UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE CIÊNCIAS

DEPARTAMENTO DE COMPUTAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

EDUARDO RODRIGUES DUARTE NETO

UMA ABORDAGEM DE PRIVACIDADE DE DADOS EM SERVIÇOS BASEADOS EM LOCALIZAÇÃO

FORTALEZA

2019

EDUARDO RODRIGUES DUARTE NETO

UMA ABORDAGEM DE PRIVACIDADE DE DADOS EM SERVIÇOS BASEADOS EM LOCALIZAÇÃO

Dissertação apresentada ao Curso de do Programa de Pós-Graduação em Ciência da Computação do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Ciência da Computação. Área de Concentração: Banco de Dados

Orientador: Prof. Dr. Javam de Castro Machado

FORTALEZA

2019

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

D871a Duarte Neto, Eduardo Rodrigues.
Uma Abordagem de Privacidade de Dados em Serviços Baseados em Localização / Eduardo Rodrigues Duarte Neto. – 2019.
75 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Ciências, Programa de Pós-Graduação em Ciência da Computação, Fortaleza, 2019.
Orientação: Prof. Dr. Javam de Castro Machado.

1. Privacidade de Dados. 2. Ofuscação. 3. Localizações Falsas. 4. Serviços de localização. 5. k-anonimato. I. Título.

CDD 005

EDUARDO RODRIGUES DUARTE NETO

UMA ABORDAGEM DE PRIVACIDADE DE DADOS EM SERVIÇOS BASEADOS EM LOCALIZAÇÃO

Dissertação apresentada ao Curso de do Programa de Pós-Graduação em Ciência da Computação do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Ciência da Computação. Área de Concentração: Banco de Dados

Aprovada em: 11 de Março de 2019

BANCA EXAMINADORA

Prof. Dr. Javam de Castro Machado (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. César Lincoln Mattos
Universidade Federal do Ceará (UFC)

Prof. Dr. Daniel Cardoso Morais de Oliveira
Universidade Federal Fluminense (UFF)

Aos meus pais, meus amigos e a toda minha família que, com muito carinho e apoio, não mediram esforços para que eu chegasse até esta etapa de minha vida.

AGRADECIMENTOS

A Deus por ter me dado saúde e força para superar as dificuldades.

Aos meus pais, Eduardo Duarte e Samia Duarte, pelo amor, incentivo e apoio incondicional.

Ao meu orientador, Prof. Dr. Javam Machado, por me orientar e aconselhar em minha carreira acadêmica.

Ao Prof. Dr. Cesar Lincoln Mattos e Prof. Dr. Daniel Cardoso por, generosamente, aceitar meu convite e compor a banca.

À banca e amigos pelo trabalho e empenho que tiveram em revisar este trabalho.

Aos amigos Bruno Leal, Daniel Praciano, André Mendonça, Edvar Filho, Felipe Timbó, Iago Chaves, Isabel Lima, Israel Vidal, Leonardo Linhares e pedro Ramyres pelo companheirismo ao longo do desenvolvimento deste trabalho.

A todos os colegas do Laboratório de Sistemas e Banco de Dados (LSBD) pela agradável convivência diária.

A todos os familiares que, nos momentos de minha ausência, sempre continuaram a me incentivar e apoiar.

À Universidade Federal do Ceará (UFC) pelo ambiente criativo e amigável que proporciona.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo aporte financeiro para a viabilização deste trabalho.

Ao LSBD por ter proporcionado uma estrutura adequada para o desenvolvimento deste trabalho e aporte financeiro para eventos científicos.

A todos que direta ou indiretamente fizeram parte da minha formação.

“O insucesso é apenas uma oportunidade para recomeçar com mais inteligência.”

(Henry Ford)

RESUMO

A crescente popularidade de dispositivos móveis com conectividade com a Internet e a presença de sistemas de posicionamento global (GPS) como componente padrão têm proporcionado o surgimento cada vez maior de serviços baseado em localização (LBS). Como consequência, tem sido gerado uma grande quantidade de dados de localização. Tais dados podem ser utilizados para diversas finalidades, tais como análise de fluxo de tráfego, planejamento de infraestrutura, entendimento do comportamento humano, etc. Entretanto, acessar dados pessoais de localização pode levantar severas questões de privacidade para a maioria dos usuários, especialmente quando o próprio provedor do serviço é um potencial agente malicioso. Dados de localização por sua natureza são altamente correlacionados a outros tipos de informações, que combinadas, podem ser utilizadas para inferência de dados sensíveis dos indivíduos associados aos dados. Muitas das abordagens já existentes envolvendo anonimização de dados de localização em serviços baseado em localização consideram o provedor do serviço como uma entidade confiável, responsável por realizar o processo de anonimização dos dados dos usuários. Neste trabalho propomos o PrivLBS, um algoritmo de ofuscação que utiliza a técnica de seleção de localizações falsas, cujo objetivo é proteger a localização do usuário entre outras localizações presentes na requisição, sem perca de qualidade do serviço. Nós também propomos um algoritmo de ataque que procura identificar a localização real do usuário em requisições enviadas ao provedor de serviço. Resultados de avaliação experimental demonstram que o nosso algoritmo de ataque obteve uma alta taxa de identificação da localização real do usuário quando aplicado sobre requisições anonimizadas por estratégias adotadas por trabalhos relacionados, enquanto em condições iguais o PrivLBS foi capaz de proteger as requisições dos usuários.

Palavras-chave: Privacidade de Dados. Ofuscação. Localizações falsas. Serviços de localização. k-anonimato

ABSTRACT

The increasing popularity of mobile devices with Internet connectivity and the presence of global positioning systems (GPS) as a standard component have provided the increasing emergence of location-based services (LBS). As a consequence, a large amount of location data has been generated. Such data can be used for a variety of purposes, such as traffic flow analysis, infrastructure planning, understanding of human behavior, etc. However, accessing personal location data can pose severe privacy issues for most users, especially when the service provider itself is a potential malicious agent. Location data by their nature are highly correlated to other types of information, which combined can be used to infer sensitive data from individuals associated with the data. Many of the existing approaches involving location anonymization in location-based services consider the service provider as a trusted entity, responsible for performing the process of anonymizing user data. In this work we propose the PrivLBS, an obfuscation algorithm that uses the dummy location technique, whose objective is to protect the user's location among other locations present in the request, without loss of service's quality. We also propose an attack algorithm that seeks to identify the actual location of the user in requests sent to the service provider. Experimental results demonstrate that our attack algorithm obtained a high rate of identification of the real location of the user when applied to anonymized requests by strategies used by related works, whereas under equal conditions PrivLBS was able to protect the user's requests.

Keywords: Data Privacy. Obfuscation. Dummy locations. Location based services. k-anonymity.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ALGORITMOS

LISTA DE ABREVIATURAS E SIGLAS

DLP	*Dummy Location Privacy*
DLS	*Dummy Location Selection*
GPS	*Global Positioning System* / Sistema de Posicionamento Global
lat	latitude
LBS	*Location Based Service* / Serviço Baseado em Localização
lon	longitude
PD	Privacidade Diferencial
POI	*Point of Interest* / Ponto de Interesse

LISTA DE SÍMBOLOS

ε	Limite de privacidade do modelo de Privacidade Diferencial
L	Conjunto de localizações cobertas pelo serviço baseado em localização
R	Requisição atual do usuário
R'	Requisição anterior do usuário
$max_{\Delta s}$	Máxima distancia alcançável
l_r	localização real do usuário
r_l	resposta à localização de consulta l enviada pelo serviço baseado em localização
d_{l_1,l_2}	Distância mínima entre as localizações l_1 e l_2
c_{l_1,l_2}	Correlação entre as localizações l_1 e l_2
v	Velocidade do usuário
Δt	Intervalo de tempo entre duas requisições consecutivas

SUMÁRIO

1 INTRODUÇÃO

Recentemente, devido ao enorme crescimento de dispositivos que utilizam tecnologia de *Global Positioning System* / Sistema de Posicionamento Global (GPS), serviços baseados em localização tornaram-se cada vez mais comuns em domínios sociais e empresariais (HU *et al.*, 2013). Estes numerosos serviços, tais como navegação, redes sociais, serviços de recomendação, jogos de realidade ampliada, entre outros, têm sido desenvolvidos e integrados às atividades diárias das pessoas, provendo informações úteis sobre seus arredores e sendo capazes de responder perguntas do dia a dia como: qual a melhor rota a ser percorrida para um determinado endereço? Quais os pontos turísticos mais próximos da minha localização atual? Em quanto tempo o táxi que eu solicitei irá demorar para chegar em meu apartamento? Tais perguntas podem ser respondidas facilmente por meio de serviços baseados em localização e suas informações geradas.

O uso das informações geradas por estes serviços pode beneficiar várias aplicações. De fato, muitas empresas e agências governamentais têm obtido conhecimento sobre os dados associados às atividades praticadas nas localizações, seja para melhorar o serviço prestado, para o lançamento de um novo produto, ou até mesmo para gerar uma nova política pela empresa. Este tipo de prática é essencial para a melhoria dos serviços prestados, seja na área de transporte, saúde, economia, comércio, ou outra área. Entretanto, acessar dados de localizações pessoais de usuários desses serviços, mesmo que com permissão, levanta severas preocupações de privacidade para a maioria dos usuários. Dessa forma, a utilização de serviços baseados em localização pode levar a sérios riscos de violação de privacidade devido a provedores de serviços não confiáveis (LI *et al.*, 2014; NIU *et al.*, 2015), que podem expor os dados de localização de seus usuários ou até mesmo vender suas informações de localizações a terceiros (ZHU *et al.*, 2013). De posse dessas informações, os dados obtidos por terceiros são utilizados para descoberta de dados sensíveis dos usuários, *i.e.*, dados de saúde, crenças religiosas, ideologias políticas, questões raciais, preferências sexuais, dentre várias outras. Um estudo publicado pelo site de notícias *NBC* revelou que o *Facebook* é capaz de identificar a preferência sexual de seus usuários através de uma rápida análise sobre os *"likes"* dados por estas pessoas em sua plataforma (ARMUS, 2017).

A Figura 1 ilustra o exemplo em que o usuário Bob ao longo do tempo realiza várias requisições a um serviço baseado em localização. No tempo t_1 Bob estava próximo ao pronto socorro de um hospital, já em um tempo t_2 ele realiza uma consulta próxima a um laboratório

de patologia, em outros dois momentos sua localização também esta próximo a localizações associadas a área de saúde. Considerando que é de conhecimento do provedor do serviço as requisições feitas pelos usuários, logo, o próprio provedor facilmente consegue inferir, com alta probabilidade, que Bob possui algum tipo de doença em razão das localizações enviadas por ele ao provedor, revelando uma informação sensível do usuário. Desta forma, considerando que o provedor de serviço pode não ser confiável, o risco de uma violação de privacidade, portanto, é bastante alto, deixando o usuário exposto.

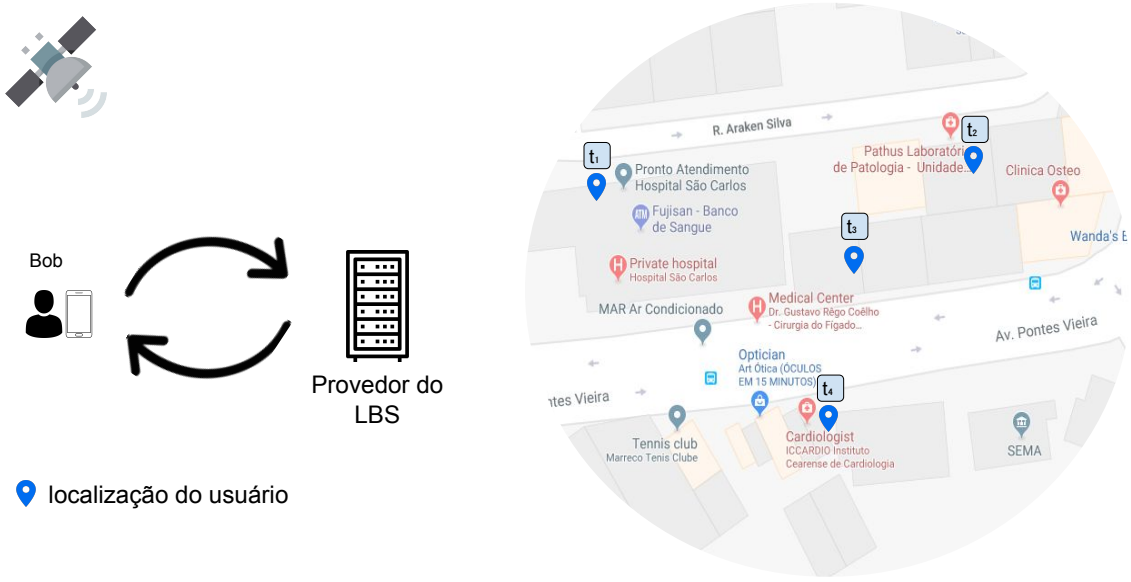

Figura 1 – Exemplo de requisições realizadas próximas a hospitais e clínicas, permitindo inferências de dados sensíveis do usuário.

A aplicação de um modelo de privacidade sobre requisições de usuários é imprescindível para evitar que as localizações dos indivíduos não sejam identificadas pelos provedores no uso destes serviços. Todavia, grande parte dos modelos de privacidade acaba provocando mudanças nos dados, afetando diretamente a sua utilidade, com impacto direto na qualidade do serviço. Portanto, gerenciar essa solução de compromisso (*trade-off*) entre privacidade dos indivíduos e utilidade dos seus dados se torna um outro grande desafio. Diversas técnicas têm sido propostas com o objetivo de garantir a privacidade dos usuários no uso destes serviços, a fim de permitir a sua utilização e assegurar um nível de garantia de privacidade (NIU *et al.*, 2016; TSOUKANERI *et al.*, 2016; ULLAH; SHAH, 2016; SUN *et al.*, 2017b). Algumas dessas técnicas utilizam o modelo de ofuscação (HUBAUX *et al.*, 2011; GHINITA, 2013), reduzindo a precisão da localização real enviada na requisição com o objetivo de preservar a privacidade do usuário. Uma forma clássica de aplicar o modelo de ofuscação é através da adição de ruído à

localização real do usuário. As técnicas de ofuscação de localização e privacidade diferencial são exemplos de técnicas que aplicam o modelo de ofuscação com adição de ruído. Entretanto, a adição de ruído diminui a qualidade do serviço prestado, por diminuir a precisão da localização de consulta.

Uma alternativa à abordagem de ofuscação é a técnica de seleção de localizações falsas (*Dummy Locations*), que garante a privacidade do usuário através da adição de localizações falsas a serem enviadas juntamente com a localização real na requisição feita ao provedor do serviço. Nessa abordagem, $k-1$ localizações falsas são geradas e adicionadas à consulta realizada pelo usuário ao provedor do *Location Based Service* / Serviço Baseado em Localização (LBS), a fim de confundir a localização real do indivíduo que realizou a consulta. Por exemplo, no momento em que um usuário desejar obter o centro comercial mais próximo de sua localização atual, ao especificar o valor de k, outras $k-1$ localizações falsas serão automaticamente geradas e enviadas ao provedor de serviço. O provedor retornará ao usuário os centros comerciais mais próximos para cada uma das k localizações presentes na requisição, dentre eles, o centro comercial mais próximo da localização real do usuário.

Embora a técnica de seleção de localizações falsas garanta bons níveis de privacidade sem afetar a qualidade do serviço, identificamos nos trabalhos existentes na literatura (KIDO *et al.*, 2005; VU *et al.*, 2012; NIU *et al.*, 2014; SUN *et al.*, 2017a) uma potencial fragilidade contra ataques que explorem de conhecimentos adversários sobre o conjunto de dados, os chamados ataques de conhecimento. Esta fragilidade aumenta bastante quando o usuário está realizando consultas contínuas ao provedor do serviço, ou seja, várias consultas consecutivas ao longo do tempo.

Como forma de demonstrar esta vulnerabilidade presente no processo de anonimização destes trabalhos que utilizam a técnica de seleção de localizações falsas, propomos um novo tipo de ataque de conhecimento, o ACon, que busca revelar a localização real do usuário presente nas requisições. Nosso ataque procura explorar o conhecimento sobre a distância *manhattan* entre as localizações de consultas consecutivas enviadas ao provedor de serviço, e a sua correlação com as localizações presentes na resposta à requisição anterior realizada pelo usuário.

Assim, em um cenário em que o próprio provedor é o atacante, como garantir a privacidade de localização dos usuários ao utilizarem serviços baseados em localização sem que haja uma perda de qualidade do mesmo?

Para solucionar o problema propomos uma nova técnica baseada no modelo de ofuscação, denominada PrivLBS, capaz de assegurar que as localizações dos indivíduos que utilizam serviços baseados em localização não sejam facilmente reidentificadas. O PrivLBS procura anonimizar a requisição, utilizando do conhecimento adversário do próprio provedor para garantir a privacidade de localização dos usuários.

1.1 Objetivos

Nesta seção apresentaremos os objetivos deste trabalho. Primeiramente apontaremos os objetivos mais gerais, e em seguida, indicaremos os objetivos específicos. Por fim, iremos destacar as contribuições alcançadas e as produções científicas decorrentes deste trabalho.

1.1.1 Objetivo geral

Diante do cenário apresentado na motivação, onde o provedor do serviço baseado em localização é um potencial atacante, o objetivo geral deste trabalho consiste em produzir uma solução que preserve a privacidade de localização dos indivíduos ao utilizarem LBSs, sem comprometer a utilidade dos dados e por consequência a qualidade do serviço.

1.1.2 Objetivos específicos

Como forma de atender ao objetivo geral deste trabalho, estabelecemos os seguintes objetivos específicos:

- Dado um modelo de privacidade, definir um método de anonimização que atenda a este modelo, mantendo a utilidade dos dados de forma a não afetar a qualidade do serviço prestado;
- Definir a arquitetura do LBS de forma a garantir que o processo de anonimização não afete a eficiência do serviço;
- Propor um método de ataque sobre as requisições feitas ao provedor de serviço como forma de avaliar a solução proposta;
- Avaliar a eficiência da solução proposta utilizando dados de localizações reais em termos de preservação de privacidade.

1.2 Contribuições

Como resultado desta dissertação, primeiramente, escrevemos um algoritmo de ataque, denominado ACon, essencial para demonstrar a eficiência de nossa solução para proteger a privacidade de localização do usuário. Ele procura explorar o conhecimento adquirido sobre as localizações do conjunto de dados, e o histórico de requisições dos usuários, a fim de identificar a localização real presente na requisição.

Como contribuição principal deste trabalho, nós definimos e implementamos o PrivLBS, um algoritmo de ofuscação, aplicado no próprio dispositivo do usuário, cujo objetivo é anonimizar a requisição a ser enviada ao provedor do LBS sem que haja perda de qualidade do serviço, protegendo a localização do usuário.

1.2.1 Produção científica

As contribuições científicas apresentadas neste trabalho possibilitaram a publicação do seguinte artigo:

- Neto, Eduardo R. D., André L. C. Mendonça, Felipe T. Brito and Javam C. Machado. "PrivLBS: uma Abordagem para Preservação de Privacidade de Dados em Serviços baseados em Localização". SBBD (2018). Este artigo recebeu o prêmio "Best Student Paper Award" desta edição do SBBD.

1.3 Estrutura da dissertação

Esta dissertação está organizada da seguinte forma: No Capítulo 2 são apresentados conceitos e definições fundamentais sobre preservação de privacidade em LBS. O Capítulo 3 ressalta e discute os trabalhos relacionados mais relevantes, caracterizando modelos de privacidade anteriormente pesquisados. Em seguida, o Capítulo 4 apresenta a nossa solução proposta para preservação de privacidade em dados de localização. O Capítulo 5 apresenta os resultados obtidos por uma série de experimentos realizados, utilizando um conjunto de dados real. Finalmente, o Capítulo 6 conclui o trabalho apresentando um resumo dos resultados alcançados e mostrando direções de pesquisa futuras.

