Re-identificação de trajetórias de veículos baseada na caracterização das preferências de caminho

Ekler Paulino de Mattos[1,2], Augusto C. S. A. Domingues[1], Antonio A. F. Loureiro[1] *

[1]Departamento de Ciência da Computação - Universidade Federal de Minas Gerais
Belo Horizonte, MG

[2]Universidade Federal de Mato Grosso do Sul - Campus de Coxim
Coxim, MS

{ekler.mattos, augusto.souza, loureiro}@dcc.ufmg.br

Abstract. *Vehicular mobility traces are datasets of vehicles' location in a region with high spatio-temporal precision. Access to this sensitive information can threaten the safety and privacy of drivers, such as analyzing this data makes it possible to discover other contextual and latent information, such as users' daily home routes or workplace's address. In this way, many obfuscation and anonymization techniques have been proposed to mitigate the problem of user location privacy. In this work, we analyze an anonymization technique called mix-zone, where selected urban regions promote the simultaneous anonymization of vehicles by changing their pseudonym. We show how information about drivers' behavior in a city, such as their road preferences, can be used to re-identify their trajectories. We present a simple and efficient re-identification technique that uses* ***only two geo-referenced points*** *as input data. We validate our technique with a real dataset of taxicabs, being able to reidentify up to 95% of anonymised trajectories.*

Resumo. *Traces de mobilidade veicular são conjuntos de dados de localização dos veículos em uma região com alta precisão espaço-temporal. O acesso a essas informações sensíveis podem ameaçar à segurança e privacidade dos motoristas, dado que a análise desses dados torna possível a descoberta de outras informações contextuais e latentes, como suas rotas diárias para casa ou o endereço do local de trabalho. Desta forma, muitas técnicas de ofuscação e anonimização têm sido propostas para mitigar o problema de privacidade de localização de usuários. Neste trabalho, analizamos uma técnica de anonimização chamada mix-zone, onde regiões urbanas promovem a anonimização simultânea de veículos pela mudança de seu pseudônimo. Mostramos como as informações sobre o comportamento dos motoristas em uma cidade, com suas preferências de caminho, podem ser usadas para re-identificar as suas trajetórias. Apresentamos uma técnica de re-identificação simples e eficiente que usa* ***apenas dois pontos geo-referenciados*** *como entrada. Validamos a nossa técnica com um dataset real dos traces de táxis da cidade de São Francisco, EUA, na qual re-identificamos até 95% das trajetórias anonimizadas.*

*O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

1. Introdução

Na era da computação ubíqua, a quantidade massiva de dados de mobilidade produzidos por entidades móveis, como *smartphones*, despertam interesse de empresas e pesquisadores. Estes dados podem ser usados para compreender o comportamento humano e desenvolver diversos serviços na área de engenharia de tráfego como, por exemplo, no monitoramento de congestionamento de veículos, controle de fluxo, planejamento de infraestruturas, entre outros (Tan et al., 2017; Chang, Li, Zhu, Lu, & Li, 2018).

No entanto, a compilação desses dados de trajetórias trazem sérios riscos à privacidade dos usuários. Agentes maliciosos podem explorar informações encontradas nas trajetórias, mesmo quando submetidas a mecanismos de proteção à privacidade (*Location Privacy Protection Mechanisms* – LPPMs), para gerar ataques a entidades de diversas formas (Matheson, 2018). Por exemplo, para identificar pontos de interesse, como residência e local de trabalho, prever a ausência ou presença de um usuário em um local (Liu, Zhou, Zhu, Gao, & Xiang, 2018), rastrear e localizar entidades móveis (Tan et al., 2017), e associar uma identidade para cada trajetória anonimizada (ataque de re-identificação de trajetórias) (Krumm, 2009; Wernke, Skvortsov, Dürr, & Rothermel, 2014).

A re-identificação é uma das abordagens de ataque à privacidade que parte do princípio de identificar trajetórias anonimizadas e, consequentemente, conhecer as identidades a partir de um conjunto limitado de informações sobre a entidade-alvo (Primault, Boutet, Mokhtar, & Brunie, 2018). Esse é um ataque que também pode ser a "porta de entrada" para outros que têm como alvo uma entidade específica. As informações compreendem pontos geo-localizados, segmentos de trajetórias obtidos a partir de registros históricos em servidores de *Location-Based Services* (LBS) e características intrínsecas das trajetórias. Essas informações podem apresentar assinaturas e serem usadas para inferir o trajeto de veículos e localizar usuários.

Na literatura existem diversas abordagens de re-identificação baseadas na caracterização de *traces* de mobilidade, que utilizam técnicas como aprendizagem de máquina (Zan, Sun, Gruteser, & Ban, 2013; Sui, Wo, Tianyu, & Li, 2013) e inferência estatística (Tan et al., 2017). Porém, a maioria delas requer considerado custo computacional e conjuntos de dados de treinamento, que nem sempre podem estar disponíveis.

Neste artigo, propomos uma abordagem de re-identificação de trajetórias simples, eficiente e de custo computacional na ordem de $\mathcal{O}(E + V\log V)$, onde E representa a quantidade de arestas (interseções) no grafo das vias da cidade, e V a quantidade de nós (ruas). Apresentamos a re-identificação de trajetórias baseada na caracterização das preferências de caminho que ocorrem em ambientes urbanos, em particular no caso dos táxis. Assumindo a premissa de que os veículos, em especial os táxis, tendem as seguir o caminho mais curto entre dois pontos, a ideia é construir o caminho a partir de dois pontos geo-localizados (início e fim da trajetória) e, em seguida, compará-lo com as trajetórias anonimizadas, como uma forma de re-identificar veículos. A abordagem elimina a necessidade da coleta de dados históricos ou a aplicação de conjuntos de treinos para o cálculo da re-identificação.

