Anonimização de Dados de Trajetórias em Grupos para Disponibilização à Pesquisa Universitária

Fernanda O. Gomes[1] **Julia Baldissera**[1] **,**
Bruno M. Agostinho[1] **, Douglas Simões** [1] **,**
Jean E. Martina[1]

[1]Universidade Federal de Santa Catarina
Departamento de Informática e Estatística
Campus Universitário – Florianópolis – SC – Brasil

fernanda.gomes@posgrad.ufsc.br, julia.baldissera@grad.ufsc.br,

bruno.agostinho@posgrad.ufsc.br, douglas.simoes@posgrad.ufsc.br,

jean.martina@ufsc.br

Abstract. *The use of mobile devices connected to the campus Wi-Fi network allows the capture of student trajectory data. The union of these data with the information of the students present on the systems of the universities makes possible the creation of semantic trajectories with personal quasi-identifiers. However, such data can put people's privacy at risk. For this reason, we demonstrate a new trajectory anonymization technique, called Mix β-k-anonymity. This technique provides a set of possible trajectories of a group of people with similar quasi-identifiers. The goal is to apply this method to solve the data privacy problem and make the publication of these datas possible. This work shows that the academic community can have access to data with privacy and quality for operational mobility research on campus.*

Resumo. *A utilização de dispositivos móveis conectados à rede Wi-Fi de câmpus universitários permite a captura de dados de trajetória dos alunos. A união destes dados em conjunto com as informações dos alunos cadastradas nos sistemas das universidades torna viável a criação de trajetórias semânticas com quase-identificadores pessoais. No entanto, esses dados podem colocar a privacidade das pessoas em risco. Por essa razão, neste artigo é demonstrada uma nova técnica de anonimização de dados de trajetória chamada Mix β-k-anonymity. A técnica disponibiliza pontos espaço-temporais que formam um conjunto de possíveis trajetórias realizadas por grupos de pessoas com quase-identificador semelhante. O objetivo é aplicar este método para resolver o problema de privacidade de dados tornando a publicação destes possível. Os resultados deste trabalho demonstram que a comunidade acadêmica pode ter acesso a dados de qualidade para pesquisa operacional em mobilidade no campus onde a relevância demográfica é atendida com a manutenção da privacidade dos usuários.*

1. Introdução

As universidades capturam diariamente uma quantidade gigantesca de dados sobre seus usuários que podem servir de base para pesquisas operacionais. Dados de autenticação

para utilização da Wi-Fi do campus são um exemplo onipresente e permitem rastrear a mobilidade das pessoas através de trajetórias. É possível juntar este tipo de informação com os dados que já existem sobre o aluno nos sistemas da universidade, como dados cadastrais, histórico escolar, grade de horários e entre outros. Essas informações são conhecidas como quase-identificadores. Nesse contexto a disponibilização destes dados pode gerar grandes pesquisas operacionais a um custo razoavelmente pequeno e com um grau de confiabilidade de coleta considerável. Projetos como disposição de pontos de ônibus dentro do campus universitário, criação de novos caminhos para pedestres e ciclistas, criação de locais para integração podem usar esses dados como fonte de dados.

A coleta desses dados começa com a conexão à internet realizada por um dispositivo móvel produzindo um ponto de uma trajetória, ou seja, um par de coordenadas e informações temporais. No entanto, é possível agregar uma informação contextual a cada ponto e criar uma trajetória semântica. Uma trajetória semântica é uma sequência de *stops* (paradas) e *moves* (movimentos) de uma pessoa durante o seu trajeto [Location Privacy in Pervasive Computing]. Um exemplo de trajetória semântica é a sequência de lugares visitado por um indivíduo em movimento, como a sequência Biblioteca, Lanchonete, Sala de Aula, Restaurante Universitário e Ponto de Ônibus [Monreale et al. 2011].

Por definição as trajetórias semânticas representam os locais mais importantes do trajeto realizado pela pessoa. Com a divulgação desse tipo de informação a privacidade do indivíduo é violada. Os dados de localização permitem inferências intrusivas, que podem revelar hábitos, costumes sociais, preferências religiosas e sexuais de indivíduos [Abul et al. 2008]. Caso pessoas mal-intencionadas tenham acesso a estes dados, perseguições a pessoas podem se tornar muito fáceis, assim como apoio operacional para a realização de crimes, ocasionando uma ameaça a segurança das pessoas.

O conjunto de dados de trajetórias são propensos a ataques de ligação à fontes externas, mesmo após a anonimização, uma vez que os atributos de espaço e tempo são quase-identificadores daquele que realizou o trajeto. Pesquisas mostram que os dois lugares mais visitados por uma pessoa tem grandes chances de serem sua casa e seu trabalho [Golle and Partridge 2009]. Existem também os ataques de conhecimento de parte da trajetória de um indivíduo, que se encaixam melhor no contexto deste trabalho, seja por meio de observação ou por divulgação da localização pelo próprio indivíduo. Com relação aos dados pessoais do indivíduo que realizou a trajetória, é possível identificar pessoas ligadas a registros supostamente anonimizados mesmo com a remoção de informações de identificação pessoal [Sweeney 2002]. Isso se deve ao fato de que os dados divulgados podem ser vinculados a outras fontes de dados, assim como nas trajetórias, por meio de um conjunto de atributos quase-identificadores comuns as duas bases de dados. Nos Estados Unidos, a combinação de código postal de cinco dígitos, sexo e data de nascimento é única para 87% dos cidadãos [Sweeney 2002] e com a adição de novos atributos a esse conjunto de quase-identificadores esta porcentagem tende a aumentar [Nergiz et al. 2007].

