La Publicación de Trayectorias: un Estudio sobre la Protección de la Privacidad

Patricia Guerra-Balboa
KASTEL Security Research Labs
Instituto Tecnológico de Karlsruhe
patricia.balboa@kit.edu

Àlex Miranda-Pascual
Dept. Ingeniería Telemática
Universidad Politécnica de Cataluña
alex.miranda.pascual@upc.edu

Javier Parra-Arnau
Dept. Ingeniería Telemática
Universidad Politécnica de Cataluña
javier.parra@upc.edu

Jordi Forné
Dept. Ingeniería Telemática
Universidad Politécnica de Cataluña
jordi.forne@upc.edu

Thorsten Strufe
KASTEL Security Research Labs
Instituto Tecnológico de Karlsruhe
thorsten.strufe@kit.edu

***Resumen*—El análisis de las trayectorias encierra numerosas promesas, desde mejoras en la gestión del tráfico hasta recomendaciones de ruta, o incluso en el desarrollo de infraestructuras. Sin embargo, conocer los lugares en los que uno ha estado es extremadamente invasivo. Por ello, surge la necesidad de anonimizar bases de datos de trayectorias, preservando las estadísticas globales útiles para el análisis, mientras que la información específica y privada de los individuos permanece inaccesible.**

En este trabajo analizamos el estado del arte en la publicación de trayectorias con garantías de privacidad, revisando nociones, mecanismos y métricas de utilidad. De este análisis concluimos limitaciones de las propuestas actuales y teniendo en cuenta tanto los problemas de privacidad como los de utilidad, esbozamos oportunidades de investigación para el desarrollo de mecanismos eficaces bajo una protección específica y rigurosa.

***Index Terms*—privacidad de trayectorias, anonimización, nociones sintácticas y semánticas, utilidad, privacidad diferencial**

I. Introducción

Día a día, el valor e interés de los datos de trayectoria se vuelven más notables, no solo en nuestras vidas, sino también entre las empresas de análisis de datos. Al mismo tiempo, la capacidad de los dispositivos personales (como los *smartphones*) y de los sistemas de navegación para recoger, procesar y analizar con precisión estos datos está creciendo a un ritmo nunca visto, gracias a los recientes avances tecnológicos. La gestión del tráfico, la planificación urbanística, el diseño de sistemas de transporte, la predicción de rutas o la seguridad pública son solo algunas de las muchas aplicaciones que se benefician del análisis de trayectorias [1].

A pesar del bien económico y social que supone este análisis, las tensiones relativas a los riesgos para la privacidad son cada vez mayores [2], [3].

Las trayectorias son secuencias de coordenadas espaciotemporales (localizaciones y tiempos). Dada la cantidad de información almacenada en ellas, las trayectorias suponen un gran riesgo de privacidad. Por ejemplo, delatan fácilmente cuándo y durante cuánto tiempo un individuo desarrolla una actividad o visita un lugar, lo que permite a un atacante inferir circunstancias y tendencias que afectan a aspectos privados de su vida, como su estado de salud, sus creencias religiosas, sus relaciones sociales, o sus preferencias políticas o sexuales.

Por otra parte, anonimizar las trayectorias no es tarea fácil, como observaremos en las siguientes secciones. Métricas y técnicas bien conocidas en el campo de la privacidad de datos, como el k-anonimato [4] o la privacidad ϵ-diferencial (ϵ-DP, por sus siglas en inglés) [5], no son aplicables de forma inmediata a estos conjuntos de datos secuenciales y de gran dimensión, y las garantías de privacidad que prometen en el campo a menudo son poco claras.

Asimismo, la singularidad de los desplazamientos humanos hace que, con poco conocimiento previo sobre los objetivos (como su lugar de residencia o trabajo), los adversarios puedan mejorar sus ataques contra los algoritmos de protección [6], [7]. Además, se pueden reconstruir las trayectorias originales utilizando mapas de carreteras, límites de velocidad o modelos simples de correlación espaciotemporal incluso tras aplicar algunos procesos de anonimización, como ofuscaciones. En este contexto, las investigaciones demuestran que conocer solo cuatro puntos espaciotemporales a baja resolución es suficiente para identificar de forma única al $95\,\%$ de los individuos de una base de datos de escala nacional [8].