Verificamos a eficiência de nosso modelo de ataque contra um esquema de privacidade chamado mix-zones. As mix-zones são regiões urbanas que promovem a anonimização simultânea de veículos através da mudança de seus pseudônimos. Em

nosso experimento, alteramos o nível de anonimização das mix-zones de modo a validar a capacidade de nossa técnica em re-identificar trajetórias de veículos. Validamos a nossa técnica de re-identificação em um *dataset* de táxis da cidade de São Francisco, EUA, em cerca de 132.645 viagens na qual re-identificamos até 95% das trajetórias anonimizadas.

O restante deste trabalho está organizado conforme segue: A Seção 2 descreve os trabalhos relacionados sobre re-identificação que utilizam a caracterização de *traces* de mobilidade. A Seção 3 formaliza a definição do problema. A Seção 4 apresenta o algoritmo de re-identificação de veículos baseado em caminho mínimo. Apresentamos e discutimos os resultados na Seção 5. Por fim, concluímos e indicamos futuras direções de pesquisa na Seção 6.

2. Trabalhos Relacionados

Na literatura existem diversos trabalhos sobre re-identificação de veículos e usuários em que utilizam características extraídas de trajetórias, como discutido a seguir.

2.1. Trabalhos de Re-identificação

Sui et al. (2013) estudaram as ameaças e preservação à privacidade durante a publicação de *traces* de táxis. Eles propuseram o ataque de pontos de estacionamento, no qual o adversário considera os hábitos de estacionamento de taxistas, extraídos dos *traces* de mobilidade dos táxis para re-identificar vítimas. Como contramedida, apresentaram um esquema de proteção que consiste em trocar sub-trajetórias dos táxis de pontos de estacionamentos mais relevantes. Uma das restrições desse ataque é que o atacante necessita do conhecimento prévio dos adversários e hábitos dos taxistas para inferir os pontos de estacionamento.

Zan et al. (2013) desenvolveram um modelo de ataque à privacidade de veículos, a partir da classificação de *traces* de mobilidade, ao esquema de privacidade Mix-Zones. Nesse ataque, os segmentos de *traces* anonimizados foram classificados por tipos de veículo: carro, moto e caminhão. Os autores defenderam a hipótese que os veículos de diferentes tipos possuem perfis de aceleração/desaceleração distintos e produzem perfis de *traces* de mobilidade diferentes, tornando-se possível classificá-los. A partir dessa classificação, tornou-se possível identificar assinaturas nos grupos de segmentos de *traces* e associá-los aos veículos correspondentes. Para isto, usaram a classificação por aprendizado de máquina e simulação para produzir o experimento, baseado em rodovias em vez de um ambiente urbano, que é menos desafiador para realizar a re-identificação.

Um dos primeiros trabalhos sobre quantificação da singularidade em traces de mobilidade humana foi proposto por De Montjoye, Hidalgo, Verleysen, and Blondel (2013). Com base na alteração da granularidade dos dados espaço-temporais do *dataset*, os autores apresentaram uma fórmula para calcular a singularidade da mobilidade humana. Propuseram também um modelo de inferência no qual quatro pontos espaço-temporais são suficientes para identificar exclusivamente 95% dos indivíduos de um *dataset* de telefonia celular contendo dados de mobilidade de pessoas de 1,5 M com 15 meses de dados.

Rossi, Walker, and Musolesi (2015) apresentaram uma técnica de re-identificação de usuários que explora a singularidade dos dados de GPS, mesmo não presentes no conjunto de dados de mobilidade. Especificamente, dado um conjunto de pontos geolocalizados da vítima, a técnica calcula a distância mínima entre esses pontos e os pontos dos *traces* anonimizados a partir de uma versão adaptada da distância de Hausdorff

(Dubuisson & Jain, 1994). O *trace* da vítima é aquele que possui a distância mínima entre os pontos. Esse estudo utilizou três *datasets* do mundo real: CabSpotting, CenceMe e GeoLife. Os autores concluíram que foram necessários a partir de três pontos espaciais para identificar quase 100% dos usuários. Porém, os autores não apresentaram detalhes sobre qual LPPM os *datasets* foram submetidos.

Tan et al. (2017) também exploraram a singularidade dos dados de GPS. Além disso, identificaram diferenças sobre as questões de privacidade entre LBSs e a sua derivação para veículos, *Vehicular Location-Based Service* (VLBS). Os autores afirmaram que os veículos são restritos a estradas, por este motivo as suas trajetórias são únicas e possíveis de serem re-identificadas. O modelo heurístico proposto foi capaz de re-identificar trajetórias de veículos anonimizadas com até 95% de acerto, a partir de quatro pontos espaço-temporais desses veículos e de mapas urbanos. Porém, os autores consideraram nesse ataque que o atacante teria total acesso ao servidor de dados de VLBS e que os ataques ocorreriam pela coleta dos pontos espaços-temporais do veículo da vítima.

Sekara, Mones, and Jonsson (2018) mostraram que é possível capturar o comportamento de usuários a partir de dados de uso de aplicativos coletados de *smartphones* levando em consideração o tempo. O comportamento de uso foi usado como uma assinatura digital para re-identificar usuários. Os autores também identificaram sazonalidades na singularidade de re-identificação e que as assinaturas digitais variam com o tempo a uma taxa constante média. Os dados foram coletados do *google play* com cerca de 12 meses de dados de 3,5 milhões de usuários. A técnica re-identificou 91.2% dos usuários, usando a estratégia inspirada em (De Montjoye et al., 2013).

Chang et al. (2018) assumiram que as trajetórias têm indicadores de perfis dos usuários, como preferências e comportamentos usuais, que são exclusivos e pouco mutáveis no tempo para re-identificar veículos. Os indicadores encontrados foram as paradas de interesse (como shopping e postos de combustível) e preferências de segmento da estrada. Eles re-identificaram trajetórias de vitimas comparando os indicadores encontrados em segmentos de trajetórias, coletados pela observação da vítima, e os dos históricos de mobilidade anonimizados. Nesse estudo, utilizaram os *datasets* de táxis de Shanghai e Shenzhen, nos quais foram testados a precisão em re-identificar trajetórias com subtrajetórias de diversos tamanhos. Os resultados mostraram que foi possível re-identificar as trajetórias anônimas com 96.64% e 77.03% de acerto para Shanghai e Shenzhen, respectivamente.

