Universidad del Bío-Bío Sede Concepción
Facultad de ciencias empresariales

UNIVERSIDAD DEL BÍO-BÍO
FACULTAD DE CIENCIAS EMPRESARIALES

“Simulación de LBS consciente de la privacidad en una red P2P inalámbrica”

Informe de proyecto de titulación para optar al título de Ingeniero Civil en Informática

Alumno: José Astudillo
Profesor Guía: Patricio Galdames

1. Resumen

1.1 Resumen

Por la naturaleza de los LBS (Location-based service) y su gran tasa de utilización, quienes mantienen estos servicios conservan, administran y retroalimentan la información de las consultas geográficas almacenando estos registros en big data, exponiendolos a minería de datos. Estos estudios podrían llevar a violaciones de privacidad al concluir mediante reconocimiento de patrones datos sensibles de usuarios finales que pudiera utilizarse para fines no autorizados por estos. A modo de respuesta, en el siguiente proyecto se implementa un software de simulación de una red P2P inalámbrica, utilizando técnicas de caching propuestas en otros trabajos paralelos, donde los datos sean almacenados y divulgados por los mismos usuarios para los usuarios. Los cuales generan constantemente consultas y respuestas entre sí evadiendo la necesidad de consultar a un servidor LBS, con el fin de proteger su privacidad. Pero, si fuese necesario, también ofreciendo la posibilidad de formular una consulta directamente a un servidor LBS empleando una técnica de enmascaramiento de consulta, que protege tanto la componente semántica como la ubicación de la consulta. El objetivo de la simulación es controlar condiciones iniciales (como número de usuarios, rango posible de conexión entre ellos, etc) y registrar variables de métricas de evaluación a través del tiempo para apoyar el estudio de las técnicas de caching de estos trabajos.

1.2 Abstract

Due to the nature of LBS (Location-based services) and their extensive use, those who maintain those services keep, manage and give data feedback to location-based queries storing all records on big data exposing them to data-mining. Those studies could lead to privacy issues when concluding via pattern-recognition sensitive data of end-users that could be used in non-authorized ways. As a response, this work implements a simulation software of a P2P wireless network, using caching techniques described in other parallel works, where the data is maintained by users for users. They constantly submit queries and responses between them, eluding the need of submitting a query to a LBS server, with the purpose of protecting their privacy. But, if needed, offering the possibility to submit a query directly to a LBS server. implementing query masking techniques that protects both semantic components and location data of said query. The objective of this simulation is to control initial conditions (like number of users, maximum range of connection, etc) and recording metrics control variables on time to support the study of caching techniques of those works.

2. Conceptos preliminares

2.1 Dato de ubicación

Puede ser un término amplio por lo que en este caso en particular estudiaremos el dato de ubicación que es generado por el uso de aplicaciones móviles. la Privacy and Electronic Communications Regulations (PECR) [8], implementando el GDPR define un dato de ubicación como: Cualquier dato procesado en una red de comunicación electrónica o por un servicio de comunicación electrónico indicando la posición geográfica de un equipo terminal propiedad de un usuario de un servicio de comunicación electrónica pública. Incluyendo datos relacionados a:

a) La latitud, longitud o altitud del equipo terminal;
b) la dirección del viaje del usuario; o
c) el tiempo y la ubicación fueron registrados.

2.2 Servicio basado en ubicación.

Los servicios basados en ubicación (LBS) son aplicaciones que utilizan la posición de un usuario final, animal o cosa basada en un dispositivo (a mano, en prenda o implantado) para un propósito en particular.

3.Introducción

3.1 Servicios basados en ubicación y privacidad.

La introducción de los servicios basados en la ubicación (desde ahora LBS, por su sigla en inglés de su nombre, Location Based Service), se ha vuelto fundamental en el uso diario del usuario promedio de dispositivos móviles quienes se retroalimentan de este de manera ininterrumpida. Muchos usuarios de dispositivos móviles utilizan de manera constante datos de ubicación a través de la simple ejecución cotidiana de aplicaciones populares que lo monitorean constantemente. Como resultado, la gran mayoría de los dispositivos móviles está enviando y recibiendo de este tipo de datos, segundo a segundo. Estos datos son procesados en un servidor LBS, quien registra tanto las consultas enviadas por los usuarios y las respuestas que les entrega. Este último reporte de información es acumulado colosalmente en big data de manera persistente.