Además, las nuevas trayectorias podrían seguir siendo "semánticamente" idénticas, de manera que la información sensible del usuario siguiese estando expuesta. Por ejemplo, tras añadir ruido a las coordenadas del usuario encontramos que las nuevas coordenadas siguen estando dentro del mismo *parking* de un centro comercial, luego, pese a la modificación numérica, la semántica sigue idéntica después de la anonimización, con lo que no se ha proporcionado ninguna protección eficaz.

Por último, numerosas aplicaciones de análisis de datos de trayectorias requieren de la publicación continuada y secuencial de datos, como el control y gestión del tráfico a tiempo real. Sin embargo, garantizar la privacidad en este escenario es una tarea desafiante. Los métodos de anonimización sintáctica no pueden ofrecer privacidad en un escenario de actualizaciones y republicaciones de la base de datos, ya que, aunque cada publicación sea, por ejemplo, k-anónima, el acceso al historial de publicaciones permite contrastar y romper la k-anonimidad. La privacidad diferencial, aunque goza de la propiedad de composición y preserva (hasta cierto punto) la garantía de privacidad después de actualizaciones

repetidas de datos, viene lamentablemente al coste de una degradación significativa de la utilidad de los datos [9], [10].

En este artículo, examinamos el estado del arte sobre la publicación de trayectorias con garantías de privacidad y utilidad. Nuestro análisis de la tecnología de anonimización actual abarca las nociones sintácticas y semánticas de privacidad, y se organiza en métricas de privacidad, utilidad, y mecanismos de anonimización. A partir de este análisis, establecemos varios retos, derivados de las ideas y también de las limitaciones de las propuestas existentes en la anonimización de trayectorias, identificando oportunidades para futuras investigaciones.

El resto del artículo se organiza de la siguiente manera. En primer lugar, se presenta el estado del arte de la anonimización de datos de trayectorias. A continuación, se exponen las limitaciones y los problemas identificados en nuestro análisis de la literatura. Por último, se analizan oportunidades y soluciones, y se formulan algunas observaciones finales.

II. Trayectorias y Bases de Datos

Hay diversos tipos de trayectorias. Las más sencillas consisten en una secuencia ordenada de puntos espaciotemporales: $T = (x_1, y_1, t_1) \to \cdots \to (x_n, y_n, t_n)$. Existen representaciones más complejas denominadas *trayectorias semánticas*, en las que se considera adicionalmente la dimensión categórica. En estas, cada punto es un *punto de interés* (PDI), es decir, coordenadas dotadas de un significado semántico, como un nombre o una descripción, y posiblemente otra información como el número de visitantes u horarios de apertura.

Las *bases de datos* de trayectorias consisten en múltiples trayectorias de diferentes individuos (u *objetos en movimiento*) sobre una región común. Sin embargo, existen notables diferencias entre ellas. Algunas bases de datos consisten en trayectorias de igual longitud, algunas de las cuales, además, están recogidas periódicamente (es decir, cada trayectoria tiene un punto cada t tiempo) [11]; mientras que otras son menos regulares, con puntos que solo aparecen cuando el usuario llega (o permanece) en un lugar notable [12].

III. Midiendo la Privacidad y Utilidad

Métricas de privacidad. El objetivo del *control de divulgación estadístico* (SDC, por sus siglas en inglés) es permitir la extracción de estadísticas globales útiles sobre toda la población, pero evitando que se pueda aprender nueva información sobre algún usuario en particular. Existen dos familias conocidas de nociones de privacidad en este campo [13]: las nociones *sintácticas* y *semánticas* [14].

En el caso sintáctico, los representantes clásicos son el k-*anonimato* [4] y sus extensiones (como l-*diversidad* [15] y t-*cercanía* [16]). Se han hecho varios intentos de adaptar estas nociones para los datos de trayectoria. Por ejemplo, se dice que un conjunto de datos satisface (k, δ)-*anonimato* [17] si, para cualquier trayectoria, existen $k-1$ otras trayectorias tales que en cada paso de tiempo las localizaciones correspondientes no están a más de $\delta/2$ de distancia. Asimismo, un conjunto de datos es k^m-*anónimo* [18] si cada subtrayectoria de longitud como máximo m está contenida en al menos k trayectorias diferentes. Otras nociones, como el $k^{\tau,\epsilon}$-*anonimato* [19] o la $(K, C)_L$-*privacidad* [20], consideran y acotan la información adicional que se le permite aprender al atacante.