2.2. Discussão

Os trabalhos apresentados acima extraem características dos *traces* de mobilidade das vítimas para então realizarem ataques de re-identificação. Algumas abordagens utilizam aprendizagem de máquina, que necessitam de dados de treinamento e de custo computacional considerado. Outras abordagens necessitam de informações adicionais sobre a vítima e de contexto para que tenham sucesso no ataque. Os trabalhos sobre inferência de singularidade, apesar de apresentarem resultados expressivos como, por exemplo, os trabalhos de (Rossi et al., 2015; Sekara et al., 2018; Tan et al., 2017), são sensíveis à resolução espaço-temporal que comprometem diretamente a taxa de re-identificação. Esse fato é comprovado em (De Montjoye et al., 2013) cuja taxa de re-identificação chega a degradar em até 50% com a mudança da resolução espaço-temporal. Além disso, certas aborda-

gens utilizaram simuladores e *datasets* de rodovias, que pouco refletem o comportamento de *traces* de veículos em ambientes urbanos.

Diferente das abordagens anteriores, a nossa proposta re-identifica veículos com o mínimo de informação conhecida, sendo necessário basicamente dois pontos geo-localizados e de custo computacional na ordem de $\mathcal{O}(E+V\log V)$. Este trabalho difere do trabalho de (Rossi et al., 2015), em que o ataque consiste em calcular a distância mínima entre um conjunto de pontos (observados da vítima) e os pontos dos *traces* de mobilidade anonimizados. Já a nossa proposta é construir o caminho mínimo entre dois pontos geo-localizados (próximos ao início e fim da trajetória) e, em seguida, compará-lo com as trajetórias anonimizadas, como uma forma de re-identificar veículos.

Uma das limitações da proposta de (Rossi et al., 2015) é que taxa de sucesso do ataque pode ser degradada em um cenário onde os veículos compartilham segmentos de trajetórias, como por exemplo, táxis que transitam em uma mesma autoestrada ou na via principal de acesso a um aeroporto. Se os pontos observados da vítima estão em uma autoestrada na qual a probabilidade de existir mais de um táxi é alta, o desempenho do algoritmo pode ser degradado. Principalmente se os segmentos foram submetidos ao esquema de ofuscação para omitir localizações (*location hiding*). Este fato não ocorre em nossa abordagem, pois a re-identificação é feita pela correlação entre trajetórias (o caminho mínino e as trajetórias anonimizadas) e não apenas entre seus segmentos. Maiores detalhes sobre a abordagem serão definidos na Seção 5.

3. *Background* e Definição do Problema

Esta seção define as terminologias e notações usadas no decorrer deste trabalho. Iremos descrever a estrutura geral do sistema de rastreamento de táxis da cidade de São Francisco, o esquema de privacidade mix-zones e o ataque de re-identificação de trajetórias.

O projeto Cabspotting tem por função rastrear os táxis da cidade de São Francisco. Os táxis amarelos são dotados de um sistema de geo-posicionamento que calcula a localização atual do táxi. O conjunto de pontos de localização representam trajetórias ou viagens que são encaminhadas para uma central de recepção (*Yellow Cab server*) uma vez a cada minuto e, então, despachadas para um servidor central (*Cabspotting server*) no qual estes dados são anonimizados (Hoque, Hong, & Dixon, 2012). Posteriormente estes dados são publicados em outros repositórios para pesquisa e desenvolvimento de novos serviços (Piorkowski, Sarafijanovic-Djukic, & Grossglauser, 2009). O problema é que essas trajetórias podem ser usadas por usuários mal intencionados para revelar identidade e localização de veículos e pessoas.

3.1. Mix-zones: Um esquema de proteção à privacidade

Com o desenvolvimento dos sistemas de rastreamento, a privacidade de localização de entidades móveis passou a ser uma questão crítica. Diante deste fato, diversas técnicas de proteção à privacidade têm sido propostas, como os algoritmos de anonimização baseados em pseudônimos. Um pseudônimo é um mecanismo de proteção a privacidade que consiste em substituir dados reais e sensitivos por identificadores que não tenham nenhuma associação com estes dados reais (Primault et al., 2018). Por exemplo, durante a publicação dos dados de localização de um veículo, as informações sensíveis como as

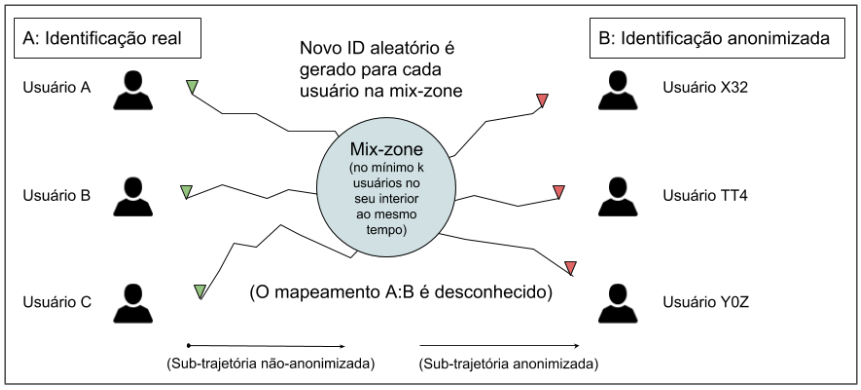

Figura 1. Conceito e funcionamento de uma mix-zone, com *K* = 3. Os usuários A, B e C entram em uma mix-zone e ao saírem, recebem novos identificadores (X32, TT4 e Y0Z, respectivamente) sem nenhuma associação com os identificadores anteriores, tornando-se difícil revelar as respectivas identidades (adaptado de (Beresford & Stajano, 2004).

placas dos veículos são substituídas por identificadores com o intuito de desassociar a identidade real do veículo[1].

