UNIVERSIDADE FEDERAL DE SANTA CATARINA
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA

Fernanda Oliveira Gomes

PRIVACY PRESERVING ON SEMANTIC TRAJECTORIES: APPLICATION ON WI-FI CONNECTIONS OF A UNIVERSITY CAMPUS

Florianópolis

2019

Fernanda Oliveira Gomes

PRIVACY PRESERVING ON SEMANTIC TRAJECTORIES: APPLICATION ON WI-FI CONNECTIONS OF A UNIVERSITY CAMPUS

Dissertação submetida ao Programa de Pós-Graduação em Ciência da Computação para a obtenção do Grau de Mestre.
Orientador: Prof. Dr. Jean Everson Martina
Universidade Federal de Santa Catarina

Florianópolis

2019

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Oliveira Gomes, Fernanda
PRIVACY PRESERVING ON SEMANTIC TRAJECTORIES: APPLICATION ON WI-FI CONNECTIONS OF A UNIVERSITY CAMPUS / Fernanda Oliveira Gomes ; orientador, Jean Everson Martina, 2019.
92 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós Graduação em Ciência da Computação, Florianópolis, 2019.

Inclui referências.

1. Ciência da Computação. 2. Privacidade. 3. Anonimização. 4. Trajetórias. 5. Wi-Fi. I. Martina, Jean Everson. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Ciência da Computação. III. Título.

Fernanda Oliveira Gomes

PRIVACY PRESERVING ON SEMANTIC TRAJECTORIES: APPLICATION ON WI-FI CONNECTIONS OF A UNIVERSITY CAMPUS

Esta Dissertação foi julgada aprovada para a obtenção do Título de “Mestre”, e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciência da Computação.

Florianópolis, 15 de fevereiro 2019.

Prof. Dr. José Luís Almada Güntzel
Universidade Federal de Santa Catarina
Coordenador do Curso

Prof. Dr. Jean Everson Martina
Universidade Federal de Santa Catarina
Orientador

Banca Examinadora:

Prof^{a}. Dr^{a}. Carla Merkle Westphall
Universidade Federal de Santa Catarina

Dr^{a}. Chiara Renso
ISTI Institute of National Research Council (Videoconferência)

Prof. Dr. José Ripper Kós
Universidade Federal de Santa Catarina

Dedico este trabalho à minha família, em especial a minha mãe, que com muita luta proporcionou a melhor educação para mim e minha irmã. Também dedico ao meu namorado que esteve sempre ao meu lado me apoiando. E aos amigos pela paciência durante esta caminhada.

ACKNOWLEDGEMENTS

Ao Professor Jean Martina pela confiança, pela experiência compartilhada no decorrer dos últimos 6 anos e principalmente por acreditar em mim em momentos complicados e me orientar com muita paciência durante este trabalho.

Agradeço a participação dos membros da banca Profª. Drª. Carla Merkle Westphall, Drª. Chiara Renso, Prof. Dr. Jose Kos e os suplentes Profª. Drª Vania Bogorny, Prof. Dr. Ricardo Custodio e Prof. Dr. Werner Kraus.

Aos Professores, Ricardo Custodio, Carlos Becker Westphall, Alexandre Gonçalves Silva, Vania Bogorny, Elder Rizzon Santos e Jean Martina, pelas aulas ministradas e conhecimento compartilhado.

Aos colegas do LabSec, principalmente Douglas e Taciane, pela ajuda no desenvolvimento da escrita em inglês do trabalho e pelo convívio diário onde compartilhamos conhecimentos e experiências.

A minha família, pelo apoio e por uma vida inteira de dedicação, possibilitando que eu pudesse superar mais esta etapa da minha vida. Principalmente a minha mãe, que sem ela não poderia estar concretizando mais este sonho.

Agradeço a meu namorado, Bruno Machado Agostinho, por sempre estar junto comigo, me apoiando em todos os momentos, me ajudando, acreditando em mim, sendo meu companheiro dessa caminhada e fazendo eu ser sempre a melhor versão de mim.

A Capes pelo apoio a realização desta pesquisa.

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

ABSTRACT

With the increased use of mobile devices, which capture information from users' locations using GPS or cell phone signal, the number of trajectory data created from their spatio-temporal evolution has grown considerably. Information about the person who made the route and about the places visited can be added to the trajectory, thus creating a semantic trajectory. It is also possible to create trajectories with the Wi-Fi connections of a network by considering that each connection represent a point of the trajectory. In universities, or large building complexes, it is common to find many access points widely distributed throughout the campus aiming to cover most of the area with Wi-Fi signal. Each time a connection between a device and an access point is performed, the data generated (e.g. location, users identification, and time) are stored in a log file. This file allows us to track back the trajectories of the users inside the universities by using the access points location and time of the connection. With the identification attribute of the users, it is possible to associate it with quasi-identifiers present in the university data systems. This data can be useful for several areas of knowledge, ranging from security, urban planning, public transport management, to epidemic prevention (MONREALE et al., 2011). Publication of such data may put the privacy of university users in risk. If malicious persons have access to this data, stalking can be facilitated, as well as operational support for committing crimes, leading to a threat to the safety of people. Given that, our work proposes a method of anonymization called Mix β-k-anonymity. It anonymizes semantic trajectories by grouping the trajectory of people who share the same quasi-identifier. This approach provides a set of possible trajectories for a group of people with the same quasi-identifiers. We evaluated the method with Wi-Fi data from undergraduate students of UFSC, campus Trindade, and discussed the effectiveness of the approach in relation to the prevention of threats. After the application of this method, we showed that with the right choice of the quasi-identifier it is possible to anonymize semantic trajectories, making feasible the release of anonymized semantic trajectory data, with quality, for research in several areas of knowledge.

Keywords: privacy, anonymization, trajectories, Wi-Fi

RESUMO

O recorrente aumento da utilização de dispositivos móveis permite que mais dados de trajetórias dos usuários sejam coletados diariamente. Tais dados podem vir a ser obtidos via GPS, sinais de antena de celular e entre outros meios. Nesse viés, ao agregar informações acerca do local visitado e da pessoa que realizou o percurso pode-se transformar a trajetória em uma trajetória semântica. Uma trajetória semântica é uma sequência de *stops* (paradas) e *moves* (movimentos) de uma pessoa durante o seu trajeto. Outro modo de se obter dados de trajetória são por meio das conexões Wi-Fi. Cada conexão realizada é considerada como ponto espaço-temporal de um trajeto. Diante disso, as redes wireless das universidades, ou de grandes complexos, são comumente compostas por diversos pontos de acesso os quais são distribuídos amplamente em torno do campus de modo a cobrir uma extensa área com o sinal Wi-Fi. Nesse ínterim, toda vez que se realiza uma conexão entre um dispositivo móvel e um ponto de acesso, dados como localização, identificação do aluno e hora são gerados e armazenados em um arquivo de log. Tal arquivo permite rastrear as trajetórias dos usuários dentro das universidades por meio da localização dos pontos de acesso e também do horário no qual a conexão foi efetuada. Com isso, através do atributo de identificação do aluno, torna-se possível associá-lo à quase-identificadores pessoais presentes na base de dados da universidade. Essas informações podem ser úteis para diversas áreas do conhecimento, variando desde o planejamento urbano, gestão do transporte público até a prevenção de epidemias (MONREALE et al., 2011). No entanto, a publicação desses dados podem vir a colocar em risco a privacidade dos estudantes e funcionários, uma vez que pessoas mal-intencionadas possam vir a ter acesso à esses dados, perseguições podem ser facilitadas, assim como, realização de crimes, criando uma ameaça à segurança das pessoas. Frente à isso, a fim de divulgar essas informações, foi proposta uma técnica de anonimização para dados de trajetórias semânticas criadas a partir de conexões a redes Wi-Fi ou de outras fontes com pontos espaçados, chamados *Mix β -k-anonymity.* Tal abordagem utiliza um quase-identificador pessoal para agrupar pessoas e suas trajetórias. Para avaliar o método, foram criados modelos de ameaça adaptados para o cenário universitário. Com relação à escolha da variável quase-identificadora, foi realizado um estudo sobre o impacto de sua escolha. Infere-se com esse estudo que a escolha do atributo quase-identificador

é crucial para a preservação da privacidade. E por fim, o método foi avaliado com dados de Wi-Fi de alunos de graduação da UFSC e foi discutida a eficácia da abordagem com relação à prevenção de ameaças. Nessa perspectiva, a aplicação do método em um campus universitário demonstrou que a comunidade acadêmica pode ter acesso a dados de qualidade para realização de pesquisas de diversas áreas do conhecimento no campus mantendo a privacidade dos usuários.
Palavras-chave: privacidade, anonimização, trajetórias, Wi-Fi

RESUMO ESTENDIDO

INTRODUÇÃO

A anonimização é considerada um estado da privacidade (WESTIN; RUEBHAUSEN, 1967). A ideia central da anonimização é garantir que a pessoa não seja identificada, alcançada e rastreada (WAREKAR; PATIL, 2014). Outro conceito diz que a anonimização garante que um indivíduo seja indistinguível dentre outros em um lugar público (WESTIN; RUEBHAUSEN, 1967). O método de anonimização mais citado na literatura é o k-anonymity. Esse método de anonimização tem como objetivo tornar um dado indistinguível dentre pelo menos outros k - 1 que apresentam os mesmos atributos quase-identificadores, atributos que podem ser combinados com dados externos e expor o indivíduo, fazendo com que as chances de reidentificação de um indivíduo seja reduzida a 1/k (SWEENEY, 2002). As duas principais técnicas utilizadas para que uma base de dados atinja o k-anonimato são a generalização e a supressão dos dados. Um tipo de dado pessoal que vem sido cada vez mais coletado e analisado são os dados de trajetórias.

O recorrente aumento da utilização de dispositivos móveis permite que mais dados de trajetórias dos usuários sejam coletados diariamente. Tais dados podem vir a ser obtidos via GPS, sinais de antena de celular e entre outros meios. Nesse viés, ao agregar informações acerca do local visitado e da pessoa que realizou o percurso pode-se transformar a trajetória em uma trajetória semântica. Uma trajetória semântica é uma sequência de *stops* (paradas) e *moves* (movimentos) de uma pessoa durante o seu trajeto. Outro modo de se obter dados de trajetória são por meio das conexões Wi-Fi. Cada conexão realizada é considerada como ponto espaço-temporal de um trajeto. Diante disso, as redes wireless das universidades, ou de grandes complexos, são comumente compostas por diversos pontos de acesso os quais são distribuídos amplamente em torno do campus de modo a cobrir uma extensa área com o sinal Wi-Fi.

Nesse ínterim, toda vez que uma conexão entre um dispositivo móvel e um ponto de acesso é realizada, dados como localização, identificação do aluno e hora são capturados e armazenados em um arquivo de log. Tal arquivo permite rastrear as trajetórias dos usuários dentro das universidades por meio da localização dos pontos de acesso e também do horário no qual a conexão foi efetuada. Com isso, através do atributo de identificação do aluno, torna-se possível associá-lo à quase-identificadores pessoais presentes na base de dados da universidade. Essas informações podem ser úteis para diversas áreas do conhecimento, variando desde o planejamento urbano, gestão do transporte público até a prevenção de epidemias (MONREALE et al, 2011). No entanto, a publicação desses dados podem vir a colocar em risco à privacidade dos estudantes e funcionários, uma vez que pessoas mal-intencionadas possam vir a ter acesso à esses dados, perseguições podem ser facilitadas, assim como, realização de crimes, criando uma ameaça à segurança das pessoas.

Frente à isso, a fim de divulgar essas informações, foi proposta uma técnica de anonimização para dados de trajetórias semânticas criadas a partir de conexões a redes Wi-Fi ou de outras fontes com pontos espaçados, chamado Mix β-k-anonymity. Tal abordagem utiliza um quase-identificador pessoal para agrupar pessoas e suas trajetórias. Para avaliar o método, foram criados modelos de ameaça adaptados para o cenário universitário.

OBJETIVOS

O principal objetivo do presente trabalho é o desenvolvimento de um novo algoritmo para anonimizar dados de trajetórias semânticas com pontos escassos, de modo a possibilitar sua divulgação para pesquisa, visando melhorar a mobilidade, segurança e planejamento urbano dos locais onde os dados foram coletados, no nosso caso, universidades. Para atingir este objetivo, os seguintes objetivos específicos devem ser cumpridos:

- Desenvolver um algoritmo para anonimizar dados de trajetória semântica com pontos escassos que contenham um quase-identificador pessoal dos proprietários das trajetórias;
- Fazer uma análise da escolha da variável quase-identificador e seus impactos no nível de privacidade;
- Definir um modelo de ameaça que represente possíveis atacantes e ataques que possam ser executados com o objetivo de divulgar informações privadas de um data set anonimizado com o algoritmo proposto neste trabalho.

METODOLOGIA

Nas primeiras etapas do trabalho foi realizada uma revisão do estado da arte de métodos de anonimização de dados pessoais, trajetórias e trabalhos que utilizam conexões Wi-Fi a fim de criar trajetórias dentro de universidades. Diante disso, realizou-se um estudo acerca do funcionamento da coleta e armazenamento dos dados Wi-Fi, utilizando o serviço Eduroam, na Universidade Federal de Santa Catarina. Nesse viés, criou-se um sistema de extração, transformação e carga (ETL) dos dados de conexões Wi-Fi coletadas na UFSC, visto que o método de anonimização sugerido utiliza quase-identificadores pessoais, os quais são agregados aos dados dos indivíduos que realizarem conexões. Nessa perspectiva, tais dados são encontrados nos serviços da universidade e após compreendê-los, criou-se um método de anonimização de dados de trajetórias semânticas criadas a partir de conexões a redes Wi-Fi ou a outras fontes com pontos espaçados, chamado Mix β-k-anonymity. Esse método agrupa as trajetórias de pessoas que possuem o mesmo quase-identificador e utiliza conceitos similares aos de stops and moves e k-anonymity. Nesse ínterim, foi definido um modelo de ameaça a um data set anonimizado com Mix β-k-anonymity, o qual mostra as características do atacante e seus conhecimentos. Por conseguinte, realizou-se uma análise da escolha do quase-identificador de forma a mostrar os impactos que uma escolha errada pode trazer ao nível de privacidade da base de dados. Por conseguinte foram realizados experimentos a partir de dados de conexões Wi-Fi realizados por estudantes de graduação da UFSC. Com isso, os resultados desse experimento foram analisados em comparação às possíveis ameaças, mostrando como o algoritmo as trata.

RESULTADOS

A aplicação do método de anonimização Mix β-k-anonymity em dados de estudantes de graduação coletados em um campus universitário mostrou-se eficiente na prevenção de ameaças apresentada no modelo de ameaça. Com a avaliação do método foi concluído que quanto maior o k mais chances de próximos lugares distintos serem visitados. Outro fator que se mostrou impactante nos resultados é a escolha da variável quase-identificador. Com esse estudo, mostramos que a escolha do atributo quase-identificador é crucial para a preservação da privacidade. Nessa perspectiva, a aplicação do método em um campus universitário demonstrou que a comunidade acadêmica pode ter acesso a dados de qualidade para realização de pesquisas de diversas áreas do conhecimento no campus mantendo a privacidade dos usuários.

Palavras-Chave: Privacidade. Anonimização. Trajetórias. Wi-Fi.

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS LIST

UK	United Kingdom	27
EU	European Union	27
GDPR	General Data Protection Regulation	27
LGPD	Lei Geral de Proteção de Dados	27
UFSC	Universidade Federal de Santa Catarina	31
GIC	Group Insurance Commission	38
NWA	Never Walk Alone	43
EDR	Edit Distance on Real sequences	43
SUDA	Special Uniques Detection Algorithm	62
MSUs	Minimal Sample Uniques	62

CONTENTS

1 INTRODUCTION AND MOTIVATION

One of the highlight privacy events in early 2018 was the leakage of data from Facebook. The newspapers of New York Times and The Guardian announced that Cambridge Analytica had misused data from more than 50 million Facebook users (GUARDIAN, 2018)(TIMES, 2018). Cambridge Analytica was an advertising company that provided a data analysis service used to construct a strategic marketing plan for both corporate and election campaigns. The company had access to this data through a researcher, the psychology professor Aleksandr Kogan from the University of Cambridge, who claimed to collect it only for academic purposes. The professor had permission from Facebook to conduct a personality test through an application called "thisisyourdigitallife". The condition was that this data should be used only for academic purposes (GUARDIAN, 2018).

In response to the data leakage, Facebook banned Cambridge Analytica from advertise on its social network. Most of the data captured was from US citizens. This enabled Cambridge Analytica to conduct analysis of US voters' data in 2016 and deliver data analysis to one of their clients, Donald Trump, in his election campaign. The company was able to predict and influence the voters' vote by analyzing their profiles in the social network, helping Donald Trump with a strategic marketing plan based on this data. Another event that was supported by the company was the Brexit Referendum - name given for the impending withdrawal of the United Kingdom (UK) from the European Union (EU) - where the Cambridge Analytica supported the official campaign group in favor of leaving the EU. Due to all this data leakage problem, several countries substituted their data protection laws to new ones (TIMES, 2018).

The General Data Protection Regulation (GDPR) is a regulation created in Europe that came into force in May 2018. GDPR comes as a substitute for the Data Protection Directive 95/46/EC and intends to harmonize data privacy laws through Europe, protect and enable data privacy of EU citizens, and reshape the way that data privacy is addressed in organizations (GDPR, 2018). Similar to GDPR, the Brazilian Congress approved in July 2018 the PLC 53/18, *Lei Geral de Proteção de Dados* (LGPD) . The LGPD aims to protect the rights of freedom and privacy of citizens, as explained in Article 1 (CIVIL, 2018). Both GDPR and LGPD require pseudonymization in order to publish personal data. If the data is anonymized, GDPR and LGPD are not applicable.

Anonymity is considered a state of privacy (WESTIN; RUEBHAUSEN, 1967). The central idea of anonymization is to ensure that the person is not identified, reached and tracked by any means (WAREKAR; PATIL, 2014). Anonymization ensures that one individual is indistinguishable from others in a public place (WESTIN; RUEBHAUSEN, 1967). One of the most cited method of anonymization in literature is k-anonymity. This method of anonymization makes a data item indistinguishable from at least other $k - 1$ that have the same quasi-identifiable attributes, making the chance of re-identification of an individual reduced to 1/k (SWEENEY, 2002). The two main techniques used to achieve k-anonymity in a data set are generalization and suppression. One type of personal data that has been increasingly collected and analyzed is trajectory data.