Tendo em vista os problemas citados, o objetivo desse trabalho é criar um algoritmo de anonimização que permita serem divulgados dados de trajetória semântica, realizadas em grupos representados por quase-identificadores, sem que seja possível identificar unicamente o par trajetória-pessoa. Pesquisas recentes, como [Rajesh and Abraham 2017], [Gramaglia et al. 2017], [Terrovitis et al. 2017], trabalham em anonimizar dados de trajetória para que esses possam ser divulgados. Todavia, to-

dos os trabalhos até o momento concentraram-se em anonimizar trajetórias e seus quase-identificadores, espaço e tempo, não vinculando essa anonimização as informações da pessoa que realizou a trajetória. Apenas um trabalho se preocupa com a anonimização de dados gerados dentro de um campus universitário [Ma et al. 2017], porém este não traz nenhuma nova solução para anonimização desses dados.

As principais contribuições deste trabalho são:

- Um sistema automatizado de extração, transformação e limpeza de dados de trajetória capturados em um campus universitário.
- Um algoritmo de anonimização de dados de trajetória em grupos utilizando uma técnica de agrupamento, conceitos semelhantes aos de *stops* e *moves* e *k-anonymity* que tem como objetivo prevenir os ataques mencionados e poder realizar a divulgação com uma porcentagem máxima de privacidade.

Este artigo está organizado da seguinte maneira: A seção 2 explica os conceitos básicos. A seção 3 traz os trabalhos correlatos. Na seção 4 é exibido o modelo de extração, transformação e captura dos dados. Em seguida é apresentada a seção 5 que apresenta modelos de ameaça a bases de trajetórias anonimizadas dentro de um campus universitário. A seção 6 propõe o método *Mix* β*-k-anonymity* de anonimização e a seção 7 mostra os resultados obtidos com a aplicação do método a base de dados. Por último, na seção 8, são apresentadas as considerações finais com algumas observações e indicações para trabalhos futuros.

2. Conceitos Básicos

Para melhor conhecimento da proposta de anonimização de dados é importante a definição dos conceitos de *k-anonymity* (2.1) e Trajetórias (2.2).

2.1. *k-anonymity*

Para que um conjunto de dados tenha a propriedade de *k-anonimity*, para cada registo formado por quase-identificadores presente nesse conjunto, devem existir pelo menos outros k - 1 registos idênticos a ele. As técnicas mais utilizadas para anonimizar um determinado conjunto de dados são generalizações e supressões. Essas técnicas também são chamadas de não perturbativas, pelo fato de não distorcerem os dados [Domingo-Ferrer and Torra 2001]. A generalização substitui os valores de atributos por valores mais genéricos. A supressão é uma técnica que exclui valores de atributos do conjunto de dados anonimizados.

Foram encontrados dois problemas na solução de *k-anonymity*. *Homogeneity Attack*, que consiste na grande quantidade de atributos sensíveis (ex. saldo bancário, doença, voto, entre outras informações confidenciais) iguais dentro de um conjunto de dados. O outro problema foi chamado de *Background Knowledge Attack*, onde um atacante tem algum conhecimento sobre um indivíduo, sabe que este está na base anonimizada e procura por pessoas com perfil anonimizado semelhante, para descobrir as informações sensíveis pertencentes a pessoa [Machanavajjhala et al. 2006]. Com isso, surgiram trabalhos afim de resolver esse problemas relativos ao *k-anonymity* tais como [Machanavajjhala et al. 2006] e [Li et al. 2007].

Embora tenha sido demonstrado por [Meyerson and Williams 2004] e [Aggarwal et al. 2005] que encontrar um *k-anonymity* ótimo é NP-difícil e que

esse apresenta algumas limitações, ele continua sendo usado como base de diversos trabalhos de anonimização, incluindo de trajetórias. Nesse trabalho o *k-anonymity* será utilizado para a base da anonimização das trajetórias e dos dados pessoais para criação de grupos. Visto que não serão utilizados atributos sensíveis, serão mostrados outros tipos de ataques, que buscam revelar a trajetória de uma pessoa.

2.2. Trajetórias

Uma trajetória é representada como uma sequência discreta de pontos. Esses pontos representam uma evolução espaço-temporal da posição de um objeto em movimento, ou seja, esse objeto está se movendo no espaço durante um determinado intervalo de tempo para atingir um determinado objetivo. Tanto o conceito de trajetória quanto trajetória semântica são baseados no artigo [Monreale et al. 2011].

Definição 1 (Trajetória)**.** Uma trajetória é uma lista de pontos espaço-temporais $p_0 < x_0, y_0, t_0 >, ..., p_n < x_n, y_n, t_n >$, onde $x_i, y_i \in R, t_i \in R^+$ para $i = 0, 1, ..., n$ e $t_0 < t_1 < ... < t_n$.

Objetos em movimento não necessariamente se movem continuamente durante uma trajetória. Sendo assim, as trajetórias podem ser semanticamente divididas. Com isso, podem ser definidas como uma sequência temporal de subintervalos de tempo onde, alternativamente, a posição do objeto pode ou não mudar [Spaccapietra et al. 2008].

Definição 2 (Trajetória Semântica)**.** Dada uma lista de locais importantes I, uma trajetória semântica $p_1, p_2, ..., p_n$ com $p_i \in I$ é uma sequência de lugares importantes visitados por objetos em movimento.