2 LBS E PRIVACIDADE DE DADOS

Este capítulo consiste na fundamentação teórica necessária para o entendimento deste trabalho, incluindo os problemas conhecidos na literatura e as técnicas utilizadas para o desenvolvimento da solução proposta. Uma visão geral acerca de serviços baseados em localização é apresentada na Seção 2.1. Em seguida, na Seção 2.2, apresentamos uma noção básica da privacidade de dados, destacando sua necessidade e importância. Na Seção 2.3 iremos falar de privacidade de localização. Na Seção 2.4 introduziremos o conceito de conhecimento adversário. Na Seção 2.5 apresentamos os modelos de privacidade mais conhecidos na literatura, com ênfase na técnica de ofuscação conhecida como localizações falsas, que tem servido de base para a realização desta pesquisa.

2.1 Serviços baseados em localização (LBS)

O GPS é um componente padrão na maioria de dispositivos móveis, tais como *smartphones* e *wearables*. Está combinação tem impulsionado o crescimento da popularidade dos serviços baseados em localização. Estes serviços utilizam a posição ou localização dos usuários, compartilhado através de um dispositivo, tais como *smartphones*, a fim de integrar com outras informações para prover uma funcionalidade. Por exemplo, um usuário pode fazer a seguinte requisição a um LBS "Encontre o hospital mais próximo da minha localização". Como resposta à requisição, o serviço irá buscar em um conjunto de dados, mantido ou não pelo próprio provedor do serviço, o *Point of Interest* / Ponto de Interesse (POI) requisitado, tendo como referência a localização passada na consulta.

Definição 1 *Pontos de interesse são dados de localizações carregados de informações complementares sobre as mesmas.*

Conforme a definição de POI acima. Estes dados representam localizações, como hospitais, supermercados e cinemas. Estes dados, em geral, são carregados de informações, tais como, coordenadas geográficas, horário de funcionamento, comentários de frequentadores, etc.

A Tabela 1 contém algumas das mais populares aplicações em LBS. Apesar dos mais comuns LBS serem relacionados à navegação e busca de POIs, os LBS vão muito além. Por exemplo, nos últimos anos, a popularidade dos *smartphones*, cada vez mais potentes em termos de capacidade de processamento e memória têm impulsionado o mercado de jogos para estes

Aplicação	Check-in	Postagem geo referencial	Localizar amigos próximos	Navegação	Localizar POI	Desempenho esportivo	Jogos de realidade aumentada
Foursquare	✓		✓		✓		
Facebook	✓		✓		✓		
Twitter		✓					
Google Maps				✓	✓		
Wechat		✓	✓				
Pokemon Go			✓				✓
Fitbit						✓	
Yelp	✓				✓		

Tabela 1 – Aplicações Baseadas em Localização

dispositivos. Pokémon Go é um jogo que surgiu como um dos primeiros grandes sucessos a utilizarem a tecnologia de realidade aumentada. Pokémon Go usa o GPS do dispositivo móvel para capturar, batalhar e treinar criaturas chamadas Pokémons, que aparecem como se estivessem no mundo real. Em Maio de 2018, o jogo já havia sido baixado mais de 800 milhões de vezes (HARRIS, 2018) e gerado um lucro de mais de 3 bilhões de dólares (SUPERDATA, 2019). Outros LBSs bem comuns são serviços de rastreamento esportivos, onde dados de usuários realizando esportes são coletados através de dispositivos móveis, tais como pulseiras eletrônicas de monitoramento esportivo, ou relógios inteligentes. Utilizado tanto por atletas profissionais como por pessoas comuns em seu cotidiano, este serviço permite que seus usuários obtenham várias informações coletadas durante a prática esportiva, como velocidade, força de arranque, distância percorrida, ou até mesmo informações de saúde dos usuários, como a sua pressão e seus batimentos cardíacos. Estes são só alguns poucos exemplos para mostrar o potencial gigantesco destes serviços.

A Figura 2 ilustra um típico LBS, cujos principais componentes presentes são:

- **Usuários:** são participantes que irão usufruir do serviço baseado em localização prestado. Através de dispositivos, *i.e.*, *smartphones*, *notebooks*, *wearables*, dentre outros, os usuários se conectam ao meio de comunicação e enviam requisições ao provedor do serviço.
- **Sistema de posicionamento:** permite determinar a localização dos objetos envolvidos, *i.e.*, POIs, usuários, ou outra entidade qualquer. O GPS é o mais popular sistema de posicionamento. Ele é um mecanismo de posicionamento por satélite

que fornece a um aparelho receptor a sua posição.

- **Provedor do LBS:** É o responsável por receber as requisições dos usuários e prestar o serviço baseado em localização de acordo com sua natureza, seja para encontrar um POI, seja para auxiliar na navegação do usuário, ou um outro tipo de serviço qualquer que utiliza a informação de localização enviada na requisição.
- **Rede:** é o meio através do qual acontece o tráfego de informações entre os participantes. Normalmente o meio utilizado é a Internet.

Os usuários, através do GPS de seus dispositivos, obtém sua posição, e usufruem do serviço se comunicando com o provedor do serviço através do envio de requisições pela Internet. Estas requisições podem ser simples, onde o usuário realiza requisições de forma esporádica, ou contínuas, onde um usuário realiza ao longo do tempo várias requisições. As requisições contínuas são aquelas mais sujeitas a quebra de privacidade, justamente pelo alto grau de correlação entre as localizações presentes em requisições consecutivas.

Um exemplo de uma aplicação baseada em localização cujas requisições, podem ser tanto simples como contínuas é o Uber. Quando o usuário adiciona a localização de destino no aplicativo do Uber em seu dispositivo e envia essa localização ao provedor do serviço, ele recebe como resposta uma estimativa de preço. Neste caso o usuário realizou uma consulta simples. Entretanto, quando a viagem começa, há uma troca constante de informações entre os participantes da viagem: motorista, usuário e provedor do serviço. Nesta situação são realizadas requisições contínuas.

Garantir a privacidade dos usuários ao utilizarem LBSs é um desafio. Especialmente quando estamos falando na preservação de privacidade de localização do usuário. Um dado sensível cuja precisão tem um impacto direto na qualidade do serviço prestado.

2.2 Privacidade de Dados

Nos últimos anos temos visto uma revolução na quantidade de dados coletados diariamente por governos, grandes corporações e instituições ao redor do mundo. As redes sociais, o comércio virtual, aplicativos dos mais diversos tipos e finalidades têm contribuído para a exposição dos dados de seus usuários. Estes dados são extremamente valiosos, muitas vezes sendo alvos de comércio dado às infinidades de uso dos mesmos. Por exemplo, dados de localização são bem úteis para empresas de publicidade, uma vez que permitiriam oferecer propaganda de serviços ou produtos de acordo com o perfil de seus usuários, analisando os locais

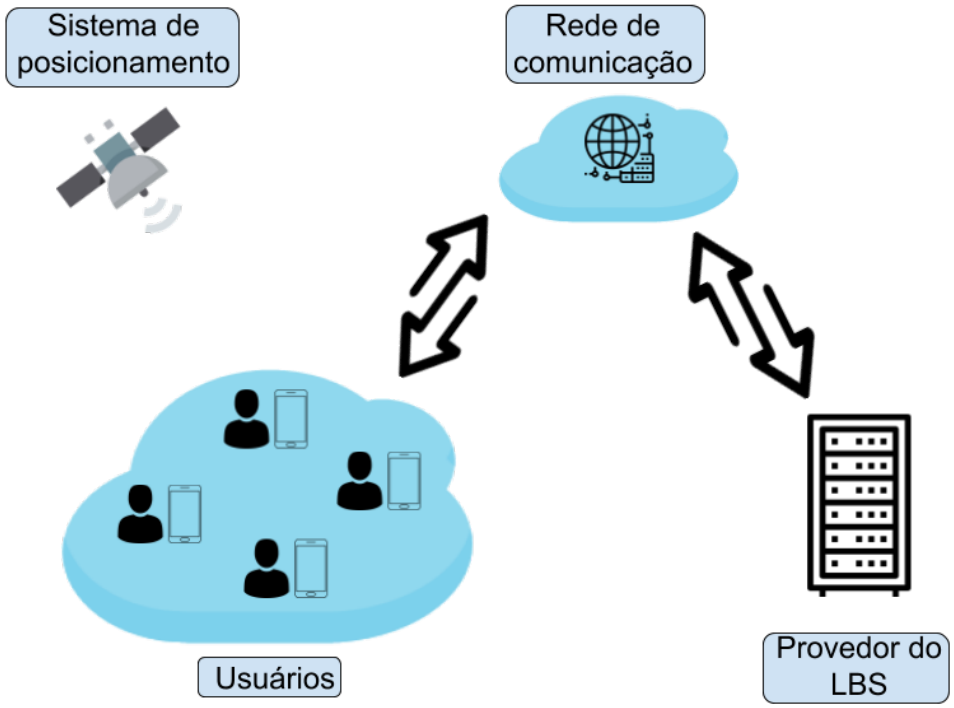

Figura 2 – Arquitetura de um típico modelo de serviço baseado em localização.

que cada indivíduo costuma frequentar. Já dados de mobilidade urbana podem ser usados por empresas de transporte público para disponibilizar mais rotas que atendem melhor à população baseada no perfil de deslocamento dos usuários. Análise de perfil, estudo de correlação, análises estatísticas e descoberta de padrões são só algumas das inúmeras aplicações que têm sido realizadas sobre estes dados coletados.

Entretanto, a exposição dos dados do usuário em seu formato original pode colocar em risco sua privacidade se caírem na posse de indivíduos maliciosos (também conhecidos por adversários ou atacantes). O que se pretende proteger é a ligação entre alguns tipos de atributos e a identidade do usuário, não impedindo portanto análises que tenham um caráter não malicioso. Em um âmbito de dados relacionados, um indivíduo é representado por um registro que contém os seguintes tipos de atributos:

- **Identificadores:** são atributos que identificam unicamente indivíduos, tais como "nome", "CPF", "e-mail", etc. e são sempre removidos antes de serem publicados;
- **Semi-identificadores (SI):** são todos aqueles atributos que não são identificadores explícitos mas podem potencialmente identificar um indivíduo, especialmente quando agrupados. São exemplos de semi-identificadores em dados relacionais

"data de nascimento"e "CEP".

- **Atributos sensíveis:** contém informações sensíveis sobre indivíduos, tais como "doença", "salário", etc.
- **Atributos não sensíveis:** é qualquer tipo de atributo que não se enquadra em nenhuma das categorias anteriores.

Desta forma o que se pretende é evitar que indivíduos maliciosos façam inferências sobre dados sensíveis, tais como: salário, doenças, crenças, preferências sexuais, entre outras. Assim, como forma de proteger efetivamente a privacidade de indivíduos, o dono dos dados precisa garantir que eventuais descobertas não ocorram no conjunto de dados publicado. Com este objetivo os dados devem ser anonimizados antes de serem publicados.

A anonimização de dados é uma técnica de preservação de privacidade que busca modificar valores dos atributos dos dados do usuário com o objetivo de ocultar a identidade e/ou informações sensíveis de indivíduos. A escolha da técnica de anonimização é fundamental, uma vez que esta modificação pode acarretar em perda de informação, o que implica na diminuição da utilidade dos mesmos. Privacidade e utilidade são tidas como duas grandezas inversamente proporcionais, como ilustra a Figura 3, podendo ser também tratado como um problema de otimização (ASIKIS; POURNARAS, 2018. Disponível em: http://www.sciencedirect.com/science/article/pii/S0167739X17322252). Sendo assim quanto maior o grau de privacidade garantido pela anonimização, menor é a utilidade dos dados, o que dado a natureza do serviço, pode vir a diminuir consideravelmente a qualidade do mesmo. Dessa forma o desafio na prestação de serviços com garantia de privacidade dos dados dos envolvidos é anonimizá-los de tal forma que a privacidade dos indivíduos seja protegida, enquanto a utilidade dos dados é mantida.

2.3 Privacidade de Localização vs Privacidade de Dados

Privacidade de localização é uma subcategoria em Privacidade de Dados em virtude da natureza peculiar dos dados de localização. São algumas características deste tipo de dado:

- **Dados massivos:** o uso de LBS gera uma quantidade enorme de dados de localização, com formato variado.
- **Alta correlação:** dados de localização são altamente correlacionados. Esta correlação pode revelar informações além do esperado.
- **Dinamicidade:** estes dados podem mudar rapidamente ao longo do tempo.

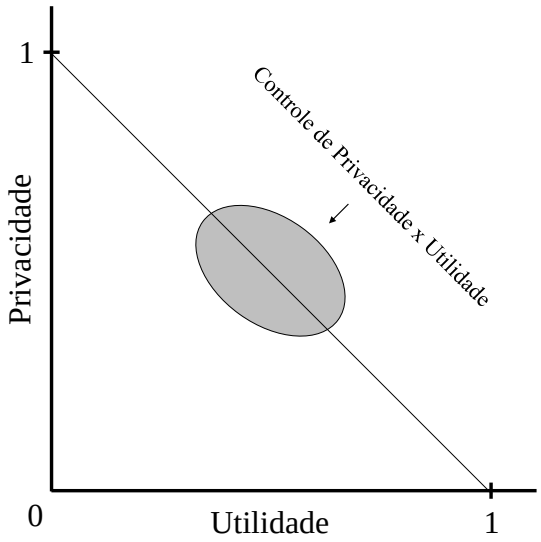

Figura 3 – Balanceamento entre utilidade e privacidade.

O vazamento da informação de localização dos usuários pode permitir uma série de ataques de indivíduos maliciosos, que vão desde vigilância física e perseguição, até roubo de identidade. Outro risco é o de inferências sobre informações sensíveis. Quando se fala sobre dados sensíveis é consenso o risco de se expor registros financeiros e médicos de indivíduos, dado sua natureza. Entretanto, este risco também existe associado a exposição de dados de localização, uma vez que estes tipos de dados em específico estão extremamente ligados ao comportamento dos indivíduos. Desta forma, através da análise sobre os dados de localização é possível traçar o perfil de indivíduos, o que pode levar a inferências sobre dados sensíveis. Por exemplo, um usuário que usualmente frequenta POIs associados a atividade financeira, como bancos, casas de crédito, dentre outros, ao expor sua localização, pode vir a expor sua situação financeira. Da mesma forma, com dados de saúde, permitindo que sejam feitas inferências sobre seu estado de saúde.

2.3.1 Correlação

Em virtude da natureza dos dados de localização, altamente correlacionados, vamos falar um pouco sobre análise de correlação.

A análise de correlação é um método estatístico largamente aplicado em pesquisas científicas de dados. Frequentemente usada para determinar a existência de relação entre duas diferentes variáveis, medindo a força dessa relação. Até por isso, muitas vezes a análise de correlação é usada para medir o grau de similaridade entre dois objetos, através da comparação entre os atributos destes objetos.

Esta correlação pode ser medida através de diferentes coeficientes. Os dois mais populares são os coeficientes de *Pearson* e *Spearman* (HAUKE; KOSSOWSKI, Jun 2011), cujo grau de correlação varia entre $[1, -1]$.

O coeficiente de *Pearson* é uma métrica de força da relação linear entre duas variáveis. Se o coeficiente entre duas variáveis for igual a 0, implica que não existe uma relação linear entre as variáveis. Para qualquer outro valor entre -1 e 1 indica o mesmo grau de correlação entre as variáveis. O sinal $+$ ou $-$ vai indicar a direção dessa relação. Em caso positivo, implica se o valor de uma variável aumenta, a outra variável também aumenta, caso contrário o sinal é negativo. Em caso de o coeficiente ser igual a $+1$ tem-se uma correlação perfeita positiva, onde os valores das variáveis são proporcionais. Quando o coeficiente de correlação é -1 a correlação é dita perfeita negativa. Neste caso, os valores das variáveis são inversamente proporcionais.

O coeficiente de *Spearman* por sua vez é uma medida não-paramétrica da força da relação entre duas variáveis que não pode ser medida quantativamente. Portanto, adequado tanto para variáveis lineares como não lineares. Ele avalia o quão bem uma função monotônica arbitrária é capaz de descrever a relação entre duas variáveis. Um coeficiente igual a 1 indica uma alta correlação e -1 uma baixa correlação entre as variáveis.

O uso dos coeficientes de *Pearson* e *Spearman* para análise estatística de correlação espacial entre dados de localização foi examinado por (GRIFFITH; CHUN, 2014), onde foi demonstrado que o coeficiente de *Spearman* é mais adequado para dados de localização.

2.4 Conhecimento Adversário

Nosso trabalho considera que uma ocorrência de uma violação de privacidade de localização ocorre por meio de um ataque, *i.e.*, quando um adversário consegue associar a identidade de um usuário, à sua localização presente em uma requisição enviada ao provedor de um serviço. Muitas vezes, o ataque utiliza do conhecimento previamente adquirido pelo agente malicioso para identificar a localização real. A força desse conhecimento está diretamente ligada ao nível de privacidade alcançada na anonimização das informações de localização do usuário.

Definição 2 *Conhecimento adversário é todo conhecimento sobre os dados da requisição que pode ser usado pelo adversário para realizar um ataque.*

Podemos citar alguns exemplos de conhecimentos adversários:

- O número de usuários de uma área em determinado horário.

- A existência de restrições físicas de qualquer natureza sobre as localizações, como se há ou não uma rodovia entre duas localizações.
- A popularidade das localizações.

Um mecanismo de privacidade adequado deve ser capaz de proteger a privacidade de localização do usuário, independente do conhecimento adversário que o atacante possua.

2.5 Modelos de Privacidade

Conforme mencionado anteriormente, a publicação de dados pode levar a sérios riscos de violação de privacidade devido à existência dos semi-identificadores. Isso pode acarretar em consequências graves por causa do uso não autorizado de informações sensíveis pertencentes aos indivíduos. Com o objetivo de preservar a privacidade de localizações, (HUBAUX *et al.*, 2011) divide os modelos de privacidade em dois grupos: modelos de anonimização e de ofuscação.

2.5.1 Modelos de anonimização

Neste modelo as técnicas de privacidade visam quebrar a ligação entre a identidade do usuário e a informação de localização do mesmo (LIU *et al.*, 2018). As principais técnicas utilizadas são o *k*-anonimato e zona mista.

2.5.1.1 k-anonimato

Considerado o mais conhecido dos modelos de privacidade, o *k*-anonimato foi proposto por (SWEENEY, 2002) com o objetivo de prevenir ataques de ligação ao registro. Em sua forma original, esse modelo assegura que, para cada combinação de k atributos semi-identificadores, existem, pelo menos, k registros distintos no conjunto de dados publicado, formando uma classe de equivalência. O *k*-anonimato atua, portanto, sobre o princípio da indistinguibilidade, onde é garantido que cada registro em um conjunto de dados *k*-anônimo, *i.e.*, que satisfaz as propriedades do modelo *k*-anonimato, é indistinguível a, no mínimo, outros $k-1$ registros no conjunto de dados, em relação ao conjunto de semi-identificadores. Em privacidade de localização, um sujeito é tido *k*-anônimo se sua localização é indistinguível da localização de outros $k-1$ usuários. Portanto, a probabilidade de um usuário malicioso violar a privacidade de um indivíduo através de um ataque de ligação ao registro não poderá ser maior do que $\frac{1}{k}$.

O parâmetro *k* do modelo é responsável por balancear a utilidade e a privacidade dos dados. Assim, quanto maior o valor de *k*, maior será a privacidade dos dados e, consequentemente, menor a utilidade dos dados, e vice-versa. É importante ressaltar que não existem abordagens analíticas para determinar um valor ótimo para o parâmetro *k* (DEWRI *et al.*, 2008), sendo este um problema NP-difícil (MEYERSON; WILLIAMS, 2004). Dessa forma, cabe aos *dataholders* esta complexa tarefa quando da aplicação do processo de anonimização por k-anonimato sobre um conjunto de dados.

2.5.1.2 Zona mista

Diferente do *k*-anonimato que necessita das informações de identidade do usuário, as zonas mistas podem ser usadas sem estas informações. Elas consistem de regiões geográficas onde o movimento dos usuários é supostamente não rastreável pela aplicação (Ying *et al.*, 2013). Esta abordagem busca esconder a identidade do usuário através do uso de pseudo-ids. Sempre que um usuário entra em uma zona mista, ele recebe um novo pseudo-id que permite o usuário utilizar o serviço, impedindo que sua identidade fique associada as localizações da zona mista.