En el caso semántico, la privacidad diferencial (DP) [5] es probablemente la noción más conocida. Bajo un marco matemático formal, esta noción establece una cota superior ϵ sobre la posibilidad de éxito de un atacante que intenta inferir la información real de un usuario a partir del output del mecanismo. Formalmente, un mecanismo aleatorio $\mathcal{M}$ es ϵ-*diferencialmente privado* (ϵ-DP) [5] si para todo par de bases de datos vecinas D, D' (i.e., que difieren en una única entrada) y todo $\mathcal{S} \subseteq \mathrm{Range}(\mathcal{M})$,

$$\mathrm{P}\{\mathcal{M}(D) \in \mathcal{S}\} \leq \mathrm{e}^{\epsilon} \cdot \mathrm{P}\{\mathcal{M}(D') \in \mathcal{S}\}. \qquad \text{(III.1)}$$

DP es una mera garantía matemática sin ninguna semántica asociada, por lo tanto, es importante especificar exactamente qué información está protegida por ella. La noción original de DP (*user-level*) pretende proteger la existencia completa de los registros de un usuario en una base de datos, es decir, toda contribución de un usuario es indetectable viendo el *output* de un mecanismo DP.

Un gran número de variantes de DP adaptan el concepto *vecindad* de bases de datos: Este concepto determina lo que se considera una única entrada en la base de datos, y, por tanto, que es lo que quedará protegido bajo un mecanismo DP. El *nivel de granularidad* se refiere a esta definición de vecindad en una noción de privacidad. El primer nivel aparece con la privacidad *event-level* [21], [22]. En el mundo de los datos de trayectorias, la privacidad *user-level* protege todo el historial de la trayectoria de cualquier usuario, mientras que la *event-level* protege cada punto espaciotemporal (es decir, un evento). De este modo, el atacante puede saber si un usuario pertenece a la base de datos (ya que cambiar todos los datos de su trayectoria produce un efecto no acotado por ϵ), pero no debería ser capaz de inferir si en un momento determinado estuvo en unas coordenadas o no.

Kellaris *et al.* [23] introducen un punto medio entre de ambas nociones, la privacidad w-*event*, que, en cambio, protege w-ventanas de eventos secuenciales (véase también la Fig. 1). Esta noción se convierte en privacidad *event-level* con $w = 1$ y *user-level* cuando w es la longitud máxima de una trayectoria en la base de datos. De esta forma, el atacante puede seguir infiriendo la presencia de un usuario en la base de datos, pero es incapaz de determinar si una secuencia de w localizaciones pertenece a su trayectoria o no.

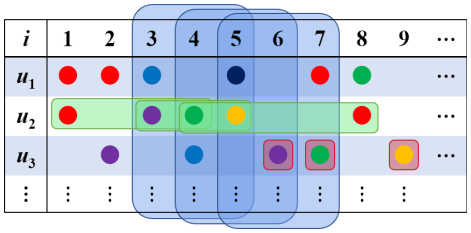

Figura 1. Un ejemplo de una base de datos no recogidas periódicamente, donde los puntos de colores representan diferentes localizaciones. Los recuadros redondeados representan el alcance de la privacidad *event-level* (rojo), w-*event* (azul) y ℓ-*trajectory* (verde), para $w = \ell = 3$. Observe que los cuadros azules (w-ventanas) siempre abarcan w intervalos de tiempo, independientemente de cuántos puntos incluyan, y que los cuadros verdes (ℓ-trayectorias) incluyen siempre ℓ puntos, independientemente del número de intervalos de tiempo que abarcan.

Como mejora de la privacidad w-*event*, Cao y Yoshikawa [12] adaptan la noción de ϵ-DP específicamente para

trayectorias no recogidas periódicamente en el tiempo, obteniendo la privacidad ℓ-*trajectory* (véase la Fig. 1). Aquí se cambia la noción de w-ventana por la de ℓ-trajectoria, siendo esta una secuencia de ℓ localizaciones consecutivas visitadas por un usuario.