Esse conceito é baseado em stops (paradas) e moves (movimentos) introduzido nos trabalhos de Alvares et al. (2007) e Spaccapietra et al. (2008). As paradas são as partes mais importantes e acontecem quando um objeto em movimento permanece por um período mínimo de tempo em um ponto importante. Movimentos são as sub-trajetórias que descrevem o deslocamento entre duas paradas consecutivas. Um conjunto de lugares importantes caracteriza uma trajetória semântica [Alvares et al. 2007]. Outra característica da trajetória semântica é que esta pode receber informações contextuais, como o meio de transporte que à executou ou o nome dos locais visitados [Monreale et al. 2011].
Definição 3 (Stop)**.** Um stop $stop_x$ é uma parte da trajetória que começa no tempo $t_{inicialx}$ e termina em t_{finalx}. A diferença entre $t_{finalx} - t_{inicialx} \neq \emptyset$. O objeto deve permanecer parado em apenas um lugar nesse período de tempo ($\Delta t = t_{finalx} - t_{inicialx}$),e cada $stop_1 \cap stop_2 \cap ... \cap stop_n = \emptyset$ (e não pode existir interseção entre os stops).
Definição 4 (Move)**.** Um *move* é a linha espaço-temporal, $move_x$, que é delimitado por um $stop_{inicial}$ e um $stop_{final}$, tendo como $t_{inicialx}$ o t_{final} do $stop_{inicial}$ e t_{finalx} o $t_{inicial}$ do $stop_{final}$. $\Delta t = t_{finalx} - t_{inicialx}$.

Neste trabalho serão implementados junto a proposta de anonimização os conceitos semelhantes aos de *stops and moves*.

3. Trabalhos Correlatos

Essa seção será dividida em duas partes. A primeira apresentará trabalhos correlatos que também apresentaram técnicas de anonimização de trajetórias semelhantes a usadas nesse trabalho. A segunda parte mostrará trabalhos que utilizaram esquemas parecidos de captura de dados Wi-fi em campus universitários.

3.1. Anonimização de trajetórias

Foram desenvolvidos muitos trabalhos com o objetivo de propor um método de anonimização de dados de trajetória. A primeira técnica de generalização de trajetórias estáticas utilizando o agrupamento de pontos próximos no espaço e tempo foi proposta por [Abul et al. 2008]. A abordagem *Never Walk Alone* (NWA) baseada em cluster, que aproveita a incerteza inerente da localização de um objeto em movimento, introduziu o conceito de *(k, δ)-anonymity*.

Para alcançar o *k-anonymity*, cada trajetória é atribuída a um grupo de pelo menos *k-1* outras trajetórias usando um algoritmo de agrupamento guloso. Então, as trajetórias de cada cluster são traduzidas espacialmente, de modo que todas elas se encontrem inteiramente dentro do mesmo cilindro (área de incerteza) do raio $\delta/2$. A similaridade deste trabalho, com o em questão, está no fato de ambos trabalharem com uma área de incerteza e *k-anonymity*. Outros trabalhos seguiram a mesma linha de generalização de trajetórias tais como [Nergiz et al. 2008], [Gramaglia et al. 2017], [Lu et al. 2017] e [Mahdavifar et al. 2012].

Neigiz et al. (2008) em [Nergiz et al. 2008] adotam a noção de *k-anonymity* para trajetórias e propõem uma abordagem baseada em generalização para a anonimização da trajetória. Também foi apresentado um algoritmo de reconstrução baseado em "randomização" para liberar dados de trajetória anonimizados. O trabalho [Gramaglia et al. 2017] se difere dos demais pois este propõe uma variação do *k-anonymity* que leva em consideração sub-trajetórias que podem ser descobertas por um atacante e garantem que existam pelo menos *k - 1* outras sub-trajetórias iguais.

Os artigos [Lu et al. 2017] e [Mahdavifar et al. 2012] trabalham com diferentes níveis de privacidade. Em [Lu et al. 2017] é proposto uma estrutura que fornece serviços de preservação da privacidade. As trajetórias são agrupadas em *clusters*, de maneira que satisfaçam uma restrição de privacidade com base nos requisitos de privacidade pessoal do usuário e que apresentem no máximo *maxR* de distorção. Este trabalho tem como objetivo anonimizar as trajetórias, para que seja possível publicá-las sem violar a privacidade dos indivíduos. Assim como em [Lu et al. 2017], o artigo [Mahdavifar et al. 2012] apresenta a ideia de requisitos de privacidade não uniformes utilizando *k-anonymity*. Cada trajetória está associada ao seu próprio nível de privacidade, sendo assim, cada trajetória tem um número *k* de trajetórias similares, onde a mesma é indistinguível entre elas. O método começa dividindo as trajetórias em grupos dependendo do seu nível de privacidade. Se uma trajetória tem um alto nível de privacidade (um *k* alto), provavelmente será parte de um grande *cluster* e com isso sofrerá uma grande perda de informação. Clusters grandes são muito generalizados e adicionam uma anonimização excessiva o que impacta na qualidade dos dados.

A proposta de [Rajesh and Abraham 2017] visa proteger a privacidade das trajetórias ocultando os *stops*. Essa ideia é relevante, pois os stops contêm os principais locais de uma trajetória. Por exemplo, se um objeto em movimento permanece ou visita um hospital frequentemente, então o adversário pode inferir que essa pessoa está tendo sérios problemas de saúde. Se este objeto passar apenas na frente do hospital, o adversário não pode inferir nada. Então o algoritmo proposto torna os *stops* em zonas. Na versão anonimizada das trajetórias, os *moves* são mantidos na sua forma original e caso passem por uma zona estes se tornam parte dela. Assim como em [Rajesh and Abraham 2017] o

trabalho [Terrovitis et al. 2017] também utiliza a técnica de supressão, dividindo as trajetórias em sub-trajetórias e utilizando o método de *l-diversity*.