With the increased use of mobile devices, which capture information from users' locations using GPS or cell phone signal, the number of trajectory data created from their spatio-temporal evolution has grown considerably. In addition to the movement of the individual, information about the person who made the route and about the places visited can be added to it, thus creating a semantic trajectory. A semantic trajectory is a trajectory that has been enhanced with semantic information and/or one or more complementary segmentation (PARENT et al., 2013). With this data, it is possible to extract information, through analysis and mining, on human mobility. This data can be useful for several areas of knowledge, ranging from security, urban planning, public transport management, to epidemic prevention (MONREALE et al., 2011).

In our work, instead of using GPS data, we create our trajectories data set with Wi-Fi connections in a university campus. Every time a mobile device performs an authenticated connection to the Wi-Fi system, several data, including location data, is generated. If all these connections are made at the same network, it is possible to track the steps of an individual by considering that each point of their trajectory is one connection with the Wi-Fi and all these connections are ordered by time. The ever growing use of mobile devices results in more people connecting to Wi-Fi networks and, consequently, leaving their traces to potential privacy violations.

Universities are a good example of a place with a lot of access points spread all over their campus. They capture a huge amount of data from their students, professors and employees daily. The data generated by the authentication process on campus Wi-Fi is an example. A wireless connection performed by a mobile device generates a log archive on the university system. This log shows: the user that was connected, MAC address of the mobile device and access point, time and location of the access point. Every connection on the access point can be used as a point in a path. A point is a pair of coordinates with temporal information. It is possible to aggregate contextual information at each point and create a semantic trajectory. An example of a semantic trajectory is the sequence of places visited by an object in movement, such as the sequence: Library, Coffee shop, Classroom, University restaurant and Bus stop. This kind of data allows any person to track people's mobility.

Since the log file keeps the users' identification, it is possible to combine this information with those users' data which already exists in university systems (e.g. registration information, school record, class schedule and many others). The availability of this data, known as quasi-identifiers, can generate a large data set for operational research. Projects such as arranging bus stops inside the university campus, creating new paths for pedestrians and cyclists, and areas for integration, are examples that can use this data.

The trajectory data set is prone to attacks from external sources even after anonymized. These attacks can be made through observation or disclosure of individual's location. An attacker can discover pieces of his victim trajectory, through spontaneous disclosure on social media or by observation of the victim daily behavior. It is also possible to re-identify the owner of the trajectory through their personal quasi-identifiers on the semantic trajectory. Sweeney in (SWEENEY, 2002) showed that it is possible to identify people linked to supposedly anonymized records. They proved that the removal of personal identifiable information is not sufficient to protect privacy. This is due to the fact that the disclosed data can be linked to other data sources through a set of quasi-identifier attributes common to both databases. For example, in the United States, the five-digit zip code combination, gender and date of birth is unique to 87% of citizens. In addiction, the disclosure of this data runs into privacy problems due to the fact that location and personal data allow intrusive inferences, which may reveal habits, social behavior, religious and sexual preferences of individuals (ABUL; BONCHI; NANNI, 2008). If malicious people have access to this data, stalking can be facilitated, as well as operational support for committing crimes, leading to a threat to the safety of people.

Recent research such as those in (ABUL; BONCHI; NANNI, 2008),(NERGIZ; ATZORI; SAYGIN, 2007),(GRAMAGLIA et al., 2017) and (SALAS; MEGÍAS;

TORRA, 2018) work on anonymizing trajectory data to release it. The works of (MONREALE et al., 2011),(RAJESH; ABRAHAM, 2017), (TERROVITIS et al., 2017), and (TU et al., 2018) anonymize semantic trajectories but only with quasi-identifiers of the trajectory itself (e.g. time, space, and name of the place). Only one paper is concerned with the anonymization of data generated within a university campus (MA et al., 2017), but their work does not provide any new solution for anonymizing this data.

In our work, we go one step further to existing approaches which only anonymize semantic trajectories with the quasi-identifiers of the trajectory, making the following contributions: (i) An algorithm for anonymizing semantic trajectory data with sparse points, that contains a personal quasi-identifier of the owner of the trajectory, by grouping the trajectories of people that have the same quasi-identifier; (ii) An analysis of the choice of the quasi-identifier variable and its impacts on privacy; (iii) a threat model that represents possible attacks that can be done in order to disclose privacy information of our anonymized semantic trajectory data set.

1.1 OBJECTIVES

The main objective of the present work is the development of a new algorithm to anonymize semantic trajectory data with sparse points in order to enable its disclosure for research, aiming to improve mobility, security and urban planning of the places in which the data was collected, in our case, universities. To achieve this goal, the following specific objectives must be fulfilled:

- To develop an algorithm for anonymizing semantic trajectory data with sparse points that contains a personal quasi-identifier of the owners of the trajectories;

- Make an analysis of the choice of the quasi-identifier variable and its impacts on privacy;

- To define a threat model that represents possible attacks that can be made in order to disclosure private information in our anonymized semantic trajectory data set, adapted to the university scenario.

1.2 METHODOLOGY AND STRUCTURE

The following tasks will be performed to achieve the objectives of this thesis:

1. Review of the state-of-the-art of trajectories and semantic trajectories anonymization.
2. Study of the main algorithms for anonymization of trajectory data.
3. Study of the Wi-Fi network of our university and the data that is stored about students, professors and employees.
4. Development of a system to extract, transform and load the Wi-Fi connections data collected by our university.
5. Conception and development of an algorithm for anonymizing semantic trajectory data in groups, using similar concepts of stops and moves and k-anonymity.
6. Analysis of the choice of the quasi-identifier variable and its impacts on privacy.
7. Definition of a threat model for our anonymized semantic trajectory database in order to identify the probable attacker's profile, his capabilities when he is in action and the threat scenarios.
8. Evaluate our anonymization algorithm with the data of undergraduate students collected by the Wi-Fi network of the UFSC (Universidade Federal de Santa Catarina), Trindade campus.

The remaining of our document is organized as follows: Chapter 2 explains the basics concepts. The 3 summarizes related work. Next, Chapter 4 proposes the Mix β-k-anonymity method, followed by the threat model and an analysis of the choice of the quasi-identifier. Chapter 5 shows the application scenario and the results obtained with the application of the method on the data collected by the Wi-Fi and services of UFSC. Finally, in Chapter 6 we conclude our work.

2 BACKGROUND

In this chapter we will present the main concepts of our work. We will define the concepts of privacy and data protection (2.1), k-anonymity (2.2) and trajectories (2.3).

2.1 PRIVACY AND DATA PROTECTION

In this subsection we bring the definitions of Privacy (§2.1.1), Anonymity (§2.1.1.1), Pseudononymity (§2.1.1.2), Unlinkability and Unobservability (§2.1.1.3) and Data Protection (§2.1.2).

2.1.1 PRIVACY

In the past few decades there have been several debates about the precise definition of privacy (YANES, 2014). The work of Solove (SOLOVE, 2008) select and categorize some concepts of privacy, as follows:

- The right to be let alone: The United States jurists Samuel D. Warren and Louis Brandeis wrote the article "The Right to Privacy". They describe in their work privacy as "right to be let alone" (WARREN; BRANDEIS, 1890).
- Limited access: Privacy can be described as the option that a person has in limiting the access to their personal information (SOLOVE, 2008).
- Control over information and States of privacy: Alan Westin (1967) argues that "privacy is the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others" (WESTIN; RUEBHAUSEN, 1967) and define four states/experiences of privacy: solitude, intimacy, anonymity, and reserve.
- Secrecy: Some authors define privacy as the right of people to hide their personal information, that can be misused by others (SOLOVE, 2008).
- Personhood and autonomy: Jeffrey Reiman (1976) conceptualized privacy as the "recognition of one's ownership of his or her

physical and mental reality and a moral right to his or her self-determination" (SOLOVE, 2008) (REIMAN, 1976).

- Self-identity and personal growth: privacy can be described as a prerequisite for the advancement of a sense of self-identity (SOLOVE, 2008).

- Intimacy: James Rachels (1975) affirms that there is a connection between the people that have access of our information and social relationships that we have with these people (RACHELS, 1975).

The Common Criteria for Information Technology Security Evaluation is an international standard (ISO/IEC 15408) for computer security certification. According to this standard, privacy is the "user protection against discovery and misuse of identity by other users" (CRITERIA, 2017), and this is the definition we adopt in this work. In order to achieve privacy, they list four requirements: Anonymity, Pseudonymity, Unlinkability, and Unobservability. The first three concepts are related to our work.

2.1.1.1 ANONYMITY

Anonymity is related to the identity of the user. Alan Westin (1967) said that anonymity is considered a state of privacy (WESTIN; RUEBHAUSEN, 1967). To Ruchira Warekar and Savitri Patil (2014) the central idea of anonymization is to ensure that the person is not identified, reached and tracked (WAREKAR; PATIL, 2014). Another definition says that anonymization ensures that one individual is indistinguishable from others in a public place (WESTIN; RUEBHAUSEN, 1967). To achieve the anonymity of a subject, it is necessary to have a set of subjects with potentially the same attribute. For an attacker, anonymity means that he is not able to identify the subject within a set of subjects (PFITZMANN; HANSEN, 2010). Another concept argues that anonymity is the "property that guarantees user's identity from being disclosed without consent" (YANES, 2014). According to GDPR, an anonymous information can not be linked to an identifiable natural person or to personal data in such a way that the owner of the information is no longer identifiable, as described in Recital 26 (GDPR, 2018).

2.1.1.2 PSEUDONONYMITY

The pseudonym identifies one or more individuals without reveling their true names (MAY, 1992). Most of the people that have a pseudonym use it in order to be anonymous, but to achieve anonymity is not an easy task and is often fraught with legal issues (POST, 1996). The CC standard considers that pseudonymity guarantee the use of a service or resource by a user without revealing their identity, yet retaining them responsible (CRITERIA, 2017). To the GDPR, pseudonymization "means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data is not attributed to an identified or identifiable natural person" (GDPR, 2018).

2.1.1.3 UNLINKABILITY

Unlinkability of user's information within a particular context is used to ensure the privacy of the user's identity (YANES, 2014). To Gerrit Bleumer (2011), Unlinkability "of two events occurring during a process under observation by an attacker is the property that the two events appear to the attacker after the process exactly as much related—or unrelated—as they did before the process started" (BLEUMER, 2011). In other words, we say that two items of interest (IOIs), from the attacker's perspective, are more or less related depending on his background knowledge. By considering that IOIs stand for a sender and receiver of messages, we can consider that both are anonymous if there is unlinkability between the IOIs and the individuals (PFITZMANN; HANSEN, 2010).

Our work aims to publish anonymized semantic trajectory data preserving the privacy of subjects who owns it. In order to achieve this goal, we propose an anonymization method and present a threat model to it based on the threats of likability. We also explain how our method achieves the anonymity, ensuring the privacy of the subjects' data, and avoid these threats.

2.1.2 DATA PROTECTION

When information about people's daily activities is available, it can easily be used by malicious individuals to monitor private activities (OLESHCHUK, 2009). In Brazil, the Lei Geral de Proteção de Dados (General Law on Data Protection) was approved on July 10th 2018. It aims to protect the fundamental rights of freedom and privacy and the free development of the personality of the natural person (CIVIL, 2018). The Law was based on the recent European data protection regulations, the General Data Protection Regulation (GDPR). The GDPR is a regulation created in Europe and came into force in May 2018. GDPR comes as a substitute for the Data Protection Directive 95/46/EC, aiming to harmonize data privacy laws through Europe, protect and enable data privacy of EU citizens, and reshape the way that data privacy is addressed in organizations (GDPR, 2018). Both GDPR and LGPD require that for personal data to be disclosed it should be pseudononymized, and follow the instructions of these laws in order to ensure privacy, as we can see in the article 13 of the LGPD and recital 156 of the GDPR.

> Art. 13. - In conducting public health studies, research agencies may have access to personal databases, which will be treated exclusively within the agency and strictly for the purpose of carrying out studies and researches and kept in a controlled and safe environment, in accordance with specific regulation and which include, wherever possible, the anonymization or pseudonymization of the data, as well as the due ethical standards related to studies and research.

> Recital 156 - [...] processing data which do not permit or no longer permit the identification of data subjects, provided that appropriate safeguards exist (such as, for instance, pseudonymization of the data).

If the data is anonymized, GDPR and LGPD are not applied, as shown in the article 12 of the LGPD and recital 26 of the GDPR.

> Recital 26 - [...] The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer

> identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes (GDPR, 2018).

Accordingly, some concerns about personal data are raised, e.g. how data is collected, stored, used, and if its use affects user's privacy. The personal data can be classified, according to a data privacy model, into three categories (CLARKE, 1999):

- Identifiers: Attributes that accurately identify the person, as for example, registration code, identity number, and passport number.
- Quasi-identifiers or Semi-identifiers: Attributes that, combined with other(s), can re-identify the person. In this case, they fall into: date of birth, course, department, gender, schedule, address, and others.
- Sensitive Attributes: These attributes contain confidential data about the person. Examples would be salary, vote, school history, and medical data.

Our anonymization approach uses a personal quasi-identifier to group people and then group their trajectories in order to confuse the attacker. We consider the trajectory as a sensitive information, and we do not use sensitive personal attributes. We also suppress all the identifiers. Since our method anonymizes the semantic trajectories GDPR and LGPD are not applied.

2.2 K-ANONYMITY

The k-anonymity concept was first introduced by Sweeney and Samarati in 1998 (SAMARATI; SWEENEY, 1998). They argue that a data set has the property of k-anonymity if for each record formed by quasi-identifiers attributes, there are at least other $k - 1$ records identical to it. In this context, the data in the data set refers to personal information that is conceptually organized as a table of rows and columns, that can be denominated tuple or record, and fields respectively (SWEENEY, 2002). So, we can also say that the data of a person is represented by a tuple. A tuple is a finite ordered list of attributes, and the union of the attributes' values represents a person. There are some attributes, different of those that uniquely identify a person that (e.g. name, passport number, and e-mail) when combined can also identify a person, as birth date and gender. These attributes are defined as quasi-identifier (DALENIUS, 1986).

Definition 1 (Attributes) *Let B($A_1, A_2, .., A_n$) be a table with a finite number of tuples. The finite set of attributes of B are ($A_1, A_2, .., A_n$).*

Definition 2 (Quasi-identifier) *Given a population of entities U, an entity-specific table T($A_1, A_2, .., A_n$), $f_c : U \rightarrow T$ and $f_g : T \rightarrow U'$, where $U \in U'$. A quasi-identifier of T, written Q_T, is a set of attributes $A_i, .., A_j \in A_1, A_2, .., A_n$ where: $\exists p_i \in U$ such that $f_g(f_c(p_i)[Q_T]) = p_i$.*

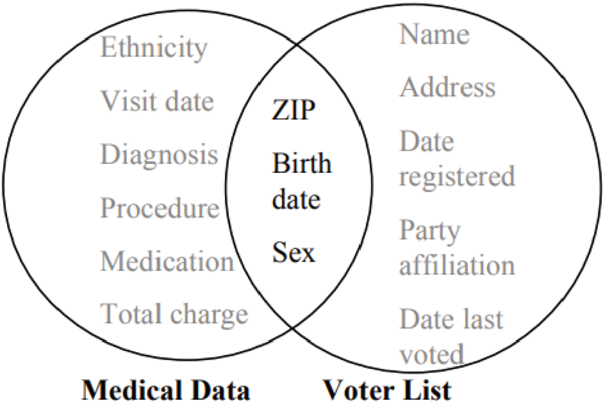


Figure 1 Linking to identify data (SWEENEY, 2002)

Sweeney conducted a study using the data collected by the Group Insurance Commission (GIC) of Massachusetts, responsible for hiding health insurance for the state employees. Figure 1 in the left hand side of circumference we see some examples of attributes present in this data set, they being ethnicity, visit date, diagnosis, ZIP code and others. GIC collects data from about 135,000 state employees and their families. At that time, they believed that this data was anonymized. So, they released this medical data to researchers, and sold to industry (SWEENEY, 2002). With this data in hands, Sweeney acquired another data set with the voter registration list for Cambridge Massachusetts. The right circumference of the Venn diagram, showed in Figure 1, shows the attributes of the voters' data set (e.g name, address, ZIP code, gender, and others). The intersection between both of each circumference represents the attributes that both data sets have in common. With these attributes, it is possible to link the information of the data sets with the attributes ZIP code, birth data, and gender in order to try to re-identify the individuals. To prove that it was possible to re-identify the people involved, Sweeney got these three information about William Weld (the governor of Massachusetts by the time). She concluded that six people had the same birth date as him, the gender of three of them were men, and he was the only one with his ZIP code (SWEENEY, 2002).

In order to solve this problem, Sweeney proposed the use of k-anonymity. The techniques most used to achieve the k-anonymity in a data set are generalization and suppression. These techniques are also called non-perturbative because they do not modify the data (DOMINGO-FERRER; TORRA, 2001). Generalization overrides attribute values by more generic values. Suppression is a technique that excludes attribute values from the anonymized data set.

There are two problems on k-anonymity solution however. The first one is called The Homogeneity Attack, which consists on the homogeneity of the sensitive attribute values. The other problem is called The Background Knowledge Attack and it is focused on the knowledge that the attacker has on an individual whose information is on the anonymous data set. With that knowledge, he can search on the anonymized data set for people with similar profiles, to discover the sensitive information attached to the person (MACHANAVAJJHALA et al., 2006). Thus, Sweeney believed that the k-anonymity was enough to solve the Background Knowledge Attack but later works shown that this was not enough to achieve anonymity, such as (MACHANAVAJJHALA et al., 2006) and (LI; LI; VENKATASUBRAMANIAN, 2007).

Table 1 shows a data set with k-anonymity using k equals to 4. As we can see, even using generalization and suppression if we know the ZIP code and age of the person that we are trying to re-identify it is feasible to figure out the sensitive attribute of this person. By knowing that the ZIP code starts with 130 and the age with 3 we can infer that this person has cancer. This is a good example of both The Background Knowledge Attack and The Homogeneity Attack. If we do not know any particular information of an individual, it is also possible to use external information sources, like related data sets, to re-identify the individuals.