2.5.2 Ofuscação

Os mecanismos de ofuscação buscam reduzir a precisão das informações de localização do usuário a fim de preservar a privacidade de localização dos mesmos (LIU *et al.*, 2018). As principais técnicas de ofuscação são: ofuscação de localização, Privacidade Diferencial (PD) e as técnicas de localizações falsas.

2.5.2.1 Ofuscação de Localização

Esta abordagem propõe diminuir a precisão da localização enviada na requisição a fim de preservar a privacidade de localização. Basicamente a localização do usuário estaria protegida pela imprecisão da informação de localização enviada. A abordagem clássica de ofuscação de localização envia uma área circular no lugar da localização real do usuário ao LBS (LIU *et al.*, 2018). Outras abordagens adicionam algum ruído a posição geográfica do usuário seguindo algum critério e esta nova posição é enviada como informação de localização.

2.5.2.2 Privacidade Diferencial

Proposta por (DWORK, 2006), a privacidade diferencial consiste em um modelo matemático que oferece sólidas garantias de privacidade. É um modelo semântico que tem como objetivo fornecer informações estatísticas sobre um conjunto de dados sem comprometer a privacidade dos indivíduos envolvidos, através de um algoritmo aleatório, geralmente chamado de mecanismo. Esse mecanismo é responsável por introduzir aleatoriedade e proteger os resultados das consultas sobre o conjunto de dados. Em geral, os modelos sintáticos, como os modelos de anonimização, são vulneráveis a ataques com conhecimento adversário prévio. A privacidade diferencial, por sua vez, procura garantir um certo grau de privacidade atribuído pelo *dataholder* independente do conhecimento prévio que um atacante possa vir a ter.

Inicialmente, a Privacidade Diferencial foi projetada para atuar em um ambiente interativo, onde usuários submetem consultas a um conjunto de dados e o ambiente, por sua vez, responde às consultas através de um algoritmo de anonimização. No entanto, mostrou-se que também é possível aplicar o modelo sobre um conjunto de dados de forma a publicar sua versão anonimizada, ou seja, com os dados perturbados. A privacidade de localização em geral é garantida baseada numa área de indistinguibilidade geo-referencial (ANDRÉS *et al.*, 2013) de raio r, cujo grau de privacidade depende de r. O grau de privacidade é alcançado pela adição de ruído à localização real do usuário. Assim, dado o raio de indistinguibilidade r, o mecanismo irá calcular o ruído necessário que será adicionado a localização do usuário. Só então o usuário irá realizar uma requisição passando a localização de consulta contendo um ruído, diminuindo, portanto, a precisão da localização do usuário.

2.5.2.3 Localizações Falsas

As técnicas de localização falsa procuram mascarar a localização real do usuário ao enviar uma requisição com informação de localização contendo múltiplas localizações falsas ao LBS juntamente a localização real, conforme Figura 4. O objetivo é garantir que a probabilidade de se identificar a localização real dentre aquelas presentes na requisição seja inferior a $\frac{1}{k}$, onde k é o grau de privacidade desejado para uma requisição com uma quantidade de k localizações presentes.

Como as localizações falsas são selecionadas aleatoriamente através do dispositivo móvel do usuário, este método não requer nenhum servidor confiável, sendo realizada pelos

Figura 4 – Técnica de Localizações Falsas.

próprios dispositivos de seus usuários, diminuindo o risco de exposição da localização real. Desta forma, esta técnica é conhecida por atingir bons níveis de privacidade sem perda de qualidade do serviço, uma vez que a localização real está presente na requisição, embora mascarada entre outras localizações. Como o provedor do serviço irá responder a requisição em razão de cada uma das localizações presentes, o usuário irá ter sua requisição atendida com a precisão dá localização real enviada.

O grau de privacidade k é garantido através da indistinguibilidade das localizações enviadas na requisição. Entretanto, as localizações podem vir carregadas de informações, tais como coordenadas, popularidade, horário de funcionamento se for um estabelecimento, além de várias outras. Isto tem um impacto na escolha destas localizações. Uma escolha totalmente aleatória, certamente não garante uma indistinguibilidade das localizações presentes na requisição. Sendo assim é necessário realizar escolhas que tornem essas localizações falsas e a localização real do usuário, similares entre si aos olhos de um atacante. Além disto é necessário escolhas realísticas, principalmente quando no deparamos com um cenário de requisições contínuas, onde o usuário realiza várias requisições em sequência ao longo do tempo. Nestes cenários, apontamos como principais informações de conhecimento adversário do provedor de serviço, o histórico de requisições do usuário, bem como a resposta a cada uma das requisições feitas, a popularidade das localizações, a distância física entre as localizações e por fim mas não menos importante a correlação entre as localizações.

2.6 Conclusão

Neste capítulo primeiramente caracterizamos um LBS típico, bem como apresentamos alguns dos seus mais populares serviços. Em seguida, apresentamos noções gerais sobre privacidade de dados, bem como a busca pelo equilíbrio perfeito entre utilidade e privacidade. Definimos privacidade de localização como um subconjunto de privacidade de dados, apresentando noções básicas sobre dados de localização e riscos de privacidade ocasionados pela existência de conhecimento adversário. Como forma de garantir a privacidade dos dados de localização dos usuários, apontamos os modelos de privacidade mais adequados. Dentre os modelos de privacidade demos mais destaque aos modelos de ofuscação, principalmente às técnicas de localizações falsas, por serem objeto deste trabalho, apontando alguns conhecimentos adversários com altos riscos de quebra de privacidade, como a correlação entre as localizações.

3 TRABALHOS RELACIONADOS

Diversas soluções foram propostas com o objetivo de garantir a privacidade de usuários ao utilizarem serviços baseados em localização e, assim, impedir que suas informações sensíveis fossem descobertas. Em sua grande maioria, as soluções são divididas em abordagens baseadas em anonimização de localizações (GEDIK; LIU, 2008; YING; MAKRAKIS, 2014; DUCKHAM; KULIK, 2005; BAMBA *et al.*, 2008), criptografia (LU *et al.*, 2014) e ofuscação (ANDRÉS *et al.*, 2013; WANG *et al.*, 2017). Nestas últimas, na Seção 3.4, daremos destaque à técnica de seleção de localizações falsas, *dummy locations* (NIU *et al.*, 2014; NIU *et al.*, 2015; SUN *et al.*, 2017a).

3.1 Anonimização de localização

O trabalho em (GEDIK; LIU, 2008) propõe uma técnica personalizada de k-anonimato utilizando a estratégia de camuflagem. Nesse trabalho, os autores utilizam um servidor de anonimização confiável que considera o *trade-off* entre a privacidade da localização e a qualidade do serviço para anonimizar a localização dos usuários. Na solução, uma região de camuflagem contendo outros $k-1$ usuários, geograficamente distribuídos, é formada e, somente então, a consulta é submetida ao serviço baseado em localização. Também utilizando a estratégia de camuflagem, o trabalho em (YING; MAKRAKIS, 2014) assegura a privacidade dos usuários ao construir uma região de camuflagem contendo, pelo menos, k usuários e l segmentos de rua. Estes trabalhos têm como pontos negativos o fato de limitarem a qualidade do serviço ao adicionar um ruído na consulta, e de utilizar um participante centralizado responsável por realizar a anonimização, o que pode levar a riscos tanto de segurança, como de ponto única de falha, e assim afetar a continuidade do serviço.

Já o trabalho em (DUCKHAM; KULIK, 2005) propõe um *framework* que procura garantir um equilíbrio entre as necessidades de privacidade dos usuários e a qualidade do serviço, enviando ao provedor do LBS apenas as informações necessárias para a prestação de um serviço satisfatório de acordo com as necessidades do próprio usuário. Em (BAMBA *et al.*, 2008) é proposto uma abordagem de preservação que utiliza servidores de anonimização a fim de proteger a localização real do usuário. Ele permite que os usuários definam seus níveis de privacidade desejado com o objetivo de estabelecerem o equilíbrio desejado entre privacidade e qualidade de serviço.

3.2 Técnica de criptografia

O trabalho proposto em (LU *et al.*, 2014) apresenta um *framework*, denominado PLAM, para a preservação de privacidade em redes sociais de área local. Esse *framework*, além de atender ao modelo de privacidade *k*-anonimato, também assegura o modelo *l*-diversidade (MACHANAVAJJHALA *et al.*, 2006), considerando casos em que um adversário pode inferir informações sensíveis sobre indivíduos mesmo sem identificá-los. Todavia, o servidor de anonimização confiável encontrado em (GEDIK; LIU, 2008; YING; MAKRAKIS, 2014) é substituído por uma técnica de criptografia, denominada pseudo-ID, a qual não mantém a utilidade dos dados para fins de análise. Assim como a proposta anterior, esse modelo adiciona ruído que diminui a qualidade do serviço.

3.3 Técnica de ofuscação

Os trabalhos propostos em (ANDRÉS *et al.*, 2013; WANG *et al.*, 2017) definem uma área de indistinguibilidade de raio *r*, onde é adicionado um ruído à localização do usuário dentro desta área. A quantidade de ruído necessária para garantir a privacidade do usuário é calculada através do mecanismo de PD. Entretanto, não há uma garantia que a localização anonimizada não seja revelada em um cenário de consultas contínuas feita pelo provedor (HUBAUX *et al.*, 2011). Isto acontece especialmente pela natureza dos dados de localização, que são extremamente correlacionados, fazendo com que os modelos de privacidade diferencial tradicionais, não consigam estimar de forma correta a quantidade de ruído necessária para garantir a privacidade do usuário. Além do mais, devemos lembrar que a precisão da localização enviada na requisição é fator que impacta diretamente na qualidade do serviço. Logo, o ruído adicionado para garantir a privacidade pode tornar os dados inúteis e afetar completamente o serviço (ZAKHARY *et al.*, 2017).

3.4 Localizações Falsas

Nosso trabalho tem como principal referência, os trabalhos que utilizam a técnica de seleção de localizações falsas (*Dummy Location*). Conforme já discutido na Seção 2.5.2.3, nesta técnica são adicionadas $k-1$ localizações falsas que, juntamente à localização real, compõem a requisição a ser enviada ao provedor de serviço. As principais virtudes desta técnica são a ausência de perda de utilidade dos dados da requisição, e a utilização de um modelo distribuído,

onde cada usuário, através de seu dispositivo móvel, é responsável por anonimizar sua consulta.

3.4.1 Dummy Location Selection

Uma escolha aleatória seria a forma mais ingênua de selecionar as localizações falsas que irão compor a requisição. Entretanto, ao considerar o conhecimento prévio que o atacante possa vir a ter, esta estratégia de seleção pode não vir a ser eficaz na garantia de privacidade. O trabalho em (NIU *et al.*, 2014) propõem o *Dummy Location Selection* (DLS), um algoritmo de seleção de localizações falsas em função da entropia da requisição (SERJANTOV; DANEZIS, 2003), o qual mede o grau de incerteza sobre um conjunto de localizações selecionadas.

Nesse trabalho, os autores apresentaram um modelo de LBS no qual o provedor do serviço é responsável por coletar e disponibilizar aos usuários dados estatísticos sobre as consultas. Tais dados dizem respeito às probabilidades nas quais as requisições são demandadas ao LBS, calculadas através da Equação 3.1.

$$p_i = \frac{\textit{número de requisições sobre } l_i}{\textit{número total de requisições}} \tag{3.1}$$

Como exemplo, suponha um cinema localizado nas coordenadas $(3.8, 4.7)$. Ao longo de um período de tempo t foram realizadas 103 requisições a um provedor de serviço de localização passando este cinema como localização de consulta. Neste mesmo período foram realizados no total 1000 requisições ao provedor. Pela equação 3.1, obtemos que a probabilidade de serem realizadas requisições sobre este cinema é de $10,3\%$. Assim, tendo como referência as probabilidades das localizações estarem presentes nas requisições, é necessário que a escolha destas localizações garanta que a probabilidade de se identificar a localização real não seja superior a $\frac{1}{k}$, para garantir o grau de privacidade k desejado.

O DLS assegura a privacidade dos usuários, garantindo o grau de privacidade das técnicas de k-anonimato, ao submeter uma consulta contendo a localização real do usuário e de outras $k-1$ localizações falsas, cujas probabilidades de estarem presente em uma requisição são semelhantes a probabilidade da localização real.

O algoritmo do DLS (1), recebe como parâmetros de entrada:

- k: grau de privacidade definido pelo usuário. É a quantidade de localizações que estarão presentes na requisição, que por consequência define também a máxima probabilidade de re-identificação desejada da localização real.

- L: a lista de localizações cobertas pelo provedor LBS, e suas respectivas informações de probabilidade.
- l_r: localização real do usuário, onde $l_r \in L$.
- m: quantidade de conjuntos criados a partir das localizações com probabilidades similares à da localização real.

O grau de privacidade k e o parâmetro m possuem um impacto direto tanto na proteção ao usuário quanto no desempenho do algoritmo. Quanto maior k e m, maior é a privacidade garantida, entretanto, maior é o custo computacional da anonimização da requisição.

Durante o processo de seleção das localizações falsas, é criado um subconjunto contendo $2k$ localizações cujas probabilidades sejam as mais próximas da localização real (linha 2 do algoritmo). A razão para a escolha de $2k$ localizações é para aumentar o grau de privacidade da requisição, podendo este valor aumentar de acordo com a necessidade do usuário. A partir deste subconjunto de $2k$ localizações são construídos m conjuntos candidatos à requisição (linha 3). Cada um destes m conjuntos possuem $k-1$ localizações escolhidas de forma aleatória entre as $2k$ localizações previamente separadas, em adição da localização real (linhas 5 e 7). O objetivo é encontrar o conjunto ótimo em termos de entropia, assim, dentre os m candidatos é escolhido aquele com maior entropia, dado pela Equação 3.2 (linha 10), onde $q_i = p_i / \sum_{i=1}^{k} p_i$ é a probabilidade normalizada da localização i estar presente em uma requisição feita ao provedor do LBS.

$$H = -\sum_{i=1}^{k} q_i . \log_2 q_i \tag{3.2}$$

Neste caso, a entropia máxima possível é $H_{max} = \log_2 k$, quando todas as localizações selecionadas na requisição possuem exatamente a mesma probabilidade p.

Por fim, tem-se uma requisição com k localizações, cujas probabilidades são similares, tornando a localização real indistinguível das outras localizações em termo de probabilidade de estar presente em uma requisição.

É importante destacar que o DLS só procura anonimizar a informação de localização contida na requisição, não realizando qualquer tratamento no conteúdo de requisição. Outro ponto a destacar é a vulnerabilidade do DLS em um cenário de consultas contínuas, por não considerar critérios mais adequados para esse cenário, como a distância e a correlação entre as localizações.

Algoritmo 1: DLS

Entrada: k, m, L, l_r
Saída: R

Ordena L em ordem crescente de probabilidades;
$P \leftarrow$ *Seleciona* $2k$ *localizações cujas probabilidades são as mais similares de* l_r;
$M \leftarrow$ *Deriva-se* m;
para $i = 1 \rightarrow i = m$ **faça**
    $M_i \leftarrow l_r$;
    para $j = 1 \rightarrow j \leq k-1$ **faça**
        $M_i \leftarrow$ *Seleciona-se aleatoriamente uma localização* $l \in P$;
    fim
fim
$R \leftarrow$ *Conjunto M com maior entropia*;
retorna R

3.4.2 Dummy Location Privacy

O trabalho em (SUN *et al.*, 2017a), propõe um algoritmo de ataque, ADLS, capaz de identificar a localização real do usuário em uma requisição anonimizada por técnicas de seleção de localização falsas, e propõe uma nova estratégia de seleção, o *Dummy Location Privacy* (DLP) que, em comparação ao DLS, obtém uma maior proteção contra o algoritmo de ataque proposto.

3.4.2.1 ADLS

O objetivo do ADLS é identificar a localização real do usuário em uma requisição que foi anonimizada através da técnica de seleção de localização falsas que procura gerar localizações com probabilidade similar à da localização real. Mais especificamente, o ADLS procura explorar o processo de seleção do DLS.

O algoritmo do ADLS (2) recebe como parâmetros de entrada:

- k: grau de privacidade definido pelo usuário. É a quantidade de localizações que estarão presentes na requisição, que por consequência define também a máxima probabilidade de re-identificação desejada da localização real.
- L: a lista de localizações cobertas pelo provedor LBS, e suas respectivas informações de probabilidade.
- R: A requisição enviada pelo usuário.

Primeiramente, o ADLS identifica o grau de privacidade k em função da quantidade de localizações presentes na requisição. Assim, para a *i-esima*$(1 \leq i \leq k)$ localização presente na requisição, o algoritmo ADLS seleciona de forma gulosa outras $k-1$ localizações falsas

pertencentes ao conjunto de localizações cobertas pelo provedor do serviço, com base na entropia, e então obtém o conjunto de localizações falsas C_i. Para cada conjunto de localizações falsas C_i $(1 \leq i \leq k)$, o ADLS calcula a variância entre o conjunto C_i e a requisição do usuário, e determina a localização real. Assim, se a variância entre o conjunto C_i e a requisição é a menor entre todos os outros conjuntos, o ADLS infere que a localização real é a localização i.

Algoritmo 2: ADLS

Entrada: k, L, R
Saída: l

Ordena L em ordem crescente de probabilidades;
Ordena R em ordem crescente de probabilidades;
para $i = 1 \rightarrow i \leq k$ **faça**
    Conjunto $C_i \leftarrow$ *Seleciona uma localização* $l \in R$ *que não tenha sido selecionada antes*;
    Conjunto $D_i \leftarrow$ *Seleciona-se* $k-1$ *localizações imediatamente antes da localização* l *e* $k-1$ *imediatamente depois*;
    para $j = 1 \rightarrow j \leq k$ **faça**
        $p_{max} \leftarrow max(C_i)$;
        $p_{min} \leftarrow min(C_i)$;
        $p_{min-max} \leftarrow$ *localização presente em* D_i*, que tenha a maior probabilidade menor que* p_{min};
        $p_{max-min} \leftarrow$ *localização presente em* D_i*, que tenha a menor probabilidade maior que* p_{max};
        se $H(C_i, p_{max-min}) > H(C_i, p_{min-max})$ **então**
            $C_i \leftarrow p_{max-min}$; *Remove-se* $p_{max-min}$;
        senão
            $C_i \leftarrow p_{min_max}$; *Remove-se* $p_{min-max}$;
        fim
        ordena os elementos de C_i *em ordem crescente de probabilidades*;
        $S_i \leftarrow \sum_{j=1}^{k}(r_i - c_i)^2$;
    fim
fim
$l \leftarrow r_i \in R$, *onde* S_i *é o menor*
retorna l

3.4.2.2 DLP

Como uma alternativa ao DLS, o trabalho em (SUN *et al.*, 2017a) propõe o DLP que, assim como o DLS, utiliza a técnica de *dummy locations*, utilizando como critério de seleção a probabilidade das localizações estarem presentes nas consultas feitas ao LBS. O DLP busca alcançar o conjunto ótimo de localizações falsas ao utilizar uma abordagem gulosa de seleção,

onde são selecionadas $k-1$ localizações de forma sucessiva, buscando garantir que a entropia atual é a maior possível, ou seja, caso o algoritmo tenha já selecionado i localizações (i<k), quando for selecionar a *(i+1)-esima* localização, ele deve garantir que H_{i+1} é a maior para todas as localizações restantes. H_{i+1} é calculada segundo a equação de entropia 3.3, onde p_j é a probabilidade de uma requisição sobre a localização j.

$$H_{i+1} = -\sum_{j=1}^{i+1} \frac{p_j}{\sum_{l=1}^{i+1} p_l} \log_2 \frac{p_j}{\sum_{l=1}^{i+1} p_l} \tag{3.3}$$

O algoritmo do DLP (3) recebe como parâmetros de entrada:

- k: grau de privacidade definido pelo usuário. É a quantidade de localizações que estarão presentes na requisição, que por consequência define também a máxima probabilidade de re-identificação desejada da localização real.
- L: a lista de localizações cobertas pelo provedor LBS, e suas respectivas informações de probabilidade.
- l_r: localização real do usuário, onde $l_r \in L$.

Primeiramente o usuário define o grau de anonimização k. Um k maior implica um maior grau de anonimização, entretanto aumenta o *overhead* gerado em virtude do custo de selecionar as localizações falsas.