Mix-zone é uma das abordagens para prover a privacidade para um grupo de entidades a partir da mudança simultânea de pseudônimos em áreas específicas (Beresford & Stajano, 2003). Esse conceito foi inicialmente aplicado na anonimização da comunicação em uma rede, denominada mix rede (Chaum, 1981). Consequentemente, foi estendido para os *Location-Based Services* (LBSs) (Chow & Mokbel, 2011).

Uma mix-zone é uma área geográfica de k-anonimidade na qual os usuários, no caso os veículos, passam e seus pseudônimos são modificados (Chen, Fu, Zhang, Zhang, & Li, 2018). Os veículos mudam de pseudônimo na mix-zone se e somente existem no mínimo K veículos presentes nela (Beresford & Stajano, 2003) (Beresford & Stajano, 2004). Um veículo pode transitar por várias mix-zones em seu trajeto e ter seu pseudônimo alterado, resultando em várias sub-trajetórias delimitadas por pseudônimos diferentes (vide Figura 1). A seguir, formalizamos a definição de mix-zone e o processo de mudança de pseudônimos:

- Seja um mapa geográfico urbano W no qual possui um tempo global associado $W.t$. Todos os objetos pertencentes a W observam $W.t$.
- Seja um conjunto de veículos $S = (v_1, v_2, v_3, ..., v_i, ..., v_n)$, $1 \leq i \leq n$ e $|S| = n$. Além disso $S \in W$.
- Um veículo v_i possui os seguintes atributos: um pseudônimo $v_i.alias$, que é distinto entre os demais n veículos. Todo v_i contém uma trajetória T.
- Uma trajetória (ou viagem) T de um veículo v_i é formado por uma sequência temporal de pontos espaço-temporais $T = (p_1, p_2, p_3, ..., p_m)$, $1 \leq i \leq m$. Um ponto $p_i = (x, y, t)$ no qual (x, y) representa a localização de um veículo (latitude e longitude, respectivamente) em um tempo t. Um conjunto de trajetórias de diferentes veículos é denotada por $T_a = (v_1.T, v_2.T, \ldots, v_j.T)$, $1 \leq j \leq n$ onde $T_a \subset S$.

[1]Este processo pode ser feito pelas *infrastructure of roadside units* (RSUs) que executam tais algoritmos que produzem e atribuem os pseudônimos aos veículos em seu domínio (Freudiger, Raya, Félegyházi, Papadimitratos, & Hubaux, 2007)

Consideramos que cada veículo v_i possui uma única trajetória T correspondente.

- Uma trajetória anonimizada T' é uma trajetória T a qual foi submetida a alguma função de anonimização $T' = anon[T]$ ao passar por uma mix-zone $M_j \in Mx$ e sofreu a mudança de pseudônimo em um algum $W.t$. Um conjunto de trajetórias anonimizadas é denotada por $T_p = (v_1.T', v_2.T', \ldots, v_j.T')$, $1 \leq j \leq n$ onde $T_p \subset S$.
- Uma mix-zone M_i é uma área geográfica que possui dimensões $M_i.r$ e requer o nível mínimo de anonimato K onde $K > 1$. O tempo de permanência de v_i em M_i é randômico. Um conjunto de mix-zones $Mx = (M_1, M_2, \ldots, M_p)$, onde $p = |Mx|$ e $Mx \subset W$.
- A mudança de pseudônimo de T_a veículos em uma mix-zone M_i é uma função $M_i(T_a)$, que ocorre em um tempo corrente $W.t$ se e somente se $\forall v_i$, $\exists$ algum $T.p_j \in M_i.r$ e $|T_a| \geq K$.

3.2. O modelo de ataque à privacidade

Dentro do contexto de mobilidade urbana, o tempo é considerado um fator valioso que reflete na dinâmica de mobilidade. As pessoas, ao se locomoverem em ambientes urbanos, tipicamente procuram levar o menor tempo possível para trafegar de uma origem a um destino. Ou seja, as pessoas tendem a evitar trajetos longos e duradouros e procuram por caminhos mais curtos e consideram as condições do trânsito para evitar rotas congestionadas, vias em obras ou com acidentes. Domingues, Silva, and Loureiro (2018) apresentaram uma caracterização de mobilidade do *trace* de táxis de São Francisco, mostrando que cerca de 70% de todas as viagens tendem a seguir o caminho mínimo entre dois pontos. O caminho mínimo é definido como aquele que apresenta a menor distância entre dois pontos, considerando a infraestrutura viária como base. Essa definição permanecerá no restante deste trabalho. Adicionalmente, as viagens que não seguem o caminho mínimo tendem a realizar desvios curtos quanto ao mesmo, com cerca de 5% de acréscimo à distância mínima.

Baseado nesse princípio, propomos as seguintes hipóteses para construir o modelo de adversário:

Hipótese 1: *A maioria dos veículos, com destaque para os táxis, escolhem caminhos mínimos para concluírem suas rotas*.

Se considerarmos a Hipótese 1 verdadeira, então temos:

Hipótese 2: *É possível re-identificar trajetórias anônimas*.

O objetivo do adversário é re-identificar o maior número possível de trajetórias anonimizadas. Assim, consideramos que o adversário tem acesso aos seguintes dados de entrada para gerar o ataque:

- trajetórias anonimizadas pela mix-zone, que foram publicadas em um servidor de acesso público;
- pontos geo-localizados - a partir da observação prévia da vítima ou pela análise das trajetórias anonimizadas, o adversário obtém dois pontos geo-localizados $inicio$ e $fim \in T$, que correspondem aos pontos próximos ao início e ao fim de uma trajetória da vítima v_i. Os pontos $inicio$ e fim podem ser identificados como os pontos com o menor e maior valor de $W.t$ na trajetória, respectivamente.