The works of Meyerson and Williams (2004) (MEYERSON; WILLIAMS, 2004) and Aggarwal et al. (2005) (AGGARWAL et al., 2005) showed that finding an optimal k-anonymity is NP-hard. Nevertheless, even with these limitations, the method continues to be used as the basis of sev-

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	<30	*	Heart Disease
2	130**	<30	*	Heart Disease
3	130**	<30	*	Viral Infection
4	130**	<30	*	Viral Infection
5	1485*	>= 40	*	Cancer
6	1485*	>= 40	*	Heart Disease
7	1485*	>= 40	*	Viral Infection
8	1485*	>= 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Table 1 – Table with medical information data with 4-anonymity (MACHANAVAJJHALA et al., 2006)

eral anonymizing techniques, including trajectories. In our work, k-anonymity will be used as a base for the anonymization of trajectories and personal data. The techniques generalization and suppression are used to achieve it.

2.3 TRAJECTORIES

According to Monreale et al. (2011) a trajectory is described as a discrete sequence of points and formally defined on Definition 4 and graphically on figure 2. A point is a tuple that contains spatio coordinates and timestamp (BOGORNY et al., 2014), as described in Definition 3. These points represent a space-time evolution of the position of a moving object. It means that this object is moving in space, during a certain time interval, to reach a certain goal. A segment of the trajectory is considered a sub-trajectory. We can also consider a sub-trajectory itself as a trajectory (BOGORNY et al., 2014). Definition 5 presents the concept of sub-trajectory.

Definition 3 (Point) *A point p is a tuple (x,y,t), where x and y are the spatio coordinates that symbolize a place, and t is the timestamp that represents the time in which the point was collected.*

Definition 4 (Trajectory) *A trajectory is a list of spatio-temporal points* $p_0 < x_0, y_0, t_0 >, ..., p_n < x_n, y_n, t_n >$*, where* $x_i, y_i \in R$*,* $t_i \in R^+$ *for* $i = 0, 1, ..., n$ *and* $t_0 < t_1 < ... < t_n$*.*

Definition 5 (Sub-trajectory) *A sub-trajectory* s *of* T *is a list of points* $p_0 < x_0, y_0, t_0 >, ..., p_n < x_n, y_n, t_n >$*, where* $p_k \subset T$ *and* $k \geq 1$*, and* $l \leq n$*.*

A trajectory can be semantically divided in a temporal sequence of sub-intervals that represents the changing of positions of a moving object or a pause of it (SPACCAPIETRA et al., 2008). In this context, we can consider a semantic trajectory as a trajectory that has been enhanced with semantic information and/or one or more complementary segmentation (PARENT et al., 2013). A semantic trajectory is based on the concept of stops and moves where the user can enrich trajectories with semantic information of their application domain (BOGORNY; WACHOWICZ, 2008), as presented in Definition 6 and Figure 2.

Definition 6 (Semantic Trajectory) *A semantic trajectory* S *is a finite sequence* $I_1, I_2, ..., I_n$*, where* I_k *can be a stop or a move.*

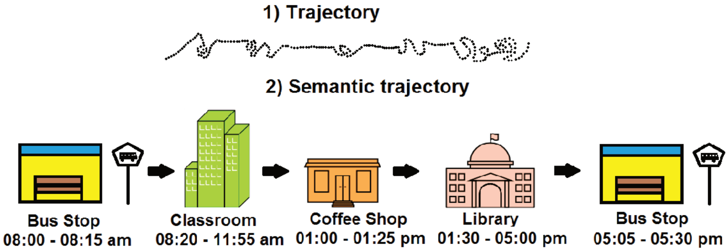


Figure 2 Example of a trajectory and a semantic trajectory

Definition 7 (Stop) *A stop is a sub-trajectory that starts by the time* $t_{inicial}$ *and ends at* t_{final}*. The differene between* $t_{final} - t_{inicial} \neq \emptyset$*. The moving object must to remain in the given position for a minimum period of time* ($\Delta t = t_{final} - t_{inicial}$)*, and each* $stop_1 \cap stop_2 \cap ... \cap stop_n = \emptyset$*.*

Definition 8 (Move) *A move is a spatio-temporal line,* $move_x$*, delimited for a* $stop_{inicial}$ *and a* $stop_{final}$*, having as* $t_{inicialx}$ *the* t_{final} *of* $stop_{inicial}$*, and as* t_{finalx} *the* $t_{inicial}$ *of* $stop_{final}$*.* $\Delta t = t_{finalx} - t_{inicialx}$*.*

This concept of semantic trajectory based on stops and moves was first introduced by Spaccapietra et al. (2008) in (SPACCAPIETRA et al., 2008) and Alvares et al. (2007) (ALVARES et al., 2007). Stops are the most important parts of a trajectory. They have a start and an end time, and happen when a moving object remains for a minimum period at an important place, as defined in Definition 7. Moves are the sub-trajectories, composed by the sample points, that describe the displacement between two consecutive stops, as presented in Definition 8. Another characteristic of the semantic trajectory is that it can receive contextual information, such as the means of transportation that were used or the name of the visited places (MONREALE et al., 2011).

In our work we will implement similar concept of stops and moves. Since we are working with sparse points, and we grouped the access points in areas, this spatial generalization made the users stay for a relevant period at the same area. So, we consider each point as a stop and the movement between two stops as the displacement.

3 RELATED WORK

In this chapter, we present related works that propose anonymization methods to trajectories in section 3.1. Section 3.2 presents some works that use Wi-Fi data of a university campus to construct trajectories. We finish this section summarizing our literature review and contextualizing our work among those presented at subsection 3.3.

3.1 TRAJECTORIES' ANONYMIZATION

The first technique that uses trajectories generalization by clustering near points in space and time was proposed by Abul et al. (2008) (ABUL; BONCHI; NANNI, 2008). The cluster-based Never Walk Alone ($\mathscr{NWA}$) approach takes advantage of the inherent uncertainty of the location of a moving object, and introduced the concept of *(k, δ)-anonymity.* In this approach the location of an object at a given moment is not a space-time point, but a circle of radius δ. The object can be anywhere within this limitation. To achieve k-anonymity, each trajectory is assigned to a group of at least k other trajectories using a greedy clustering algorithm. The concept of trajectory changes in this approach. Before, it was a sequence of points, and now it is a cylinder composed by consecutive circles. Then, the trajectories of each cluster are spatioly translated, so they all lie entirely within the same cylinder (area of uncertainty) of radius $\delta/2$. Figure 3 shows an example of an anonymity set formed by two co-localized trajectories, their uncertainty volumes, and the central cylindrical volume of radius $\delta/2$ that contains these two trajectories.

The work of Abul et al. (2008) (ABUL; BONCHI; NANNI, 2008) has an improved version of (ABUL; BONCHI; NANNI, 2010) called Wait-4-Me ($\mathscr{W}4\mathscr{M}$), made by the same authors, which uses the Edit Distance on Real sequences (EDR) distance instead of the euclidean distance. The problems with the Euclidian distance are that it can only use trajectories of exactly the same length, and it does not recognize similar trajectories with local shifts (e.g. similar trajectories that sometimes have sub-trajectories a little shifted in time) (ABUL; BONCHI; NANNI, 2010). The similarity of these works to ours lies in the fact that both work with an area of uncertainty and use k-anonymity. Other works followed similar concepts of trajectories generalization as in (NERGIZ; ATZORI; SAYGIN, 2008), (GRAMAGLIA; FIORE, 2015), (GRAMAGLIA et al., 2017), (TU et al., 2018), (LU et al., 2017) and (MAHDAVIFAR et al., 2012).

The paper of Nergiz et al. (2008) adopts the notion of k-anonymity for trajectories, and proposes a generalization-based approach for trajectory anonymization. They also presented a reconstruction algorithm based on randomization to release anonymized trajectory data (NERGIZ; ATZORI; SAYGIN, 2008). In Gramaglia and Fiore (2015) work (GRAMAGLIA; FIORE, 2015), they create an approach very similar with the Abul et al. (2010) work (ABUL; BONCHI; NANNI, 2010), but the major difference is the nature of the data used, in their case mobile fingerprints. The $\mathscr{W}4\mathscr{M}$ algorithm of (ABUL; BONCHI; NANNI, 2010) was created to anonymize trajectories with a lot of points, in other words, with a high frequency collection. When the authors, Gramaglia and Fiore, of (GRAMAGLIA; FIORE, 2015) test this algorithm with a data set where the trajectories have sparse points, the result showed that their proposed algorithm, GLOVE, has a better performance.

The work of Gramaglia et al. (2017) differs from the others because they propose a variation of k-anonymity. This method takes into account sub-trajectories that can be discovered by an attacker, and guarantees that there are at least *k-1* other identical sub-trajectories.

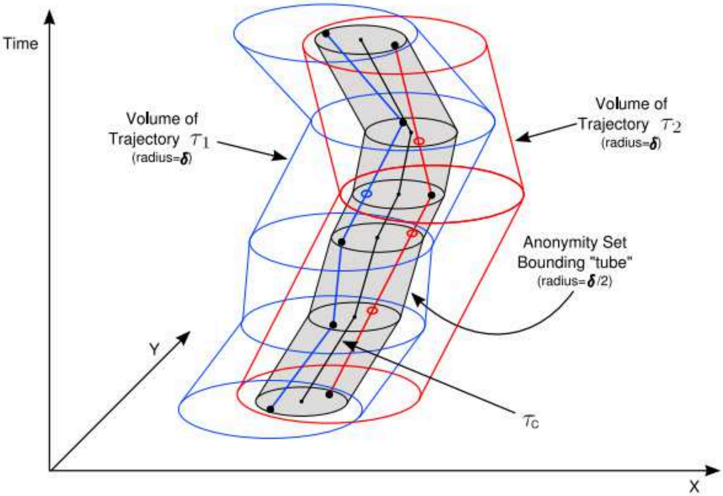


Figure 3 – Example of an anonymity set with (2, δ)-anonymity (ABUL; BONCHI; NANNI, 2008)

Their work presents two new algorithms. The first one is called *k-merge* and it accomplishes the generalization of the trajectories. The second is the implementation of the algorithm *kte-hide* that executes algorithms *k-merge* and the modified k-anonymity over the data set (GRAMAGLIA et al., 2017). Tu et al. (2018) in (TU et al., 2018) use a similar technique of spatio-temporal generalization to merge the trajectories as the one proposed in (GRAMAGLIA; FIORE, 2015), but also taking into account the semantic information of the points (in their case, the points are the base stations). They considered l-diversity and t-closeness to the merging process of spatio-temporal points, that recognize the semantic information related to each place where the point is located. A table is said to have l-diversity if there are at least "l" values for the sensitive attribute (MACHANAVAJJHALA et al., 2006). A table is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class is no more than "t" (LI; LI; VENKATASUBRAMANIAN, 2007).

The articles of Lu et al. (2017) (LU et al., 2017) and Mahdavifar et al. (2012) (MAHDAVIFAR et al., 2012) work with different levels of privacy. Lu et al. (2017) proposed a framework that provides privacy preservation services based on users' personal privacy requirements, and establishing a value maxR that represents the maximum distortion that the data set can have. Trajectories are grouped into clusters to satisfy a privacy restriction based on the user's personal privacy requirements. Their work anonymizes the trajectories, making possible to publish them without violating the privacy of individuals. As in (LU et al., 2017), Mahdavifar et al. (2012) introduces the idea of non-uniform privacy requirements using k-anonymity. Each trajectory is associated with its own level of privacy, so each trajectory has a number k of similar trajectories making all these trajectories indistinguishable among themselves. The method begins by dividing the trajectories into groups depending on their level of privacy. If a path has a high level of privacy (a high *k*), it will probably be part of a large cluster, and will suffer a great loss of information. Large clusters are very generalized, impacting on data quality (MAHDAVIFAR et al., 2012).

The proposal of Rajesh and Abraham (2017) aims to protect the privacy of trajectories by hiding stops. For example, if a moving object stays or visits frequently a hospital then the attacker may infer that this person is having serious health problems. In case of this object passes only in front of the hospital, the adversary can not infer anything. They proposed an algorithm that makes stops become part of zones. In the anonymized version of the trajectories, moves are kept in their original form (RAJESH; ABRAHAM, 2017). As in Rajesh and

Abraham (2017), Terrovitis et al. (2017) also uses the suppression technique. They divide the trajectories into sub-trajectories and use the *l-diversity* method (TERROVITIS et al., 2017). The work of Primault et al. (2015) proposes to hide by suppressing the places most visited by the person, like home or work. Their proposed method uses the k-anonymity concept (PRIMAULT et al., 2015).

Cicek, Nergiz and Saygin (2014) propose a method called p-confidentiality. They limit in p the probability of a place being visited. This value denotes that the chance of a user visiting a given place can not be more of p-per-cent, meaning that their method aims at the diversity of places visited (CICEK; NERGIZ; SAYGIN, 2014). Hu et al. (2010) proposed that at least k people be in a sensitive place together, the idea of k-anonymity (HU et al., 2010). The work of Monreale et al. (2011) classifies the stop as sensitive or not sensitive. The proposed algorithm is based on the generalization driven by the taxonomy of the place, thus providing a way to preserve the semantics of the trajectories (e.g. tourist point, a beach, Canasvieiras beach). A probabilistic limit, c-safity, is set regarding the chance of an adversary correctly inferring any sensitive place visited by a person, using his knowledge of part of their trajectory (MONREALE et al., 2011). The authors Lin et al. (2018) proposed a method of anonymization that randomly excludes some points of the trajectories in order to make harder the re-identification (LIN et al., 2018). The SwapMob method proposed by Salas, Megías and Torra (2018) presents a perturbative anonymization method based on the exchange of trajectory segments (SALAS; MEGÍAS; TORRA, 2018).

Most of the papers presented anonymize the trajectories using quasi-identifiers of the trajectories themselves. They also create clusters with similar trajectories in space and time. Other works use the suppression of points to anonymize the trajectories. Our work uses a quasi-identifier that represents a group of similar people. In our proposal, a stop is only disclosed if there are at least $k-1$ people of the same group in this stop within the same period of time. The displacement to the next location is displayed only if its start stop has other β moves made by people from the same group, as explained in Chapter 4.

3.2 TRAJECTORIES CREATED BY WI-FI CONNECTIONS INSIDE UNIVERSITIES

The first behavioral study based on trajectory data created by Wi-Fi connections on mobile devices was conducted by the Stanford University in 1999 (TANG; BAKER, 2000). The authors performed the monitoring of internet connections made by 74 volunteers using their laptops for 3 months in one of the university buildings. Data was collected using three different techniques (tcpdump, SNMP polling and authentication logs). A lot of similar works were made after that such as (SCHWAB; BUNT, 2004), (HUTCHINS; ZEGURA, 2002) and (WANG; ZHU; MIAO, 2017).

The article of Schwab and Bunt (2004) (SCHWAB; BUNT, 2004) presents a study of the University of Saskatchewan Campus WLAN. The campus is composed of 40 buildings that include public spaces, classrooms, library, laboratories, and offices. Traffic was tracked for a week using EtherPeek, a software package that allows physical address registration and provides network traffic information. The work of Hutchins and Zegura (2002) (HUTCHINS; ZEGURA, 2002) analyzed the WLAN for the Georgia Tech Campus for five months. They extracted information about user behavior from the authentication logs in the *firewall*. In the paper of Wang, Zhu and Miao (2017) (WANG; ZHU; MIAO, 2017) data were collected from Wi-Fi for a six-month period at a university. Their work presents a measure of similarity of semantic trajectories and estimates the level of intimacy of people. The main similarity found between these works and ours is that they all extract information of Wi-Fi connections in a university campus and turn it into trajectories.

The only work found that raises the question of privacy and anonymization of the trajectories created by Wi-Fi connections on a university campus is (MA et al., 2017). However, they do not present any proposal of anonymization method to solve the privacy problem. This article shows the risk of the dissemination of trajectory data using the anonymity method. They demonstrate that, by relating classes schedule to the trajectory data of the students, there is a possibility of identifying an individual.

3.3 LITERATURE REVIEW SUMMARY

In order to classify the related work and to contextualize the coverage and application of our work we divided them into the following categories:

(a) Generalization of points or trajectories

(b) K-anonymity

(c) Users' personal privacy requirements

(d) Suppression of points or sub-trajectories

(e) Trajectories created by Wi-Fi in universities

(f) Sparse points

(g) Introduction of personal quasi-identifiers

(h) Semantic Trajectories

Table 2 show the characteristics of the related work analyzed in the previous sections, as well as how the compare to our work.

Paper/Characteristics	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
ABUL; BONCHI; NANNI[2008]	x	x		x				
ABUL; BONCHI; NANNI[2010]	x	x		x				
NERGIZ; ATZORI; SAYGIN[2008]	x	x		x				
GRAMAGLIA; FIORE[2015]	x	x		x		x		
GRAMAGLIA et al.[2017]	x	x		x		x		
TU et al.[2018]	x	x		x		x		x
LU et al.[2017]	x	x	x	x				
MAHDAVIFAR et al.[2012]	x	x	x					
RAJESH; ABRAHAM[2017]	x							x
TERROVITIS et al.[2017]	x	x		x		x		x

Paper/Characteristics	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
PRIMAULT et al.[2015]	x	x		x				x
CICEK; NERGIZ; SAYGIN[2014]	x			x				x
MONREALE et al.[2011]	x		x	x				x
LIN et al.[2018]				x		x		
SALAS; MEGÍAS; TORRA[2018]								x
TANG; BAKER[2000]					x	x		
HUTCHINS; ZEGURA[2002]					x	x		
SCHWAB; BUNT[2004]					x	x		
WANG; ZHU; MIAO[2017]					x	x		
MA et al.[2017]		x			x	x		x
Our work	x	x	x	x	x	x	x	x

Table 2 – Related Work summary

As can be seen in the above table, most of the works we analyzed are related with (b) k-anonymity, specially because they either cover (a) Generalization or (d) Suppression as ways of rendering data private. The major difference in our work to the ones we reviewed is that we focus not only on (a), (b) and (d) but we also take into consideration (c) users' privacy requirements, as does Lu et al. in (LU et al., 2017). Our work complements Lu et al. (2017) because we deal with (h) Semantic Trajectories and (f) Sparse points.

Another important feature to notice on Table 2 regards the use of (e) trajectories created by Wi-Fi. Most of the works in this area do not take any consideration regarding privacy of the released data or the problems it creates. An exception to that is Ma et al. work (MA et al., 2017). Our work differentiates from Ma et al. (2017) because they only identify the problem and do not propose any solution for that.

An important characteristic of our work that was not found in others related works is (g) the use of personal quasi-identifiers to group users. We also bring a discussion about the better choice of such quasi-identifiers and how this can impact the security of the released data set.

4 PROPOSAL

In this chapter we present our proposal for capturing people's movement data from wireless associations, in our case within the university campus, and anonymizing this data to make it available for operational mobility research. In order to achieve this goal, we first show in Section 4.1 our proposed method, called Mix β-k-anonymity. In Section 4.2, we present the threat model of our scenario. Section 4.3, we explain how to choose a quasi-identifier and its impacts on the privacy level of the data.