Para selecionar as localizações que garantem uma entropia ótima, o primeiro passo é selecionar aquelas que possuem uma probabilidade p_i de estarem na requisição, semelhante à da localização real (linha 2 do algoritmo). Denota-se por $\overline{k}$ a quantidade de localizações nesta condição. Se $\overline{k} \geq k$ então seleciona-se aleatoriamente $k-1$ localizações cujas probabilidades são iguais a da localização real (linha 5). Caso contrário, seleciona-se as $\overline{k}-1$ localizações cujas probabilidades são iguais às da localização real. As outras $k-\overline{k}$ localizações são selecionadas sucessivamente entre as localizações remanescentes que possuem maior similaridade de probabilidade da localização real, representadas por um conjunto S, cujo tamanho é $2(k-\overline{k})$ (linha 9). Esta seleção se dá de forma gulosa, onde em cada seleção é verificado se a entropia é a maior possível para todas as localizações residuais do conjunto S.

Assim como o DLS, o DLP só procura proteger a informação de localização da requisição, não aplicando qualquer processo de anonimização ao conteúdo de requisição. Da mesma forma, identificamos que o DLP é vulnerável em um cenário de consultas contínua, justamente por não considerar critérios de seleção mais apropriados, como a distância e a correlação entre as localizações.

Algoritmo 3: DLP

Entrada: k, L, l_r
Saída: R

Ordena L em ordem crescente de probabilidades;
$P \leftarrow$ *Seleciona as localizações em L que tenham a mesma probabilidades de* l_r;
$\overline{k} \leftarrow tamanho(P)$;
se $\overline{k} \geq k$ **então**
    $R \leftarrow$ *Seleciona-se aleatoriamente k localizações incluindo a localização real* l_r;
senão
    se $k/4 < \overline{k} < k$ **então**
        $R \leftarrow H$;
        $S \leftarrow$ *Seleciona-se* $2(k-\overline{k})$ *localizações cujas probabilidades são similares a de* l_r;
        $R \leftarrow l_r$;
        para $j = 1 \rightarrow j \leq k-\overline{k}$ **faça**
            Seleciona-se uma localização $l \in S$*, cuja entropia H(R,l) é a maior do conjunto S*;
            $R \leftarrow l$;
            Remove-se l de S;
        fim
    senão
        $S \leftarrow$ *seleciona-se* $4k-\omega-\varepsilon)$ *localizações cujas probabilidades são similares a de* l_r;
        Aleatoriamente seleciona-se uma localização $l \in S$;
        $R \leftarrow H + l$;
        para $j = 1 \rightarrow j \leq k-2$ **faça**
            Seleciona-se uma localização $l \in S$*, cuja entropia H(R,l) é a maior do conjunto S*;
            $R \leftarrow l$;
            Remove-se l de S;
        fim
    fim
fim
retorna R

3.5 Discussão

Neste trabalho, consideramos o cenário em que o próprio provedor do serviço é um potencial atacante, seja de forma direta, quando ele utiliza os dados coletados para proveito próprio, ou indireta, quando ele venha a vazar informações sensíveis dos usuários.

Desta forma, a utilização de uma estratégia distribuída, em que os próprios usuários são responsáveis pela anonimização de suas requisições e, ao mesmo tempo, garanta a qualidade do serviço, tanto em consultas simples como em consultas contínuas, são fundamentais para assegurar a garantia de privacidade dos usuários ao utilizarem serviços baseados em localização.

Em virtude destes fatores, analisamos profundamente a técnica de seleção de localizações falsas (*dummy locations*). Os algoritmos DLS e DLP desenvolveram uma técnica de seleção de localizações falsas que utiliza-se do conhecimento prévio sobre o conjunto de dados durante o processo de seleção das localizações falsas que irão compor a requisição, justamente para evitar que a real localização do usuário seja descoberta em virtude deste conhecimento adversário.

Apesar disto, os algoritmos de seleção do DLS e DLP apresentam falhas quando aplicados em um cenário onde são feitas consultas contínuas ao provedor de serviço. Isto acontece, principalmente, porque eles não consideram o deslocamento do usuário entre duas consultas enviadas em sequência ao provedor de serviço, ficando sujeitos a ataques que exploram esta falha.

Outro ponto importante que muitas vezes passa sem qualquer tratamento neste tipo de serviço é a proteção de privacidade sobre o conteúdo da requisição. A requisição é composta da localização do usuário e do conteúdo da requisição, que é a requisição propriamente dita do usuário, i.e. "Qual o cinema mais próximo?". Assim, caso não haja qualquer tipo de proteção a este conteúdo, será possível fazer inferências sobre o mesmo, expondo informações sensíveis sobre o usuário.

A importância de se antecipar possíveis conhecimentos que o atacante possa vir a ter sobre o conjunto de dados na utilização destas técnicas é, portanto, fundamental para a garantia do k-anonimato. Alguns conhecimentos que podem ser utilizados para identificar a localização real do usuário em uma requisição são: distância física entre as localizações, popularidade das localizações e correlação entre as localizações.

Quanto aos conhecimentos de **distância física**. Os ataques baseados em distância buscam recuperar informação da distância física entre as localizações, ou seja, se a distância entre

elas pode ser usada para inferir algum tipo de informação. Um exemplo para este tipo de ataque são localizações que se encontram em uma área com um certo nicho de empreendimento, i.e. uma área industrial, ou uma área com muitos hospitais e clínicas médicas. Como as localizações são sempre próximas a certos tipos de estabelecimentos, é possível traçar o perfil do usuário baseado nestas informações. Uma estratégia para evitar este tipo de inferência é a adição de ruído à localização de consulta do usuário, ou seja, diminuindo a precisão de localização do usuário, o que pode levar a uma resposta imprecisa do LBS, prejudicando a qualidade do serviço prestado. Por isto, a necessidade de se escolher a melhor estratégia de privacidade de acordo com o tipo de serviço desejado.

As técnicas que utilizam seleção de localizações falsas, em geral, estão protegidas neste cenário. Além de evitar selecionar localizações muito próximas entre si, a localização real do usuário está presente na requisição sem qualquer adição de ruído a ela, não afetando, portanto, a qualidade do serviço. Apesar disto, estas técnicas ainda podem ficar vulneráveis quando o usuário realiza consultas contínuas ao provedor LBS. Os algoritmos DLS e DLP têm como critério de seleção de localizações unicamente a probabilidade destas localizações. Assim, o atacante que tenha acesso a duas requisições feitas em sequência pelo usuário, poderá calcular a distância entre as localizações destas duas requisições e descartar aquelas que não estejam dentro de uma raio razoável, calculado pelo atacante em função da sua velocidade média e do tempo decorrido entre as duas requisições. Quanto mais preciso for o cálculo de distância entre as localizações, mais preciso é um ataque que procure explorar esta vulnerabilidade. Evitando, caso o cálculo de distância entre as localizações fossem mais precisos, descartar erroneamente, localizações que não estariam fora da área de alcançabilidade, ou deixar de descartar, localizações que estariam fora da área de alcançabilidade. Sendo assim, é fundamental para o atacante escolher a melhor forma de se estimar a distância entre as localizações: distância euclidiana, *manhattan*, ou de redes de rua.

Quanto aos conhecimentos de **popularidade**. Os ataques de popularidade buscam identificar a localização do usuário baseado na probabilidade do usuário estar em uma determinada localização. Como sabemos, certas localizações podem ser bem mais frequentadas, ou que geram maior interesse do que outras. Por exemplo, um *shopping center* é bem mais frequentado do que um cemitério. Assim, baseado neste conhecimento, é possível identificar a localização do usuário. Uma requisição anonimizada através da técnica de seleção de localizações falsas, cujas localizações foram escolhidas de forma aleatória, seleciona localizações com diversas

popularidades, não garantindo, portanto, que a probabilidade de se identificar a localização real do usuário entre as localizações presentes na requisição seja inferior a $\frac{1}{k}$, onde k é a quantidade de localizações presentes na requisição.

Já sobre conhecimento de **correlação**. Um outro tipo de ataque é o de correlação entre as localizações. Uma localização possui uma série de atributos, tais como posição geográfica, tipo de localização (residencial, comercial, clínica, etc), popularidade, entre outros. Assim a correlação entre as localizações pode revelar mais sobre as mesmas. Por exemplo, em uma consulta que o usuário requisita uma localização, ao fazê-la, ele está mostrando seu interesse sobre a mesma. Sendo assim, pode existir uma correlação direta entre a localização requisitada e as próximas localizações do usuário. Em uma requisição, esta informação pode ser útil ao atacante para revelar a localização real do usuário.

Em uma consulta realizada usando um algoritmo de seleção de localizações falsas, o atacante pode usar a correlação das localizações presentes nas repostas enviadas ao usuário na consulta anterior e as atuais localizações presentes na nova requisição para identificar se o usuário está se deslocando, ou se deslocou em direção a uma das localizações presentes nas respostas e, assim, identificar a localização real do usuário.

A Tabela 2 contém uma comparação entre os algoritmos de anonimização *DLS* e *DLP*, que utilizam a técnica de seleção de localizações falsas, juntamente com duas versões do PrivLBS, desenvolvidas como fruto desse trabalho. Comparamos a ausência ou presença de proteção ao conteúdo de requisição, bem como o uso dos conhecimentos adversários já citados nesta seção como critério de seleção das localizações falsas.

O DLS e DLP não possuem qualquer mecanismo de proteção ao conteúdo da requisição. Os dois trabalhos apenas procuram proteger a informação de localização, ou seja a localização do usuário enviada na consulta. Com este objetivo, os algoritmos utilizam somente a popularidade das localizações, medida em razão da probabilidade destas localizações estarem presentes na requisição, como critério de seleção. Desta forma, não sendo capazes de proteger a localização em um cenário de consultas contínuas.

A primeira versão do PrivLBS, que chamamos de PrivLBSv1, assim como o DLS e DLP, não protege o conteúdo da requisição, enviando apenas um conteúdo por requisição. Ele utiliza como critérios de seleção, a popularidade e a distância entre as localizações. Garantindo uma proteção parcial à localização do usuário. Já a versão final do PrivLBS protege o conteúdo de requisição, ao gerar um conteúdo de requisição para cada localização falsa presente na

Abordagens	Popularidade	Distância	Correlação	Proteção ao conteúdo
DLS	✓			
DLP	✓			
PrivLBSv1	✓	✓		
PrivLBS	✓	✓	✓	✓

Tabela 2 – Comparação entre as técnicas de seleção de localizações falsas.

requisição. Ele utiliza como critérios de seleção, a popularidade, a distância e a correlação entre as localizações, garantindo uma maior proteção contra ataques de conhecimento.

3.6 Conclusão

Neste capítulo, apresentamos as abordagens mais relevantes em privacidade de dados de localização. Destacamos a necessidade de se preservar a privacidade do indivíduo sem que isto afete a qualidade do serviço prestado. Desta forma, considerando que o provedor de serviço não é confiável, propomos uma nova estratégia de anonimização sem perda de utilidade dos dados, onde procuramos proteger não apenas a localização real do usuário enviada no momento da requisição, como o conteúdo da requisição, impedindo a realização de inferências sobre a requisição como um todo.

Utilizamos a técnica de seleção de localizações falsas com a adição de conhecimentos adversários no processo de seleção, a fim de proteger a privacidade de localização do usuário contra ataques de conhecimento. Para isto, utilizamos três critérios de seleção das localizações. A popularidade das localizações. A distância percorrida pelo usuário entre duas consultas consecutivas. E por último, a correlação entre as localizações presentes na requisição a ser realizada e as localizações presentes na resposta à consulta anterior realizada pelo usuário. Além disto, com o objetivo de proteger o conteúdo da requisição, procuramos gerar falsos conteúdos de requisição para cada localização falsa selecionada.

4 PRIVACIDADE DE INDIVÍDUOS EM LBS

Neste capítulo apresentaremos nossa solução para o problema de garantia de privacidade de indivíduos em serviços de localização. Nosso objetivo é garantir que os dados sensíveis de localização dos indivíduos não sejam expostos. Apresentaremos também o ACon, um algoritmo de ataque de conhecimento em LBS que procura identificar a localização real do usuário em requisições que utilizam a técnica de seleção de localizações falsas.

Nossa abordagem, chamada de PrivLBS, faz uso de uma técnica de ofuscação com seleção de localizações falsas, preservando a privacidade de localização do usuário sem perda de utilidade dos dados, permitindo portanto que o LBS possa responder precisamente à consulta do usuário apesar da garantia de privacidade. O grau de privacidade é medido pela garantia de k-anonimato e a alta entropia alcançada. Primeiramente iremos descrever o modelo de LBS adotado, definindo a função de cada entidade que participa do processo.

4.1 Modelo de sistema do LBS

A Figura 5 ilustra a estrutura do LBS adotado. Ela contém os seguintes componentes:

- **Sistema de posicionamento:** um sistema de posicionamento que permita determinar a localização dos objetos, como por exemplo GPS.
- **Usuários:** os usuários são os clientes do modelo que irão fazer uso do serviço prestado pelos LBSs através de dispositivos móveis, tais como *smartphones*, dispositivos *wearables* ou veiculares.
- **Rede:** meio através do qual acontece o tráfego de informações entre os participantes do modelo. Os dados são usualmente transmitidos através da Internet.
- **POI:** um ponto de interesse é uma localização cuja informação está disponível. Em geral, as informações sobre os POIs vão desde suas coordenadas geo-espaciais, latitude (latitude (lat)) e longitude (longitude (lon)), bem como o horário de funcionamento, ou o tipo de localização, *i.e.* hospital, supermercado, escola, etc. O termo localização e POI neste trabalho é permutável.
- **Provedor do LBS:** é o responsável por prestar o serviço aos usuários, respondendo suas requisições. Ele contém informações sobre os POIs, armazenadas na tabela de localizações.
- **Tabela de localizações:** é uma lista de POIs contendo suas informações de loca-

lização. São informações armazenadas pela tabela de localização: a popularidade de cada POI, medido pela equação 3.1, o tipo de localização, e a lista de vizinhos dos POIs, contendo a menor distância até ele. Em nosso trabalho consideramos duas localizações vizinhas, quando existe pelo menos um caminho que permita um deslocamento entre estas localizações, conforme Definição 3. Essa tabela é armazenada pelo provedor do LBS e disponibilizada para os usuários.

Definição 3 *Uma localização l_2 é dita vizinha de uma localização l_1 quando existe pelo menos uma rota $l_1 \rightarrow l_2$, passando ou não por outras localizações.*

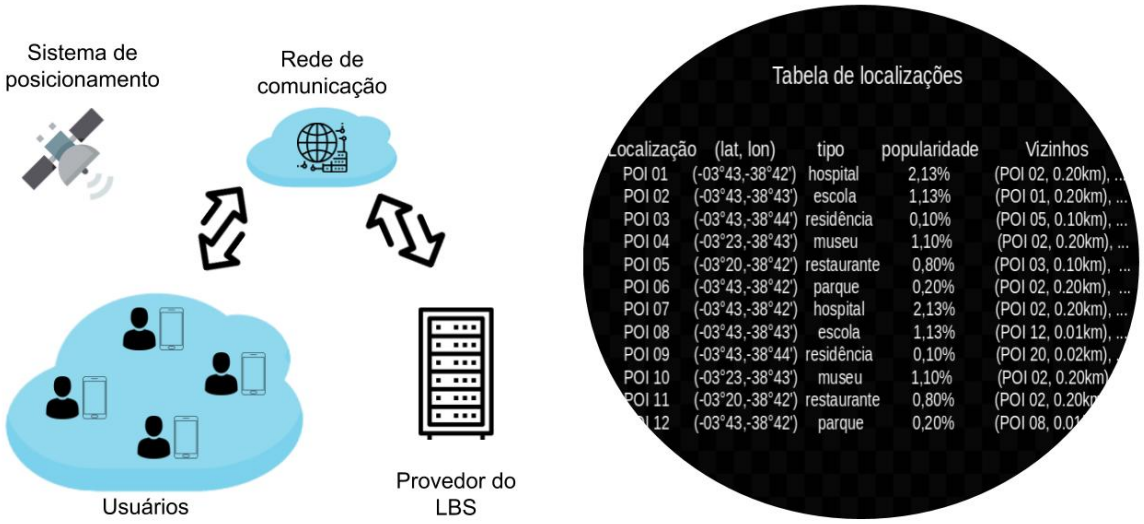

Figura 5 – Modelo de sistema do LBS.

O provedor do LBS é responsável por coletar dados estatísticos sobre os POIs e calcular a probabilidade de cada POI estar presente em uma requisição feita a ele. Este cálculo é feito através da Equação 3.1. Esta informação é adicionada às informações básicas de cada POI, tais como coordenadas e tipo, para formar a tabela de localizações, contendo assim informações complementares sobre cada localização presente. Dessa forma uma localização l no nosso modelo é definida pela tripla $l = \{coordenadas,\ tipo,\ probabilidade\}$. Por fim, a tabela de localização contém para cada localização a lista de vizinhos e a distância para seus vizinhos. Esta distância diz respeito a distância física entre as localizações, *i.e.* distância euclidiana, *manhatan*, rede de ruas, ou outra métrica.

A Tabela de localizações disponibilizada pelo provedor do serviço é ponto chave no processo de anonimização. Ela garante uma transparência do serviço, permitindo que o usuário tenha o mesmo tipo de conhecimento que o próprio provedor possui sobre os dados armazenados. Sendo assim, permitindo que estratégias sejam criadas para garantir a privacidade dos indivíduos que utilizem estes serviços. Em nosso modelo, o usuário poderá enviar em cada requisição, até k pares (localização, conteúdo de requisição), onde a localização é um POI passado como referência, e o conteúdo de requisição é a consulta propriamente dita que será feita ao provedor LBS, como por exemplo, "O hospital mais próximo do POI de referência".

Desta forma, uma requisição na verdade é um conjunto de k consultas, tendo cada uma, uma localização de referência e um conteúdo de requisição. O provedor ao receber a requisição, envia k respostas, uma para cada um dos pares (localização, conteúdo de requisição) presentes na requisição. Podemos observar este fluxo na figura 6.

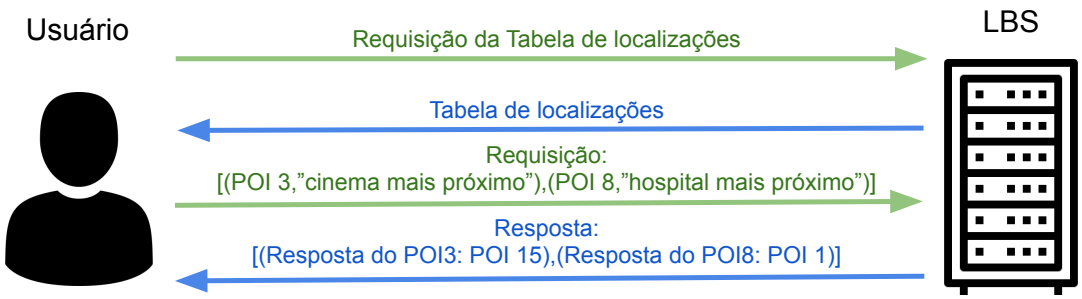

Figura 6 – Fluxo de informações do PrivLBS.

O processo de anonimização é essencial para a garantia de privacidade tanto da localização, como do conteúdo presente na requisição. Desta forma, o PrivLBS procura utilizar em seu favor o conhecimento disponibilizado pelo provedor do serviço sobre as localizações, garantindo uma requisição anônima contra ataques de conhecimento adversário. Com o nosso algoritmo de ataque, ACon, procuramos demonstrar a importância de se proteger contra estes conhecimentos que estão disponíveis, apontando a fragilidade dos principais trabalhos anteriores que utilizam a técnica de seleção de localizações falsas, e demonstrando a garantia de privacidade do PrivLBS contra este modelo de ataque.

4.2 Ataque de conhecimento em LBS (ACon)

Nesta seção iremos apresentar o ACon, nosso ataque de conhecimento, que foi proposto com o objetivo de identificar a localização real do usuário em requisições feitas ao LBS.