A ideia geral do algoritmo de re-identificação é construir o caminho mínimo de v_i ($\min[v_i.T]$) gerado a partir dos pontos $inicio$ e fim fornecidos de cada trajetória. O caminho é calculado com a aplicação de um algoritmo de caminho mínimo em grafos (e.g., algoritmo de Dijkstra), em um grafo composto das vias da cidade, representadas pelos vértices, e suas interseções, representadas pelas arestas. Os comprimentos das vias são representados pelos pesos das arestas. Após construir o caminho mínimo, o passo seguinte é compará-lo com a trajetória do indivíduo na tentativa de encontrar uma correlação entre elas. O adversário terá êxito se encontrar alguma trajetória de maior correspondência ao caminho mínimo. Maiores detalhes do algoritmo são apresentados na Seção 4.

4. A re-identificação de veículos a partir do caminho mínimo

Esta seção detalha a abordagem proposta de ataque à privacidade que utiliza o caminho mínimo para re-identificar trajetórias.

4.1. Algoritmo de re-identificação

O Algoritmo 1 representa os passos para re-identificar trajetórias através do caminho mínimo. Ele recebe como entradas os pontos geo-localizados das trajetórias reportados anteriormente à passagem dos veículos por alguma *mix-zone* $M_i(T_a)$, e os pontos geo-localizados das trajetórias reportados após a passagem dos veículos pela *mix-zone* (T_p). Adicionalmente, recebe também o grafo (G) contendo as vias e as interseções da cidade. Como resultado, o algoritmo retorna um mapeamento bijetor Φ das trajetórias em T_a sobre as trajetórias em T_p.

Algoritmo 1: Reidentificação de trajetórias

Data: Trajetórias anteriores à mix-zone T_a,
Trajetórias posteriores à mix-zone T_p,
Grafo G de vias e interseções
Result: Mapeamento $\Phi : T_a \rightarrow T_p$

```
1  custos ← Matriz(T_a linhas, T_p colunas);
2  for Trajetória i ∈ T_a do
3      for Trajetória j ∈ T_p do
4          inicio ← arg min_p f(p) | f(p) = p.t;
5          fim ← arg max_p f(p) | f(p) = p.t;
6          caminho-minimo ← Dijkstra(G, inicio, fim);
7          trajetoria-candidata ← ⟨i, j⟩;
8          erro ← DTW(trajetoria-candidata, caminho-minimo);
9          custos[i, j] ← erro;
10     end
11 end
12 Φ ← minimize(custos);
13 return Φ;
```

O algoritmo funciona como segue: a Linha 1 define uma matriz de custos, responsável por armazenar os resultados de cada correspondência possível das trajetórias de T_a em T_p. Para cada trajetória em T_a (Linha 2), itera-se sobre cada uma das possíveis trajetórias em T_p (Linha 3), extraindo os pontos inicial e final da trajetória (Linhas 4 e 5).

Em seguida, calcula-se o caminho mínimo a partir de tais pontos (Linha 6). A Linha 7 define a trajetória candidata como a junção da trajetória i anterior à mix-zone e a trajetória j posterior à mix-zone. Ou seja, assume-se que a vítima da trajetória i é também a mesma da trajetória j anonimizada.

Por fim, calcula-se a correlação entre a trajetória candidata e o caminho mínimo, através do algoritmo *Dynamic Time Warping* (DTW), que calcula a correlação não linear ótima de duas séries temporais. Em nossa implementação, o DTW retorna o nível de correlação representado por um erro. Se o erro for baixo, significa que houve alta correlação entre as duas trajetórias, caso contrário não. O erro é armazenando em uma matriz de custos (Linhas 7 a 9). O mapeamento das trajetórias T_a e T_p é calculado através da solução do problema da minimização de custos a partir da matriz de custos:

$$\Phi : \min\{\textit{custos}[i, j] : i \to j, i \in T_a, j \in T_p\}. \tag{1}$$

Ao assumir que os veículos tendem a seguir o caminho mínimo entre dois pontos, esperamos que, quanto menor a distância entre a trajetória candidata e o caminho mínimo, mais provável que aquela trajetória seja um caminho mínimo e, consequentemente, o motorista a seguiu. Em outras palavras, trajetórias candidatas que possuem um erro grande não representam um caminho mínimo entre seus pontos de origem e destino, e portanto possuem uma pequena probabilidade de serem a trajetória real escolhida.

4.2. Análise de complexidade do algoritmo de re-identificação

O Algoritmo 1 possui complexidade

$$\mathcal{O}(C(E + VlogV) + D^3) \approx \mathcal{O}(E + VlogV) \tag{2}$$

onde $C = |T_a||T_p|MN$, $|\cdot|$ representa a cardinalidade do conjunto de trajetórias que estão em uma mix-zone, M representa o tamanho da maior trajetória em T_a, N representa o tamanho da maior trajetória em T_p, D representa a dimensão da matriz de custos, $|T_a| = |T_p| = D$.

Essa complexidade é derivada do algoritmo de caminho mínimo de Dijkstra (Linha 6), que é usado pelo algoritmo proposto para construir a rota a ser comparada às trajetórias anonimizadas. O algoritmo DTW (Linha 8) possui complexidade $\mathcal{O}(MN)$, e a minimização (Linha 12) pode ser resolvida por métodos de atribuição linear, com complexidade $\mathcal{O}(D^3)$. Porém, como $|T_a|, |T_p|, D, M, N \ll E \approx V$, a complexidade do algoritmo 1 pode ser representada pela aproximação na Equação 2.

5. Experimentos

Os experimentos produzidos avaliam a eficiência do algoritmo de re-identificação em diferentes níveis de privacidade (K).