4.1 MIX β-K-ANONYMITY

In this section, we propose an algorithm for anonymizing semantic trajectory data called Mix β-k-anonymity. This approach uses a personal quasi-identifier to group people and then group their trajectories in order to confuse the attacker. Another important characteristic of this algorithm is that it was designed to work with trajectories that have sparse points. It happens because the wireless access points are located in buildings, and we group the buildings in areas which we consider as stops. All the semantic trajectories' data sets that have sparse points and personal quasi-identifiers can use this algorithm as well, such as data collected by telecommunications companies. To those that does not have sparse points, it will be necessary to generalize the points in areas.

4.1.1 PARAMETERS AND CONCEPTS

There are three main concepts in this method: β value, k value, and group. The β value represents the minimum number of next possible places that a point must have in order to show these displacements. Figure 4 shows a point called "Biblioteca Central Universitária" representing the library of UFSC. This point has 5 next possible locations that people that were there went after being there. If we establish β equals to 5, all the points that have 5 or more next places visited will show the displacements to them (represented by an arrow in Figure 4). In cases where the β value is lower then 5, these displacements are suppressed.

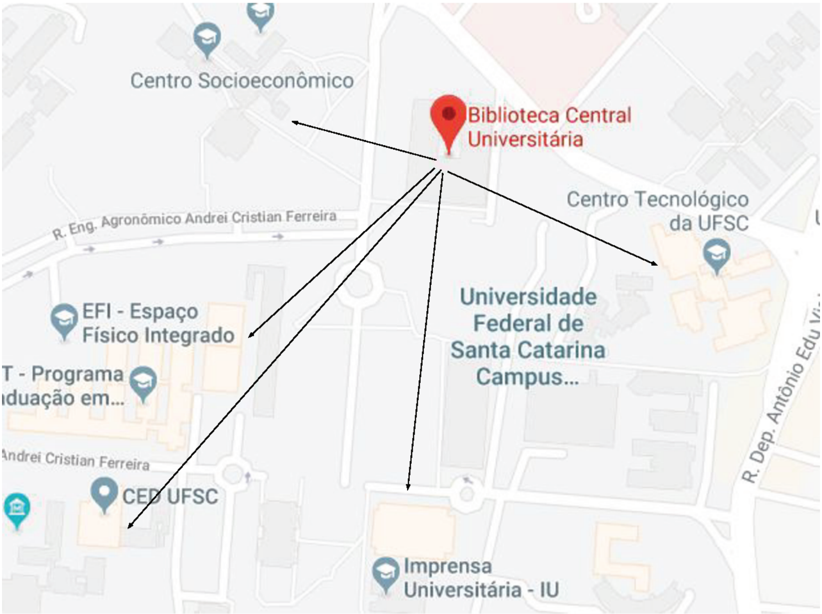
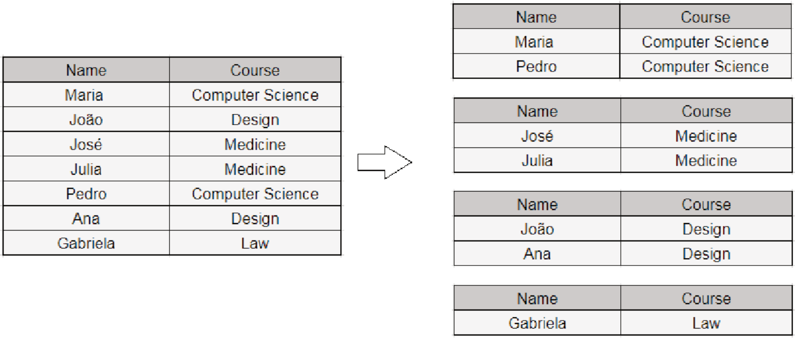

Centro Socioeconômico
Biblioteca Central Universitária
Centro Tecnológico da UFSC
R. Eng. Agronômico Andrei Cristian Ferreira
EFI - Espaço Físico Integrado
T - Programa ıduação em...
Universidade Federal de Santa Catarina Campus...
Andrei Cristian Ferreira
CED UFSC
R. Dep. Antônio Edu Vi...
Imprensa Universitária - IU

Figure 4 Map with a point and its 5 next points options

A group represents people that have the same quasi-identifier in common. Figure 5 shows an example of grouping the people by the quasi-identifier course. If we group the first table by gender, we will have as result 2 groups: the men's group (João, José e Pedro) and the women's group (Maria, Julia, Ana, Gabriela). In Section 4.3, we will show how to choose a quasi-identifier.

The k value stands for the minimum quantity of people of the same group that must be at the same point (represented by time range). Figure 6 presents an example where it is applied a k equals 2 to the first table. As a result, the second table contains only the records where the date, time range, location and course are the same. Since we consider that each tuple is a point, we can say that only the points that have more than 2 others identical to it of the same group are showed at the data set.

After introducing these concepts, we are able to prepare the data in order to anonymize it. The first thing that is needed to do before using the algorithm is to choose a quasi-identifier to group the people, and then group their data separately. The algorithm will be executed



Name	Course
Maria	Computer Science
João	Design
José	Medicine
Julia	Medicine
Pedro	Computer Science
Ana	Design
Gabriela	Law

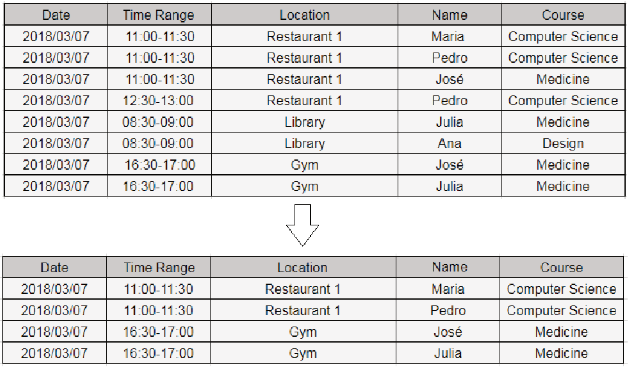
Name	Course
Maria	Computer Science
Pedro	Computer Science

Name	Course
José	Medicine
Julia	Medicine

Name	Course
João	Design
Ana	Design

Name	Course
Gabriela	Law

Figure 5 Example grouping people by their university course



Date	Time Range	Location	Name	Course
2018/03/07	11:00-11:30	Restaurant 1	Maria	Computer Science
2018/03/07	11:00-11:30	Restaurant 1	Pedro	Computer Science
2018/03/07	11:00-11:30	Restaurant 1	José	Medicine
2018/03/07	12:30-13:00	Restaurant 1	Pedro	Computer Science
2018/03/07	08:30-09:00	Library	Julia	Medicine
2018/03/07	08:30-09:00	Library	Ana	Design
2018/03/07	16:30-17:00	Gym	José	Medicine
2018/03/07	16:30-17:00	Gym	Julia	Medicine

Date	Time Range	Location	Name	Course
2018/03/07	11:00-11:30	Restaurant 1	Maria	Computer Science
2018/03/07	11:00-11:30	Restaurant 1	Pedro	Computer Science
2018/03/07	16:30-17:00	Gym	José	Medicine
2018/03/07	16:30-17:00	Gym	Julia	Medicine

Figure 6 Example of the data after the k-anonymization

for each group individually. After that, the time information of the points must be divided into time ranges such that the quantity of points per time range is enough to promote encounters of people with the same quasi-identifier at the same area. The idea of creating time ranges is to increase uncertainty and avoid attacks on the data set. Another benefit of the time range is that the bigger the time range the bigger the chances of people of the same group make connections at the same place. In order to execute the algorithm, it is required that the data set contains these attributes: unique identifier, user name, attribute quasi-identifier, location name, time, time range, latitude, and longitude.

4.1.2 ALGORITHM

We can divide the algorithm into 2 parts in order to make it easier. In the first part we remove the points of the trajectories of the individuals that do not have at least *k* people at, remove those points that were collected out of the operation time of the university, and for each point set the next point (place) that was visited by the person. We suppress the points that are out of the operational time of the university to avoid cases where there are enough people grouped in a time range at the same place but that are not a lot of people at the university at this time. If these people start to be together more often an attacker can get this pattern and commit a crime.

Table 3 shows an example of data set. Table 4 represents this data set after the first part of the algorithm being executed using k equals 2. As we can see, the points (C, 19:00-19:30) and (B, 17:00-17:30) were unique and for that reason they were suppressed, and now we have a column named next point that was set with the next point information.

Time Range	Location	Name
11:00-11:30	A	Bruno
12:00-12:30	C	Bruno
13:00-13:30	B	Bruno
18:00-18:30	D	Bruno
07:30-08:00	D	Eduarda
08:30-09:00	B	Eduarda
11:00-11:30	A	Eduarda
15:00-15:30	A	Eduarda
17:00-17:30	B	Eduarda
18:00-18:30	D	Eduarda
07:30-08:00	D	Fernanda
11:00-11:30	A	Fernanda
13:00-13:30	B	Fernanda
13:00-13:30	B	Fernanda
15:00-15:30	A	Fernanda
08:30-09:00	B	Maria
11:00-11:30	A	Maria
12:00-12:30	C	Maria
18:00-18:30	D	Maria
07:30-08:00	D	Pedro

Time Range	Location	Name
12:00-12:30	C	Pedro
13:00-13:30	B	Pedro
18:00-18:30	D	Pedro
19:00-19:30	C	Pedro

Table 3 – Example of data set

Point (Location, Time Range)	Next Point	Name
(A, 11:00-11:30)	(C, 12:00-12:30)	Bruno
(C, 12:00-12:30)	(B, 13:00-13:30)	Bruno
(B, 13:00-13:30)	(D, 18:00-18:30)	Bruno
(D, 18:00-18:30)		Bruno
(D, 07:30-08:00)	(B, 08:30-09:00)	Eduarda
(B, 08:30-09:00)	(A, 11:00-11:30)	Eduarda
(A, 11:00-11:30)	(A, 15:00-15:30)	Eduarda
(A, 15:00-15:30)	(D, 18:00-18:30)	Eduarda
(D, 18:00-18:30)		Eduarda
(D, 07:30-08:00)	(A, 11:00-11:30)	Fernanda
(A, 11:00-11:30)	(B, 13:00-13:30)	Fernanda
(B, 13:00-13:30)	(B, 13:00-13:30)	Fernanda
(B, 13:00-13:30)	(A, 15:00-15:30)	Fernanda
(A, 15:00-15:30)		Fernanda
(B, 08:30-09:00)	(A, 11:00-11:30)	Maria
(A, 11:00-11:30)	(C, 12:00-12:30)	Maria
(C, 12:00-12:30)	(D, 18:00-18:30)	Maria
(D, 18:00-18:30)		Maria
(D, 07:30-08:00)	(C, 12:00-12:30)	Pedro
(C, 12:00-12:30)	(B, 13:00-13:30)	Pedro
(B, 13:00-13:30)	(D, 18:00-18:30)	Pedro
(D, 18:00-18:30)		Pedro

Table 4 – Data set from Table 4 after the execution of the first part of the algorithm

The second part of the algorithm group the trajectories. All the points that represent the same place at the same time range are grouped. To group a point, it is necessary to also grouping the next point of each point. Table 5 represents the data set after grouping the points. It can be seen that all the points that represented (A, 11:00-

11:30) were grouped in one single point. The next point of each of them was put in the next point list. So now, we have a point and a list of next possible places that were visited after. The last part is to suppress the list of next points of the points that do not have at least β distinct next places visited. The result can be seen in Table 5, where it was applied a β value of 2. The points (B, 08:30-9:00) and (A, 15:00-15:30) do not show their next places because they have only one distinct next place. The point (D, 18:00-18:30) does not have any next place with at least k people.

Point	Next Points	Total Next Points
(A, 11:00-11:30)	(C, 12:00-12:30), (A, 15:00-15:30), (B, 13:00-13:30), (C, 12:00-12:30)	3
(C, 12:00-12:30)	(B, 13:00-13:30), (D, 18:00-18:30), (B, 13:00-13:30)	2
(B, 13:00-13:30)	(D, 18:00-18:30), (A, 15:00-15:30), (D, 18:00-18:30)	2
(D, 18:00-18:30)		0
(D, 07:30-08:00)	(B, 08:30-09:00), (A, 11:00-11:30), (C, 12:00-12:30)	3
(B, 08:30-09:00)	(A, 11:00-11:30), (A, 11:00-11:30)	1
(A, 15:00-15:30)	(D, 18:00-18:30)	1

Table 5 – Data set after grouping the points

Our proposed algorithm, showed in Algorithm 1, anonymizes the semantic trajectory data of a group in the daily granularity. We work with time range that represent the generalization of the connection time to an access point. This time range turns into the temporal information of a point in the path. The idea of creating time ranges is to increase uncertainty and avoid attacks on the data set. Another benefit of the time range is that the bigger the time range the bigger the chances of people of the same group make connections at the same place. Then it is necessary to group the data by: group (quasi-identifier attribute common to a large group of people), place and time range. With this, the number of records that present the same attribute is counted. The number of people from a certain group who were in the same place and in a certain time range is represented by x.

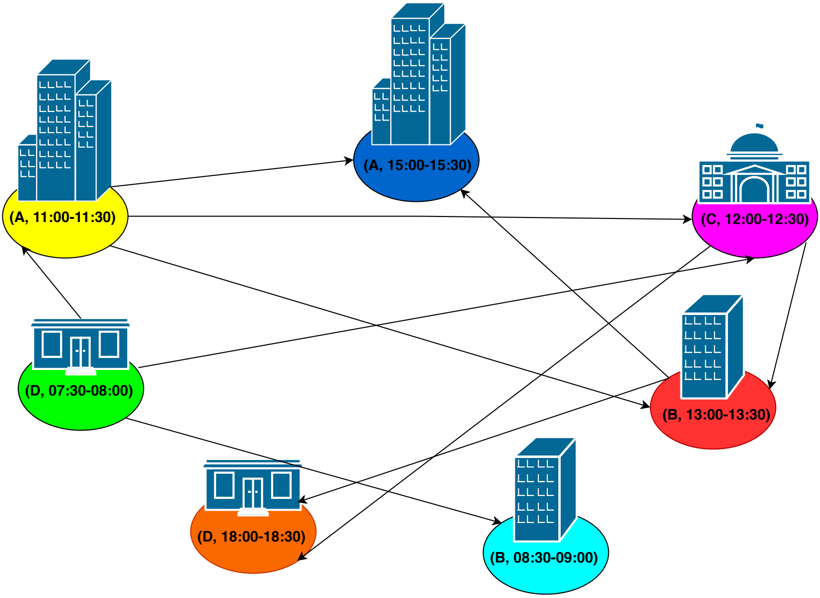


Figure 7 – Data set after the complete execution of the algorithm

Algorithm 1 receives a list of trajectory points from people of the same group. Each point contains the attributes: id, location name, time, time range, latitude, longitude, number of people who were also in this location within this time range, user, an empty list of next points, and a variable of type Boolean, named as grouped, starting with "false" value. The algorithm must run once for each group. In lines 1-2, points that are not among the operational time of the university, and do not concentrate at least an amount of k people, are removed.

On lines 3-12, the list is traversed bottom up as the next point visited is always below the current point in the list. There is no next point when it represents a user's last connection. Line 5 checks if there is no next point for that current point. It also checks whether the user of the current point and the next point are not the same. This verification is made because if the comparison between users returns false, the current point is the last point of this user's trajectory, or there are no next points.

Line 8 checks if the location and time range of the current points are the same. This happens when multiple connections are made in the same location over a short and continuous time interval. So this

Algorithm 1: Mix β-k-anonymity

```
    Input  : An array $T$ of size $l$ in ascending order sorted by time and user.
             Four integers: $\beta$, $k$, $openHour$ and $closeHour$ .
    Output: $T$ anonymized.
 1  FilterByOperatingHours($openHour$,$closeHour$);
 2  RemovePointWithoutKMin($T$,$k$);
 3  for $y \leftarrow l - 1$ to $0$ do
 4      point $\leftarrow T[y]$;
 5      if $y == l - 1$ or point.$user \neq$ nextPoint.$user$ then
 6          nextPoint $\leftarrow$ point;
 7      else
 8          if nextPoint.$place ==$ point.$place$ and nextPoint.$timeRange ==$
              point.$timeRange$ then
 9              point.grouped $\leftarrow$ true ;
10          else
11              AddOnNextPointsList(point, nextPoint);
12              nextPoint $\leftarrow$ point;
13  for $i \leftarrow l - 1$ to $0$ do
14      for $x \leftarrow i$ to $0$ do
15          if $T[i].timeRange \neq T[x].timeRange$ and $T[i].time > T[x].time$
              then
16              $x \leftarrow -1$;
17          if $x > -1$ and $T[i].place == T[x].place$ and
              $T[i].range_time == T[x].range_time$ and $x \neq i$ and $!T[i].grouped$
              and $!T[x].grouped$ then
18              foreach $nextPoint \in T[x].listOfNextPoints$ do
19                  if !AlreadyExistsOnTheNextPointsList ($nextPoint$,
                      $T[i].nexts$) and $T[x].id \neq T[i].id$ then
20                      AddOnNextPointsList($T[i]$, $next$);
21              $T[x]$.grouped $\leftarrow$ true;
22      UpdateListofNextPoints($T[i]$);
23  RemoveGroupedPoints($T$);
24  foreach $point \in T$ do
25      if $point.listOfNextPoints.length < \beta$ then
26          RemoveListofNextPoints ($point$);
```

trajectory is grouped to the previous one, changing only the *flag* of grouped to "true". Lines 10-12 represent the situation where the current point has a next point that is not situated in the same location and time range. In this case, the next point is added to the list of next points visited from the current point, and the next point receives the current point.

Each point has a list of next places that have been visited by more than k people of that group, and each next point has its next points and so on. Hence, several possible trajectories that can be carried out by a person of that group are created. There is a chance of connecting a person to a path depending on the value of β and k as will be shown in the 5th chapter.

The second part of the algorithm lies between lines 13-26. In this step, all the points are grouped by the same place in the time range in a single point. The next points of each grouped point are added at the point representing all of them. In order to add this next point into the list, we verify whether it is not already there and if it is not the same point in question, as shown in line 20.

After grouping, the points that were grouped are removed. Finally, the amount of next points of each point is checked. Those who do not have at least β points have their next points removed. Only the group attributes, location name, time range, latitude, longitude, and next points list (if it exists) are released. The other attributes are suppressed. The displacement between a stop and its next points is what our work regards as move.