Nosso algoritmo de ataque busca identificar a localização real do usuário explorando o conhecimento adquirido sobre as localizações presentes na requisição. Ele é um algoritmo específico para requisições feitas ao provedor LBS que utiliza a técnica de seleção de localizações falsas, *Dummy Locations*. O ACon explora os seguintes conhecimentos adversários:

- **Histórico de requisição:** o provedor do LBS é a parte responsável por receber e responder às requisições feitas pelos usuários, logo é de seu conhecimento o teor destas, sendo capaz portanto de manter um histórico sobre as mesmas. Isto permite que ele possa utilizar desse conhecimento em seu favor e assim identificar a localização real do usuário.
- **Distância:** a distância entre as localizações é outra propriedade de fácil conhecimento, tanto do provedor do serviço, como de outros possíveis atacantes. Embora o conhecimento da distância entre as localizações sozinho possa não vir a ser uma ameaça à violação de privacidade, quando combinado a outros conhecimentos, como o histórico de requisições de um usuário, pode tornar-se a ser. E é justamente combinado ao histórico de requisição do usuário que o ACon age. Nosso algoritmo de ataque toma como referência a requisição anterior mais recente do usuário e procura eliminar as localizações falsas da nova requisição que não são alcançáveis por nenhuma das localizações desta requisição anterior, baseado numa velocidade média a ser estimada pelo atacante e o intervalo de tempo entre esta consulta anterior e a atual requisição. Sendo assim, a velocidade média estimada é um ponto extremamente sensível, uma vez que ela vai ter um impacto direto na capacidade do ataque em identificar as localizações alcançáveis. Entretanto, não faz parte do escopo deste trabalho o estudo sobre a estimativa da velocidade média dos usuários.
- **Correlação:** nosso algoritmo de ataque procura identificar a localização dentro da requisição que tenha a maior correlação com as localizações dos POIs presentes na resposta da requisição anterior enviada ao provedor LBS. O provedor do LBS tem acesso às consultas anteriores do usuário e a resposta enviadas

a estas requisições, assim ele é capaz de medir o grau de correlação entre as localizações presentes na requisição e as presentes na resposta enviada ao usuário na requisição anterior, aumentando bastante a probabilidade de identificação da localização real.

- **Popularidade:** a popularidade diz respeito à probabilidade de uma localização estar presente na requisição. Este conhecimento pode ser facilmente adquirido pelo provedor do LBS através da equação 3.1, que diz respeito a razão entre a soma de requisições feitas sobre uma dada localização e o total de requisições já feitas ao provedor. O ACon procura identificar entre as localizações da requisição que ainda são consideradas possíveis localizações reais, aquela com maior probabilidade de realmente ser a real.

O ACon procura explorar estes conhecimentos a fim de identificar a localização real do usuário entre as k localizações presentes na requisição.

O primeiro passo do LBS é descartar as localizações que não são alcançáveis por nenhuma localização da requisição anterior. Assim, o atacante primeiramente deve estimar a velocidade média do usuário. Só então é possível calcular a distância máxima de deslocamento do usuário de acordo com a equação $max_{\Delta s} = v \times \Delta t$, onde v é a velocidade média estimada do usuário e Δt é o intervalo de tempo entre a consulta anterior e a atual. Sendo assim, as localizações da atual requisição que estiverem fora da área de alcançabilidade de todas as localizações da consulta anterior são descartadas.

Em seguida, verifica-se qual das localizações remanescentes da atual requisição possui a maior correlação com as localizações presentes na resposta à requisição anterior feita pelo usuário. Entre as localizações com maior grau de correlação, o algoritmo identifica como localização real, aquela com maior probabilidade de estar presente em uma requisição.

A Figura 7 ilustra o funcionamento do nosso algoritmo de ataque, onde o provedor do serviço é responsável por guardar o histórico de requisições dos usuários, contendo não apenas as requisições em si, mas também as respostas que foram enviadas. Assim, ao receber uma requisição, o primeiro passo é verificar que localizações presentes na consulta são alcançáveis a partir de cada uma das localizações de consulta presentes na requisição anterior. Na figura, a localização l_4 é descartada, justamente por estar a uma distância inalcançável para qualquer das localizações da consulta anterior (localizações em cinza). Dentre as localizações que são alcançáveis é verificada o grau de correlação delas com as localizações presentes nas respostas

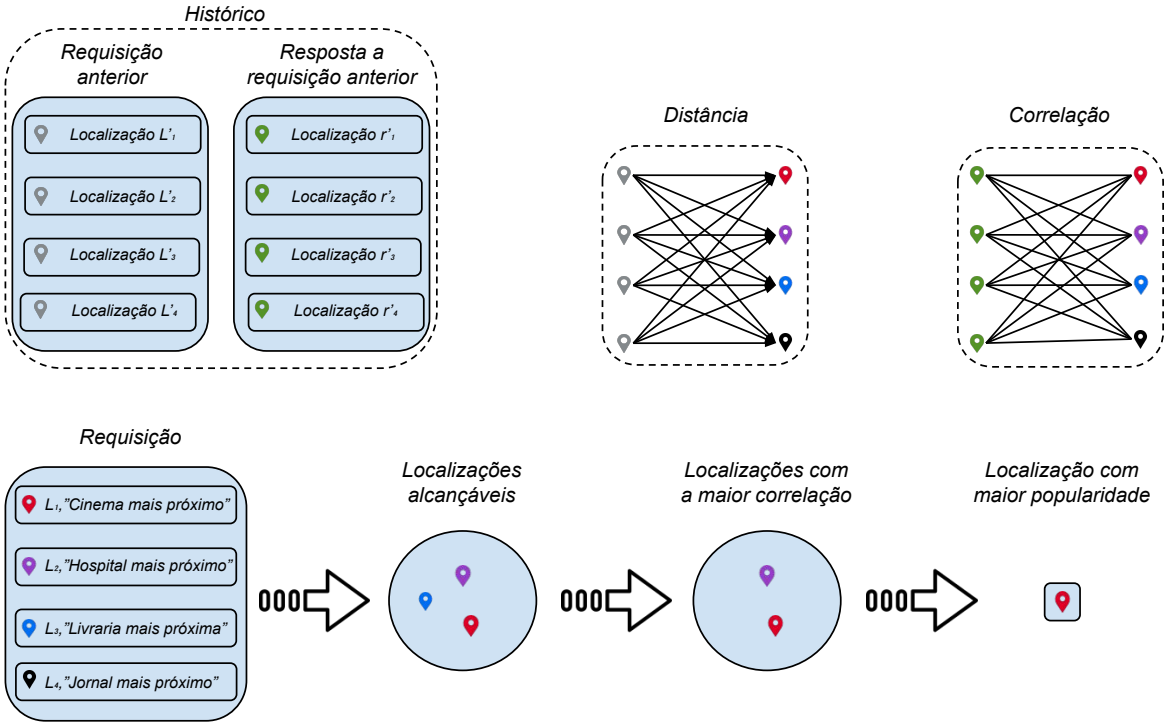

Figura 7 – Fluxo de execução do ACon

da consulta anterior (localizações em verde), e separadas, justamente, aquelas com o maior grau de correlação. Na figura, as localizações com maior correlação com as localizações de resposta foram as localizações l_1 e l_2. Dentre essas localizações, o ACon identifica como localização real, a localização com maior popularidade, ou seja, com maior probabilidade de estar presente na requisição. Na figura, a localização selecionada foi l_1, em vermelho.

O ACon recebe como parâmetros de entrada:

- R: requisição atual recebida pelo provedor do serviço.
- $limite$: refere-se à máxima distância alcançável pelo usuário, calculada baseada através da fórmula $max_{\Delta s} = v \times \Delta t$, onde v é uma estimativa da velocidade média do usuário e Δt é o intervalo de tempo entre a consulta anterior e a atual.
- L: a lista de localizações cobertas pelo provedor LBS, e suas respectivas informações complementares, *i.e.* popularidade, e lista de vizinhos com suas respectivas distâncias até a localização.
- R': última requisição enviada ao provedor do serviço e sua respectiva resposta.

Caso R' esteja vazio (linha 1), ou seja não há registro de requisição anterior do usuário, o algoritmo de ataque irá identificar como localização real a localização $l_i \in R$ cuja probabilidade de estar em uma requisição feita ao LBS é a maior dentre as localizações em R.

Se houver uma requisição anterior feita pelo usuário, R' não está vazio (linha 3), o

algoritmo irá para cada localização $l'_i \in R'$, identificar que localizações $l_i \in R$ estão dentro de suas áreas de alcançabilidade, ou seja estão a uma distância inferior ao parâmetro de entrada $limite$ (linha 6). Estas localizações são consideradas localizações candidatas à localização real do usuário e são adicionadas ao conjunto de localizações alcançáveis D (linha 7). As outras localizações são descartadas.

Dentre as localizações remanescentes, verifica-se qual delas possui a maior correlação com a resposta à consulta anterior R' (linha 14). Se houver mais de uma localização com maior correlação, é escolhida como localização real, aquela com maior popularidade (linha 19).

Algoritmo 4: ACon

```
   Entrada: R' , R, limite, L
   Saída: l_r
 1 se R' == ∅ então
 2 |  l_r ← Localização l ∈ L com maior probabilidade p_l;
 3 senão
 4 |  para cada location l ∈ R faça
 5 |  |  para cada location l' ∈ R' faça
 6 |  |  |  D ← r;
 7 |  |  fim
 8 |  fim
 9 |  Res ← localizações presentes na resposta à requisição R';
10 |  para cada location l ∈ D faça
11 |  |  para cada location l' ∈ Res faça
12 |  |  |  se c_{l',l} == maior correlação encontrada entre l e l' então
13 |  |  |  |  C ← l;
14 |  |  |  fim
15 |  |  fim
16 |  fim
17 |  l_r ← Localização l ∈ C com a maior probabilidade p_l;
18 fim
19 retorna l_r
```

4.3 PrivLBS

Nesta seção iremos analisar detalhadamente o PrivLBS, nossa abordagem de anonimização em serviços baseados em localização, demonstrando como ele garante a privacidade de localização do usuário ao usar esse tipo de serviço.

O PrivLBS é um algoritmo de preservação de privacidade que utiliza a técnica de seleção de localizações falsas, procurando ofuscar a localização real do usuário dentre as outras $k-1$ localizações falsas presentes na requisição. A ideia central do PrivLBS é garantir que o

provedor do serviço, mesmo possuindo conhecimento das requisições anteriores enviadas pelo usuário e sobre informações complementares das localizações, tais como a probabilidade da localização estar presente numa requisição, ou os tipos destas localizações, ao receber uma nova requisição, não consiga distinguir uma localização das outras $k-1$ localizações presentes na requisição, não permitindo, portanto, que a localização real seja identificada. Entretanto, o PrivLBS não procura proteger apenas as informações de localização do usuário, ele procura proteger também o conteúdo da requisição, gerando uma requisição contendo até k conteúdos de requisição diferentes, aumentando a garantia de privacidade do usuário.

4.3.1 Critérios de seleção

Um desafio dos modelos de privacidade é como proteger a requisição anonimizada de possíveis conhecimentos adversários, ou seja, conhecimentos prévios que um atacante possa usar para expor a privacidade de dados dos usuários. Desta forma, considerando o provedor do LBS um potencial atacante, identificamos três conhecimentos adversários disponíveis em um típico serviço de localização e que portanto devem ser considerados no processo de seleção das localizações falsas: a distância, a probabilidade, e a correlação entre as localizações.

4.3.1.1 Distância

A distância entre as localizações é essencial para garantir a privacidade do usuário, especialmente quando consideramos o cenário de consultas contínuas, onde o usuário realiza várias requisições em curtos intervalos de tempo. Nestas situações, observando a distância entre as localizações de requisições consecutivas podemos identificar quais deslocamentos são possíveis e quais não são. A Figura 8 ilustra essa situação. No primeiro quadro observamos as localizações enviadas pelo usuário em uma requisição, estando em azul a localização real e em vermelho as localizações falsas. No momento seguinte, o usuário se desloca para a nova localização real, conforme o quadro seguinte, onde temos em cinza as localizações da consulta anterior e em azul a nova localização real. Podemos observar ainda em tracejado a área máxima de deslocamento do usuário para cada uma das localizações presentes no momento da consulta anterior, caso o usuário estivesse em qualquer destas localizações, considerando o intervalo de tempo decorrido entre estes dois momentos e a velocidade do usuário. O processo de anonimização, no momento da seleção das novas localizações falsas, poderá ou não considerar esta distância. O terceiro quadro da figura mostra justamente a situação onde a distância não é

critério de seleção das novas localizações falsas. Como podemos observar, apenas a localização real se encontra dentro de uma destas áreas de alcançabilidade. Ficando totalmente exposta, uma vez que as outras localizações falsas não apresentam um deslocamento plausível, podendo portanto serem desconsideradas.

Desta forma, o PrivLBS toma como referência as localizações da consulta anterior no momento da seleção das novas localizações, cujo objetivo é a escolha de localizações que sejam alcançáveis por pelo menos uma localização da consulta anterior. Esta ideia de alcançabilidade é calculada em função da velocidade média do usuário e do intervalo de tempo entre duas requisições feitas ao provedor, através da equação 4.1, onde calculamos a máxima distância que o usuário é capaz de percorrer. De posse desta informação são escolhidas localizações que estejam dentro desse raio de alcance.

$$max_{\Delta_s} = v \times \Delta t \tag{4.1}$$

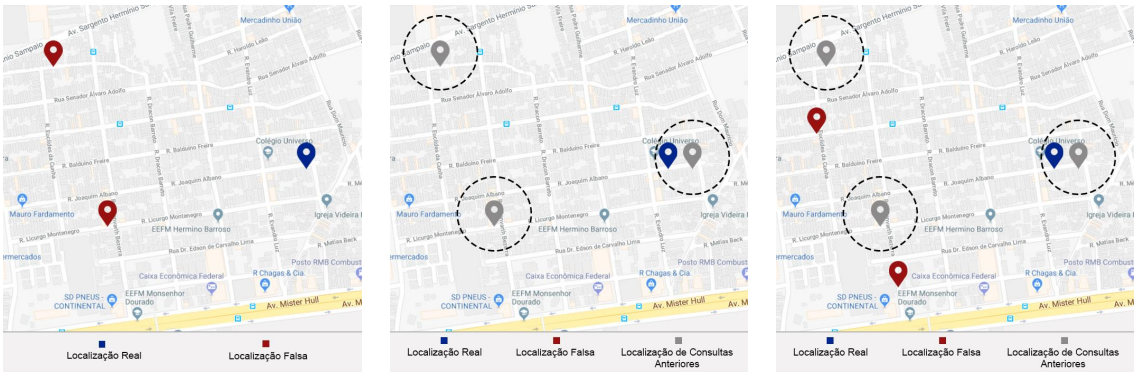

Figura 8 – Distância entre as localizações em consultas contínuas.

Embora a escolha de localizações alcançáveis seja fundamental durante o processo de seleção, eventualmente, este conjunto pode vir a ser muito pequeno, prejudicando a anonimização através dos outros critérios de seleção. Duas medidas foram tomadas para contornar este problema.

A primeira é uma medida de prevenção. Considerando que as localizações selecionadas para compor a requisição atual serão usadas como referências na anonimização de requisições futuras. Para evitar situações excepcionais em que estas localizações escolhidas possuam uma quantidade pequena de localizações que estejam dentro de suas áreas de alcançabilidade, o PrivLBS procura evitar a escolha de localizações falsas que possuam uma quantidade pequena de vizinhos. Esta situação só é permitida, quando a localização real também possui uma quantidade

limitada de vizinhos. Desta forma, como o objetivo é sempre buscar um comportamento de escolha de localizações similar ao da localização real, caso esta localização possua uma quantidade limitada de vizinhos, será admitido a escolha de localizações falsas que também tenham esta condição. Através da taxa de isolamento identificamos se uma localização possui uma quantidade limitada de vizinhos. Ela mede o grau de isolamento da localização em função de sua quantidade de vizinhos e da quantidade mínima de vizinhos considerada adequada para garantir uma seleção de localização falsa que preserve a privacidade do usuário.

A segunda medida é expandir de forma controlada a área de alcançabilidade durante o processo de seleção de novas localizações falsas quando a quantidade de localizações alcançáveis não atingir uma quantidade mínima. Esta quantidade mínima é definida em função dos requerimentos de privacidade desejados.

4.3.1.2 Probabilidade

Como já explicado na Seção 3, a popularidade, calculada em razão da probabilidade de uma localização estar presente em uma requisição feita ao provedor LBS é um conhecimento que pode expor a privacidade do usuário. Considerando o cenário em que o próprio provedor é o atacante, a importância de se proteger contra ataques que procuram explorar este tipo de conhecimento ganha mais destaque, uma vez que ele tem acesso direto às requisições feitas sobre cada localização.

O PrivLBS como forma de evitar um ataque que procure explorar a popularidade das localizações presentes na requisição, procura selecionar localizações que tenham probabilidades semelhante de estar na requisição. Para isto é verificada a probabilidade de a localização real estar em uma requisição, e então é escolhida outras $k-1$ localizações cujas probabilidades mais se aproximam da real. Ao final, a requisição terá localizações cujas probabilidades de estarem na requisição são semelhante entre si.

4.3.1.3 Correlação

O último conhecimento adversário que devemos considerar no processo de seleção é a correlação entre as localizações. Como falamos na seção 2.3, os dados de localização são extremamente correlacionados. Em serviços baseados em localização, uma requisição feita sobre uma localização específica pode revelar muito sobre o usuário, especialmente se analisarmos a interação deste usuário com a resposta enviada pelo provedor.

Assim, ao realizar consultas contínuas a um provedor LBS e receber respostas a estas requisições, podemos identificar dois comportamentos distintos do usuário. Ele pode se deslocar em direção a esta localização presente na resposta à requisição ou tomar um deslocamento que não tenha muita relação com esta resposta. Uma forma de se identificar este comportamento do usuário é através do cálculo de correlação entre a localização real atual e a localização de resposta da requisição anterior. Como medimos a correlação entre duas localizações em razão de suas coordenadas, da popularidade e do tipo destas localizações, uma correlação alta indica que o usuário está se deslocando a uma localização com atributos semelhantes. Já uma correlação baixa indica o oposto, que o usuário não está se deslocando para uma localização com atributos semelhantes, ou seja esta se deslocando em outra direção.

Como uma forma de tentar capturar este comportamento do usuário e reproduzir na seleção de localizações falsas procuramos utilizar a correlação entre as localizações como critério de seleção de localizações falsas. Dessa forma, o PrivLBS primeiramente identifica o grau de correlação entre a localização real atual e a localização dada como resposta pelo provedor LBS à localização real da requisição anterior e passa a escolher localizações falsas que também possuam uma correlação com as respostas às localizações falsas da consulta anterior semelhante.

A Figura 9 ilustra a situação. As localizações r_1, r_2 e r_3 na figura são as localizações dos POIs requisitados tendo como localizações de consulta l'_1, l'_2 e l'_3 respectivamente. No momento seguinte o usuário irá realizar uma nova requisição, estando na localização l_2. Como l_2 e r_2 são próximos, a correlação entre elas deve ser alta, por apresentarem atributos semelhantes, especialmente em termos de coordenadas. Espera-se, portanto, que as correlações de l_1 e l_3, com r_1 e r_3 respectivamente, também sejam altas. Assim, todas as localizações falsas escolhidas demonstram um possível comportamento de deslocamento do usuário semelhante ao real.

Calculamos o grau de correlação através do coeficiente de *Spearman*, por ser mais adequado à natureza dos atributos dos dados de localização que são não lineares, conforme já discutido na Seção 2.3.1. Os valores de correlação são medidos por uma função monotônica arbitrária, cujos valores variam entre -1 indicando uma baixa correlação e $+1$ indicando uma alta correlação.