Métricas de utilidad. Se han propuesto diversas métricas para cuantificar la utilidad de las trayectorias anónimas. Dado que las técnicas de anonimización siempre llevan asociadas una pérdida de utilidad, uno de los principales objetivos de los mecanismos es minimizar esta pérdida al máximo.

Una forma de medir la utilidad en el número o la proporción de datos que quedan inalterados tras el saneamiento. La *preservación de localizaciones* [24] es un buen ejemplo: Alta utilidad se preserva cuando las trayectorias saneadas incluyen las localizaciones presentes en las trayectorias originales, y no falsas. Asimismo, se puede medir la utilidad como la minimización del número de localizaciones no descartadas.

No obstante, normalmente, mover las coordenadas de la trayectoria unos pocos metros no altera la utilidad. Por lo tanto, la preservación también puede definirse en función de un factor de similitud. Una forma popular de cuantificar este factor es mediante el uso de *métricas de similitud*, funciones que cuantifican la diferencia entre dos trayectorias. Estas incluyen, por ejemplo, las medidas clásicas como la distancia euclidiana [17] o de Hausdorff [11], [25], [26], o EDR [27], entre otras. Por otro lado, en [28] se considera una medida que hace una media geométrica de las distancias espacial, temporal y categórica; teniendo así en cuenta cada una de las dimensiones de las trayectorias semánticas.

Otras métricas menores se basan en la preservación de propiedades específicas, como son la *preservación de la longitud de trayectorias* [29], [30], *secuencias frecuentes* [31] y *lugares más visitados* [29], [28]. Esta información se obtiene mediante funciones de consulta q, y, por lo tanto, se puede medir la preservación usando una *función de error para consultas* (Ec. III.2) [17], [31], [32], [33], [30], que compara los datos anonimizados D' frente a los datos originales D:

$$\text{error}(q) = \frac{|q(D) - q(D')|}{\text{máx}\{q(D), b\}}, \qquad \text{(III.2)}$$

donde b es una cota usada para funciones de consulta extremadamente selectivas.

Finalmente, una última categoría de métricas de utilidad se basaría en asegurar resultados realistas, es decir, que eviten localizaciones consecutivas inalcanzables en el tiempo dado o en lugares geoespacialmente incoherentes [24], [28].

IV. Mecanismos: Logrando Privacidad

En esta sección, examinamos los mecanismos de anonimización de trayectorias más relevantes que cumplen las nociones de privacidad mencionadas. En primer lugar, describimos los mecanismos que ofrecen garantías sintácticas, para después abarcar los que garantizan las nociones semánticas.

Privacidad sintáctica. Existen tres técnicas principales para proporcionar privacidad sintáctica [34]: *supresión*, la eliminación de aquellas localizaciones, o trayectorias enteras, que presentan un riesgo de reidentificación; *generalización*, que hace que los registros sean indistinguibles de otros reduciendo la precisión de las trayectorias o agrupando los datos en grupos más grandes; y el *enmascaramiento (perturbativo)*, que comprende una multitud de técnicas incluyendo la *perturbación* de los datos, basada en la adición de ruido; el *clustering* o *microagregación* de localizaciones; y la *generación de trayectorias falsas*, entre otras muchas. La gran mayoría de las tecnologías de anonimización combina varias de estas técnicas. A continuación, describimos sucintamente los trabajos más relevantes.

El primer mecanismo que utiliza k-anonimato para abordar la anonimización de trayectorias es *Never Walk Alone* (NWA) [17]. Consiste en un algoritmo voraz que agrupa las trayectorias en *clusters* y luego realiza una translación espacial (enmascaramiento) para lograr (k, δ)-anonimato. También suprime las localizaciones atípicas. Los mismos autores introducen posteriormente variaciones mejoradas como W4M [27].

Otros enfoques basados en el enmascaramiento incluyen un método perturbador descrito en [24]. Este método agrupa las trayectorias con microagregación y luego permuta las localizaciones usando la generalización de atributos sensibles y la supresión local. Poulis *et al.* [18] proponen métodos en los que las localizaciones más cercanas se fusionan en pares hasta que se satisface k^m-anonimato. Otros métodos que agrupan y suprimen trayectorias incluyen TOPF [35].