A maioria dos trabalhos apresentados anonimiza as trajetórias utilizando os quase-identificadores das próprias trajetórias criando *clusters* com trajetórias semelhantes no espaço e no tempo. Outros trabalhos utilizam a supressão de pontos para anonimizar as trajetórias. O trabalho proposto utiliza um quase-identificador que representa um grupo de pessoas semelhantes. Nessa proposta um *stop* é divulgado apenas se existirem pelo menos *k - 1* pessoas de um mesmo grupo em um *stop* no mesmo período de tempo. Um *move* é exibido apenas se o *stop* de partida dele tiver outros β moves, realizados por pessoas de um mesmo grupo, distintos iniciando nesse *stop*, como será explicado na seção 6.

3.2. Captura de dados de Wi-fi em campus universitários

O primeiro estudo comportamental baseado em dados de trajetórias, criados por conexões Wi-fi em dispositivos móveis, foi realizado pela Universidade de Stanford em 1999 [Tang and Baker 2000]. Os dados foram coletados usando três técnicas diferentes (tcpdump, SNMP polling e logs de autenticação). O artigo [Schwab and Bunt 2004] apresenta um estudo sobre o uso da WLAN do Campus da Universidade de Saskatchewan. O tráfego foi rastreado durante uma semana usando EtherPeek, um pacote de software que permite ao registro de endereços físicos e informações de tráfego. O trabalho [Hutchins and Zegura 2002] analisou a WLAN do Georgia Tech Campus durante cinco meses. Com isso, eles extraíram informações sobre o comportamento do usuário a partir dos *logs* de autenticação no *firewall*.

Em [Wang et al. 2017] foram coletados dados de Wi-fi durante um seis meses em uma universidade. O trabalho apresenta uma medida de similaridade de trajetórias semânticas e estima o nível de intimidade das pessoas. A diferença deste trabalho para o artigo em questão é que não existe uma autenticação muito menos conexão à internet, além de não apresentar preocupação com a anonimização dos dados. A principal semelhança encontrada entre esses trabalhos e o proposto é que todos extraem informações de trajetórias criadas a partir de conexões de dispositivos móveis em pontos de acesso Wi-fi dentro de um campus universitário. Cada conexão gera um ponto em uma trajetória.

O único trabalho encontrado que levanta a questão da anonimização dos dados de Wi-fi capturados em um campus universitário em consideração é o [Ma et al. 2017]. Este porém, não apresenta nenhuma proposta de método de anonimização para solucionar o problema da anonimidade. O artigo mostra o risco da divulgação de dados de trajetória utilizando como método de anonimização o *k-anonymity*. Eles mostram que ao relacionar a grade curricular com os dados de trajetória dos alunos existe uma possibilidade de se identificar um indivíduo unicamente.

4. Coleta e Limpeza dos Dados

A proposta consiste na captura de dados de movimento de pessoas a partir de associações à rede sem fio dentro do campus universitário e na posterior disponibilização anonimizada desses dados para pesquisa operacional de mobilidade. Os dados são coletados através das associações de dispositivos móveis de usuários à pontos de acesso sem fio da rede acadêmica da universidade. Os dispositivos móveis precisam estar a certo raio de alcance

do sinal de rádio destes pontos de acesso para a associação. Estes pontos estão espalhados pelo campus com sua localização geográfica conhecida, para que as pessoas tenham conectividade a Internet. Para associar-se o dispositivo deve realizar uma autenticação no serviço Eduroam. O serviço foi desenvolvido para a comunidade internacional de educação e pesquisa para oferecer acesso sem fio à Internet de forma simples, rápida e segura.

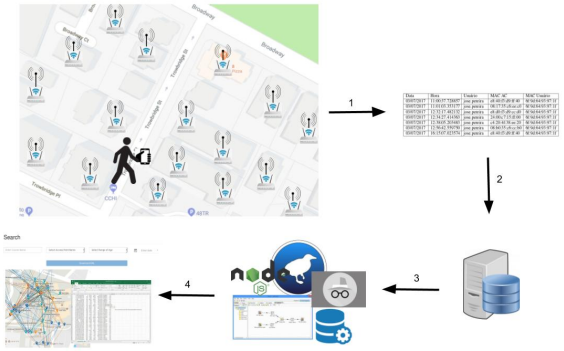

Figura 1. Desenho do modelo de extração, limpeza e disponibilização.

O local físico onde os pontos de acesso são instalados tem suas coordenadas conhecidas. Para realizar uma conexão é necessário que o um identificador único e uma senha. O identificador é utilizado para o acesso a todos os serviços da universidade. Um arquivo de *logs* é atualizado toda vez que uma conexão é realizada. Esse arquivo contém as seguintes informações: Data, Hora, Identificador do Usuário, MAC Access Point (Ponto de Acesso), MAC Usuário e confirmação de sucesso na conexão. Para encontrar os quase-identificadores, coordenadas e nome da localização do ponto de acesso, é realizada uma consulta a uma tabela contendo essas informações buscando pelo MAC. As informações sobre a pessoa relacionada ao identificador são acessadas via requisição os servidores dos serviços fornecidos pela universidade. O acesso a esses serviços é restrito a universidade, pessoas autorizadas por esta e ao aluno relacionado com o identificador.

A extração, transformação e limpeza dos dados foram realizadas utilizando a ferramenta de ETL (*Extract Transform Load*) Kettle. Os dados foram armazenados no banco de dados não relacional MongoDB versão 3.4.

5. Modelo de Ameaça

Ao divulgar uma base de dados anonimizada T* de trajetórias realizadas por grupos G em um campus universitário, ataques de inferência podem ser realizados. Esses ataques são realizados com conhecimento prévio dos atacantes de sub-trajetórias de um indivíduo que está em algum grupo da base de dados. Para que seja possível realizar as inferências o atacante deve ter conhecimento de alguns dados.