Una vez en big data, estos datos son comparados entre sí y son sometidos a constante análisis y estudio a través de minería de datos. Como resultado, se obtiene información sobre la información del usuario, es decir, como producto se obtienen patrones, directrices y componentes basados en el comportamiento del usuario. Estos patrones son utilizados para engrosar aún más la funcionalidad del servicio, ahondando ahora en contenido personalizado al comportamiento en específico del usuario final: Sus aplicaciones filtran su contenido en torno a sus preferencias, puede corregir o depurar la experiencia de usuario en torno a su perfil, es identificable como público objetivo de publicidad relevante para el usuario, etc. Sin embargo, muchas veces esta práctica termina en usos cuestionables. La práctica más común es la de User Profiling, que consiste en identificar el comportamiento y preferencias de un conjunto amplio de individuos en base a su huella digital. Estos datos han sido históricamente utilizados como moneda de cambio o como objetivo de filtraciones y robos de información.

Esta tendencia (que crece cada vez más sofisticada y compleja) se abre a un consumo ampliado y masivo en un mundo interconectado, donde los datos se vuelven un recurso clave. La ubicación con la que trabajan estos servicios resulta una pieza importante. Entre la mezcla de información que se amalgama en big data, los patrones en los datos de ubicación (en especial) pueden vulnerar información sensible de un usuario individual.

3.2 El adversario

Existen muchos precedentes de vulneración de privacidad y fuga de datos, el más grande (desde el punto de vista de usuarios afectados, entradas de datos filtradas y número de cuentas comprometidas) es la fuga de Yahoo (2013) que con un exorbitante número de tres mil millones de cuentas afectadas[1] encendió las alarmas en el globo con respecto a la seguridad de los datos y los riesgos embebidos en la seguridad individual de cada usuario involucrado. El segundo caso más grande es el de la base de datos de registro de identificación indio Aadhar (2018) con más de 1.1 mil millones de cuentas afectadas [2], incluyendo información personal, bancaria y perfiles biométricos. El caso más emblemático de filtraciones en la cultura popular es el caso de la venta y filtración de información que conlleva el escándalo Facebook-Cambridge Analytica donde la cifra alcanzó los 87 millones de usuarios afectados. Si bien los números parecen más acotados que en los casos anteriores, los datos fueron utilizados para influenciar las elecciones presidenciales del 2016 en los EEUU y el referéndum del brexit en el mismo año [3].

¿Son estas brechas el punto de partida? En los registros de Statista [4], plataforma de datos de negocios, vemos que dista mucho de serlo. Las brechas estudiadas por la plataforma datan desde el 2005, donde se enumeraron 157 brechas con un número de 66,9 millones de registros de datos comprometidos.

Existe una explosión en el aumento de estas cifras de manera anual desde entonces donde el 2016 alcanza un valor de 1106 brechas, superando el umbral de las 1000 brechas anuales, número del cual no logran caer las cifras hasta ahora. Se registra un pico de 1632 brechas con un número de 197,1 millones de registros filtrados en el 2017 y con un pico de 471,23 millones de registros filtrados con un número de brechas de 1257 brechas en el año 2018. Estas cifras, si bien decaen abruptamente el año 2020 debido al Reglamento General de Protección de Datos de la Unión Europea (GDPR) que entró en vigor en mayo del 2018, no muestran signos de atenuarse a niveles aceptables a la fecha.