5.1. Configuração do experimento

Neste estudo, utilizamos o *dataset* dos *traces* de mobilidade de táxis da cidade de São Francisco, EUA, que contém dados de aproximadamente 500 táxis coletados em um período de 30 dias (Piorkowski et al., 2009). Porém, consideramos as trajetórias dos

Mix-zone	Latitude	Longitude	Raio (m)
mixzone0	37,614350	−122,395635	500
mixzone1	37,633000	−122,419134	300
mixzone2	37,628569	−122,432339	300
mixzone3	37,768201	−122,406079	500
mixzone4	37,769199	−122,453495	500
mixzone6	37,635672	−122,403605	500
mixzone7	37,735237	−122,406974	500
mixzone8	37,768914	−122,406881	500

Tabela 1. Configurações das Mix-zones

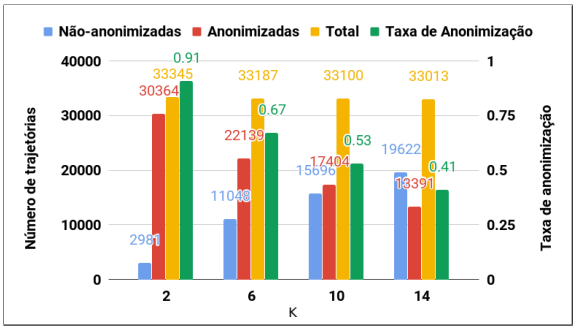

Figura 2. Trajetórias anonimizadas por K e a taxa de anonimização.

táxis nas quais efetivamente passaram por alguma mix-zone previamente definida. Logo o período de duração dos experimentos foi cerca de 25 dias (entre 17 de maio e 10 de junho de 2008). Os pontos geo-localizados das trajetórias possuem espaçamento temporal de 10 segundos entre si.

O experimento está divido em duas fases. A primeira é a anonimização das trajetórias do *dataset*. A segunda fase é o ataque de re-identificação das trajetórias pelo algoritmo proposto.

Na fase de anonimização, o *dataset* de (Piorkowski et al., 2009) foi anonimizado pelo mecanismo de privacidade mix-zones. Foram inseridas oito mix-zones estrategicamente posicionadas nos cruzamentos de maior tráfego de táxis. O objetivo era posicionar as mix-zones de modo que atingissem o maior número possível de táxis durante os 25 dias de experimento. Para isso, realizamos um estudo prévio sobre o fluxo de táxis da cidade. As informações sobre as mix-zones estão representadas na Tabela 1. Algumas mix-zones possuem raios distintos das outras para evitar possíveis sobreposições.

Ao todo foram realizados cinco experimentos que totalizam 132.645 trajetórias analisadas que atravessaram as mix-zones: 83.298 viagens foram anonimizadas e 49.347 não anonimizadas (vide Figura 2). Em cada experimento foi alterado o nível de privacidade (*K-anonimity*) das mix-zones. Em cada nível K, as trajetórias que passavam por uma ou mais mix-zones e atendiam aos seus requisitos eram anonimizados.

Podemos observar que a cobertura de veículos anonimizados é inversamente proporcional ao valor de K. Isto acontece porque o anonimato de n veículos ocorre apenas se existe simultaneamente $n \geq K$ dentro de uma mix-zone. Logo, se existir um K pequeno, mais veículos mudarão de pseudônimo. Porém, a chance de re-identificá-lo será maior. Consideramos também que não foram anonimizadas as trajetórias dos veículos que entraram mas não saíram da mix-zone e não atingiram o valor mínimo de K. Na Figura 2 está a taxa de anonimização para os diferentes níveis de privacidade K, no qual para $K = 2, 6, 10$ e 14 foram obtidas, para todas as mix-zones 91.06%, 66.70%, 52.58% e 40.56% de trajetórias anonimizadas, respectivamente.

Na fase de re-identificação das trajetórias comparamos a estratégia proposta (Algoritmo 1) com a abordagem de re-identificação aleatória, que serviu de *baseline* nos experimentos (Zan et al., 2013). Neste algoritmo, a trajetória candidata é formada pela

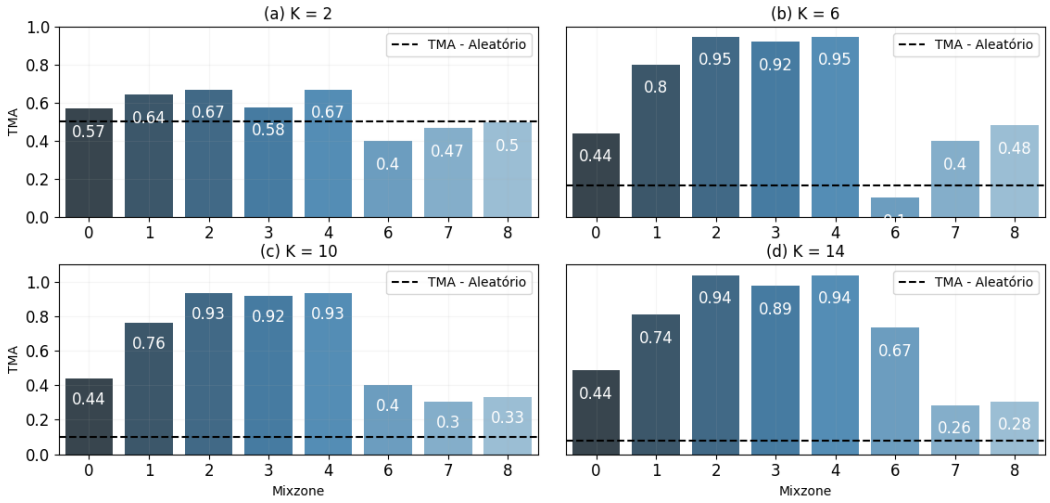

Figura 3. A distribuição do TMA para os diferentes níveis de anonimização para cada uma das mix-zones definidas

junção de uma sub-trajetória de T_a e de T_p que são selecionadas aleatoriamente.

5.2. Validação da eficiência

Utilizamos uma métrica, denominada *Trajectory Matching Accuracy* (TMA) para medir a eficácia dos ataques de re-identificação (Chang et al., 2018), que é definida como:

$$\text{TMA} = \frac{N_{reid}}{|T_p|} \tag{3}$$

onde $N_{reid} \in [0, |T_p|]$ e $|T_p|$ indicam o total de trajetórias re-identificadas e o total de trajetórias anonimizadas que passaram em alguma das mix-zones, respectivamente. O TMA está contido no intervalo de 0 a 1, com valores próximos a 0 indicando uma baixa acurácia, e valores próximos a 1 indicando uma alta acurácia na re-identificação das trajetórias.