Definição 5 (Conhecimentos do Atacante)**.** Dada uma base de dados de trajetórias de grupos anonimizada *T** o atacante pode saber: 1) Se dada pessoa tem uma trajetória anonimizada *t** onde $t* \subset T*$. 2) Todos os grupos *g* que pertencem a *G*. 3) Se dado grupo $g \in G$ uma pessoa $p \in g$. 4) Uma sub-trajetória, realizada por uma pessoa *p*, $st \subset T*$.

Com esses conhecimentos é possível realizar ataques sobre a base. Os modelos de ataque foram criados baseados no cenário onde os dados são capturados: um campus universitário.

Definição 6 (Divulgação espontânea ou por Observação)**.** Um atacante pode perseguir pessoas nas redes sociais (ex. *Facebook*, *Instagram*, *Twitter*, entre outros) ou observar o comportamento (ex. o seguindo) de indivíduo alvo *i* e descobrir lugares visitados por *i* em um certo horário formando conjuntos de sub-trajetórias *st* e descobrir a trajetória completa de um indivíduo buscando por *st* em *T**. Esse ataque pode ser comparado ao de *Background Knowledge Attack* proposto por [Machanavajjhala et al. 2006].

Definição 7 (Grade de Horário)**.** Esse ataque foi proposto por [Ma et al. 2017], onde podem ser criados conjuntos de sub-trajetórias *st* possíveis de certo grupo de pessoas, a partir da grade de horário dos cursos e comparar com as trajetórias *T** divulgadas.

Levando em consideração que este trabalho tem como objetivo anonimizar trajetórias utilizando a noção de grupos de pessoas , o trabalho mostra que é possível garantir a privacidade com no mínimo $1/\beta$ chances de se descobrir a real trajetória de um indivíduo de um certo grupo. Para evitar os ataques definidos acima, o trabalho propõe o método *Mix* β*-k-anonymity*.

6. Mix β-k-anonymity

Primeiramente, os dados devem chegar ao algoritmo limpos e já processados no formato que o algoritmo exige. O algoritmo anonimiza os dados de um grupo na granularidade diária, ou seja, o log de um dia de conexões. Ele trabalha com faixas de horários que representam a generalização do horário da conexão a um *access point*, que por conseguinte vira a informação temporal de um ponto na trajetória. A ideia de criar faixas de horários, visa aumentar a imprecisão e dificultar ataques a base. Visto isso, é necessário realizar o agrupamento dos dados por: grupo (variável quase-identificadora comum a um grupo grande de pessoas), local e faixa de horário. Com isso, contabiliza-se a quantidade de registros que apresentam as mesmas variáveis. Essa quantidade representa x pessoas de um certo grupo que estavam no mesmo lugar em uma certa faixa de horário.

O algoritmo recebe uma lista de pontos que pertencem a trajetórias realizadas por pessoas de um mesmo grupo. Cada ponto contém os atributos: id, nome do local, horário, faixa de horário, latitude, longitude, quantidade de pessoas que também estavam nesse local nesta faixa de horário, usuário, uma lista de próximos vazia e uma variável do tipo *boolean*, nomeada como agrupado, iniciada com o valor falso. O algoritmo deve ser rodado uma vez para cada grupo. Na linha 1 os pontos que não concentram pelo menos uma quantidade de k pessoas são removidos.

Nas linhas 2-11, a lista é percorrida de baixo para cima, visto que o próximo ponto visitado está sempre abaixo do ponto atual na lista. Um ponto não tem próximo quando ele representa a última conexão de um usuário. A linha 4 verifica se é o primeiro ponto a ser analisado ou se o usuário do ponto atual e do próximo ponto não são os mesmos. Essa verificação é feita para que no caso de a comparação entre os usuários retorne falso, o ponto atual é o último ponto da trajetória deste usuário e portanto não deve possuir um próximo.

Caso não entre na primeira condição então é verificado se o local e faixa de horário dos pontos são os mesmos. Isso acontece quando são feitas várias conexões em uma

mesma localização em um intervalo curto e contínuo de tempo. Nessa caso essa trajetória é agrupada à anterior mudando apenas a *flag* de agrupado para verdadeiro. As linhas de 9-11 representam a situação onde o ponto tem um próximo que não é situado na mesma localização ou faixa de horário que ele ou o ponto não tem próximo. Então ponto próximo é adicionado na lista de próximos do ponto atual e o ponto atual vira o próximo ponto.

Algorithm 1: Mix β-k-anonymity

```
   Input  : Um array T de tamanho l ordenado em ordem crescente por horário e usuário, dois inteiros β, k.
   Output: T anonimizado.
1  ExluiPontosSemKMinimo(T,k);
2  for y ← l − 1 to 0 do
3      Ponto ← T[y];
4      if y == l − 1 ou Ponto.usuario ≠ proximoPonto.usuario then
5          proximoPonto ← Ponto;
6      else
7          if proximoPonto.nomeLocal == Ponto.nomeLocal e proximoPonto.faixaHorario == Ponto.faixaHorario then
8              Ponto.agrupado ← true ;
9          else
10             AdicionaNaListaDeProximos(Ponto, proximoPonto);
11             proximoPonto ← Ponto;
12 for i ← l − 1 to 0 do
13     for x ← i to 0 do
14         if T[i].faixaHorario ≠ T[x].faixaHorario e T[i].horario > T[x].horario then
15             x ← −1;
16         if x > −1 e T[i].nomeLocal == T[x].nomeLocal e T[i].faixaHorario == T[x].faixaHorario e x ≠ i e !T[i].agrupado e
             !T[x].agrupado then
17             foreach proximo ∈ T[x].proximos do
18                 if !JaTemNaListaDeProximosPontos(proximo, T[i].proximos) e T[x].id ≠ T[i].id then
19                     AdicionaNaListaDeProximos(T[i], proximo);
20         T[x].agrupado ← true;
21     AtualizaListaDeProximosPontos(T[i]);
22 RemovePontosAgrupados(T);
23 foreach ponto ∈ T do
24     if ponto.proximos.length < β then
25         ApagaProximosDoPonto(ponto);
```

Como cada ponto tem uma lista de próximo locais que foram visitados, por mais de k pessoas desse grupo, e cada próximo ponto tem seus próximos e assim por diante, diversas trajetórias possíveis de terem sido realizadas por uma pessoa desse grupo são criadas, como pode ser visto na Figura 2. Existe a chance de conseguir ligar uma pessoa a uma trajetória dependem do valor do β e k, como será mostrado na seção 7.

