura 4. Média e IC para a métrica mais distante do vizinho

5.2. Uniqueness

As Figuras 2c, 2d mostram a distribuição da *uniqueness* dos usuários em Shanghai e Shenzhen em diferentes janelas temporais de agregação. Vale destacar que, quanto menor o valor de *uniqueness*, mais vulnerável é o usuário. Se o usuário tiver uma *uniqueness* igual a 1, isso significa que, em algum momento, apenas esse usuário realizou aquela sub-trajetória, sendo, portanto, é possível reidentifica-lo. Chamamos o conjunto de usuários com *uniqueness* igual a um de $Uniq^*$. Observe que em Shanghai (Fig 2c), quando $t_w = 1$, apenas 40% dos usuários são vulneravéis. Isso é justificado pela coleta de dados que é realizada, em média, a cada hora, o que torna a quantidade pequena a quantidade de combinações avaliadas quanto $t_w = 1$, assim, reduzindo a probabilidade de identificar usuários vulneravéis. Apesar de ser uma cidade consideravelmente menor, Shenzhen (Fig. 2d) apresenta para $t_w = 1$ uma alta vulnerabilidade dada a grande unicidade temporal de seus usuários, dado a sua coleta que é gerada por eventos. Em ambos cenários, é observável na Figura 2 que um platô é alcançado em ambas as distribuições quando

$t_w = 4$, indicando que padrões semelhantes de vulnerabilidade são capturados quando $t_w \geq 4$. Logo, para todas as nossas investigações subsequentes, utilizamos $t_w = 4$.

5.3. hipercubo x Uniqueness

	Shanghai			Shenzhen		
V	$Uniq^* - Hiper^*$	$Uniq^* \bigcap Hiper^*$	$Hiper^* - Uniq^*$	$Uniq^* - Hiper^*$	$Uniq^* \bigcap Hiper^*$	$Hiper^* - Uniq^*$
1%	0,00	0,98	0,02	0,00	0,78	0,22
5%	0,00	0,98	0,02	0,00	0,79	0,21
10%	0,09	0,9	0,01	0,06	0,79	0,15

Tabela 3. % de usuários vulneráveis capturados em diferentes subconjuntos

Nesta seção comparamos a técnica proposta do hipercubo, com a técnica da *uniqueness*. Neste cenário avaliamos os grupos de usuários capturados como vulneravéis por ambas as técnicas por meio da análise de conjuntos. Os usuários capturados apenas pela *uniqueness* ($Unique^* - Hiper^*$), os usuários capturados por ambas as técnicas ($Unique^* \bigcap Hiper^*$) e os usuários capturados apenas pelo hipercubo ($Hiper^* - Unique^*$). A Tabela 3 mostra a porcentagem de usuários capturados nos diferentes subgrupos em diferentes valores de variação, para Shanghai e Shenzhen.

Primeiramente, observe que a grande interseção entre os usuários capturados pelo hipercubo e a *uniqueness*. O mínimo de interseção capturado para ambas as cidades é de 79%, o que mostra que os usuários que possuem unicidade espaço-temporal, também apresentam grande unicidade comportamental. Isso também corrobora para o fato que o hipercubo consegue, de fato, capturar usuários que são vulneravéis.

Usuários capturados apenas pela *uniqueness*: Ao avaliarmos os usuários capturados apenas pela *uniqueness* ($Unique^* - Hiper^*$), percebemos que a porcentagem de usuários desse grupo é de no máximo 9% (Shanghai quando $v = 10\%$). Avaliando esses usuários no cenário de $v = 10\%$, visto que este é o único cenário onde $|Unique^* - Hiper^*| \neq 0$, percebemos que os usuários desse grupo em Shanghai apresentam em média $1,6$ vizinhos, e $1,2$ vizinhos em Shenzhen. Isso mostra que mesmo usuários que possuem padrões espaço-temporais únicos podem ter um padrão comportamental que o camufla em meio a multidão em certos valores de v, logo a escolha do valor de v é crucial para capturar os usuários que apresentam um comportamento vulnerável. Ilustrando este cenário, vemos usuários que visitam apenas um local, i.e., são altamente estacionários, mas que estão em locais isolados, que não são visitados por mais nenhum usuário. Neste cenário, esse usuário possui uma *uniqueness* igual a 1, porém, eles não são vulneravéis pelo comportamento, visto que há outros usuários que possuem o mesmo comportamento de alta estacionariedade. Logo, esses usuários de alta estacionariedade estão próximos uns aos outros, fornecendo proteção comportamental.