The complexity of Algorithm 1 is $O(l^2)$ in the worst case, where the l represents the size of the list of points. However, we made the algorithm more efficient by doing some optimizations. First, we require as input an array ordered by time and user. We avoid the whole execution of the *for* by breaking it in line 14. We break the *for* when the point being compared is from a different time range, avoiding unnecessary comparison once the list is ordered and all the points after that will be in another time range. Since we are grouping the points in this part of the algorithm (part 2, as we call it), we do not need to compare the current evaluated point with points that are out of its time range. We start the for in line 14 with the value of variable "i" from the for in line 13. This action also reduces needless executions, as the points that are above the current point on the list are also in another time range. Taking these actions, we reduce a lot the complexity of the algorithm.

In the next section, we will define a threat model that represents possible attacks to disclosure privacy information of our anonymized semantic trajectory data set, adapted to the university scenario. We will also describe the probable attacker's profile, his capabilities when he is in action and the threat scenarios.

4.2 THREAT MODEL

The threat model of our scenario concerns on the re-identification of anonymized semantic trajectory information. In other words, the recognition of the owners of the trajectories, since it is the knowledge that the attacker attempts to acquire about the victim. The main weapon that the attacker has is his background knowledge. We assume that the background knowledge is some spatio-temporal points of the victim's trajectory. The attacker aims to acquire individual's movement at some spatio-temporal points in the trajectory by using location and personal semantic information. In order to acquire this information, the attacker can be someone who knows the victim. If the attacker does not know the victim, he can stalk them and obtain the background information. By analyzing the behavior of the victim, the attacker can create possible trajectories that the victim make daily and compare it with the anonymized trajectories released.

The threat models were created based on the scenario where the data is captured: a university campus. An attacker who has access to an anonymized semantic trajectories data set, that we will define as ST*, and knows which anonymization method was used, may possess some background knowledge that allows him to make attacks by doing inferences on the data set. In order to operate an attack, we assume the following adversary knowledge.

Definition 9 (Knowledges of Attacker) *Given an anonymized database of semantic trajectories performed by groups ST*, the attacker may know:*

1. *The method of anonymization used.*
2. *If the person has an anonymous trajectory t*, where t*$\subset$ ST*.*
3. *All groups g that belong to G.*
4. *If a group $g \subset G$ and the person $p \subset g$.*
5. *If a sub-trajectory performed by a person p st $\subset$ ST*.*

Definition 9 asserts that for an attacker to be able to perform the attack at the anonymized semantic trajectory data set the first thing he must know is the anonymization method used. By knowing that the method is the Mix β-k-anonymity, he must be sure that the victim trajectory is on the data set. Another two important things are how people were grouped, and to which group this person belongs. And finally, he must know a sub-trajectory performed by the victim. With this sub-trajectory, the attacker can compare it with all the sub-trajectories of the anonymized data set.

Definition 10 (Spontaneous Disclosure or Observation) *An attacker can find the location of his victim on social media (e.g. Facebook,* *Instagram, Twitter, and others). He can also observe the behavior of an individual i by following him, and then figure out places visited by i at a certain time, forming a sub-trajectory set st. With this knowledge, he can discover the full trajectory of i by researching for st on an anonymized semantic data set ST*. This Threat model is an adaptation of the one that is presented in (MONREALE et al., 2011).*

The Spontaneous Disclosure or Observation attacks are shown in Definition 10. These attacks can be executed even if the attacker does not personally know the victim. If they are well executed, the attacker can use this data to infer habits, social behavior, religious and sexual preferences of individuals (ABUL; BONCHI; NANNI, 2008). This data can also be used as support for the commission of crimes, leading to a threat to the safety of the victim. In order to deal with these attacks, we develop the Mix β-k-anonymity approach in order to make their execution more difficult.

One of the most important thing regarding the privacy level of the data set, when we use the Mix β-k-anonymity method, is the choice of the quasi-identifier. A wrong choice of the quasi-identifier attribute, which will characterize the groups, may decrease the degree of privacy of the data available and make the data set more vulnerable to these attacks. In order to understand impact of this choice, we show in Section 4.3 how to choose it by describing the most important characteristics of a good quasi-identifier.

4.3 CHOOSING THE QUASI-IDENTIFIER

The availability of trajectory data made by a group of people with some common characteristics can be useful for researches of several areas of knowledge. The wrong choice of the quasi-identifier attribute, which will characterize the groups, may decrease the degree of privacy of the data available. The first characteristic that an attribute has to have is to present a significant number of individuals per group. The bigger the group, there are more chances of happening encounters in the university. As there are many people in this group, and it tends to generate many encounters, a large number of points will be displayed, increasing the quality of the trajectory data disclosed.

The second characteristic is the difficult of linking this variable with an external information base, e.g. using the neighborhood attribute to group people can facilitate the inference of the people financial situation. To circumvent this problem, the idea is to generalize neighborhoods attribute for regions with the help of clustering algorithms.

Another important point is the amount of quasi-identifiers used. The more quasi-identifiers are used, the smaller the groups formed and greater the chance of making an inference with an external base (SWEENEY, 2002). If we group by gender and course, those courses that have much less women than men will suppress their movement from the results. This happens because some combinations of values do not meet the requirements of minimum amount of people at a certain point, since few people belong to this group. Even though the privacy of the person is not violated, the quality of data is decreased since important data is suppressed. This is also true for courses where men are the minority.

The last characteristic is the quantity of unique values that an attribute quasi-identifier has and also the quantity of unique sets created when this quasi-identifier is linked with others. In order to evaluate this attributes, the sdcMicro library of the RStudio tool (R, 2018) provides statistical disclosure control methods for anonymization of microdata and risk estimation. In order to verify the impact that each variable has, on the re-identification of individuals, after the suppression and generalization, this library uses the Special Uniques Detection Algorithm (SUDA) algorithm (ELLIOT; MANNING; FORD, 2002). This method is based on the special uniques of the records. A set is defined as a special unique if it is a sample unique both on the complete set of quasi-identifiers and simultaneously has at least one Minimal Sample Uniques (MSU), unique attribute sets without any unique subsets (ELLIOT; SKINNER; DALE, 1998). SUDA is divided in two steps. In the first one, all unique attribute sets are evaluated at record level. The minimum size of the sets is established by the user. To evaluate, SUDA only considers sets that are MSUs algorithm.

	Age Range	Labor Status	Residence	Kids (<= 10 years)
1	30 - 35 years	Employed	Urban	2
2	30 - 35 years	Employed	Urban	2
3	30 - 35 years	Employed	Rural	2
4	30 - 35 years	Employed	Rural	2
5	25 - 30 years	Unemployed	Urban	1
6	20 - 25 years	Unemployed	Urban	3
7	20 - 25 years	Unemployed	Urban	2
8	40 - 45 years	Employed	Rural	1

Table 6 – Data set example

In the Table 6, one example of MSU presented in the record 8 is the set {Employed, 1}. This set is MSU because none of its subsets, {Employed} or {1}, is unique in the sample. Although {40 - 45 years, Employed, rural, 1 } is a unique attribute set, it is not a MSU because its subsets {40 - 45 years, Employed, 1} and {Employed, 1} are both unique subsets in the data set. The potential risk of the records is driven on two considerations: "1) the smaller the size of the MSU within a record, the greater the risk of the record, and 2) the larger the number of MSUs possessed by the record, the greater the risk of the record" (TEMPL et al., 2013). The evaluation is made for each MSU of size k contained in a given record. The score is computed by $\prod_{i=k}^{M}(ATT - i)$ for each MSU. The ATT variable represents the total number of attributes in the dataset. The smaller the size k of the MSU, the larger the score for the MSU. The final SUDA score for the record is the sum of the scores of each MSU. The records with more MSUs have a higher SUDA score (TEMPL et al., 2013). As an example of how SUDA scores operates, record 8 in Table 6 has two MSUs: {40 - 45 years} of size 1, and {Employed, 1} of size 2. By setting the maximum size of MSUs on 3, the score of (40 - 45 years) is computed by $\prod_{i=1}^{3}(4 - i) = 6$, and the score of (Employed, 1) is $\prod_{i=2}^{3}(4 - i) = 2$. The SUDA score for the record 8 is 8.

Variable	Contribution
center	13.00
marital status	82.15
age	38.20
gender	26.83
course	84.92

Table 7 – Table with the contribution of each categorical key variable to the SUDA scores.

We checked some quasi-identifiers of the students that are on the log file of the Wi-Fi of our university, on May 16th, using the SUDA algorithm in order to choose a quasi-identifier that had fewer unique values. Table 7 presents the contribution of each categorical key variable to the SUDA scores. The contribution of an attribute is the quantity of MSUs that have this variable compared to the total of MSUs. Table 7 shows that the quasi-identifiers marital status and course presented a contribution of 82.15% and 84.92%, respectively, and it indicates that these two variables have more unique attributes.

It happened because by inspecting the data we discovered that some courses had few people making connections with the Wi-Fi on May 16th of 2018, and in the case of marital status few widowed people made connections that day. The quasi-identifiers center, age range and gender had a score of 13.0%, 38.20% and 26.83%, respectively.

Center	Frequency	Percentage
CENTRO DE CIÊNCIAS DA SAÚDE	1810	12.57
CENTRO DE CIÊNCIAS AGRÁRIAS	718	4.99
CENTRO DE CIÊNCIAS BIOLOGICAS	550	3.82
CENTRO DE CIÊNCIAS FÍSICAS E MATEMÁTICAS	896	6.22
CENTRO DE CIÊNCIAS JURÍDICAS	576	4.00
CENTRO DE COMUNICAÇÃO E EXPRESSÃO	1602	11.13
CENTRO DE DESPORTOS	305	2.12
CENTRO DE EDUCAÇÃO	499	3.47
CENTRO DE FILOSOFIA E CIÊNCIAS HUMANAS	1240	8.61
CENTRO SOCIOECONÔMICO	1931	13.41
CENTRO TECNOLÓGICO	4272	29.67
NA	0	0.00
Sum	14399	100.00

Table 8 – Table with the distribution of users per center

We analyzed the attributes that have the smaller contribution to the SUDA score. The center attribute has a greater number of distinct values, totaling 11, and have one of the best distribution of the quantities per attribute in the data set, being behind just of age range and gender as we can see in Tables 8, 9, and 10. It makes us conclude that the uniqueness of attributes is directly linked with the bad distribution of records per value, since the attributes that have the smaller contribution to the SUDA score also have the better distribution. In Chapter 5, we will do an analysis of our method by considering these attacks and using the attribute center, which presented the better results in or analyses, to group the people. We will also present the application scenario and experimental results.

age	Frequency	Percentage
[20 - 21) years	2076	14.42
[21 - 22) years	2002	13.90
[22 - 23) years	1500	10.42
[23 - 24) years	1213	8.42
[24 - 25) years	1493	10.37
<= 19 years	3677	25.54
>= 26 years	2438	16.93
NA	0	0.00
Sum	14399	100.00

Table 9 – Table with the distribution of users per age

gender	Frequency	Percentage
Female	7018	48.74
Male	7381	51.26
NA	0	0.00
Sum	14399	100.00

Table 10 – Table with the distribution of users per gender

5 EXPERIMENTAL EVALUATION AND DISCUSSION

In this chapter, we will first present our application scenario explaining how we collect and clean the data in order to create the semantic trajectory data set. Then, we will show the experiments that we have done by applying our anonymization method, Mix β-k-anonymity, on the Wi-Fi data of UFSC. We also discuss the effectiveness of our approach in relation to the prevention of threats.

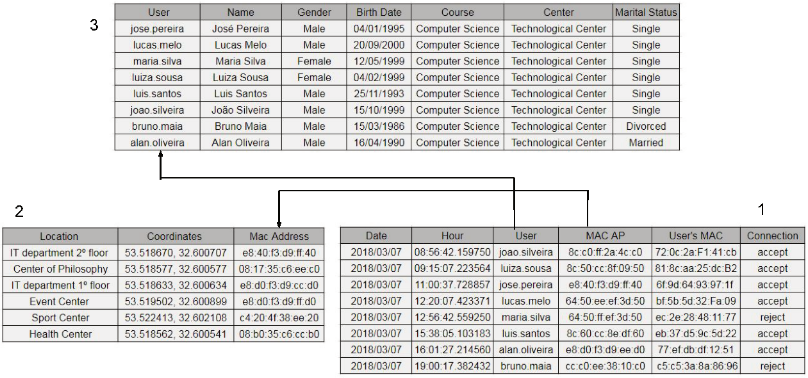
5.1 APPLICATION SCENARIO: UFSC

Every time a mobile device performs an authenticated connection to the Wi-Fi, a lot of data, including location, is generated. If all these connections are made at the same network, it is feasible to track the steps of an individual by considering that each connection with the Wi-Fi is one point of their trajectory, and all these connections are ordered by time. Universities are a good example of network with a lot of access points spread all over their campi. In our work, we use the UFSC Wi-Fi data to test our anonymization method, β-k-anonymity. In order to use this data, we studied how it is collected and stored. After that, we created an automated system to extract, transform and load this data.

Data is collected through the user's device associations to the wireless access points of the university's wireless network. Devices must be within range of a given radio to get the signal from these access points and then make the association. In order to access the Wi-Fi at UFSC, we must perform an authentication to the Eduroam service. The Eduroam service was developed for the international education and research community to provide wireless Internet access simply, quickly and securely (RNP, 2018). A lot of universities throughout the world use the Eduroam service and, it makes our research replicable to other universities.

Each access point has its own geographic coordinates representing the place that they were installed. As already said, to make a connection we must authenticate ourselves, and it requires a unique identifier and password. This identifier is used to access all university services. In the eduroam case at UFSC, every time that a connection is made, a log file is updated with this information. This file contains the following attributes: Date, Time, User ID, MAC address of the access point, MAC address of the user's device and confirmation of successful connection. In order to get the location of the access point, the university has a file that relates a MAC address of an access point to the coordinates of the location it is installed and its name (e.g. North of University Restaurant, Library First Floor). The information about the person related to the identifier is accessed via requests to the IT services of the university. Access to query these services is restricted to authorized persons, with the required authentication.

The scenario explained is represented on Figure 8 by showing the relation of 3 data sources. The first data source shows how the Wi-Fi data is stored in the log file of the university IT system. The second data source shows the information about each access point of the university including the name of the place and geographic coordinates. It is possible to join the data from the source 1 with the data from the source 2 by the number of the MAC address of the access point. The third data source shows the information about the people on the university IT systems. Data from the source 1 and 3 are linked by the user's name, that is a unique attribute. We have authorization of the University administration to access the data and their sources in order to anonymize them. We were required to sign non-disclosure agreements and be cleared in order to have access to the raw data.



User	Name	Gender	Birth Date	Course	Center	Marital Status
jose.pereira	José Pereira	Male	04/01/1995	Computer Science	Technological Center	Single
lucas.melo	Lucas Melo	Male	20/09/2000	Computer Science	Technological Center	Single
maria.silva	Maria Silva	Female	12/05/1999	Computer Science	Technological Center	Single
luiza.sousa	Luiza Sousa	Female	04/02/1999	Computer Science	Technological Center	Single
luis.santos	Luis Santos	Male	25/11/1993	Computer Science	Technological Center	Single
joao.silveira	João Silveira	Male	15/10/1999	Computer Science	Technological Center	Single
bruno.maia	Bruno Maia	Male	15/03/1986	Computer Science	Technological Center	Divorced
alan.oliveira	Alan Oliveira	Male	16/04/1990	Computer Science	Technological Center	Married

Location	Coordinates	Mac Address
IT department 2º floor	53.518670, 32.600707	e8:40:f3:d9:ff:40
Center of Philosophy	53.518577, 32.600577	08:17:35:c6:ee:c0
IT department 1º floor	53.518633, 32.600634	e8:d0:f3:d9:cc:d0
Event Center	53.519502, 32.600899	e8:d0:f3:d9:ff:d0
Sport Center	53.522413, 32.602108	c4:20:4f:38:ee:20
Health Center	53.518562, 32.600541	08:b0:35:c6:cc:b0

Date	Hour	User	MAC AP	User's MAC	Connection
2018/03/07	08:56:42.159750	joao.silveira	8c:c0:ff:2a:4c:c0	72:0c:2a:F1:41:cb	accept
2018/03/07	09:15:07.223564	luiza.sousa	8c:50:cc:8f:09:50	81:8c:aa:25:dc:B2	accept
2018/03/07	11:00:37.728857	jose.pereira	e8:40:f3:d9:ff:40	6f:9d:64:93:97:1f	accept
2018/03/07	12:20:07.423371	lucas.melo	64:50:ee:ef:3d:50	bf:5b:5d:32:Fa:09	accept
2018/03/07	12:56:42.559250	maria.silva	64:50:ff:ef:3d:50	ec:2e:28:48:11:77	reject
2018/03/07	15:38:05.103183	luis.santos	8c:60:cc:8e:df:60	eb:37:d5:9c:5d:22	accept
2018/03/07	16:01:27.214560	alan.oliveira	e8:d0:f3:d9:ee:d0	77:ef:db:df:12:51	accept
2018/03/07	19:00:17.382432	bruno.maia	cc:c0:ee:38:10:c0	c5:c5:3a:8a:86:96	reject

Figure 8 Representation of the data sources and their relations

We studied the data flow to make the ETL process feasible. The extraction, transformation and loading of the data were performed using the ETL (Extract, Transform, and Load) tool Kettle. The data was stored in the non-relational database MongoDB version 3.4. Figure 9 shows 3 important parts of the flow on Kettle. It starts in 1, where the data of the log file is loaded and the time ranges are created. The second part gets the user's id and send a request with the identifiers to get the personal information that is on the university server. In this part, it is also attached the information of the access points. The last part groups the data by location, the quasi-identifier (in our case, the center name), and the time range of the connection to create a new attribute that represents the quantity of people on the same group that were in a same place and time range. This attribute will be used to ensure that at this place there were at least k-1 people. Figure 10 presents the output of the ETL process in the database.