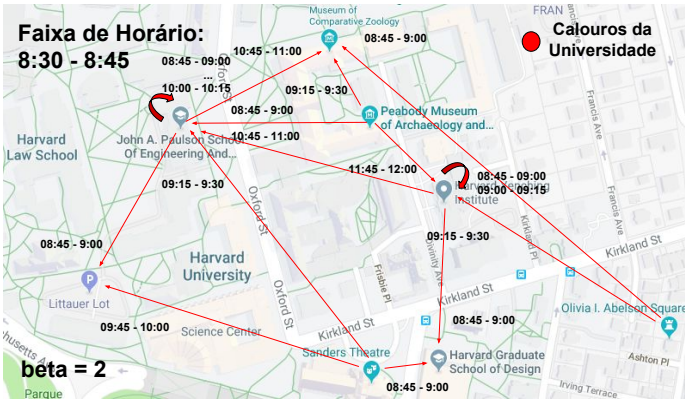

Figura 2. Exemplo da disposição dos pontos do grupo de calouros anonimizados no período de 08:30 à 08:45.

A segunda parte do algoritmo fica entre as linhas 12-25. Nessa etapa são agrupa-

dos pontos, de todas as pessoas desse grupo, que representem o mesmo local na mesma faixa de horário em um único ponto. Os próximos dos pontos de cada ponto agrupado são adicionados no ponto que representa todos eles. Para que esse próximo ponto seja adicionado na lista é feita uma verificação se este já não está lá e se esse não é o mesmo ponto em questão, como mostra a linha 22.

Como alguns pontos foram agrupados, os próximos pontos devem ser atualizados pelo ponto que agora os representa. Caso esse ponto atualizado seja o mesmo que o receberá como próximo, este é retirado da lista de próximos durante essa atualização. Com a realização dessa segunda etapa os pontos não pertencem mais a um único usuário e sim a um grupo de pessoas com alguma característica semelhante que realizaram uma conexão em um mesmo local na mesma faixa de horário.

Após o agrupamento, os pontos que foram agrupados são removidos. Por fim, é verificada a quantidade de próximos pontos de cada ponto, por exemplo, a partir do ponto *Sanders Theatre* em Figura 2, a quantidade de próximos pontos três (*Littauer Lot*, *John A. Paulson School* e *Harvard Graduate School*). Aqueles que não tiverem pelo menos β pontos tem seus registros de próximos apagados. Apenas os atributos grupo, nome do local, faixa de horário, latitude, longitude e próximos (caso existam) são divulgados os demais são suprimidos. O deslocamento entre um ponto (*stop*) seus próximos pontos (*stops*) é o que este trabalho considera como *move*.

7. Resultados

O algoritmo de anonimização foi testado em uma base de dados contendo mais de 1.5 milhões de registros de conexões realizadas por alunos da universidade no dia 15 de maio de 2018. Essa data foi escolhida pelo fato de que nesta semana não existiu feriado, não era época de férias e muito menos de recuperação, fazendo com que a movimentação dentro do campus fosse parecida com os outros dias do ano.

Os dados de *log* das conexões foram disponibilizados com autorização da superintendência da universidade, com o objetivo de incentivar a pesquisa de anonimização para que esses dados possam ser utilizados futuramente em pesquisas de outras áreas do conhecimento, garantindo assim a privacidade dos alunos. O trabalho também recebeu autorização para buscar nos sistemas da universidade, a partir dos identificadores pessoais presentes nos *logs*, os dados quase-identificadores dos alunos. Os *access points* foram agrupadas em 27 prédios que apresentam maior movimentação de pessoas dentro do campus. A escolha do quase-identificador do grupo precisar ser bem analisada. O grupo precisa ser grande para que o algoritmo não suprima muitos dados. Visto isso, quanto maior o grupo, maior a chance de encontros entre pessoas. Para esse experimento os alunos foram agrupados em 11 centros. A faixa de horário foi dividida de 15 em 15 minutos.

Como já comentado, o problema de encontrar o *k-anonymity* ideal é NP-Difícil. Por essa razão, os parâmetros β e k são escolhidos pelo usuário de acordo com o nível de privacidade que este deseja aplicar sobre os dados. Esse nível de privacidade é dado em porcentagem e representa as chances que um atacante tem de descobrir informações dentro da base de dados anonimizada. Realizando um análise sobre os dados armazenados no banco de dados, antes de serem anonimizados, foi notado que utilizando valores de k maiores que 50 alguns centros já não retornam mais dados visto que esses não tem

encontros com mais de 50 pessoas em algum lugar em uma certa faixa de horário. O primeiro gráfico da Figura 4 mostra que a maioria dos dados tem entre 2 e 70 registros indistinguíveis a ele e isso representa mais de 50% dos dados da base. Com isso, ao aplicar a métrica, deve-se balancear a perda de informação com o uso de um k muito grande e a baixa privacidade com o uso de um k muito pequeno.