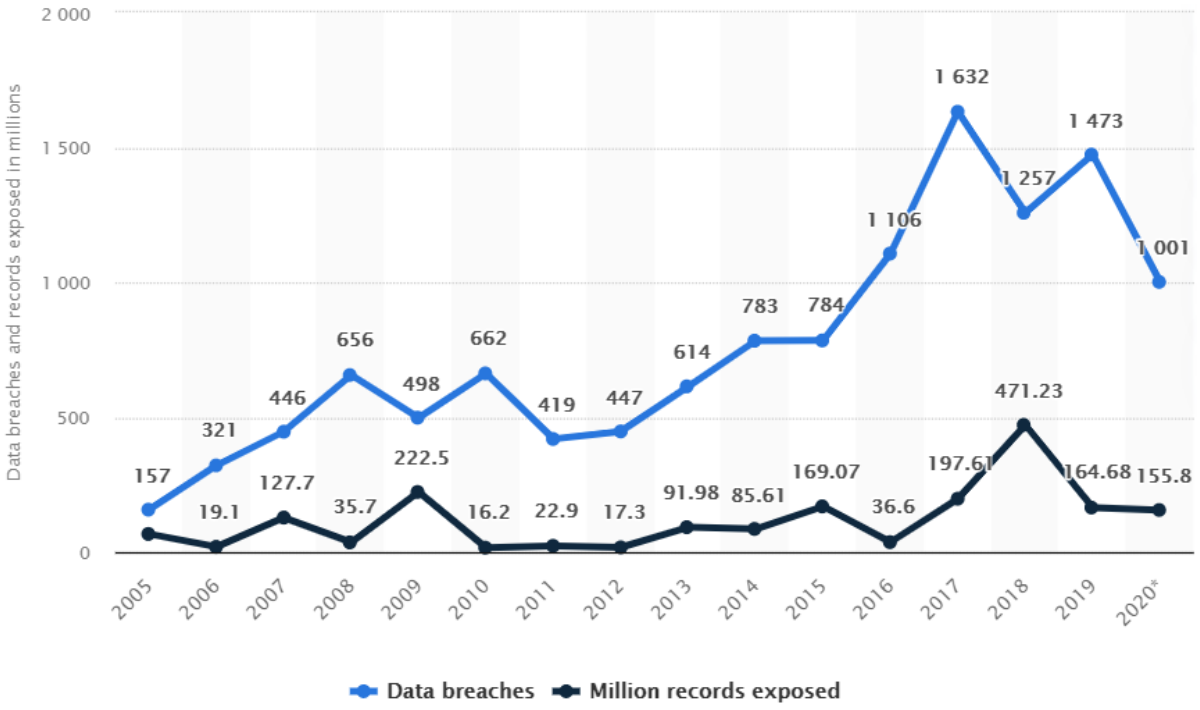

fig.1: Gráfico de número de brechas anuales y registros comprometidos, según catastro de Statista. [4]

Con respecto al aumento desmedido de estas brechas, la legislación en general ha tardado en responder a cabalidad la problemática. Existen muchos tópicos que restan por discutir sobre la ética en la privacidad, como: La responsabilidad por la disponibilidad y la precisión de la información, manejo de reportes en base a frecuencia y priorización de la ubicación de un usuario, la conciencia directa de la permeabilidad del dato de ubicación del usuario final, la libertad del usuario de permanecer o abandonar el servicio, las responsabilidades de quienes prestan servicios como cuidadores y guardianes de esta información, la transparencia sobre las transacciones de datos y la duración del almacenamiento de datos de ubicación [5]. Muchos de estos problemas son tema de discusión en cortes hoy en día, usualmente entre proveedores de estos servicios y usuarios vulnerados.

Pero, ¿Por qué la ubicación importa? Hay tres rasgos de estos datos que los hacen especialmente delicados:

1) La mayor preocupación sobre el uso de datos de ubicación es que puede ser vinculado directamente a un usuario individual. Es por eso que el anonimato es clave. Pero eliminar todos los posibles métodos en el cual un usuario puede ser vinculado e identificado a sus datos de ubicación puede ser técnicamente complejo.

2) El análisis de los datos de ubicación de un individuo puede revelar información personal sensible. Este término es usado por la GDPR para describir información que necesita protección especial. Esto incluye datos que pueden revelar distintas propiedades de una persona: Su etnicidad, su compás político, su religión, creencias filosóficas, su estado de salud, su orientación sexual, etc. esto a través del estudio de los lugares que los individuos frecuentan. Como congregación en iglesias, cercanías a otros individuos ya estudiados, compras en centros comerciales, consultas en centros médicos, etc.

3) Los patrones de información geoespacial acumulados pueden dar información crucial. Incluso si es que es anónima. Por ejemplo, en 2018 a través de la plataforma Strava (cuyo uso es estudiar rutas de trote y ejercicio físico) se triangularon varias ubicaciones de bases militares debido a que los soldados utilizaban aplicaciones de fitness para complementar sus rutinas de ejercicios, sin importar que los datos fueran anónimos. Sólo la frecuencia y los patrones de estos datos, sin la individualización de los usuarios, pueden generar estadísticas de alto valor para organizaciones y compañías, pero pueden tener usos cuestionables.

El acceso a la ubicación es generalmente concedido cuando se aceptan términos y condiciones cuando un usuario instala una aplicación. Muchas de estas aplicaciones no funcionarían sin el uso de los datos de ubicación, no siempre es claro cuál es el uso que se le da a estos datos o de porque los datos son necesarios en primera instancia.

Por ejemplo: Las aplicaciones de mapas usan datos de aplicación para mostrar la ubicación del usuario relativa a un mapa; Uber utiliza datos de ubicación para interconectar pasajeros con conductores; Las aplicaciones de fitness registran y establecen rutas de trote en base a monitoreo de datos de ubicación para ayudar al usuario a llevar una mejor inspección de sus actividades de ejercicio físico.