Usuários capturados apenas pelo hipercubo: A análise dos usuários capturados exclusivamente pelo hipercubo ($Hiper^* - Unique^*$) revela que o comportamento do usuário é um indicativo mais robusto da vulnerabilidade, visto a capacidade de identificar um maior número de usuários vulneráveis. De fato, o que percebemos ao avaliar os gráficos de Shanghai e Shenzhen é que no cenário mais permissivo ($v = 10\%$), o hipercubo é capaz de capturar no mínimo 1% e 15%, em Shanghai e Shenzhen, respectivamente, de usuários que não foram capturados pela *uniqueness*. Essa diferença na captura em cada conjunto de dados é justificada pelo fato que Shanghai é uma cidade maior, onde usuários tendem

a ser mais ativos, tendo um deslocamento espaço-temporal maior e mais único. Consequentemente, isso aumenta a interseção entre as duas técnicas, e reduz a porcentagem capturada de maneira exclusiva. Já em Shenzhen, que é uma cidade espacialmente mais restrita, os usuários possuem um comportamento espaço-temporal mais similar, visto que há menos locais para serem visitados, consequentemente tendo menos padrões singulares de mobilidade, reduzindo a quantidade de usuários capturados pela *uniqueness*, e a interseção entre ambas essas técnicas. Por sua vez, é possível ver que esses usuários são singulares em seu comportamento, o que consequentemente deixam eles expostos. Isso corrobora para que o comportamento seja avaliado juntamente com os padrões espaço temporais para a captura de padrões de vulnerabilidade. Ilustrando esse cenário, percebemos que os usuários que são vulneravéis pelo comportamento mas não vulneravéis pela *uniqueness*, são usuários mais estacionários que visam uma quantidade mais restrita de locais, porém não são completamente estacionários. Ao visitar uma quantidade de locais reduzida, estes usuários se expõe menos a padrões espaço-temporais únicos, porém, possuem um comportamento singular comparado aos seus vizinhos.

Em conclusão, o hipercubo é uma técnica que captura como uma perspectiva diferente sobre a vulnerabilidade dos usuários, logo, a ideia pode ser vista como um complemento de outras técnicas para mensurar a vulnerabilidade como a *uniqueness*. Em algum cenários, como visto em Shanghai, é o comportamento espaço-temporal pode ser considerado indicativo suficiente para mensurar a vulnerabilidade, visto que há no máximo apenas 9% de usuários que não são capturados pela *uniqueness*. Porém, em cenários como o de Shenzhen, onde os usuários se deslocam por locais similares, vemos que o comportamento é um melhor indicativo da vulnerabilidade, visto que essa, por sua vez, é capaz de capturar uma maior quantidade de usuários em um espaço multidimensional que modela diferente características da mobilidade do usuário.

6. Conclusões

Neste trabalho, propomos uma nova métrica para quantificar a vulnerabilidade em conjuntos de dados de mobilidade chamada **hipercubo**. O **hipercubo** modela através de um espaço de métricas de mobilidade, o comportamento de um usuário i para seus vizinhos, quantificando essa variação comportamental v. Tais métricas permitem que o **hipercubo** capture a unicidade no comportamento dos usuários sob diferentes perspectivas das que são capturadas por outras métricas na literatura. O **hipercubo** possui como características a **alta interpretabilidade, independência de conhecimento adversário, baixo custo computacional, e a captura de diferentes perspectivas de vulnerabilidade**. Assim, o **hipercubo** permite que pesquisadores quantifiquem a exposição geral em seus conjuntos de dados e abordem adequadamente a vulnerabilidade de acordo com as características e o nível de exposição.