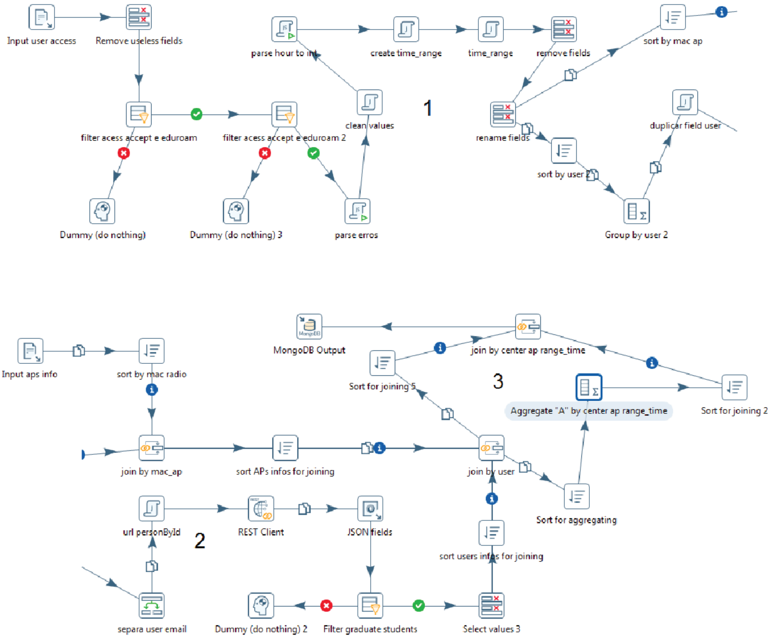
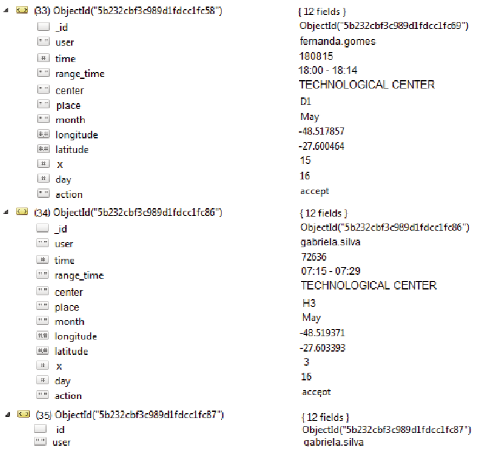


Figure 9 The ETL process made on Kettle

In the next section, we show the experiments that we have done by applying our anonymization method, Mix β-k-anonymity, on the Wi-Fi data of UFSC. We also discuss the effectiveness of our approach in relation to the prevention of threats.



```
(33) ObjectId("5b232cbf3c989d1fdcc1fc58")    { 12 fields }
    _id                                      ObjectId("5b232cbf3c989d1fdcc1fc69")
    user                                     fernanda.gomes
    time                                     180815
    range_time                               18:00 - 18:14
    center                                   TECHNOLOGICAL CENTER
    place                                    D1
    month                                    May
    longitude                                -48.517857
    latitude                                 -27.600464
    x                                        15
    day                                      16
    action                                   accept
(34) ObjectId("5b232cbf3c989d1fdcc1fc86")    { 12 fields }
    _id                                      ObjectId("5b232cbf3c989d1fdcc1fc86")
    user                                     gabriela.silva
    time                                     72636
    range_time                               07:15 - 07:29
    center                                   TECHNOLOGICAL CENTER
    place                                    H3
    month                                    May
    longitude                                -48.519371
    latitude                                 -27.603393
    x                                        3
    day                                      16
    action                                   accept
(35) ObjectId("5b232cbf3c989d1fdcc1fc87")    { 12 fields }
    id                                       ObjectId("5b232cbf3c989d1fdcc1fc87")
    user                                     gabriela.silva
```

Figure 10 Representation of the output data on MongoDB

5.2 EXPERIMENTAL WORK AND DISCUSSION

We tested the algorithm in a database containing almost 1.4 million connection records. These connections were made by university undergraduate students of Federal University of Santa Catarina, campus Trindade, on May 16th 2018. We chose this date because this was a typical day, making the movement inside the campus similar to most days of the year.

The log data of the connections was made available with authorization from the University administration. We were required to sign non-disclosure agreements in order to have access to raw data. Their purpose was encouraging anonymization research to make this data available for future research in other areas of knowledge, guaranteeing the users' privacy. We were also authorized to search quasi-identifiers on the university's systems, finding them by the personal identifier present in the logs.

Figure 11 shows how the access points were grouped. They were grouped into 32 areas in order to promote more encounters of people of the same group. This sectorization was made by the Laboratory of Urban Ecology of Architecture (LEUr) at UFSC and this measure was taken because some places had few meetings of people of the same group. After grouping, the area of interest was bigger, promoting more possibilities of people of the same group to be there connected to the

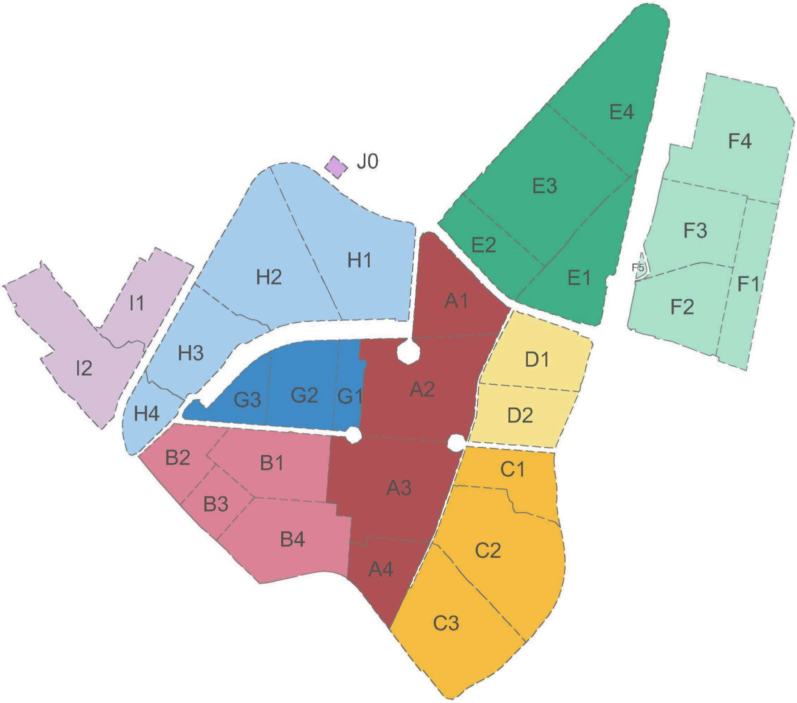


Figure 11 – A map of UFSC, sectorized by location

Wi-Fi.

The users were grouped by center and there are 11 centers. The time range was split in 15 minutes for these experiments, but this value can change. We, by analyzing the data, understand that 15 minutes was a good choice since we do not lose a lot with the generalization and it is enough time to make people of the same group being together at the same place in our campus. It is important to understand that the bigger the time range the bigger the number of people of the same group at the same place. So this decision is related to the number of records of the data set. If we have a data set with few records it is better creating bigger time ranges in order to group more people together.

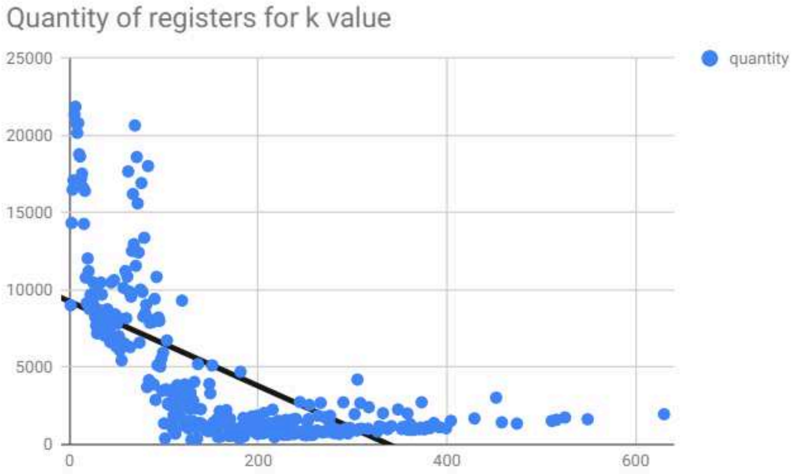


Figure 12 – Graph of quantity of records by k value

The problem of finding the ideal k for k-anonymity is NP-hard (MEYERSON; WILLIAMS, 2004). For this reason, both β and k parameters are chosen accordingly to the level of privacy the user wishes to apply to the data. This level of privacy is given in percentage and represents the chances that an attacker has of discovering some information within the disclosed database, as we will see in Section 5.2.1. Analyzing the data stored in the database before its anonymization, we noticed that, by using values of k greater than 50, some centers no longer returned data since they did not have enough students concentrated in any specific area of the campus in the same range time. Figure 12 presents the relation between the quantity of records and the k value. Most of the data has between 2 and 70 indistinguishable registers, and this represents more than 50% of the database. So, if we put a k equals to 70, we would probably loose half of the records. The graph shows a decreasing relation between the quantity of records and k. It means that in almost all cases the bigger the value of k, the more registers are suppressed. Therefore, in applying the metric, we must balance the loss of information with the use of a very large k, and the low privacy with the use of a very small k. The main value in the privacy level measure is β. This value influences the probability of an attacker being able to extract information from the anonymized database, we will discuss more about it in Subsection 5.2.1.

The three main scenarios are: 1) the attacker saw his victim *entering* a certain building at a given time and knows the group that the victim belongs to; 2) the attacker saw his victim *leaving* a certain building at a given time and knows the group that the victim belongs to; 3) the attacker knows that, at a certain time, the victim was in a certain building.

There are some considerations that can be applied to all three cases. First, even if an attacker knows that the person is in a certain place there are $k - 1$ other people indistinguishable to him who are in this same place, and at least β possibilities of next places that will be visited by people of this group. Even if the attacker knows the time range that the person entered or left a certain building, there are some uncertainties inherent in the model of trajectories created by Wi-Fi devices. The attacker can only be sure that the person possessed a mobile device next to it if at that moment he saw them carrying the object. However, the mobile device may be out of battery power, turned off, or users may have switched to 4G. The device could even be in airplane mode or even generating authentication failure.

In the first scenario, for the attacker to find out where that person came from before entering the building, there are at least k people from the same group who made connections in that time zone. Even if you have only one indicator pointing to this place, there are several possibilities for inference: some of the people in the group who entered the building came from the starting point of the indicator; others may have made their first connection in this building; the people may also have just passed near a building and made the connection; and there are also the possibilities that we have cited on Section 4.2.

Those inferences make it difficult for the attacker to be sure where the person came from. The only way to be completely sure is to follow the person everywhere they go. By analyzing our data set no single point was received, then this attack could not be easily performed.

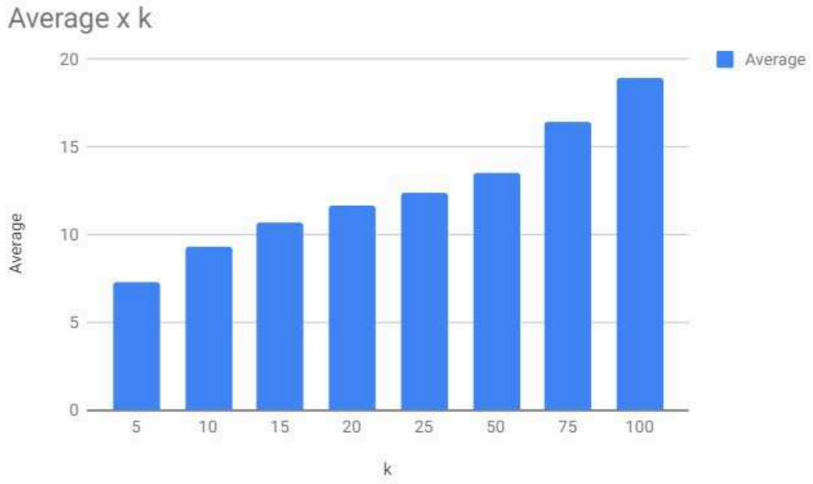


Figure 13 – Graphs of k value x average of the size of the next points list.

In the second scenario, if it is not the last connection, there are $k-1$ people in this time zone and at least β possible next places visited, letting the attacker with $1/(\beta+1)$ % of chance to hit the next place. This "plus one" represents the chance of that being the last connection of the person, it means that there is no next place visited. Figure 13 presented the relation between the k value and the quantity of next points. As we can see, the larger the k the greater the average number of next places. When we group more people together more possibilities of different next places could exist. The lesser the quantity of people the lesser the chances of create different next places. The highest value of next places found was 85. Another important relation is between the time range and the value of k. The wider the time ranges the bigger the chance of people of the same group being together. If we use narrower time ranges the chance of achieving at least $k-1$ people of the same group in a same place is lower.

The complexity of finding a person's trajectory from a known point can become exponential. Knowing one point of the trajectory, there are β possibilities of next places to be visited. Each possibility may have no next place or at least β options, and each of them also follows this pattern. If a sequence of points generates at least x next points, e.g., the first point generates x and each one of the next x points generates more x and so on until a final point appears, or that does not have a minimum β next points. In Section 5.2.1, we will demonstrate how many possibilities of finding more points of the trajectory of his victim the attacker has if he does not know any place that the person went previously. It is valid to remember that some points of the victims' trajectory may have been suppressed. So, in the anonymized data set probably there is not all the points of the victims' trajectory, because some of them may have been suppressed. If the attacker knows a point that was suppressed in fact he does not know any point.

The fact that the place does not show up next points does not mean that they do not exist. This may also mean that the next points were suppressed. In the third scenario, after applying the anonymization method the attacker has remote chances of being able to find out the points that do not belong to the place where the target was. This is due to this target being with many people similar to him, generating several possibilities of next places to be created. If the target is together with people not so similar to him, those locations will be suppressed.

5.2.1 EVALUATION METHOD

In this subsection, we will show how many possible trajectories, composed by points, are generated when we anonymize the data set with Mix β-k-anonymity. This probability indicates the chance of the attacker figuring out his victim trajectory if he does not know any point of their trajectory, but he does know the group that the victim belongs to. We will call the trajectory as path in this section. Figure 14 shows an example of Mix 2-2-anonymity. The victim could have made their first connection in any of these points. In order to evaluate the quantity of possible paths that the victim could have taken, we must get each of these points and go through all the possible paths.

Figure 15 presents all the achievable paths by starting on the point (E, 07:30-08:00). In order to find them all, we start by considering that the victim just made one connection all day long and this connection generates the point (E, 07:30-08:00). But if the victim goes a step further, the victim could have connected at (B, 08:30-09:00) or (D, 07:30-08:00). If they connected in one of these two places, the next possible places are (E, 07:30-08:00), (B, 08:30-09:00), (A, 11:00-11:30), (C, 12:00-12:30) if the person went to (D, 07:30-08:00) and no next places

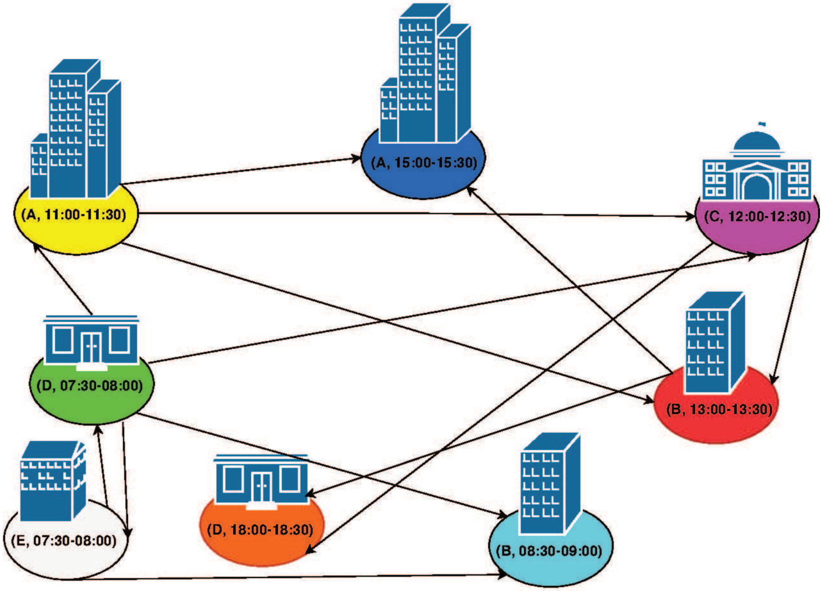
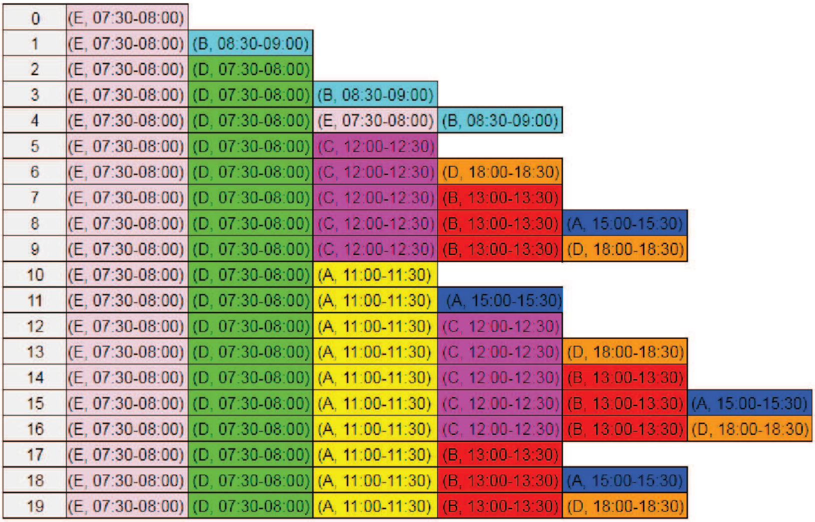


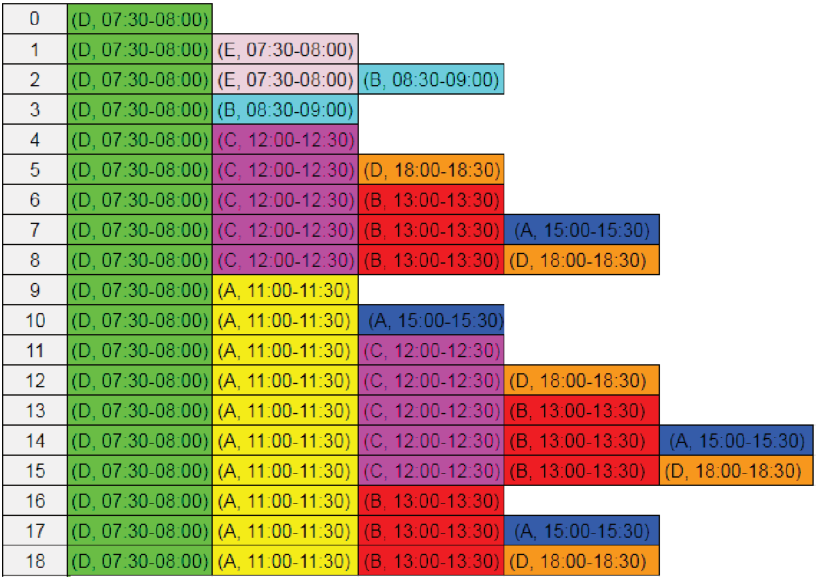
Figure 14 Example of a data set after being anonymized with Mix 2-2-anonymity