5.3. Resultados e discussões

Os resultados de re-identificação de trajetórias para cada mix-zone e cada nível de K estão representados nas Figuras 3(a)–(d). Em uma visão panorâmica, o maior TMA atingido, cerca de 95% de acerto, foi para as configurações das mix-zones onde os valores de K eram expressivos. Ou seja, para valores de $K > 2$. É evidente a discrepância com os resultados do algoritmo de re-identificação aleatório (representado pelas linhas pontilhadas nos gráficos).

Outro ponto a ser considerado é que as mix-zones de 1 a 4 apresentaram resultados melhores do que para as outras mix-zones. Isto ocorre porque estas mix-zones estão posicionadas em um ponto central em relação às trajetórias das viagens. Desta forma, os segmentos da trajetória antes e depois da passagem pelas mix-zones possuem o comprimento máximo possível, permitindo que mais pontos sejam comparados com o caminho mínimo pelo algoritmo de re-identificação, tendo como resultado análises mais detalhadas. Esta realidade é diferente para as mix-zones 0, 6, 7 e 8, que estão situadas próximas aos pontos de início e fim das trajetórias e produzem a anonimização dos dados iniciais e finais. Apesar de apresentarem um valor baixo de re-identificação, essas mix-zones não

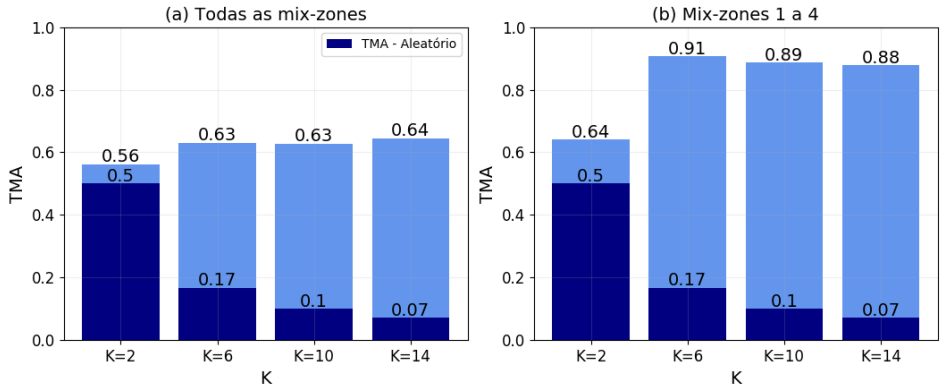

Figura 4. TMA agregado para cada nível de privacidade K

são boas soluções para um mecanismo de defesa à privacidade, visto que a anonimização ocorrida muito próxima aos locais de início e fim da trajetória, permite ao adversário inferir que estes são os verdadeiros pontos de início e fim da trajetória (Chen et al., 2018; Arain et al., 2018).

Na Figura 3(a), somente duas de oito mix-zones obtiveram taxas de re-identificação menores do que o algoritmo aleatório. Para o restante, as mix-zones 2 e 4 obtiveram os valores mais altos, com 67% de re-identificação cada uma. Quando comparado aos outros níveis K de anonimização, o nível $K = 2$ possui as menores taxas, devido à quantidade mínima de veículos presentes na mix-zone, diminuindo o espaço de busca do algoritmo. Para $K = 6$ (Figura 3(b)), observa-se que o aumento no número de veículos na mix-zone proporciona melhores resultados pelo algoritmo. Neste caso, somente uma de oito mix-zones apresentou resultados abaixo daqueles produzidos pelo algoritmo aleatório. Para o restante, o algoritmo produziu bons resultados, com destaque para as mix-zones 2, 3 e 4, com taxas de re-identificação acima de 90% das viagens. Este mesmo comportamento pode ser observado para $K = 10$ e $K = 14$ (Figuras 3(c) e 3(d), respectivamente), em que o algoritmo apresentou taxas acima de 90% para mix-zones localizadas em pontos centrais das trajetórias, e, mesmo para aquelas localizadas em pontos no início ou fim das trajetórias, foi capaz de re-identificar com precisão considerável quando comparado ao algoritmo aleatório.

Na Figura 4 está representado o TMA agregado (considerando a média simples entre todas as mix-zones (4(a)) e entre as mix-zones 1, 2, 3 e 4 (4(b)) para cada um dos níveis de privacidade K definidos. Adicionalmente, para comparação, também é ilustrado o TMA médio obtido para o algoritmo aleatório. É possível observar como o aumento no nível de privacidade diminui drasticamente o nível de re-identificação do algoritmo aleatório, sendo inviável o seu uso para qualquer valor de K maior que 2. Considerando todas as mix-zones (Figura 4(a)), o algoritmo apresenta valores semelhantes para os diferentes níveis de anonimização, i.e., 56% para $K = 2$, 63% para $K = 6$ e $K = 10$, e 64% para $K = 14$.

Apesar de ser consideravelmente mais preciso do que o algoritmo aleatório, a baixa precisão, ao re-identificar trajetórias anonimizadas por mix-zones localizadas em pontos no início ou fim das trajetórias, causa um decréscimo no TMA agregado. Devido à essa questão, a Figura 4(b) apresenta o TMA agregado considerando somente as mix-zones 1, 2, 3 e 4, cuja localização tende a ser no centro das trajetórias das viagens. É

possível observar que o algoritmo é eficiente na re-identificação das trajetórias, com valores acima de 90% de re-identificação em um dos níveis de anonimização. Esses resultados comprovam as Hipóteses I e II levantadas na Seção 3.2.