4.3.2 Processo de anonimização

O PrivLBS combina estes três critérios, distância, probabilidade e correlação, durante o processo de anonimização a fim de garantir a privacidade da localização real de con-

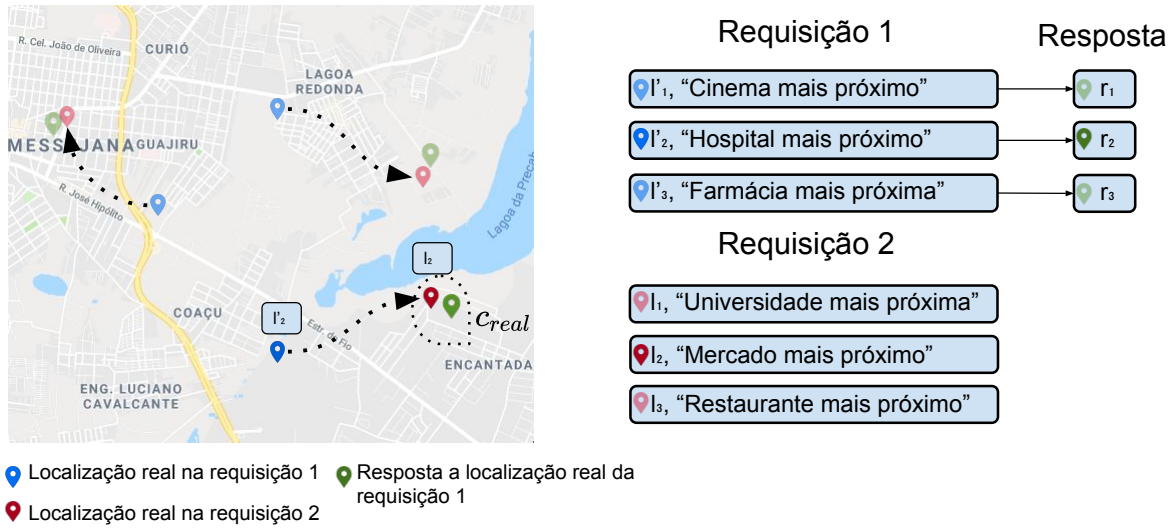

Figura 9 – Correlação como critério de seleção do PrivLBS.

sulta do usuário. Entretanto, ainda é necessário anonimizar o conteúdo de requisição, ou seja a requisição pela localização de um POI, tendo como referência a localização de consulta informada. Nos modelos anteriores de seleção de localizações falsas, em uma requisição $req = \{(l_1, l_2, ..., l_k), \textit{conteúdo de requisição}\}$, o conteúdo de requisição é único, apesar de haver várias localizações de consulta enviadas na requisição. O problema disso, é que o conteúdo de requisição está em geral associado a um tipo, assim, uma consulta realizada sem a aplicação de qualquer método de anonimização sobre este conteúdo irá permitir que sejam feitas inferências sobre este tipo, permitindo que sejam extraídos dados sensíveis do usuário. Como forma de proteger o conteúdo de requisição, o PrivLBS procura anonimizar este tipo, impedindo que o atacante identifique o conteúdo de requisição real. Assim, para cada localização falsa presente na informação de localização, será selecionado um tipo a ser adicionado ao conteúdo de requisição.

Podemos dividir o processo de anonimização do PrivLBS em três fases:

Na **primeira fase**, Figura 10, para cada localização falsa da requisição anterior enviada ao provedor do serviço, são construídos subconjuntos com localizações que estão dentro do raio de distância máxima que o usuário é capaz de percorrer no intervalo de tempo entre as consultas. Em caso de não haver uma requisição anterior a escolha da localização falsa é baseada na popularidade das localizações, procurando escolher aquelas com probabilidade de estar presentes em uma requisição, semelhante à probabilidade da localização real. Neste caso, após a seleção dessas localizações falsas, segue-se direto à fase 3 do processo de anonimização.

Na **segunda fase**, Figura 11, procura-se identificar o comportamento de deslocamento do usuário durante o intervalo de tempo entre a última requisição feita por ele e o momento

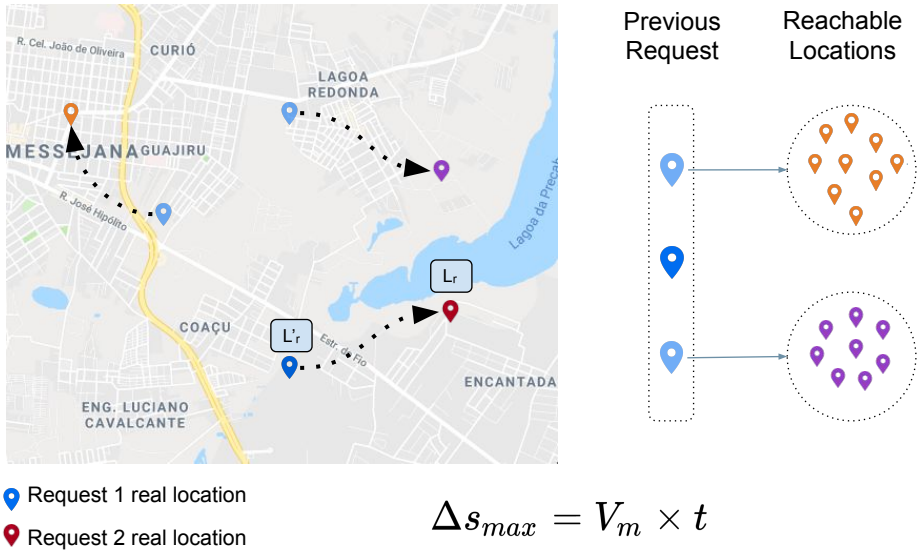

Figura 10 – Primeira fase do PrivLBS.

atual, com o objetivo de utilizar este conhecimento na seleção de localizações falsas. Desta forma, dada a localização de resposta que foi enviada para o usuário em razão da última requisição que ele fez ao provedor do serviço, qual o comportamento do usuário? Ele está se deslocando para esta localização? Ele está nesta localização? Ou ele está se deslocando para uma localização qualquer? Com esta finalidade através do coeficiente de *spearman* calculamos o grau de correlação real, c_{real}, entre a localização enviada em resposta a localização real anterior do usuário e a sua atual localização. Após o cálculo de c_{real}, se inicia o processo de seleção das localizações falsas que serão adicionadas à requisição. Na primeira fase, para cada umas das $k-1$ localizações falsas l' da requisição anterior foi gerado um conjunto de localizações que são alcançáveis a partir de l'. Destes $k-1$ conjuntos é selecionado uma localização falsa l. Esta localização falsa é escolhida em função da sua correlação com a localização da resposta que foi enviada pelo provedor de serviço em resposta à localização falsa l' passada como referência na requisição anterior. Será escolhida, portanto, a localização l cuja correlação com a localização de resposta a l' é semelhante a c_{real}. Ao final são escolhidas $k-1$ localizações falsas alcançáveis por pelo menos uma localização da consulta anterior e que apresentam um comportamento de deslocamento com as localizações falsas anteriores, semelhante ao comportamento entre a localização real anterior e a atual localização real do usuário.

Na **terceira fase**, Figura 12, procura-se anonimizar o conteúdo da requisição. Para cada localização falsa selecionada na nova requisição, é definido um tipo a ser associado ao conteúdo da requisição que será anexada à localização falsa. Para atingir uma diversidade dos

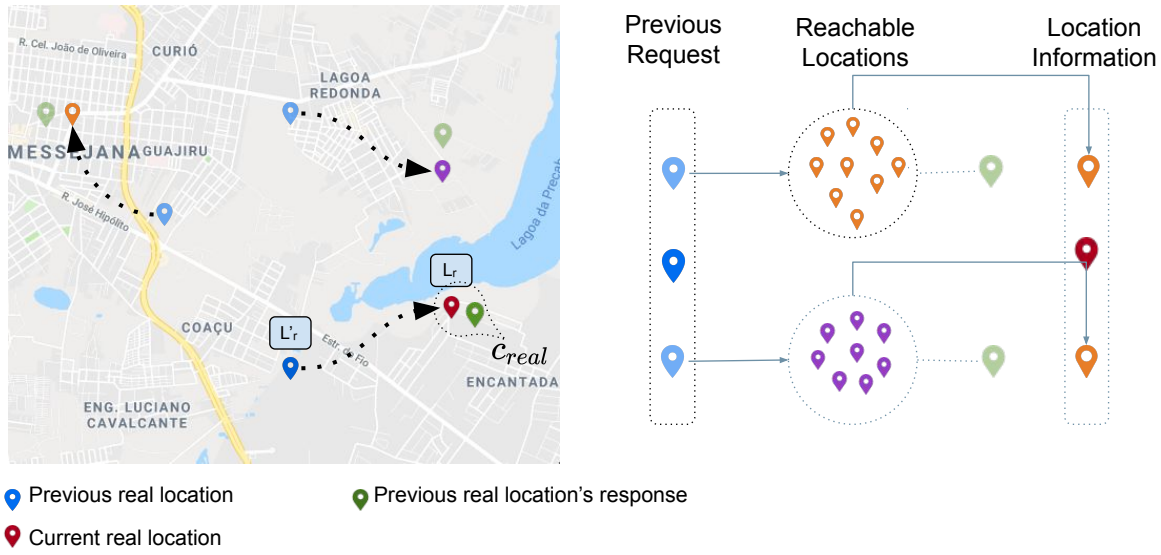

Figura 11 – Segunda fase do PrivLBS.

conteúdos de requisição que garanta a privacidade do usuário, o tipo a ser associado respeita a distribuição do conjunto de dados. Assim, se o tipo "Restaurante"estiver presente em 5% das localizações do conjunto de localizações, este mesmo tipo terá uma probabilidade de ser selecionado para compor a requisição de 5%.

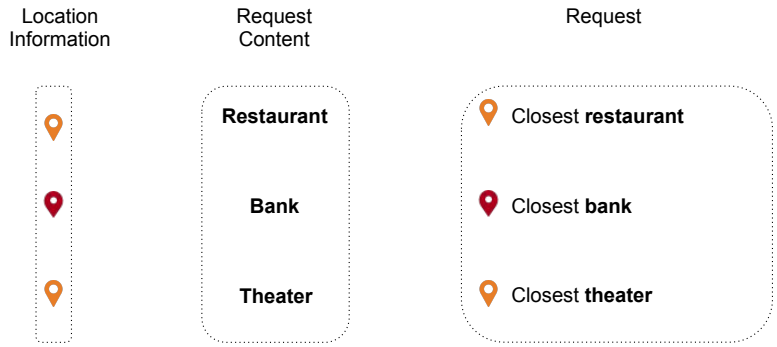

Figura 12 – Terceira fase do PrivLBS.

A Figura 13 representa fluxo completo do processo de anonimização do PrivLBS. Nela podemos observar cada uma das fases do PrivLBS. Na fase 1 é tratado o critério de distância entre as localizações. Na segunda fase é tratado de forma direta a correlação e de forma indireta a popularidade, por ser um dos atributos das localizações, assim como a distância e o tipo. Por fim, na terceira fase é tratado a anonimização do conteúdo de requisição.

O Algoritmo do PrivLBS (5) recebe como parâmetros de entrada:

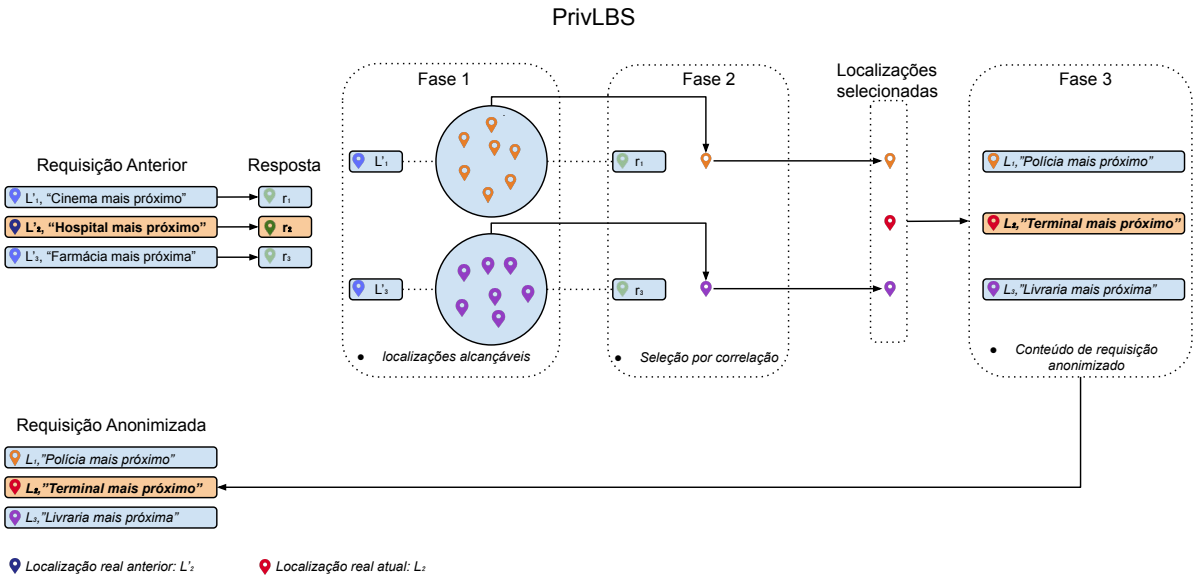

Figura 13 – As três fases de anonimização do PrivLBS.

- k: grau de privacidade definido pelo usuário. É a quantidade de localizações que estarão presentes na requisição, que por consequência define também a máxima probabilidade de re-identificação desejada da localização real.
- L: a lista de localizações cobertas pelo provedor LBS, e suas respectivas informações complementares, *i.e.* popularidade, tipo, e lista de vizinhos com suas respectivas distâncias até a localização.
- l_r: localização real do usuário, onde $l_r \in L$.
- R': última requisição enviada ao provedor do serviço e sua respectiva resposta.

O usuário é responsável por armazenar o histórico das últimas requisições enviadas ao LBS e as respostas destas requisições. Se o histórico do usuário R' estiver limpo, o que indica que o usuário está fazendo sua primeira requisição após uma autenticação válida ao LBS, utilizando ou não alguma técnica de proteção neste tipo de serviço, o parâmetro R' se encontra vazio (linha 1), linha 1 do algoritmo. Neste caso, são escolhidas $k-1$ localizações, cujas probabilidades de estarem numa requisição sejam semelhantes à da localização real l_r, medida por sua popularidade. Sendo assim, é criado o conjunto C (linha 3), contendo $2k$ localizações cujas probabilidades sejam as mais próximas de l_r. Definimos este tamanho de $2k$ com o objetivo de garantir uma representatividade de localizações que permitam uma escolha aleatória segura contra ataques que busquem identificar a localização real baseada na variância das probabilidades das localizações presentes. Entre as localizações de C serão escolhidas aleatoriamente as $k-1$ localizações que irão compor a requisição. Conforme discutido na seção 4.3.1, outro cuidado

que devemos ter neste momento é o de não escolher localizações que sejam muito isoladas de outras. Desta forma medimos a taxa de isolamento das localizações através da equação 4.2 (linha 5), que calcula a razão entre a quantidade de vizinhos de uma localização e a quantidade de vizinhos mínimo necessário para se atingir um certo grau de imprevisibilidade na escolha das localizações, dificultando a aplicação de algoritmos de ataque específicos para a estratégia de anonimização aplicada. Mais uma vez definimos essa quantidade mínima de $2k$ com o objetivo de garantir uma representatividade das localizações que permitam uma boa escolha em termos de proteção de privacidade. Este valor tem garantido um ótimo grau de privacidade em nossos experimentos. Sendo assim, serão escolhidas as localizações cuja taxa de isolamento seja maior ou igual a 1, caso em que a quantidade de vizinhos da localização é igual a $2k$, ou que seja maior que a taxa de isolamento da localização real (linha 6).

$$taxa\ de\ isolamento = \frac{número\ de\ vizinhos\ alcançáveis}{2k} \tag{4.2}$$

Caso R' não esteja vazio (linha 13), o que indica que há uma consulta anterior enviada ao LBS, é necessário ter esta requisição anterior como referência na escolha das localizações falsas. A primeira medida tomada é calcular a área de alcançabilidade para cada uma das localizações da consulta anterior. Dada a velocidade média do usuário v no intervalo de tempo entre duas requisições $\Delta t = t - t'$, com $t > t'$, definimos a máxima distância alcançável como $max_{\Delta s} = v \times \Delta t$ (linha 15). Assim, para cada localização falsa $l'_i \in R'$ definimos o conjunto de localizações alcançáveis D_i (linha 16), que denota o conjunto de localizações $l_i \in L$ cuja distância *manhattan* de l'_i, denotada por $d_{l'_i,l_i}$, não é maior que $max_{\Delta s}$. Importante, deixar claro que adotamos a distância *manhattan* por nos garantir uma maior verossimilhança com um cenário real do que a distância euclidiana, entretanto, o PrivLBS pode usar como critério de distância qualquer tipo.

Para garantir uma representatividade maior das localizações em D_i, definimos o tamanho mínimo de $2k$ localizações, adicionando as localizações mais próximas que estão fora da área de alcançabilidade até que o tamanho mínimo seja alcançado. Esta medida tende a diminuir a precisão do algoritmo em garantir que todas as localizações selecionadas serão alcançáveis por pelo menos uma localização da requisição anterior, entretanto, é fundamental para garantir uma quantidade mínima de localizações que permitam uma escolha adequada nas fases seguintes do processo de seleção. Entretanto, os experimentos têm demonstrado que o impacto desta medida não tem gerada perca de privacidade significativa.

Algoritmo 5: PrivLBS

Entrada: k, L, l_r, R'
Saída: R

1 **se** $R' == \emptyset$ **então**
2 $p_r \leftarrow$ *probabilidade da localização real estar na requisição*;
3 $C \leftarrow$*2k localizações* $l_i \in L$ *com probabilidades* p_{l_i} *mais próximas à* p_r;
4 **para** $i = 0 \rightarrow k$ **faça**
5 $ti_r \leftarrow$ *Taxa de isolamento de* l_r;
6 **se** *Existe* $l \in C$, *com* $ti_l \geq ti_r$ *ou* $ti_l \geq 2k$ **então**
7 $loc_{info} \leftarrow l$;
8 **senão**
9 $loc_{info} \leftarrow$ *Seleciona-se aleatoriamente uma localização* $l \in C$;
10 **fim**
11 $C \leftarrow C - l$;
12 **fim**
13 **senão**
14 **para cada** $l' \in R'$ **faça**
15 $max_{\Delta s} \leftarrow v \times \Delta t$
16 $D \leftarrow$ *Seleciona as localizações l em vizinhos de* l' *onde* $d_{l',l_i} <= max_{\Delta s}$;
17 $c_{l_r,r_{l'_r}} \leftarrow$ *Correlação da localização real atual e a resposta à localização real na consulta anterior*
18 $loc_{info} \leftarrow l \in D$ *com* $c_{l,r_{l'}}$ *mais próximo de* $c_{l_r,r_{l'_r}}$;
19 **fim**
20 **fim**
21 **para cada** $l \in loc_{info}$ **faça**
22 $ti \leftarrow$ *Seleciona um tipo presente em L segundo a distribuição do conjunto de dados*;
23 $req \leftarrow$ *Gera um conteúdo de requisição associado ao tipo ti*;
24 $R \leftarrow$ *Adiciona a tupla (l,req) à requisição R*;
25 **fim**
26 **retorna** R

De cada conjunto de localizações alcançáveis D_i, uma localização falsa é selecionado para compor a requisição. Esta seleção tem como critério a correlação entre as localizações do conjunto de localizações alcançáveis D_i e a resposta à localização l'_i a partir da qual D_i foi gerado. Para o cálculo da correlação entre dadas duas localizações l_i e $l_j \in L$ utilizamos o coeficiente de *Spearman*, definida por $c_{l_i,l_j} \in [-1,1]$. O objetivo é tentar capturar o comportamento do usuário ao se deslocar da localização anterior, enviada na requisição ao provedor de serviço, e a sua atual localização, uma vez que são atributos das localizações suas coordenadas, e assim tentar reproduzir este comportamento na seleção das localizações falsas. Neste processo, primeiramente calculamos a correlação entre a localização atual do usuário l_r e a resposta à localização real na consulta anterior $r_{l'_r}$ (linha 17), denotada por $c_{l_r,r_{l'_r}}$. A partir disso, para cada conjunto D_i, gerado

a partir de uma localização falsa $l'_i \in R'$, procura-se selecionar a localização $l_i \in D_i$ cuja $c_{l_i,r_{l'_i}}$ é a mais próxima de $c_{l_r,r_{l'_r}}$ (linha 18).