0	(E, 07:30-08:00)					
1	(E, 07:30-08:00)	(B, 08:30-09:00)				
2	(E, 07:30-08:00)	(D, 07:30-08:00)				
3	(E, 07:30-08:00)	(D, 07:30-08:00)	(B, 08:30-09:00)			
4	(E, 07:30-08:00)	(D, 07:30-08:00)	(E, 07:30-08:00)	(B, 08:30-09:00)		
5	(E, 07:30-08:00)	(D, 07:30-08:00)	(C, 12:00-12:30)			
6	(E, 07:30-08:00)	(D, 07:30-08:00)	(C, 12:00-12:30)	(D, 18:00-18:30)		
7	(E, 07:30-08:00)	(D, 07:30-08:00)	(C, 12:00-12:30)	(B, 13:00-13:30)		
8	(E, 07:30-08:00)	(D, 07:30-08:00)	(C, 12:00-12:30)	(B, 13:00-13:30)	(A, 15:00-15:30)	
9	(E, 07:30-08:00)	(D, 07:30-08:00)	(C, 12:00-12:30)	(B, 13:00-13:30)	(D, 18:00-18:30)	
10	(E, 07:30-08:00)	(D, 07:30-08:00)	(A, 11:00-11:30)			
11	(E, 07:30-08:00)	(D, 07:30-08:00)	(A, 11:00-11:30)	(A, 15:00-15:30)		
12	(E, 07:30-08:00)	(D, 07:30-08:00)	(A, 11:00-11:30)	(C, 12:00-12:30)		
13	(E, 07:30-08:00)	(D, 07:30-08:00)	(A, 11:00-11:30)	(C, 12:00-12:30)	(D, 18:00-18:30)	
14	(E, 07:30-08:00)	(D, 07:30-08:00)	(A, 11:00-11:30)	(C, 12:00-12:30)	(B, 13:00-13:30)	
15	(E, 07:30-08:00)	(D, 07:30-08:00)	(A, 11:00-11:30)	(C, 12:00-12:30)	(B, 13:00-13:30)	(A, 15:00-15:30)
16	(E, 07:30-08:00)	(D, 07:30-08:00)	(A, 11:00-11:30)	(C, 12:00-12:30)	(B, 13:00-13:30)	(D, 18:00-18:30)
17	(E, 07:30-08:00)	(D, 07:30-08:00)	(A, 11:00-11:30)	(B, 13:00-13:30)		
18	(E, 07:30-08:00)	(D, 07:30-08:00)	(A, 11:00-11:30)	(B, 13:00-13:30)	(A, 15:00-15:30)	
19	(E, 07:30-08:00)	(D, 07:30-08:00)	(A, 11:00-11:30)	(B, 13:00-13:30)	(D, 18:00-18:30)	

Figure 15 List with possible paths starting on point (E, 07:30-08:00)

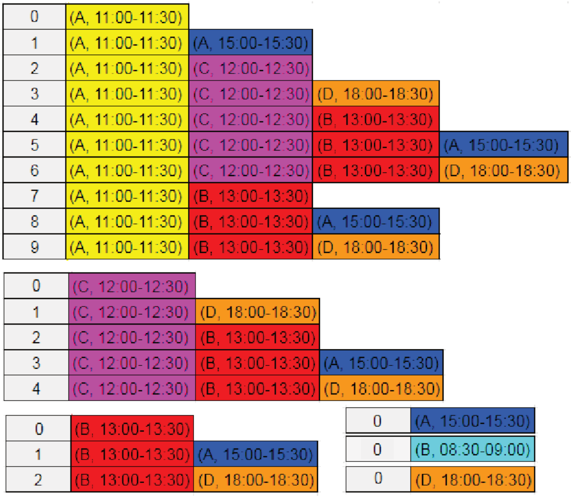
exhibited if they went to (B, 08:30-09:00). The point (B, 08:30-09:00) does not show any next places because it does not achieve β distinct next places or because no one who was there made new connections after that. By choosing one of the next places of (D, 07:30-08:00), each one of them could have more next places visited. If the victim chooses to go to (A, 11:00-11:30), and after that if the victim made another connection, it was made in one of these points: (C, 12:00-12:30), (B, 13:00-13:30) or (A, 15:00, 15:30). The idea is to follow all the possible paths achievable by a point. We map all the possible paths that the victim could have made by starting in (D, 07:30-08:00), as we can see in Figure 16. The same thing was made to all the other points as Figure 17 shows.



0	(D, 07:30-08:00)				
1	(D, 07:30-08:00)	(E, 07:30-08:00)			
2	(D, 07:30-08:00)	(E, 07:30-08:00)	(B, 08:30-09:00)		
3	(D, 07:30-08:00)	(B, 08:30-09:00)			
4	(D, 07:30-08:00)	(C, 12:00-12:30)			
5	(D, 07:30-08:00)	(C, 12:00-12:30)	(D, 18:00-18:30)		
6	(D, 07:30-08:00)	(C, 12:00-12:30)	(B, 13:00-13:30)		
7	(D, 07:30-08:00)	(C, 12:00-12:30)	(B, 13:00-13:30)	(A, 15:00-15:30)	
8	(D, 07:30-08:00)	(C, 12:00-12:30)	(B, 13:00-13:30)	(D, 18:00-18:30)	
9	(D, 07:30-08:00)	(A, 11:00-11:30)			
10	(D, 07:30-08:00)	(A, 11:00-11:30)	(A, 15:00-15:30)		
11	(D, 07:30-08:00)	(A, 11:00-11:30)	(C, 12:00-12:30)		
12	(D, 07:30-08:00)	(A, 11:00-11:30)	(C, 12:00-12:30)	(D, 18:00-18:30)	
13	(D, 07:30-08:00)	(A, 11:00-11:30)	(C, 12:00-12:30)	(B, 13:00-13:30)	
14	(D, 07:30-08:00)	(A, 11:00-11:30)	(C, 12:00-12:30)	(B, 13:00-13:30)	(A, 15:00-15:30)
15	(D, 07:30-08:00)	(A, 11:00-11:30)	(C, 12:00-12:30)	(B, 13:00-13:30)	(D, 18:00-18:30)
16	(D, 07:30-08:00)	(A, 11:00-11:30)	(B, 13:00-13:30)		
17	(D, 07:30-08:00)	(A, 11:00-11:30)	(B, 13:00-13:30)	(A, 15:00-15:30)	
18	(D, 07:30-08:00)	(A, 11:00-11:30)	(B, 13:00-13:30)	(D, 18:00-18:30)	

Figure 16 Lists with possible paths starting on point (D, 07:30-08:00)

There is a logic to evaluate the quantity of possible paths. As points (A, 15:00-15:30), (D, 18:00-18:30) and (B, 08:30-09:00) do not have next points, then the quantity of possible paths of these points stands for 1 possible trajectory. If the person made their first connection in (B, 13:00-13:30), this person could have made that connection or made another one in (A, 15:00-15:30) or (D, 18:00-18:30). So, there are 3 possible places, as Figure 17 shows. If the person made their first connection in (C, 12:00-12:30), this connection could be the only one



0	(A, 11:00-11:30)			
1	(A, 11:00-11:30)	(A, 15:00-15:30)		
2	(A, 11:00-11:30)	(C, 12:00-12:30)		
3	(A, 11:00-11:30)	(C, 12:00-12:30)	(D, 18:00-18:30)	
4	(A, 11:00-11:30)	(C, 12:00-12:30)	(B, 13:00-13:30)	
5	(A, 11:00-11:30)	(C, 12:00-12:30)	(B, 13:00-13:30)	(A, 15:00-15:30)
6	(A, 11:00-11:30)	(C, 12:00-12:30)	(B, 13:00-13:30)	(D, 18:00-18:30)
7	(A, 11:00-11:30)	(B, 13:00-13:30)		
8	(A, 11:00-11:30)	(B, 13:00-13:30)	(A, 15:00-15:30)	
9	(A, 11:00-11:30)	(B, 13:00-13:30)	(D, 18:00-18:30)	

0	(C, 12:00-12:30)		
1	(C, 12:00-12:30)	(D, 18:00-18:30)	
2	(C, 12:00-12:30)	(B, 13:00-13:30)	
3	(C, 12:00-12:30)	(B, 13:00-13:30)	(A, 15:00-15:30)
4	(C, 12:00-12:30)	(B, 13:00-13:30)	(D, 18:00-18:30)

0	(B, 13:00-13:30)	
1	(B, 13:00-13:30)	(A, 15:00-15:30)
2	(B, 13:00-13:30)	(D, 18:00-18:30)

0	(A, 15:00-15:30)

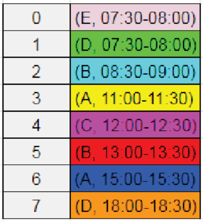
0	(B, 08:30-09:00)

0	(D, 18:00-18:30)

Figure 17 List with possible paths starting on points (A, 11:00-11:30), (C, 12:00-12:30), (B, 13:00-13:30), (A, 15:00-15:30), (D, 18:00-18:30) and (B, 08:30-09:00)

or the person could have made another one in (D, 18:00-18:30), and no more connections after that. The person could also have gone to (B, 13:00-13:30) and stopped in there or went to (A, 15:00-15:30) or (D, 18:00-18:30) and stopped in one of these two places. There is a total of 5 possible paths, as we can see in Figure 17. If we follow this pattern, we will see that the quantity of possible paths of a point is 1 plus the quantity of paths of its next places. The number of achievable places starting in (C, 12:00-12:30) are 1 plus the achievable paths of (D, 18:00-18:30), in this case 1, plus the 3 achievable paths of (B, 13:00-13:30). For point (A, 11:00-11:30), it is the sum of the achievable paths of (C, 12:00-12:30), (B, 13:00-13:30), and (A, 15:00-15:30). The total of possible paths starting in (A, 11:00-11:30) is 1 + 5 + 3 + 1, totalizing 10, as presented in Figure 17.

There is just one exception that does not follow exactly this logic. This exception happens when there is a possibility of loop. In Figure 14, we can see a loop in the points (D, 07:30-08:00) and (E, 07:30-08:00). We know that in this time range some people made a connection in (D, 07:30-08:00) and after that in (E, 07:30-08:00), and some people did the opposite sequence. In order to avoid this loop, the first point of this relation just count as a path of size 1 plus the their other achievable places. For example, if we evaluate first point (D, 07:30-08:00) we will count the achievable paths of (E, 07:30-08:00) as 2. The first one is (D, 07:30-08:00) and the second is (B, 08:30-09:00). So the total of possible paths starting in (D, 07:30-08:00) is 1 + 2 + 1 (B, 08:30-09:00) + 5 (possible paths of (C, 12:00-12:30)) + 10 (possible paths of (C, 12:00-12:30)) totalizing 19, as shown in Figure 16. Since (D, 07:30-08:00) is already evaluated then we sum 1 + 19 for the total of possible paths starting in (E, 07:30-08:00), as seen in Figure 16. The sum of these paths represents the possibilities of trajectories that could have been made by the victim, in this example it would be 20 + 19 + 10 + 5 + 3 + 1 + 1 + 1 = 60.



0	(E, 07:30-08:00)
1	(D, 07:30-08:00)
2	(B, 08:30-09:00)
3	(A, 11:00-11:30)
4	(C, 12:00-12:30)
5	(B, 13:00-13:30)
6	(A, 15:00-15:30)
7	(D, 18:00-18:30)

Figure 18 List of points ordered by time range.

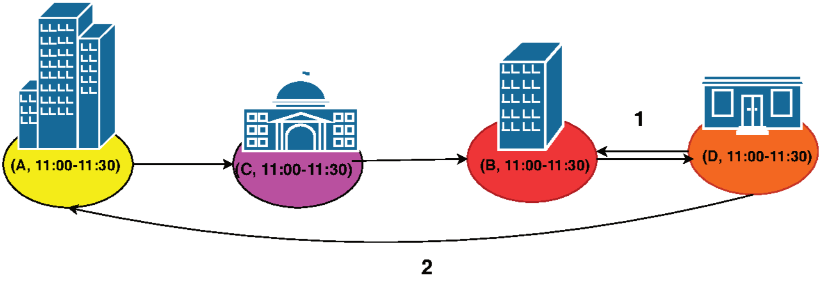


Figure 19 Example of flow with direct and indirect loops.

This evaluation is made by Algorithms 2, 3, and 4. Algorithm 2 receives an array with all points already anonymized by Mix $\beta - k$-anonymity in ascending order, as can be seen in Figure 18. For each point of this array, it evaluates the number of possible paths for each point of the array. Algorithm 3 receives the point to be evaluated. If the point does not have any next points, it is evaluated with 1 possible path, that represents just one connection at this point, as shown in lines 1 and 2. If the quantity of paths was not evaluated yet, the id of this point is added in a list of points already visited. Then, each of the next points of this point is visited in order to get its quantity of possible paths. If the next point has been already evaluated, its quantity of paths is added to the point's "*count*" variable, as shown in line 27. A point can have a next point that has not been evaluated yet, like in Figure 18 where point (D, 07:30-08:00) has point (E, 07:30-08:00) as next point not evaluated yet. If it was not evaluated, having the variable "*paths*" equals 0, as we can see in line 9, we verify the existence of a loop. A direct loop can happen when a point has a next point and this next point has that point in its next places list. Figure 19 shows two types of loop, the first one is the direct one, signalized by the number 1, and the indirect, shown in number 2. In order to avoid the loops, we created a list where we put all the points that were already visited.

Algorithm 2: QuantityOfPaths

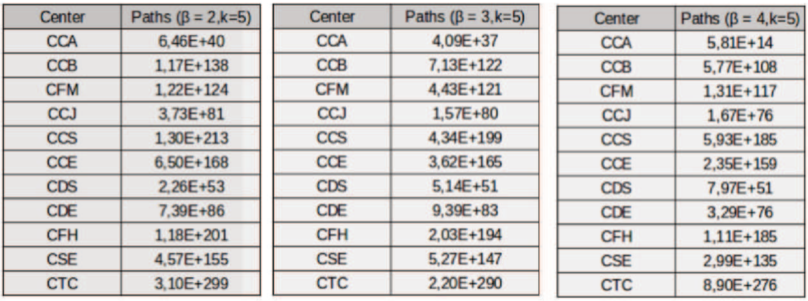
```
Input  : An array $T$ of size $l$ in ascending order.
Output: count that represents the quantity of possible paths.
1  count ← 0;
2  for $y \leftarrow l - 1$ to 0 do
3      point ← $T[y]$;
4      Evaluate (point);
5      for $x \leftarrow 0$ to point.listOfNextPoints.length do
6          nextPoint ← ponto.listOfNextPoints[$x$];
7          for $z \leftarrow 0$ to nextPoint.listOfNextPoints.length do
8              if nextPoint.listOfNextPoints[z].paths == 0
                 nextPoint.listOfNextPoints[z].id != point.id
                 nextPoint.position < point.position then
9                  nextPoint.listOfNextPoints ← [];
10 return count;
```

Algorithm 4 verifies if one of the next points of the point sent by parameter is present in the list of points already visited. If it is, this relation is counted as 1 and this point is not evaluated, as we can see in lines 11-19. Since the next point that generated the loop was not evaluated, after evaluating the next point in question, the quantity of paths is set to 0, as we can see in Algorithm 2, in lines 8-10. This happens because it will be necessary to evaluate its next point, that contains the loop, in order to have its quantity of paths (since the point needs all the quantity of paths of each next point). The final result is the sum of the possible paths for all points.