Pero para las aplicaciones cuyo uso no involucra fundamentalmente el uso de datos de ubicación para su funcionamiento, generalmente el valor del servicio se vuelve debatible. Facebook e instagram llevan un registro constante de la ubicación por defecto [6], incluso cuando las aplicaciones no están en uso directo. El registro persistente de estos datos y la falta de transparencia del uso de estos registros ha levantado las alarmas con respecto a las vulneraciones en la privacidad, tanto en usuarios como en reguladores.

Entonces, ¿Deben los usuarios renunciar al uso de LBS? Restarse del uso de una tecnología ya arraigada en el colectivo sería un esfuerzo vano. Sobre el 90% de los usuarios de dispositivos móviles utilizan aplicaciones ligadas a servicios basados en ubicación [7]. Uber alcanza los 93 millones de usuarios y Waze alcanza los 130 millones de usuarios y existen muchas alternativas competitivas a estos servicios con muchos usuarios más. Renunciar al uso de esta tecnología sería casi imposible. Lo que sí es posible es renunciar a la necesidad de un servidor LBS para el procesamiento de estas consultas. Pues el servidor, al procesar consultas y respuestas y registrar este tráfico, es el real adversario de la privacidad.

3.3 La solución

Para abordar los problemas de privacidad en el uso extensivo de LBS, existen trabajos previos que dividen la problemática en dos enfoques distintos: Enfoque de ubicación [8] y enfoque semántico. [9]. El primer enfoque busca enmascarar los datos de ubicación, es decir, las coordenadas y/o direcciones desde las cuales el usuario genera las consultas al servicio. El segundo enfoque busca enmascarar los conceptos asociados a la consulta efectuada por el usuario.

La gran mayoría de los trabajos que abordan el enfoque de ubicación utilizan la técnica de K-anonimato, que consiste en escoger K-1 ubicaciones, distintas e indistinguibles entre sí. Estas ubicaciones luego son enviadas en consultas distintas junto a la consulta que incluye la ubicación relevante para el usuario, haciendo difuso, de esta manera, la información real que resulta sensible. De manera similar, el segundo enfoque ha sido abordado ampliamente utilizando una técnica llamada ℓ-diversidad. Esta técnica emplea ℓ-1 consultas semánticamente distintas a la relevante y se envían al servidor al mismo tiempo, volviendo difusas las etiquetas semánticas con las que catalogar la consulta. Enmascarando la información sensible.

Este trabajo, generalmente, es llevado a cabo en un servidor de anonimato en una arquitectura centralizada. Como la descrita en la siguiente figura:

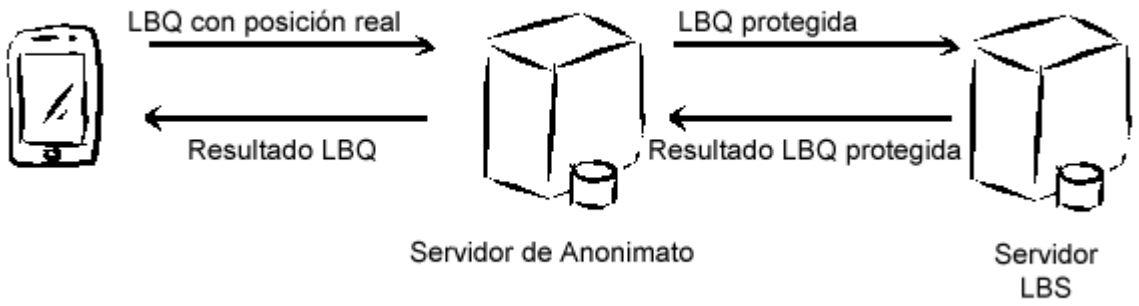

fig.2 : Diagrama representativo del funcionamiento de un servidor de anonimato en una arquitectura centralizada

Sin embargo el emplear este tipo de arquitectura. Conlleva problemas similares al uso directo de un servidor LBS. Como las brechas de datos en el servidor de anonimato. Por lo que en el trabajo paralelo, al cual se busca dar soporte, el alumno Fernando Vera junto al Profesor Patricio Galdames, abordan el problema del procesamiento de LBQs en una red móvil ad-hoc inalámbrica (MANET) que sea consciente de la privacidad de las consultas (Query Privacy). La idea es que los usuarios móviles conformarán primero una MANET y cada vez que deseen resolver una LBQ, intentarán resolverla entre ellos antes de enviarla.