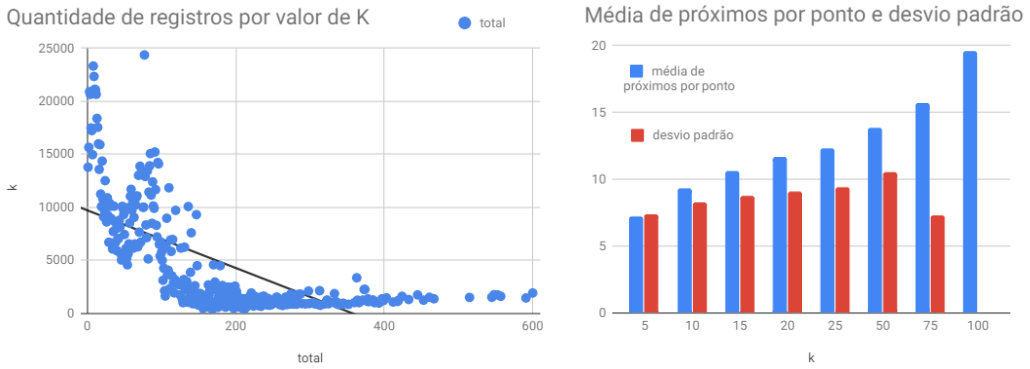

Figura 3. Gráficos de quantidade de registros por valor de k e média de próximos por ponto e desvio padrão.

O principal valor na medida de nível de privacidade é o β. Esse valor tem influência na probabilidade de um atacante conseguir extrair informações da base de dados anonimizada. Os três principais cenários são: 1 e 2) o atacante viu sua "vítima" entrando ou saindo em um certo prédio, em uma certa faixa de horário e conhece o grupo que esta pertence; 2) o atacante sabe que em uma certa faixa de horário a vítima estava em um certo prédio.

As primeiras observações servem para todos os casos. Se um atacante souber que a pessoa está em certo lugar, existem k outras pessoas indistinguíveis a ela que estão neste mesmo lugar e pelo menos β possibilidades de próximos lugares que serão visitados por pessoas desse grupo. O atacante pode saber a faixa de horário que a pessoa entrou ou saiu de certo prédio, existem algumas incertezas inerentes ao modelo de trajetórias criadas por dispositivos de Wi-fi. O atacante só poderá ter certeza que a pessoa possuía alguma dispositivo móvel junto a ela se, naquele momento, ele a viu portando o objeto. Ainda, o dispositivo móvel poderá estar sem bateria, desligado ou o usuários pode ter preferido utilizar a 4G ao invés da Wi-Fi. O dispositivo poderia estar até mesmo em modo avião ou até gerando falha na autenticação.

Na primeira situação, para que o atacante descubra de onde determinada pessoa veio antes de entrar no prédio, existem no mínimo k pessoas do mesmo grupo que fizeram conexão nessa faixa de horário. Portanto, mesmo que tenha apenas um indicador apontando para esse lugar, existem diversas possibilidades de inferência:

- Algumas das pessoas do grupo que chegaram a este lugar, vieram do lugar de partida do indicador.
- Outras pessoas podem ter feito sua primeira conexão nestes prédio.
- Também existem outras possibilidades como a conexão ter caído.
- A rede da universidade pode estar em manutenção nessa faixa de horário.

- Alunos estarem realizando prova com o celular desligado.
- A pessoa pode ter vindo por uma área aberta onde a conexão é fraca.
- Ela pode também ter apenas passado perto de um prédio e realizado a conexão.

Como foi mostrado, existem diversas possibilidades de inferências impossibilitando o atacante de ter certeza que da localização anterior. A única maneira de ter total certeza é seguindo a pessoa em todos os lugares que esta vai, fazendo com que não exista necessidade de consultar a base. Ao analisar os dados nenhum ponto recebeu apenas um indicador para que fosse possível realizar este ataque.

Na segunda situação existem no mínimo $k - 1$ pessoas com ela nessa faixa de horário e β possíveis próximos lugares visitados. Com isso, no mínimo $1/(\beta + 1)$ % de chances de acertar o próximo lugar. Esse valor um representa o caso desse local ter sido o último local que a pessoa realizou uma conexão. Foram testados diversos valores de k e a média de próximos lugares por ponto que foram gerados. O segundo gráfico exibido na Figura 4 mostra que quanto maior o k maior a quantidade média de próximos lugares. O desvio padrão se mantém entre a faixa de 7 a 11 registros. O maior valor de β encontrado foi 74.

A complexidade para se achar a trajetória de uma pessoa a partir de um ponto conhecido pode se tornar até exponencial. Tendo conhecimento de um ponto da trajetória, existem β possibilidades de próximos locais a serem visitados e cada possibilidade pode não ter nenhum próximo ponto, por não apresentar pelo menos β opções ou ser final, ou β outras possibilidades e cada uma dessas também seguem esse comportamento. Caso uma sequência de pontos gerem no mínimo x próximos, por exemplo, o primeiro ponto gera x, cada um dos próximos gera mais x e assim sucessivamente até aparecer um ponto final ou que não tenha no mínimo β próximos.

O fato de o local não apresentar próximos não significa que este não tem próximos locais. Isso pode significar apenas que nenhuma pessoa do grupo realizou novas conexões depois desse local ou que os pontos de próximo foram suprimidos. Na terceira situação, após aplicar o método de anonimização, o atacante tem chances remotas de conseguir descobrir os pontos que não pertencem ao local onde o aluno tem aula. Isso ocorre pois esse aluno terá aula com muitas pessoas semelhantes a ele fazendo com que diversos possibilidades de próximos lugares sejam criadas. Caso o aluno faça alguma matéria com pessoas não tão semelhantes ele será suprimido nessas localizações.