After anonymizing the data from Wi-Fi connections of students of UFSC (Trindade Campus), we evaluated the number of possible

Algorithm 3: Evaluate

```
Input  : A point with the list of next points.
Output: count that represents the quantity of possible paths.
if point.listOfNextPoints.lenght == 0 then
    point.quantityPaths ← 1;
else
    if point.paths == 0 then
        visited.add(point.id);
        count ← 1;
        for nextPoint ∈ point.listOfNextPoints do
            visited.add(nextPoint.id);
            if nextPoint.paths == 0 then
                result = ContainsLoop(nextPoint, visited);
                if result.length != 0 then
                    removed ← [];
                    for z ← result.length − 1 to 0 do
                        pointRemoved ←
                          nextPoint.listOfNextPoints.remove(result[z]);
                        removed.add(pointRemoved);
                    count += Evaluate (nextPoint);
                    for x ← 0 to removed.length - 1 do
                        if removed [w].id != nextPoint [x].id then
                            nextPoint [x].listOfNextPoints.push(removed
                              [w]);
                    nextPoint.paths += removed.length;
                    count += nextPoint.paths;
                    nextPoint.paths = 0;
                else
                    Evaluate (nextPoint [x]);
                    count += nextPoint [x].paths;
            else
                count += nextPoint [x].paths;
        nextPoint [x].paths ← count;
```



Center	Paths (β = 2,k=5)
CCA	6,46E+40
CCB	1,17E+138
CFM	1,22E+124
CCJ	3,73E+81
CCS	1,30E+213
CCE	6,50E+168
CDS	2,26E+53
CDE	7,39E+86
CFH	1,18E+201
CSE	4,57E+155
CTC	3,10E+299

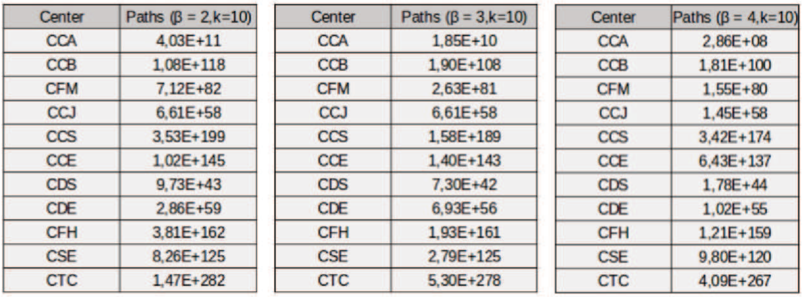
Center	Paths (β = 3,k=5)
CCA	4,09E+37
CCB	7,13E+122
CFM	4,43E+121
CCJ	1,57E+80
CCS	4,34E+199
CCE	3,62E+165
CDS	5,14E+51
CDE	9,39E+83
CFH	2,03E+194
CSE	5,27E+147
CTC	2,20E+290

Center	Paths (β = 4,k=5)
CCA	5,81E+14
CCB	5,77E+108
CFM	1,31E+117
CCJ	1,67E+76
CCS	5,93E+185
CCE	2,35E+159
CDS	7,97E+51
CDE	3,29E+76
CFH	1,11E+185
CSE	2,99E+135
CTC	8,90E+276

Figure 20 Quantity of possible paths per center with Mix 2,3,4-5-anonymity

Center	List of Points Size (k = 5)	List of Points Size (k = 10)
CCA	127	45
CCB	888	607
CFM	415	228
CCJ	251	161
CCS	1138	1029
CCE	874	516
CDS	156	101
CE	326	151
CFH	845	500
CSE	795	374
CTC	1303	1100

Table 11 Size of the list of points per course



Center	Paths (β = 2,k=10)
CCA	4,03E+11
CCB	1,08E+118
CFM	7,12E+82
CCJ	6,61E+58
CCS	3,53E+199
CCE	1,02E+145
CDS	9,73E+43
CDE	2,86E+59
CFH	3,81E+162
CSE	8,26E+125
CTC	1,47E+282

Center	Paths (β = 3,k=10)
CCA	1,85E+10
CCB	1,90E+108
CFM	2,63E+81
CCJ	6,61E+58
CCS	1,58E+189
CCE	1,40E+143
CDS	7,30E+42
CDE	6,93E+56
CFH	1,93E+161
CSE	2,79E+125
CTC	5,30E+278

Center	Paths (β = 4,k=10)
CCA	2,86E+08
CCB	1,81E+100
CFM	1,55E+80
CCJ	1,45E+58
CCS	3,42E+174
CCE	6,43E+137
CDS	1,78E+44
CDE	1,02E+55
CFH	1,21E+159
CSE	9,80E+120
CTC	4,09E+267

Figure 21 Quantity of possible paths per center with Mix 2,3,4-10-anonymity

paths for each center. Figure 20 shows the results applied in a data set with Mix 2-5-anonymity, 3-5-anonymity and 4-5-anonymity. The CCA center has few paths because we anonymized only data from the Trindade Campus, and subjects do not have a lot of classes there. The center that has more possible paths is CTC, because it is the center that has more students, so there will be more connections made by the people of this course, as we can see in Table 11. By analyzing the data from Figure 20, we can notice that the bigger the β value the fewer the quantity of possible paths. Figure 21 shows the results applied in a data set with Mix 2-10-anonymity, 3-10-anonymity, and 4-10-anonymity. By comparing the data from Figures 20 and 21, we see that the bigger the k value the lesser the quantity of paths. It happens because highest values of k suppress more points. If more points are suppressed higher the chances of the point that the attacker saw the victim be also suppressed. This values returned by the evaluation metric represents the possible paths that a person could have done, and depends on both values of k and *beta* chosen.

Algorithm 4: ContainsLoop

```
Input  : A point with the list of next points and a list with the id of
         points already visited.
Output: A list with the position of the nextPoints of the point that were
         already visited.
result ← [];
count ← 0;
for $y \leftarrow 0$ to point.$listOfNextPoints.length$ - $1$ do
    for $z \leftarrow 0$ to visited.$length$ - $1$ do
        if visited $[z] ==$ point $[y].id$ then
            result [count ] ← y;
            count $++$;
            z ← visited.$length$;
return count;
```

The evaluation method showed so far is related to the scenario where the attacker does not know any point of his victim trajectory. However, as we explained, the attacker can use some background knowledge to acquire information about the victim. Given this circumstance, we also create an evaluation metric that returns the possible trajectories that the attacker has when he knows at least one point of the victim's trajectory. First we must guarantee that the points that the victim was seen at are on the anonymized data set. If the attacker only knows one point and this point was suppressed then the attacker return to the case where he does not know any point. If the attacker knows more then one point and some of them were suppressed, then these points will be ignored. So, the bigger the k value the bigger the chances of some points of the victim's trajectory have been suppressed. And it is also important to know that for both cases, where the attacker does not know any point or knows at least one, the victim's trajectory may have been sliced. A trajectory is sliced when at least one of the points that compose this trajectory does not have at least β next points, because in this case the displacements are suppressed and the connection between the points is missed.

Algorithm 5: EvaluateWithKnownPoints

```
   Input  : A list with the points that the attacker knows.
   Output: The quantity of possible trajectories that contains these points.
1  count ← 0;
2  allPossibleTrajetories ← ReturnPossibleTrajetories ();
3  foreach trajectory ∈ allPossibleTrajetories do
4      if trajectory.containsAll (listOfKnownPoints) then
5          count++;
6  return count;
```

Algorithm 5 receives as input a list of points known by the attacker. All these points must be in the data set because if one of them is not it will be evaluated as 0, since no trajectory contains this point. A similar process than that done in Algorithm 2 is also made in Algorithm 5, in order to return the possible trajectories. The difference between them is that the method ”ReturnPossibleTrajetories”, showed in Line 2 of Algorithm 5, instead of counting the number of returning the possible trajectories, it saves each one on a list. After returning the possible trajectories, the method verifies all trajectories that contain the points known by the attacker, as we can see in Lines 3-5. For each trajectory containing each of the points, a “*count*” variable is increased in 1. Finally, the “*count*” variable is returned. This variable represents the quantity of trajectories that contains those points. For example, if we know that the victim was at point (A, 11:00-11:30), showed in Figure 14, there are 30 possible trajectories that contains this point. So the attacker has 1/30 chances of figuring out the right trajectory.

6 CONCLUSION

In the last decade, several works were proposed in order to anonymize trajectories. Few works care about semantic trajectories. The ones who do, just take into account the quasi-identifiers or semantic information of the trajectory. In this dissertation, we showed that it is possible to correlate personal information to the trajectory. The data can be collected, for example, by GPS, Wi-Fi connections, and cell phone signal. People who have access to this data probably also have access to our personal data. In universities, for instance, they have all Wi-Fi records and personal data of their contributors. This data can be useful for several areas of knowledge, ranging from security, urban planning, public transport management, to epidemic prevention (MONREALE et al., 2011).

In order to release this data, we proposed an anonymization technique for semantic trajectory data created by connections to Wi-Fi networks or another sources with sparse points, called *Mix β-k-anonymity.* Our approach uses a personal quasi-identifier to group people and their trajectories. The application of the method on data of undergraduate students collected on a university campus has shown that the academic community may have access to quality data for operational mobility research on campus with the maintenance of the users' privacy. We also adapt the threat models for the university scenario. We made a study of the choice of the quasi-identifier variable and the impact of this choice. With that study, we showed that the choice of the quasi-identifier attribute is crucial for the preservation of privacy. Finally, we evaluated the method with Wi-Fi data from students of UFSC, campus Trindade, and discussed about the effectiveness of the approach in relation to prevention of threats.

Regarding to future works, we intend to add more personal quasi-identifiers without raising the risk of re-identification. We also aim to study the spatio-temporal generalization in order to find the optimal one. Another important future work is related to the visualization of this multidimensional trajectory. And lastly, since our method was made to operate with sparse points, we intend to create a new version of the method in order to accept trajectories with more points, in other words, trajectories with moves.

As result of this work, we published two articles. The first one is entitled "Anonimização de Dados de Trajetórias em Grupos para Disponibilização à Pesquisa Universitária" (GOMES et al., 2018a) and was accepted on the SBSeg (Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais) symposium. This symposium was classified by Capes with a qualis B3. The other paper is entitled "Privacy Preserving on Trajectories Created by Wi-Fi Connections in a University Campus" (GOMES et al., 2018b), being accepted on the ISI (IEEE Intelligence and Security Informatics) conference. The qualis of this conference is B1.

REFERENCES

ABUL, O.; BONCHI, F.; NANNI, M. Never walk alone: Uncertainty for anonymity in moving objects databases. In: IEEE. *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on.* [S.l.], 2008. p. 376–385.

ABUL, O.; BONCHI, F.; NANNI, M. Anonymization of moving objects databases by clustering and perturbation. *Information Systems*, Elsevier, v. 35, n. 8, p. 884–910, 2010.

AGGARWAL, G. et al. Anonymizing tables. In: SPRINGER. *International Conference on Database Theory.* [S.l.], 2005. p. 246–258.

ALVARES, L. O. et al. A model for enriching trajectories with semantic geographical information. In: ACM. *Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems.* [S.l.], 2007. p. 22.

BLEUMER, G. Unlinkability. In: *Encyclopedia of Cryptography and Security.* [S.l.]: Springer, 2011. p. 1350–1350.

BOGORNY, V. et al. Constant–a conceptual data model for semantic trajectories of moving objects. *Transactions in GIS*, Wiley Online Library, v. 18, n. 1, p. 66–88, 2014.

BOGORNY, V.; WACHOWICZ, M. *A Framework for Context-Aware Trajectory Data Mining. Data Mining for Business Applications.* [S.l.]: Springer, 2008.

CICEK, A. E.; NERGIZ, M. E.; SAYGIN, Y. Ensuring location diversity in privacy-preserving spatio-temporal data publishing. *The VLDB Journal—The International Journal on Very Large Data Bases*, Springer-Verlag New York, Inc., v. 23, n. 4, p. 609–625, 2014.

CIVIL, P. da República do B. C. *LEI N^o 13.709.* 2018. <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm>. Acessado em 12/10/2018.

CLARKE, R. Introduction to dataveillance and information privacy, and definitions of terms. *Roger Clarke's Dataveillance and Information Privacy Pages*, 1999.

CRITERIA, C. *Common Criteria for Information Technology Security Evaluation - Part 2: Security functional components*. 2017. <https://www.commoncriteriaportal.org/files/ccfiles/CCPART2V3.1R5.pdf>. Acessado em 12/12/2018.

DALENIUS, T. Finding a needle in a haystack or identifying anonymous census records. *Journal of official statistics*, Statistics Sweden (SCB), v. 2, n. 3, p. 329, 1986.

DOMINGO-FERRER, J.; TORRA, V. Disclosure control methods and information loss for microdata. *Confidentiality, disclosure, and data access: theory and practical applications for statistical agencies*, Elsevier, p. 91–110, 2001.

ELLIOT, M.; SKINNER, C. J.; DALE, A. Special uniques, random uniques and sticky populations: some counterintuitive effects of geographical detail on disclosure risk. *Research in Official Statistics*, Statistical Office of the European Communities, v. 1, p. 53–67, 1998.

ELLIOT, M. J.; MANNING, A. M.; FORD, R. W. A computational algorithm for handling the special uniques problem. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, World Scientific, v. 10, n. 05, p. 493–509, 2002.

GDPR, E. *The EU General Data Protection Regulation (GDPR) is the most important change in data privacy regulation in 20 years.* 2018. <https://eugdpr.org/>. Acessado em 12/10/2018.

GOMES, F. O. et al. Anonimização de dados de trajet órias em grupos para disponibilização á pesquisa universitária. In: *XVIII SIMPÓSIO BRASILEIRO EM SEGURANÇA DA INFORMAÇÃO E DE SISTEMAS COMPUTACIONAIS*. [S.l.: s.n.], 2018. p. 71–84.

GOMES, F. O. et al. Privacy preserving on trajectories created by wi-fi connections in a university campus. In: IEEE. *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*. [S.l.], 2018. p. 181–186.

GRAMAGLIA, M.; FIORE, M. Hiding mobile traffic fingerprints with glove. In: ACM. *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies*. [S.l.], 2015. p. 26.

GRAMAGLIA, M. et al. Preserving mobile subscriber privacy in open datasets of spatiotemporal trajectories. In: IEEE. *INFOCOM*

2017-IEEE Conference on Computer Communications, IEEE. [S.l.], 2017. p. 1–9.

GUARDIAN, T. *Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach*. 2018. <https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election>. Acessado em 12/10/2018.

HU, H. et al. Privacy-aware location data publishing. *ACM Transactions on Database Systems (TODS)*, ACM, v. 35, n. 3, p. 18, 2010.

HUTCHINS, R.; ZEGURA, E. W. Measurements from a campus wireless network. In: IEEE. *Communications, 2002. ICC 2002. IEEE International Conference on*. [S.l.], 2002. v. 5, p. 3161–3167.

LI, N.; LI, T.; VENKATASUBRAMANIAN, S. t-closeness: Privacy beyond k-anonymity and l-diversity. In: IEEE. *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. [S.l.], 2007. p. 106–115.

LIN, C.-Y. et al. Efficiently preserving privacy on large trajectory datasets. In: IEEE. *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)*. [S.l.], 2018. p. 358–364.

LU, Q. et al. Personalized privacy-preserving trajectory data publishing. *Chinese Journal of Electronics*, IET, v. 26, n. 2, p. 285–291, 2017.

MA, M. et al. You can hide, but your periodic schedule can't. In: IEEE. *Quality of Service (IWQoS), 2017 IEEE/ACM 25th International Symposium on*. [S.l.], 2017. p. 1–6.

MACHANAVAJJHALA, A. et al. l-diversity: Privacy beyond k-anonymity. In: IEEE. *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on*. [S.l.], 2006. p. 24–24.

MAHDAVIFAR, S. et al. A clustering-based approach for personalized privacy preserving publication of moving object trajectory data. In: *International Conference on Network and System Security*. [S.l.: s.n.], 2012. p. 149–165.

MAY, T. The crypto anarchist manifesto. *High Noon on the Electronic Frontier: Conceptual Issues in Cyberspace*, 1992.

MEYERSON, A.; WILLIAMS, R. On the complexity of optimal k-anonymity. In: ACM. *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. [S.l.], 2004. p. 223–228.

MONREALE, A. et al. C-safety: a framework for the anonymization of semantic trajectories. *Trans. Data Privacy*, v. 4, n. 2, p. 73–101, 2011.

NERGIZ, M. E.; ATZORI, M.; SAYGIN, Y. *Perturbation-driven anonymization of trajectories*. [S.l.], 2007.

NERGIZ, M. E.; ATZORI, M.; SAYGIN, Y. Towards trajectory anonymization: a generalization-based approach. In: ACM. *Proceedings of the SIGSPATIAL ACM GIS 2008 International Workshop on Security and Privacy in GIS and LBS*. [S.l.], 2008. p. 52–61.

OLESHCHUK, V. Internet of things and privacy preserving technologies. In: IEEE. *Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology*. [S.l.], 2009. p. 336–340.

PARENT, C. et al. Semantic trajectories modeling and analysis. *ACM Computing Surveys (CSUR)*, ACM, v. 45, n. 4, p. 42, 2013.

PFITZMANN, A.; HANSEN, M. A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management. 2010.

POST, D. G. Pooling intellectual capital: Thoughts on anonymity, pseudonymity, and limited liability in cyberspace. *U. Chi. Legal F.*, HeinOnline, p. 139, 1996.

PRIMAULT, V. et al. Time distortion anonymization for the publication of mobility data with high utility. In: IEEE. *Trustcom/BigDataSE/ISPA, 2015 IEEE*. [S.l.], 2015. v. 1, p. 539–546.

R. *Statistical Disclosure Control Methods for Anonymization of Microdata and Risk Estimation*. 2018. <https://cran.r-project.org/web/packages/sdcMicro/sdcMicro.pdf>. Acessado em 20/12/2018.

RACHELS, J. Why privacy is important. *Philosophy & Public Affairs*, JSTOR, p. 323–333, 1975.

RAJESH, N.; ABRAHAM, S. Privacy preserved approach for trajectory anonymization through zone creation for halting points. In: IEEE. *Networks & Advances in Computational Technologies (NetACT), 2017 International Conference on.* [S.l.], 2017. p. 229–234.

REIMAN, J. H. Privacy, intimacy, and personhood. *Philosophy & Public Affairs*, JSTOR, p. 26–44, 1976.

RNP. *Eduroam.* 2018. <https://www.rnp.br/servicos/servicos-avancados/eduroam>. Acessado em 20/12/2018.

SALAS, J.; MEGÍAS, D.; TORRA, V. Swapmob: Swapping trajectories for mobility anonymization. In: SPRINGER. *International Conference on Privacy in Statistical Databases.* [S.l.], 2018. p. 331–346.

SAMARATI, P.; SWEENEY, L. *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.* [S.l.], 1998.

SCHWAB, D.; BUNT, R. Characterising the use of a campus wireless network. In: IEEE. *INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies.* [S.l.], 2004. v. 2, p. 862–870.

SOLOVE, D. Understanding privacy. 2008.

SPACCAPIETRA, S. et al. A conceptual view on trajectories. *Data & knowledge engineering*, Elsevier, v. 65, n. 1, p. 126–146, 2008.

SWEENEY, L. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, World Scientific, v. 10, n. 05, p. 571–588, 2002.

TANG, D.; BAKER, M. Analysis of a local-area wireless network. In: ACM. *Proceedings of the 6th annual international conference on Mobile computing and networking.* [S.l.], 2000. p. 1–10.

TEMPL, M. et al. Introduction to statistical disclosure control (sdc). *Project: Relative to the testing of SDC algorithms and provision of practical SDC, data analysis OG*, 2013.

TERROVITIS, M. et al. Local suppression and splitting techniques for privacy preserving publication of trajectories. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, v. 29, n. 7, p. 1466–1479, 2017.

TIMES, N. Y. *How Trump Consultants Exploited the Facebook Data of Millions.* 2018. <https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html>. Acessado em 12/10/2018.

TU, Z. et al. Protecting trajectory from semantic attack considering k-anonymity, l-diversity and t-closeness. *IEEE Transactions on Network and Service Management*, IEEE, 2018.

WANG, F.; ZHU, X.; MIAO, J. Semantic trajectories-based social relationships discovery using wifi monitors. *Personal and Ubiquitous Computing*, Springer, v. 21, n. 1, p. 85–96, 2017.

WAREKAR, R.; PATIL, S. Efficient approach for anonymizing tree structured dataset using improved greedy search algorithm. *International Journal of Science and Research (IJSR), Impact Factor*, 2014.

WARREN, S. D.; BRANDEIS, L. D. The right to privacy. *Harvard law review*, JSTOR, p. 193–220, 1890.

WESTIN, A. F.; RUEBHAUSEN, O. M. *Privacy and freedom.* [S.l.]: Atheneum New York, 1967.

YANES, A. Privacy and anonymity. *CoRR*, abs/1407.0423, 2014.