6. Conclusão e Trabalhos Futuros

Este trabalho propôs uma abordagem de re-identificação de trajetórias baseada na caracterização das preferências de caminho de veículos. Baseado na premissa de que a maioria dos veículos, com destaque aos táxis, escolhem o caminho mínimo para concluírem suas viagens, propomos duas hipóteses que serviram para a construção do algoritmo de re-identificação. Assim, elaboramos um algoritmo capaz de re-identificar trajetórias anonimizadas utilizando somente dois pontos de cada trajetória a ser re-identificada. O algoritmo é de simples implementação e não requer etapas de treinamento, como a maioria das soluções disponíveis na literatura, eliminando a necessidade de dados históricos. Finalmente, a solução também apresenta baixo custo computacional.

Os resultados mostram que o algoritmo consegue re-identificar trajetórias anonimizada com até 95% de eficiência, validando assim as hipóteses levantadas. Adicionalmente, nota-se que o algoritmo é robusto, apresentando altas taxas de re-identificação em um *trace* de mobilidade real e em larga-escala, mesmo quando existem muitos veículos (i.e., valores altos de K) na mix-zone.

Como trabalhos futuros, pretendemos avaliar a técnica em esquemas de privacidade baseados em anonimização mais refinados e também nos esquemas baseados em ofuscação (como por exemplo, *Dummy trajectories*) para explorar as suas vulnerabilidades. Em seguida, propor derivações mais robustas desses esquemas que consideram o *trade-off* entre privacidade e qualidade dos dados. Também verificar a eficiência da abordagem proposta em *datasets* de naturezas distintas, uma vez que a mobilidade de veículos tem um comportamento distinto das demais entidade móveis, como por exemplo, pessoas que usam dispositivos móveis.

Referências

Arain, Q. A., Memon, I., Deng, Z., Memon, M. H., Mangi, F. A., Zubedi, A. (2018). Location monitoring approach: multiple mix-zones with location privacy protection based on traffic flow over road networks. *Multimedia Tools and Applications*, *77*(5), 5563–5607.

Beresford, A. R., Stajano, F. (2003). Location privacy in pervasive computing. *IEEE Pervasive computing*(1), 46–55.

Beresford, A. R., Stajano, F. (2004). Mix zones: User privacy in location-aware services. In *Pervasive computing and communications workshops, 2004. proceedings of the second ieee annual conference on* (pp. 127–131).

Chang, S., Li, C., Zhu, H., Lu, T., Li, Q. (2018). Revealing privacy vulnerabilities of anonymous trajectories. *IEEE Transactions on Vehicular Technology*.

Chaum, D. L. (1981). Untraceable electronic mail, return addresses, and digital pseudonyms. *Communications of the ACM*, *24*(2), 84–90.

Chen, Z., Fu, Y., Zhang, M., Zhang, Z., Li, H. (2018). A flexible mix-zone selection scheme towards trajectory privacy protection. In *2018 17th ieee international conference on trust, security and privacy in computing and communications/12th ieee*

international conference on big data science and engineering (trustcom/bigdatase) (pp. 1180–1186).

Chow, C.-Y., Mokbel, M. F. (2011). Trajectory privacy in location-based services and data publication. *ACM Sigkdd Explorations Newsletter*, *13*(1), 19–29.

De Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, *3*, 1376.

Domingues, A. C., Silva, F. A., Loureiro, A. A. (2018). Space and time matter: An analysis about route selection in mobility traces. In *2018 ieee symposium on computers and communications (iscc)* (pp. 00958–00963).

Dubuisson, M.-P., Jain, A. K. (1994). A modified hausdorff distance for object matching. In *Proceedings of 12th international conference on pattern recognition* (pp. 566–568).

Freudiger, J., Raya, M., Félegyházi, M., Papadimitratos, P., Hubaux, J.-P. (2007). Mix-zones for location privacy in vehicular networks. In *Acm workshop on wireless networking for intelligent transportation systems (win-its).*

Hoque, M. A., Hong, X., Dixon, B. (2012). Analysis of mobility patterns for urban taxi cabs. In *Computing, networking and communications (icnc), 2012 international conference on* (pp. 756–760).

Krumm, J. (2009). A survey of computational location privacy. *Personal and Ubiquitous Computing*, *13*(6), 391–399.

Liu, B., Zhou, W., Zhu, T., Gao, L., Xiang, Y. (2018). Location privacy and its applications: A systematic study. *IEEE Access*, *6*, 17606–17624.

Matheson, R. (2018). *The privacy risks of compiling mobility data.* Retrieved 2018-12-07, from `http://news.mit.edu/2018/privacy-risks-mobility-data-1207`

Piorkowski, M., Sarafijanovic-Djukic, N., Grossglauser, M. (2009, February). *CRAWDAD dataset epfl/mobility (v. 2009-02-24).* Downloaded from `https://crawdad.org/epfl/mobility/20090224`. doi: 10.15783/C7J010

Primault, V., Boutet, A., Mokhtar, S. B., Brunie, L. (2018). The long road to computational location privacy: A survey. *IEEE Communications Surveys & Tutorials*.

Rossi, L., Walker, J., Musolesi, M. (2015). Spatio-temporal techniques for user identification by means of gps mobility data. *EPJ Data Science*, *4*(1), 11.

Sekara, V., Mones, E., Jonsson, H. (2018). Temporal limits of privacy in human behavior. *arXiv preprint arXiv:1806.03615*.

Sui, P., Wo, T., Tianyu, Z., Li, X. (2013). Privacy-preserving trajectory publication against parking point attacks. *2013 IEEE 10th International Conference on Ubiquitous Intelligence and Computing and 2013 IEEE 10th International Conference on Autonomic and Trusted Computing*.

Tan, Z., Wang, C., Fu, X., Cui, J., Jiang, C., Han, W. (2017). Re-identification of vehicular location-based metadata. *ICST Trans. Security Safety*, *4*(11), e1.

Wernke, M., Skvortsov, P., Dürr, F., Rothermel, K. (2014). A classification of location privacy attacks and approaches. *Personal and ubiquitous computing*, *18*(1), 163–175.

Zan, B., Sun, Z., Gruteser, M., Ban, X. (2013). Linking anonymous location traces through driving characteristics. In *Proceedings of the third acm conference on data and application security and privacy* (pp. 293–300).