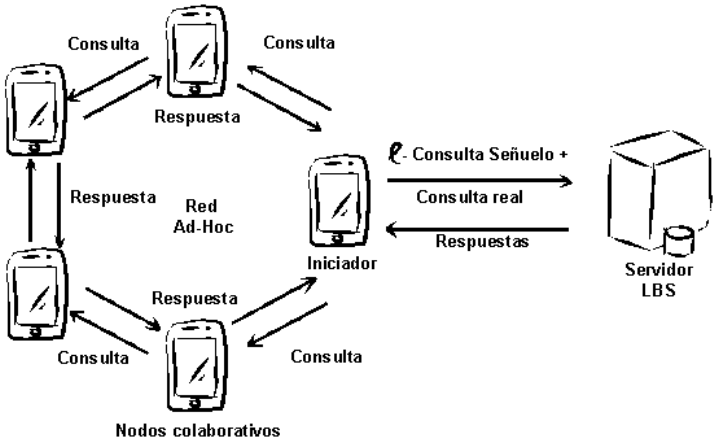

fig.3 Diagrama representativo de la red descentralizada, con posible enmascaramiento de consultas a un servidor LBS.

Para procesar una LBQ en la MANET, el objetivo es definir una estrategia de almacenamiento en caché colaborativa que sea ejecutada por los propios usuarios móviles y que explote las similitudes geográficas y semánticas entre las LBQ. Para proteger la privacidad de la consulta de un usuario cuando un usuario requiere enviar una LBQ al servidor LBS, se planea desarrollar un algoritmo distribuido para proporcionar ' ℓ-diversidad' solo cuando esta protección tenga sentido aplicar. De acuerdo a nuestro conocimiento, las técnicas de almacenamiento en caché existentes para MANET no explotan las similitudes semánticas y geográficas entre los LBQ y asumen que los usuarios móviles tienen acceso a cierta infraestructura de almacenamiento fijo para mantener la información global que permite el cálculo que requiere la ℓ-diversidad.

.Objetivos

.1 Objetivo general

Implementar software de simulación gráfica de un LBS consciente de la privacidad de ubicación y de consulta que opera sobre una red P2P inalámbrica colaborativa empleando técnicas de caché semánticas.

.2 Objetivos específicos

- Estudiar técnicas de caching semántico para redes P2P inalámbricas propuestas en proyecto de tesis en desarrollo del alumno de magíster en ciencias de la computación.
- Implementar un simulador gráfico que simula el comportamiento de usuarios de dispositivos móviles.
- Simular un algoritmo de k-anonimato y l-diversidad para enmascaramiento de consultas basadas en la ubicación que se envían al servidor LBS (Internet).
- Implementar técnicas de caching semántico en una red P2P inalámbrica propuestas en tesis de magíster.
- Proporcionar resultados de métricas de evaluación de técnicas de caching en un formato de archivo adecuado (texto, csv, gnuplot).

.Bibliografía

[1]https://www.nytimes.com/2017/10/03/technology/yahoo-hack-3-billion-users.html (All 3 Billion Yahoo Accounts Were Affected by 2013 Attack, New York Times)
[2]https://www.statista.com/statistics/290525/cyber-crime-biggest-online-data-breaches-worldwide/ (Number of compromised data records in selected data breaches as of January 2021, Statista)
[3]https://www.bbc.com/mundo/noticias-49093124 (Cambridge Analytica: la multa récord que deberá pagar Facebook por la forma en que manejó los datos de 87 millones de usuarios, BBC)
[4]https://www.statista.com/statistics/273550/data-breaches-recorded-in-the-united-states-by-number-of-breaches-and-records-exposed/ (Annual number of data breaches and exposed records in the United States from 2005 to 2020, Statista 2021)
[5]https://ro.uow.edu.au/cgi/viewcontent.cgi?article=1521&context=infopapers (Control, Trust, Privacy, and Security: Evaluating Location-Based Services. Laura Perusco Katina Michael 2007)
[6]https://www.fastcompany.com/40477441/facebook-google-apple-know-where-you-are (How—And Why—Apple, Google, And Facebook Follow You Around In Real Life, Fast Company Tech Business Media)

[7]https://geomarketing.com/overwhelming-number-of-smartphone-users-keep-location-services-open(overwhelming-number-of-smartphone-users-keep-location-services-open, Geomarketing)

[8] Galdames, P., Gutierrez-Soto, C., & Curiel, A. (2019). Batching location cloaking techniques for location privacy and safety protection. Mobile Information Systems, 2019.
[9] Liu, F., Hua, K. A., & Cai, Y. (2009, May). Query l-diversity in location-based services. In 2009 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware (pp. 436-442). IEEE.

[]:https://ico.org.uk/for-organisations/guide-to-pecr/ (Guide to Privacy and Electronic Communications Regulations, Information Comissioner's Office UK)