8. Conclusão

Este trabalho apresentou um cenário de coleta, limpeza e anonimização de dados de trajetória criados a partir de conexões a rede Wi-fi dentro de um campus universitário, assim como modelos de ameaça a este. Ao aplicar esta proposta, a instituição pode fornecer informações de mobilidade, importantes para pesquisas e tomada de decisões de outras áreas de atuação, garantindo a privacidade da trajetória dos indivíduos. Em vista disso, foi proposto um novo método de anonimização chamado de *Mix β-k-anonymity*. Este anonimiza as trajetórias individuais agrupando a de outros indivíduos que compartilham um quase-identificador em comum. A aplicação do método sobre dados coletados em um campus universitário mostrou que a comunidade acadêmica pode ter acesso a dados de qualidade para pesquisa operacional em mobilidade no campus onde a relevância demográfica é atendida com a manutenção da privacidade dos usuários. No que se refere

a trabalhos futuros, espera-se que o algoritmo seja otimizado sendo efetuadas mudanças para que este se torne mais eficiente e escalável. Outro ponto importante é o estudo do impacto da escolha errada de um quase-identificador para representar um grupo e possíveis ameaças que essa escolha errônea pode criar.

Referências

Abul, O., Bonchi, F., and Nanni, M. (2008). Never walk alone: Uncertainty for anonymity in moving objects databases. In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*, pages 376–385. Ieee.

Aggarwal, G., Feder, T., Kenthapadi, K., Motwani, R., Panigrahy, R., Thomas, D., and Zhu, A. (2005). Anonymizing tables. In *International Conference on Database Theory*, pages 246–258. Springer.

Alvares, L. O., Bogorny, V., Kuijpers, B., de Macedo, J. A. F., Moelans, B., and Vaisman, A. (2007). A model for enriching trajectories with semantic geographical information. In *Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems*, page 22. ACM.

Domingo-Ferrer, J. and Torra, V. (2001). Disclosure control methods and information loss for microdata. *Confidentiality, disclosure, and data access: theory and practical applications for statistical agencies*, pages 91–110.

Golle, P. and Partridge, K. (2009). On the anonymity of home/work location pairs. In *International Conference on Pervasive Computing*, pages 390–397. Springer.

Gramaglia, M., Fiore, M., Tarable, A., and Banchs, A. (2017). Preserving mobile subscriber privacy in open datasets of spatiotemporal trajectories. In *INFOCOM 2017-IEEE Conference on Computer Communications, IEEE*, pages 1–9. IEEE.

Hutchins, R. and Zegura, E. W. (2002). Measurements from a campus wireless network. In *Communications, 2002. ICC 2002. IEEE International Conference on*, volume 5, pages 3161–3167. IEEE.

Li, N., Li, T., and Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pages 106–115. IEEE.

Lu, Q., Wang, C., Xiong, Y., Xia, H., Huang, W., and Gong, X. (2017). Personalized privacy-preserving trajectory data publishing. *Chinese Journal of Electronics*, 26(2):285–291.

Ma, M., Zhao, K., Sui, K., Xu, L., Li, Y., and Pei, D. (2017). You can hide, but your periodic schedule can't. In *Quality of Service (IWQoS), 2017 IEEE/ACM 25th International Symposium on*, pages 1–6. IEEE.

Machanavajjhala, A., Gehrke, J., Kifer, D., and Venkitasubramaniam, M. (2006). l-diversity: Privacy beyond k-anonymity. In *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on*, pages 24–24. IEEE.

Mahdavifar, S., Abadi, M., Kahani, M., and Mahdikhani, H. (2012). A clustering-based approach for personalized privacy preserving publication of moving object trajectory data. In *International Conference on Network and System Security*, pages 149–165.

Meyerson, A. and Williams, R. (2004). On the complexity of optimal k-anonymity. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 223–228. ACM.

Monreale, A., Trasarti, R., Pedreschi, D., Renso, C., and Bogorny, V. (2011). C-safety: a framework for the anonymization of semantic trajectories. *Trans. Data Privacy*, 4(2):73–101.

Nergiz, M. E., Atzori, M., and Saygin, Y. (2007). Perturbation-driven anonymization of trajectories. Technical report, Technical Report 2007-TR-017, ISTI-CNR, Pisa.

Nergiz, M. E., Atzori, M., and Saygin, Y. (2008). Towards trajectory anonymization: a generalization-based approach. In *Proceedings of the SIGSPATIAL ACM GIS 2008 International Workshop on Security and Privacy in GIS and LBS*, pages 52–61. ACM.

Rajesh, N. and Abraham, S. (2017). Privacy preserved approach for trajectory anonymization through zone creation for halting points. In *Networks & Advances in Computational Technologies (NetACT), 2017 International Conference on*, pages 229–234. IEEE.

Schwab, D. and Bunt, R. (2004). Characterising the use of a campus wireless network. In *INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies*, volume 2, pages 862–870. IEEE.

Spaccapietra, S., Parent, C., Damiani, M. L., de Macedo, J. A., Porto, F., and Vangenot, C. (2008). A conceptual view on trajectories. *Data & knowledge engineering*, 65(1):126–146.

Sweeney, L. (2002). Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):571–588.

Tang, D. and Baker, M. (2000). Analysis of a local-area wireless network. In *Proceedings of the 6th annual international conference on Mobile computing and networking*, pages 1–10. ACM.

Terrovitis, M., Poulis, G., Mamoulis, N., and Skiadopoulos, S. (2017). Local suppression and splitting techniques for privacy preserving publication of trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 29(7):1466–1479.

Wang, F., Zhu, X., and Miao, J. (2017). Semantic trajectories-based social relationships discovery using wifi monitors. *Personal and Ubiquitous Computing*, 21(1):85–96.