Avalizamos o **hipercubo** com dois conjuntos de dados de CDR coletados nas cidades de Shanghai e Shenzhen. Os usuários em ambos os conjuntos de dados apresentam conjuntos diferentes de características, o que nos permite analisar adequadamente o **hipercubo** em cenários distintos. Nessa avaliação, foi possível perceber que os padrões de alta vulnerabilidade são diferentes nessas cidades. Além disso, também realizamos uma comparação entre o **hipercubo** e a *uniqueness*. Nossos resultados mostram que conseguimos capturar, não apenas, padrões semelhantes aos da *uniqueness*, com um custo computacional menor, mas também capturar novos usuários que apresentam alta unici-

dade em seu comportamento, mas não em suas subsequências de visitas. Como trabalhos futuros, pretendemos avaliar métricas distintas que sejam capazes de capturar perspectivas adicionais do comportamento, avaliando também outras bases de dados, em especial baseados em GPS, o que pode reduzir qualquer viés na coleta.

7. Agradecimentos

Este trabalho faz parte do projeto Mob Sci-Dat Factory (ANR-23-PEMO-0004) no âmbito do programa France 2030. Ele foi apoiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pela Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e pelo projeto CAPES-STIC-AMSUD 22-STIC-07 LINT.

Referências

Blondel, V. D., Decuyper, A., and Krings, G. (2015). A survey of results on mobile phone datasets analysis. *EPJ data science*, 4:1–55.

Cui, J., Liu, F., Janssens, D., An, S., Wets, G., and Cools, M. (2016). Detecting urban road network accessibility problems using taxi gps data. *Journal of Transport Geography*.

De Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., and Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3(1):1–5.

Esper, J. P., Viana, A. C., and Almeida, J. M. (2024). Beauty or beast: Human behavioral insights and learning power of federated mobility prediction. In *ACM SIGSPATIAL*.

Farzanehfar, A., Houssiau, F., and de Montjoye, Y.-A. (2021). The risk of re-identification remains high even in country-scale location datasets. *Patterns*, 2(3).

Gadotti, A., Rocher, L., Houssiau, F., Creţu, A., and de Montjoye, Y. (2024). Anonymization: The imperfect science of using data while preserving privacy. *Science Advances*.

Gonzalez, M. C., Hidalgo, C. A., and Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *nature*, 453(7196):779–782.

Naretto, F., Pellungrini, R., Monreale, A., Nardini, F. M., and Musolesi, M. (2020). Predicting and explaining privacy risk exposure in mobility data. In *ICDS*.

Pellungrini, R., Pappalardo, L., Pratesi, F., and Monreale, A. (2017). A data mining approach to assess privacy risk in human mobility data. *ACM TIST*.

Pyrgelis, A., Troncoso, C., and De Cristofaro, E. (2017). Knock knock, who's there? membership inference on aggregate location data. *arXiv preprint arXiv:1708.06145*.

Pyrgelis, A., Troncoso, C., and De Cristofaro, E. (2020). Measuring membership privacy on aggregate location time-series. *ACM Measure. and Analysis of Computing Systems*.

Recchi, E. and Tittel, K. (2023). The empirical study of human mobility: Potentials and pitfalls of using traditional and digital data. In *Comp. Social Science for Policy*.

Sojahrood, Z. B., Taleai, M., and Cheng, H. (2023). Hybrid poi group recommender system based on group type in lbsn. *Expert Systems with Applications*, 219:119681.

Wagner, I. and Eckhoff, D. (2018). Technical privacy metrics: a systematic survey. *ACM Computing Surveys (Csur)*, 51(3):1–38.

Zang, H. and Bolot, J. (2011). Anonymization of location data does not work: A large-scale measurement study. In *Inter. conf. on Mobile computing and networking*.