Para completar a requisição, para cada localização falsa selecionada é atribuído um conteúdo de requisição associado a um tipo de POI (linha 23), respeitando a distribuição dos tipos presente nas localizações em L. Desta forma, os conteúdos da requisição também são anonimizados, dificultando a descoberta da localização real e impedindo que sejam feitas inferências sobre os mesmos.

Como saída do algoritmo do PrivLBS, temos uma requisição

$$req = \{(l_1, conteudo_1), (l_2, conteudo_2), ...(l_k, conteudo_k)\}$$

(linha 26), onde o provedor do LBS vê todas as localizações com características semelhantes, induzindo a probabilidade de reconhecimento da localização real a $\frac{1}{k}$.

4.4 Conclusão

Neste capítulo, primeiramente, apresentamos o ACon, um algoritmo de ataque baseado em conhecimento adversário que considera o conhecimento prévio sobre os dados da requisição, procurando identificar a real localização do usuário. Em seguida, propomos o PrivLBS, nossa abordagem para preservação de privacidade em serviços baseados em localização, procurando através da técnica de localizações falsas ofuscar a localização real do usuário de potenciais atacantes, incluindo o próprio provedor do serviço.

5 EXPERIMENTAÇÃO

Nesta seção iremos apresentar os experimentos realizados com a finalidade de demonstrar a eficiência do PrivLBS em proteger a privacidade de localização do usuário em serviços LBS.

Foram implementados os algoritmos de anonimização DLS e DLP, propostos por (NIU *et al.*, 2014) e por (SUN *et al.*, 2017a) respectivamente. Implementamos também uma versão mais básica do PrivLBS, proposta em (NETO *et al.*, 2018), onde são considerados como critérios de seleção das localizações falsas, a popularidade e a distância entre as localizações. Para diferenciarmos da nova versão do PrivLBS, a esta versão mais básica chamaremos de PrivLBSv1.

Realizamos uma série de experimentos onde comparamos o desempenho destes algoritmos de anononimização com o PrivLBS. Implementamos também o algoritmo de ataque ADLS proposto em (SUN *et al.*, 2017a). Desta forma, procuramos medir a taxa de reconhecimento da localização real anonimizada por cada um dos algoritmos já citados quando sujeitos aos algoritmos de ataque ADLS e ACon. Medimos também a entropia das requisições geradas por cada um dos algoritmos de anonimização. Por fim, verificamos a taxa de identificação do tipo real presente no conteúdo das requisições.

5.1 Configuração Experimental

Antes de partimos para a análise de desempenho dos algoritmos de anonimização implementados e seus respectivos resultados, apresentaremos o ambiente de desenvolvimento e o conjunto de dados.

5.1.1 Ambiente de Desenvolvimento

Nós configuramos nossos experimentos com dez mil usuários se deslocando e executando consultas contínuas ao provedor de LBS. Cada usuário realiza 4 consultas. O intervalo de tempo entre cada consulta varia entre 30 segundos e 10 minutos. A velocidade do usuário varia entre 5 quilômetros por hora e 80 quilômetros por hora. Além disso, presumimos que o usuário possui uma probabilidade entre 10% e 20% de se deslocar em direção à localização presente na resposta da consulta anterior, alcançando-a se o intervalo de tempo e a velocidade do usuário forem suficientes para isso. Após cada requisição anonimizada pelos algoritmos,

aplicamos os algoritmos de ataque ADLS e ACon, medindo a taxa de reconhecimento da localização real presente na requisição. Para cada requisição, calculamos também a entropia da requisição, medida através da Equação 3.2.

Todos os algoritmos foram implementados na linguagem Python 2.7. Os experimentos foram executados em uma máquina *desktop*, pertencente ao Laboratório de Sistemas e Banco de Dados (LSBD/UFC), equipada com sistema operacional Ubuntu 14.04, processador Intel Core i5 de 3,2 GHz, 8 GB de RAM e disco de 500 GB.

5.1.2 Conjuntos de Dados

Utilizamos um conjunto de dados real disponibilizado pela *Chicago Transit Authoroty* (CTA) (AUTHOROTY, 2018). O conjunto de dados é composto por 11.593 estações de ônibus, nossos POIs, com informações de latitude, longitude e média de embarque diário, que foi utilizado para estimar a popularidade de cada localização. Para nossos experimentos, adicionamos sinteticamente o campo de tipo para cada localização no conjunto de dados. Foram estimados 90 tipos, em concordância com o número de tipos de localizações presentes na API do Google (GOOGLE, 2017). A distribuição dos tipos das localizações no conjunto de dados seguiu uma distribuição normal, com estes tipos variando de 1 a 90, representando os noventa tipos presentes no conjunto de dados.

5.1.3 Algoritmos implementados

Com o objetivo de avaliarmos o desempenho do PrivLBS, implementamos os algoritmos de seleção de localizações falsas, DLS, DLP, PrivLBSv1, além dos algoritmos de ataque ADLS e ACon.

5.1.4 Análise de Desempenho

Primeiramente na Figura 14, analisamos o tempo de execução, em segundos, dos algoritmos de anonimização. Como esperando, a adição de novos critérios de seleção no PrivLBS, aumentando a sua garantia de privacidade, gerou um maior *overhead*. Podemos observar que para $k = 16$ o tempo de execução do algoritmo foi em média de 0.01 segundos, contra 0.001 segundos do tempo de anonimização do *DLP*, que obteve o melhor resultado.

Em seguida avaliamos a entropia da requisição. Ela nos dá o grau de incerteza sobre

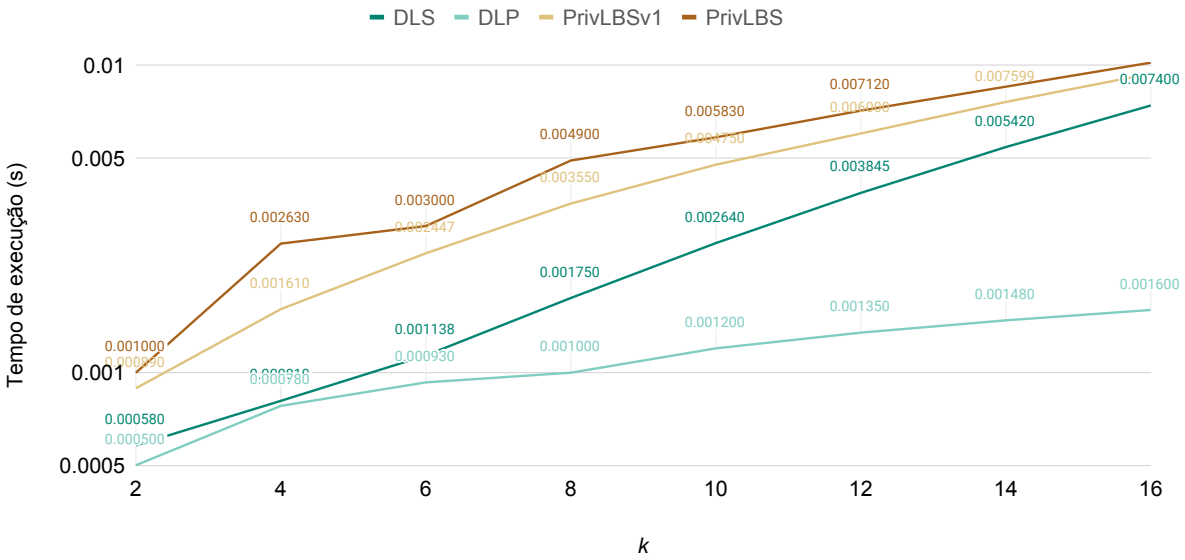

Figura 14 – Tempo de anonimização em segundos.

as localizações presentes na requisição. Quanto maior a entropia, maior é o grau de similaridade entre as localizações presentes na requisição. Neste caso, a entropia é calculada em razão da probabilidade das localizações estarem presentes em uma requisição, medida pela Equação 3.1, onde a máxima entropia possível, $\log_2 k$, é obtida quando todas as localizações presentes na requisição possuem a mesma probabilidade.

	Grau de anonimização k				
	2	**4**	**8**	**12**	**16**
DLS	0,691	1,386	2,079	2,484	2,771
DLP	0,693	1,389	2,079	2,484	2,772
PrivLBSv1	0,69	1,381	2,073	2,478	2,765
PrivLBS	0,68	1,375	2,066	2,470	2,758

Tabela 3 – Entropia calculada em função da probabilidade das localizações estarem presentes em uma requisição.

Na Tabela 3 podemos observar que os algoritmos DLS e DLP, que possuem como critérios de seleção apenas a probabilidade das requisições e que utilizam o cálculo de entropia como função de custo no momento da seleção, atingem uma entropia ligeiramente maior que as duas versões do PrivLBS. Este resultado era esperado visto que adotamos novos critérios de seleção além da probabilidade. Sendo assim, tanto o PrivLBSv1 quanto o PrivLBS, apesar de procurarem selecionar localizações cujas probabilidades sejam semelhantes, acabam por diminuir o domínio de busca por exigir localizações alcançáveis. Entretanto, a perda de entropia não é significativa, apresentando valores próximos para qualquer grau de privacidade medido.

Este comportamento acaba se refletindo na taxa de reconhecimento da localização real presente na requisição, quando aplicado o algoritmo de ataque ADLS, especialmente em relação ao PrivLBSv1 e PrivLBS. A Figura 15 apresenta os resultados obtidos ao aplicar o ADLS sobre requisições anonimizadas utilizando os quatro algoritmos. Foi medida a taxa de reconhecimento da localização real presente nas requisições com grau de privacidade $k = [2, 4, 8, 12, 16]$.

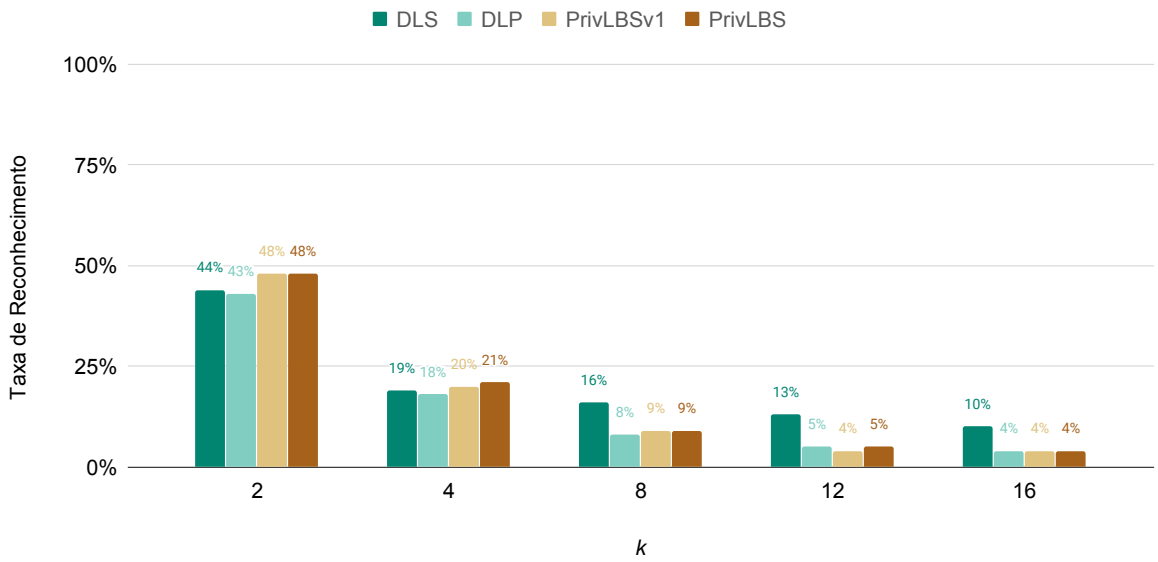

Figura 15 – Taxa de reconhecimento da localização real quando aplicado o algoritmo de ataque ADLS.

Como já discutido na Seção 3.4.2.1, o ADLS utiliza como critério de identificação da localização real a variância da probabilidade das localizações presentes na requisição, justamente como uma forma de demonstrar a vulnerabilidade do DLS no processo de seleção das localizações falsas. Como podemos observar pelos resultados, o DLP apresentou um resultado, em média, melhor que as outras abordagens, especialmente para um k pequeno. Podemos observar que à medida que o valor do k aumenta, seu desempenho passa a ser semelhante ao PrivLBSv1 e PrivLBS. Já o DLS se mostrou vulnerável, atingindo as maiores taxas de identificação para $k = 8, 12$ e 16. Desta forma, tanto o DLP quanto as duas abordagens do PrivLBS demonstraram que garantem o grau de privacidade k, sempre apresentando uma taxa de identificação abaixo de $\frac{1}{k}$. Isto só mostra que, apesar de possuir uma entropia, calculada em razão das probabilidades das localizações presentes na requisição, ligeiramente menor, o PrivLBS é capaz de garantir o grau de privacidade k das requisições mesmo sobre um ataque que procura explorar este critério, como o algoritmo de ataque ADLS.

Como já demonstrado, a entropia calculada em razão da probabilidade não é suficiente para medir o grau de incerteza das localizações presentes na requisição. Sendo assim, estendemos o cálculo da entropia para cada um dos critérios adotados. Primeiramente, como já fazíamos, calculamos a entropia em função das probabilidades das localizações presentes na requisição. Em seguida, calculamos a entropia em função da distância entre as localizações presentes na atual requisição e as localizações presentes na requisição anterior. Por último, calculamos a entropia em função da correlação das localizações presentes na atual requisição e nas respostas às localizações presentes na resposta à requisição anterior. De posse das entropias em função dos três critérios, calculamos a média entre elas. Como resultado, a Tabela 4 mostra que, para qualquer grau de privacidade k, a entropia da requisição anonimizada pelo PrivLBS é a maior entre os métodos comparados, demonstrando que as localizações selecionadas pelo nosso algoritmo possuem um maior grau de similaridade, gerando uma maior indistinguibilidade entre as localizações presentes na requisição.

	Grau de anonimização k				
	2	**4**	**8**	**12**	**16**
DLS	0,613	1,294	1,986	2,392	2,680
DLP	0,615	1,294	1,986	2,392	2,680
PrivLBSv1	0,643	1,324	1,985	2,392	2,707
PrivLBS	0,652	1,325	2,014	2,401	2,710

Tabela 4 – Entropia calculada em função da probabilidade, distância e correlação.

Uma vez feita esta análise sobre a entropia das requisições, passamos a analisar a taxa de reconhecimento quando aplicado o algoritmo de ataque proposto, ACon. Mais uma vez, variamos o grau de privacidade k, que indica a quantidade de localizações presentes na requisição e a probabilidade máxima de reconhecimento aceitável, $\frac{1}{k}$.

A Figura 16 apresenta os resultados obtidos quando aplicado o ACon sobre as requisições anonimizadas pelos algoritmos DLS, DLP, PrivLBSv1 e PrivLBS. A primeira observação é que o DLS e o DLP, por não utilizarem nenhum outro critério de seleção além da probabilidade das localizações, quando sofrem ataques de conhecimento que exploram outros conhecimentos adversários, como a distância e a correlação, não são capazes de proteger a localização real do usuário. Pelos resultados, podemos observar que, para qualquer grau de privacidade k, a taxa de reconhecimento da localização real do usuário se manteve sempre acima de $\frac{1}{k}$ nas requisições anonimizadas utilizando tanto o DLS quanto o DLP. Já quando comparamos às duas versões do PrivLBS, o PrivLBSv1, que possui como critério de seleção somente a

distância e a probabilidade das localizações, apresentou um resultado melhor que o DLS e DLP, o que era esperado, visto que seu processo de seleção é mais criterioso, garantindo uma maior proteção. Entretanto, quando comparado ao PrivLBS, que além dos dois critérios já citados, ainda considera a correlação no momento da anonimização da requisição, o PrivLBSv1 apresenta um resultado pior, expondo a localização real do usuário com uma taxa de identificação em alguns momentos acima de $\frac{1}{k}$, não garantindo, portanto, o grau de privacidade k desejado pelo usuário.

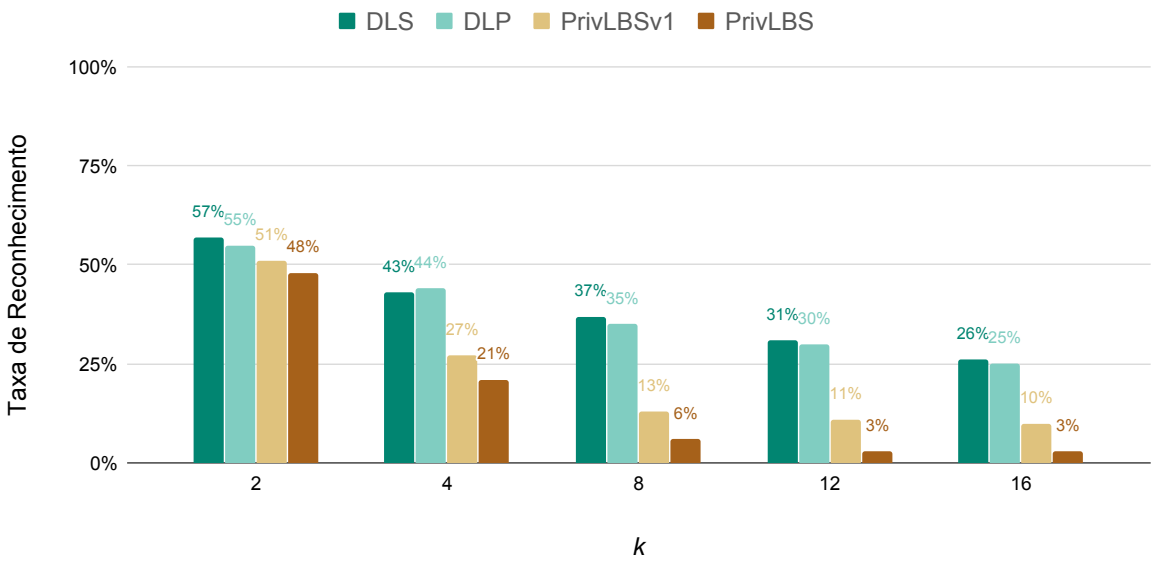

Figura 16 – Taxa de reconhecimento da localização real quando aplicado o algoritmo de ataque ACon.

A Figura 17 apresenta o comportamento dos algoritmos de ataque ADLS e ACon sobre as requisições anonimizadas pelo PrivLBS, em um cenário onde a velocidade do usuário é baixa, o que poderia resultar em um conjunto pequeno de localizações alcançáveis. Podemos observar que, para $k = 8$, a taxa de identificação da localização real se manteve constante, próxima de 6%, a medida que a velocidade aumentava, demonstrando que o PrivLBS é capaz de proteger a informação de localização do usuário mesmo em um cenário adverso, com poucas localizações alcançáveis.

Por fim, analisamos a proteção ao conteúdo de requisição. Para demonstrarmos que o processo de seleção dos tipos em razão da sua distribuição no conjunto de localizações é suficiente para garantir a sua privacidade na requisição, aplicamos um ataque sobre o conteúdo da requisição. Neste ataque, procuramos identificar o tipo presente no conteúdo de requisição associado à localização real do usuário, em razão da probabilidade deste tipo estar presente no

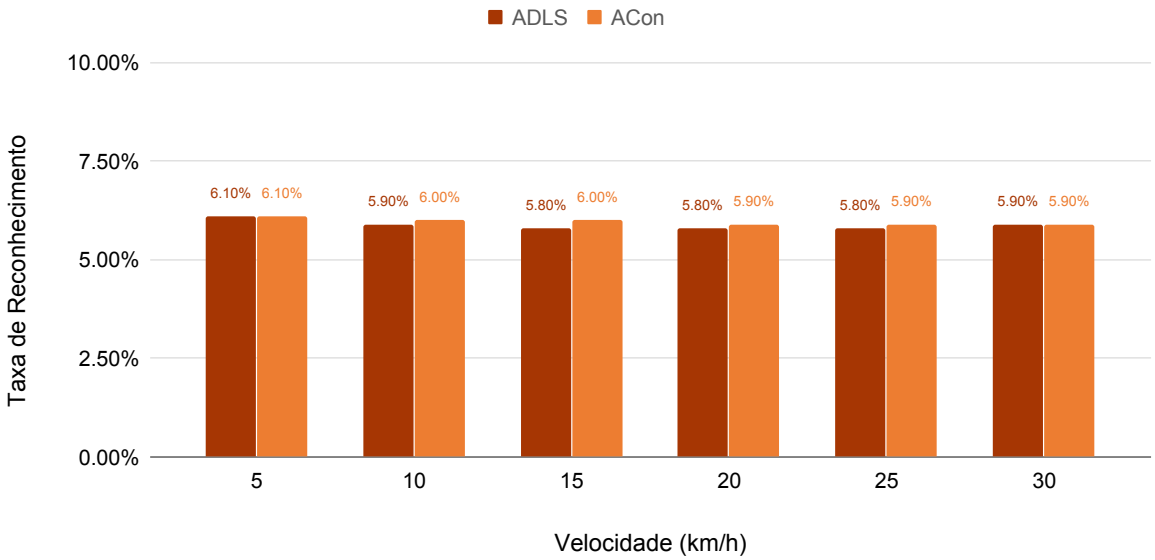

Figura 17 – Taxa de reconhecimento da localização real quando aplicado o algoritmo de ataque ACon e ADLS, em função da velocidade do usuário para um $k = 8$.

conteúdo de requisição, ou seja, buscamos identificar como o verdadeiro conteúdo de requisição, aquele que possui o tipo com a maior probabilidade de estar presente em uma requisição, dentre os tipos dos conteúdos da requisição. A Figura 18 apresenta a taxa de identificação do tipo no conteúdo anonimizado pelo PrivLBS. Os resultados demonstram que o PrivLBS é capaz de proteger com eficiência o tipo presente no conteúdo real da requisição, obtendo uma taxa de reconhecimento próximo de $\frac{1}{k}$ para qualquer grau de privacidade k, diferente das outras abordagens que não realizam nenhum tipo de anonimização do conteúdo de requisição.

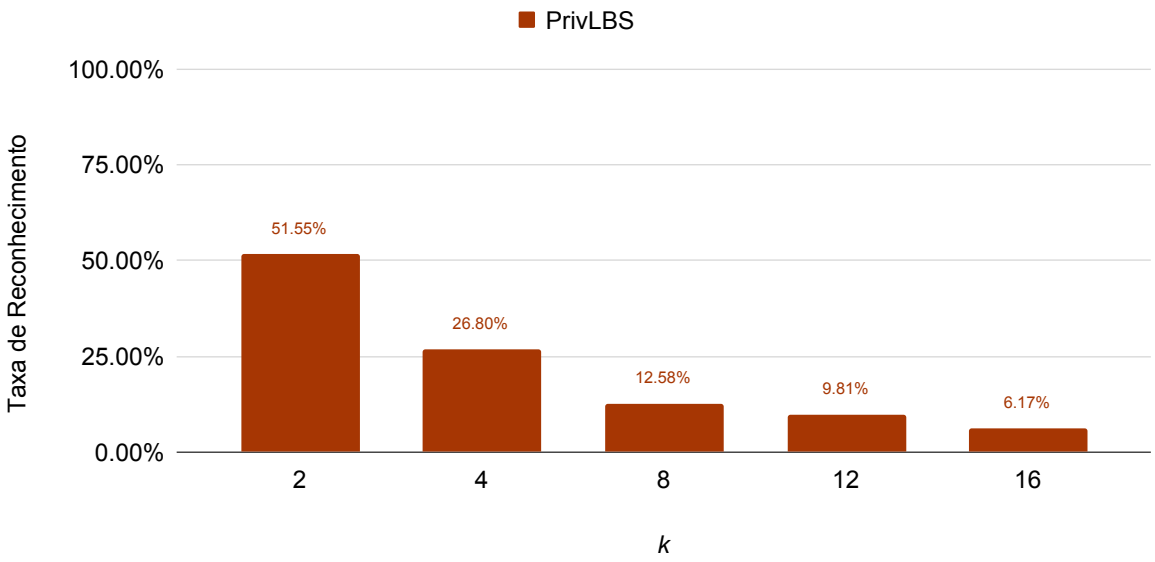

Figura 18 – Ataque sobre conteúdo de requisição anonimizada pelo PrivLBS.

5.2 Conclusão

Podemos observar, através dos experimentos, que, em um cenário de requisições contínuas, o ACon é bastante eficaz em identificar a localização real do usuário sobre requisições que utilizam a estratégia de seleção de localizações falsas, especialmente quando a estratégia não apresenta mecanismos para se proteger contra este tipo de ataque. Como visto, os algoritmos DLS e DLP demonstraram que não são capazes de garantir a privacidade do usuário contra ataques de conhecimento em um cenário de consultas contínuas.

O PrivLBSv1 foi nossa primeira proposta com o objetivo de aplicar a técnica de seleção de localizações falsas neste cenário. Para isto, procuramos proteger a localização do usuário aplicando, no processo de seleção das localizações falsas, o critério de distância entre as localizações de consultas consecutivas, além da popularidade destas localizações. Assim, em um primeiro momento, garantindo a privacidade do usuário para ataques que exploram a popularidade e a distância entre as localizações. Entretanto, os dados de localização, por sua natureza, são extremamente correlacionados. Desta forma, apenas o critério de distância e popularidade não são suficientes para proteger o usuário contra ataques. Com o objetivo de sanar o problema, propomos o PrivLBS, onde adicionamos, no processo de seleção das localizações falsas, um cálculo de correlação entre as localizações. Os experimentos mostraram que o PrivLBS é capaz de garantir a privacidade do usuário mesmo sofrendo um ataque de conhecimento como o do ACon, apresentando uma taxa de reconhecimento abaixo de $\frac{1}{k}$, onde k é o grau de privacidade desejado. Entretanto, anonimizar a informação de localização da requisição não é suficiente para garantir a privacidade do usuário. Sendo assim, procuramos anonimizar, também, o conteúdo de requisição, protegendo o tipo associado ao conteúdo real desejado pelo usuário. Demonstramos que o PrivLBS, diferente das abordagens existentes, é capaz de proteger a informação de localização e o conteúdo de requisição.

6 CONSIDERAÇÕES FINAIS

Nesta seção apresentaremos nossas considerações finais a cerca do problema de preservação de privacidade de dados de localização dos usuários em serviços baseados em localização.

6.1 Conclusão

Neste trabalho, inicialmente propomos o ACon um modelo de ataque de conhecimento que busca identificar a localização real do usuário em requisições realizadas a provedores de serviço de localização. Nosso objetivo foi demonstrar que em um cenário, onde o provedor do serviço é o próprio atacante, o ACon explora o conhecimento adversário adquirido pelo próprio provedor para identificar a localização real do usuário dentre as localizações presentes na requisição. Nosso algoritmo de ataque obteve uma alta taxa de identificação da localização real do usuário em requisições anonimizadas utilizando a técnica de seleção de localizações falsas, tais como os trabalhos propostos por (NIU *et al.*, 2015; SUN *et al.*, 2017a), demonstrando que estes trabalhos não são capazes de garantir o grau de privacidade k em um cenário de consultas contínuas.

Como solução ao problema de preservação de privacidade em serviços de localização, propomos o PrivLBS. Para garantir a efetividade de nossa solução e manter a qualidade do serviço, o PrivLBS utiliza a técnica de ofuscação chamada de seleção de localizações falsas, onde são enviadas k localizações na requisição a fim de ofuscar a localização real do usuário. Diferente das técnicas já existentes de seleção de localização falsa, o PrivLBS procura proteger não apenas a localização do usuário, mas também o conteúdo da requisição.

Para proteger a localização real do usuário, nosso algoritmo utiliza de conhecimentos adversários, tais como: a distância entre as localizações em requisições consecutivas, a popularidade das localizações e a correlação entre as localizações presentes na atual requisição e as localizações presentes na resposta à requisição anterior do usuário.

Para proteger o conteúdo da requisição, o PrivLBS procura ofuscar o tipo da localização presente no conteúdo, gerando $k-1$ tipos para cada localização falsa selecionada durante o processo de anonimização da localização real. Garantindo assim, que o atacante seja incapaz de inferir qual o tipo real presente no conteúdo de requisição. Com este objetivo, os tipos selecionados respeitam a distribuição de tipos presentes no conjunto de dados.

Utilizando um conjunto de dados reais, com a adição do campo tipo, objeto de anonimização do conteúdo da requisição, realizamos uma série de requisições anonimizadas pelos algoritmos DLS, DLP e duas versões do PrivLBS. Após cada requisição executada, calculamos a entropia da requisição e a taxa de identificação da localização real na requisição pelos algoritmos de ataque ADLS e ACon. Os resultados obtidos em todos os experimentos comprovam que o PrivLBS é capaz de proteger a privacidade de localização do usuário, garantindo o grau de privacidade k. Mostramos também, que os conteúdos de requisição presentes nas requisições apresentam a mesma distribuição de tipos do conjunto de dados, demonstrando que o conteúdo de requisição real está ofuscado pelos outros conteúdos presentes na requisição.

6.2 Desafios

Ao longo do desenvolvimento deste trabalho encontramos uma série de desafios e limitações que gostaríamos de destacar. Primeiramente, em LBS, encontrar um conjunto de dados adequado, com informações que permitam alimentar o modelo é extremamente desafiador. Muitas vezes é necessário coletar de várias fontes para obter o melhor resultado. No PrivLBS, tivemos bastante dificuldade em encontrar um conjunto de localizações que nos permitissem extrair as informações necessárias para aplicar o processo de anonimização.

Outro desafio que nos deparamos foi configurar nossa simulação com parâmetros que se aproximassem de um cenário real de interação entre usuários e provedor do serviço. Por falta de informações, quando estimamos a probabilidade de um usuário se deslocar para a resposta da consulta anterior, tivemos que presumir esta probabilidade por acreditar nesta tendência de o usuário se deslocar em razão do interesse demonstrado na requisição. Outro parâmetro crítico, foi a velocidade média do usuário no algoritmo de ataque, ACon, estimada em razão da velocidade média de deslocamento de veículos em uma zona urbana. Uma análise mais detalhada sobre dados de trajetória de usuários em serviços de localização seria necessário para uma melhor configuração destes parâmetros.

Uma limitação a ser explorada no PrivLBS é o *overhead* gerado pelo processo de anonimização. Pela análise do tempo de execução do PrivLBS, demonstra-se o problema latente de escalabilidade do algoritmo à medida que o grau de privacidade k aumenta, sendo necessário, portanto, o uso de técnicas que permitam a otimização do PrivLBS.

6.3 Trabalhos Futuros

Como trabalhos futuros, primeiramente, deixamos em aberto os desafios levantados na seção anterior. Além disso, pretendemos explorar novas técnicas mais robustas contra ataques de conhecimento, como a Privacidade Diferencial, modelo matemático de privacidade que parte do pressuposto que o atacante possui um conhecimento global sobre o conjunto de dados. Sendo fundamental, portanto, encontrar um equilíbrio entre privacidade e utilidade de dados que não gere um impacto considerável na qualidade dos serviços de localização prestados.

REFERÊNCIAS

ANDRÉS, M. E.; BORDENABE, N. E.; CHATZIKOKOLAKIS, K.; PALAMIDESSI, C. Geo-indistinguishability: Differential privacy for location-based systems. *In:* **Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security**, New York, NY, USA: ACM, p. 901–914, 2013.

ARMUS, T. **Facebook can tell whether you're gay based on a few 'likes,' study says**. 2017. Disponível em: https://www.nbcnews.com/feature/nbc-out/facebook-can-tell-if-you-re-gay-based-few-likes-n823416. Acesso em: 13 mar. 2019.

ASIKIS, T.; POURNARAS, E. Optimization of privacy-utility trade-offs under informational self-determination. **Future Generation Computer Systems**, Elsevier B.V., [*S.l.*], 2018. Disponível em: http://www.sciencedirect.com/science/article/pii/S0167739X17322252. Acesso em: 03 mar. 2019.

AUTHOROTY, C. T. **Chicago Transit Authoroty**. 2018. Disponível em: http://www.transitchicago.com. Acesso em: 16 mar. 2018.

BAMBA, B.; LIU, L.; PESTI, P.; WANG, T. Supporting anonymous location queries in mobile environments with privacygrid. *In:* **Proceedings of the 17th International Conference on World Wide Web, ACM**, New York, NY, USA, p. 237–246, 2008.

DEWRI, R.; RAY, I.; RAY, I.; WHITLEY, D. On the optimal selection of k in the k-anonymity problem. *In:* **24th ICDE International Conference on Data Engineering, IEEE**, Cancun, Mexico, p. 1364–1366, 2008.

DUCKHAM, M.; KULIK, L. A formal model of obfuscation and negotiation for location privacy. In: GELLERSEN, H. W.; WANT, R.; SCHMIDT, A. (Ed.). **Pervasive Computing**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005. p. 152–170.

DWORK, C. Differential privacy. *In:* **33rd International Colloquium on Automata, Languages and Programming, Springer**, Venice, Italy, p. 1–12, 2006.

GEDIK, B.; LIU, L. Protecting location privacy with personalized k-anonymity: Architecture and algorithms. *In:* **IEEE Transactions on Mobile Computing**, [*S.l.*], v. 7, n. 1, p. 1–18, 2008.

GHINITA, G. Privacy for location-based services. *In:* **Synthesis Lectures on Information Security, Privacy, and Trust**, [*S.l.*], v. 4, n. 1, p. 1–85, 2013.

GOOGLE. **Simple, battery-efficient APIs for location and context**. 2017. Disponível em: https://developers.google.com/location-context/. Acesso em: 03 mar. 2019.

GRIFFITH, D.; CHUN, Y. Spatial autocorrelation and spatial filtering. **Handbook of regional science**, Springer, [*S.l.*], p. 1477–1507, 2014.

HARRIS, I. **Pokemon Go captures 800 million downloads**. 2018. Disponível em: https://www.pocketgamer.biz/news/68209/pokemon-go-captures-800-million-downloads/. Acesso em: 03 fev. 2019.

HAUKE, J.; KOSSOWSKI, T. Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. **Quaestiones Geographicae**, [*S.l.*], v. 30, n. 2, p. 87–93, Jun 2011.

HU, H.; CHEN, Q.; XU, J. Verdict: Privacy-preserving authentication of range queries in location-based services. *In:* **IEEE 29th International Conference on Data Engineering ICDE**, Brisbane, QLD, Australia, p. 1312–1315, 2013.

HUBAUX, J.; THEODORAKOPOULOS, G.; BOUDEC, J. L.; SHOKRI, R. Quantifying location privacy. *In:* **IEEE Symposium on Security and Privacy(SP)**, Oakland, California, USA, v. 00, p. 247–262, 05 2011.

KIDO, H.; YANAGISAWA, Y.; SATOH, T. An anonymous communication technique using dummies for location-based services. *In:* **International Conference on Pervasive Services, 2005.**, Santorini, Greece, Greece, p. 88–97, 2005.

LI, H.; SUN, L.; ZHU, H.; LU, X.; CHENG, X. Achieving privacy preservation in wifi fingerprint-based localization. *In:* **INFOCOM, Proceedings IEEE**, [*S.l.*], p. 2337–2345, 2014.

LIU, B.; ZHOU, W.; ZHU, T.; GAO, L.; XIANG, Y. Location privacy and its applications: A systematic study. **IEEE Access**, Piscataway, NJ, USA, v. 6, p. 17606–17624, 2018. ISSN 2169-3536.

LU, R.; LIN, X.; SHI, Z.; SHAO, J. Plam: A privacy-preserving framework for local-area mobile social networks. *In:* **INFOCOM, 2014 Proceedings IEEE**, [*S.l.*], p. 763–771, 2014.

MACHANAVAJJHALA, A.; GEHRKE, J.; KIFER, D.; VENKITASUBRAMANIAM, M. l-diversity: Privacy beyond k-anonymity. *In:* **22nd International Conference on Data Engineering, IEEE**, [*S.l.*], p. 24–24, 2006.

MEYERSON, A.; WILLIAMS, R. On the complexity of optimal k-anonymity. *In:* **Proceedings of the 23rd ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems**, Paris, France, p. 223–228, 2004.

NETO, E. R. D.; MENDONçA, A. L. C.; BRITO, F. T.; MACHADO, J. C. Privlbs: uma abordagem para preservação de privacidade de dados em serviços baseados em localização. *In:* **Brazilian Symposium on Databases SBBD**, Rio de janeiro, Brazil, 2018.

NIU, B.; GAO, S.; LI, F.; LI, H.; LU, Z. Protection of location privacy in continuous lbss against adversaries with background information. *In:* **International Conference on Computing, Networking and Communications (ICNC)**, Kauai, Hawaii, USA, p. 1–6, 2016.

NIU, B.; LI, Q.; ZHU, X.; CAO, G.; LI, H. Achieving k-anonymity in privacy-aware location-based services. *In:* **INFOCOM, 2014 Proceedings IEEE**, [*S.l.*], p. 754–762, 2014.

NIU, B.; LI, Q.; ZHU, X.; CAO, G.; LI, H. Enhancing privacy through caching in location-based services. *In:* **Computer Communications (INFOCOM), IEEE**, [*S.l.*], p. 1017–1025, 2015.

SERJANTOV, A.; DANEZIS, G. Towards an information theoretic metric for anonymity. *In:* **Proceedings of the 2nd International Conference on Privacy Enhancing Technologies**, Springer-Verlag, Berlin, Heidelberg, p. 41–53, 2003.

SUN, G.; CHANG, V.; RAMACHANDRAN, M.; SUN, Z.; LI, G.; YU, H.; LIAO, D. Efficient location privacy algorithm for internet of things (iot) services and applications. **Journal of Network and Computer Applications**, [*S.l.*], v. 89, p. 3–13, 2017.

SUN, G.; LIAO, D.; LI, H.; YU, H.; CHANG, V. L2p2: A location-label based approach for privacy preserving in lbs. **Future Generation Computer Systems**, [*S.l.*], v. 74, p. 375–384, 2017.

SUPERDATA. **Market Brief — 2018 Digital Games & Interactive Entertainment Industry Year In Review**. 2019. Disponível em: https://www.superdataresearch.com/market-data/market-brief-year-in-review/. Acesso em: 03 fev. 2019.

SWEENEY, L. k-anonymity: A model for protecting privacy. **International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems**, World Scientific, [*S.l.*], v. 10, n. 05, p. 557–570, 2002.

TSOUKANERI, G.; THEODORAKOPOULOS, G.; LEATHER, H.; MARINA, M. K. On the inference of user paths from anonymized mobility data. *In:* **European Symposium on Security and Privacy (EuroS P), IEEE**, [*S.l.*], p. 199–213, 2016.

ULLAH, I.; SHAH, M. A. A novel model for preserving location privacy in internet of things. *In:* **22nd International Conference on Automation and Computing (ICAC)**, [*S.l.*], p. 542–547, 2016.

VU, K.; ZHENG, R.; GAO, J. Efficient algorithms for k-anonymous location privacy in participatory sensing. *In:* **2012 Proceedings IEEE INFOCOM, IEEE**, Orlando, FL, USA, p. 2399–2407, 2012.

WANG, L.; YANG, D.; HAN, X.; WANG, T.; ZHANG, D.; MA, X. Location privacy-preserving task allocation for mobile crowdsensing with differential geo-obfuscation. *In:* **Proceedings of the 26th International Conference on World Wide Web**, Perth, Australia, p. 627–636, 2017.

YING, B.; MAKRAKIS, D. Protecting location privacy with clustering anonymization in vehicular networks. *In:* **Computer Communications Workshops (INFOCOM WKSHPS), IEEE Conference on**, [*S.l.*], p. 305–310, 2014.

Ying, B.; Makrakis, D.; Mouftah, H. T. Dynamic mix-zone for location privacy in vehicular networks. *In:* **IEEE Communications Letters**, [*S.l.*], v. 17, n. 8, p. 1524–1527, August 2013.

ZAKHARY, V.; SAHIN, C.; GEORGIOU, T.; ABBADI, A. E. Locborg: Hiding social media user location while maintaining online persona. *In:* **Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems**, New York, NY, USA, p. 12:1–12:4, 2017.

ZHU, X.; CHI, H.; NIU, B.; ZHANG, W.; LI, Z.; LI, H. Mobicache: When k-anonymity meets cache. *In:* **Global Communications Conference (GLOBECOM), IEEE**, Atlanta, GA, USA, p. 820–825, 2013.