Otra combinación popular es la de las técnicas de supresión con las de generalización: En [36], puntos específicos son eliminados o se sustituyen por regiones en una cuadrícula de celdas, obteniendo trayectorias generalizadas. De forma similar, GLOVE [37] elimina primero todas las localizaciones periféricas, y luego emplea una generalización no uniforme, de modo que cada punto se somete a una reducción mínima para garantizar k-anonimato. Basándose en este, Tu *et al.* [38] introducen los primeros mecanismos que satisfacen k-anonimato, l-diversidad y t-cercanía al mismo tiempo. Otros métodos usando esta combinación aparecen en [39], donde las localizaciones se clasifican en áreas y luego se agrupan.

Por último, Chen *et al.* [20] definen un método basado en la supresión local, que elimina solo algunas instancias del conjunto de datos para garantizar $(K, C)_L$-privacidad, preservando al mismo tiempo los puntos espaciotemporales y las secuencias frecuentes.

Privacidad semántica. A continuación, examinamos los algoritmos de anonimización que buscan publicar bases de datos de trayectorias con garantías de privacidad diferencial.

Los *conteos ruidosos* [32], [31] es un enfoque común basado en añadir ruido al conteo de las trayectorias o secuencias de las mismas. El ejemplo fundamental es [32], que construye un árbol que almacena toda la base de datos. En cada nivel n del árbol se almacenan los conteos de las secuencias de n localizaciones (n-gramas) que se obtienen recorriendo el árbol desde la raíz hasta cada uno de los nodos de dicho nivel. Estos conteos se alteran con ruido a partir de la distribución de Laplace (incluidos los conteos que originalmente eran cero), obteniendo así DP. Además, en [31], se propone un método para generar datos sintéticos a partir de los n-gramas publicados.

Los *mecanismos basados en clustering* constituyen otro enfoque usado en la privacidad de trayectorias [11], [26], [20]. La idea es fusionar localizaciones de diferentes trayectorias en cada tiempo siguiendo una partición probabilística basada en el mecanismo exponencial.

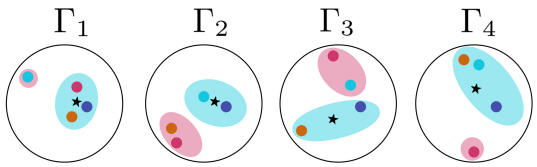

Figura 2. Ejemplo de la anonimización mediante técnicas de *clustering*. Las diferentes trayectorias se representan en diferentes colores, con puntos que corresponden a las localizaciones físicas en cada paso de tiempo. Las áreas coloreadas representan los *clusters* escogidos, y las estrellas denotan sus respectivos centroides. En este ejemplo, las trayectorias tienen longitud $|T| = 4$ y la partición seleccionada contiene $m = 2$ subconjuntos.

Más concretamente, los autores proponen una función de puntuación para medir las distancias entre las trayectorias que cruzan las localizaciones en cada instante de tiempo. Utilizando el mecanismo exponencial y esta función, se elige una de las m-particiones candidatas de Γ_i, el universo de localizaciones en el tiempo t_i. Luego, todas las localizaciones de cada subconjunto se agrupan y se sustituyen por su correspondiente centroide (véase la Fig. 2). Las nuevas trayectorias se construyen reconectando los centroides correspondientes, con su recuento atribuido siguiendo el mecanismo de Laplace.

Por último, Cunnigham *et al.* [28] introducen un mecanismo que *perturba trayectorias semánticas* y satisface ϵ-DP local (ϵ-LDP) [40]. Los autores también encuentran una forma de implementar información pública en el mecanismo para mejorar su utilidad sin afectar al valor ϵ. Utilizan esta información pública para dividir el conjunto de todos los PDI en regiones espacio-tempo-categóricas, de manera que cada una de ellas contenga un cierto número de PDI. Esencialmente, el mecanismo puede dividirse en cuatro partes: en primer lugar, se generaliza cada localización en la región correspondiente; se dividen estas nuevas trayectorias en n-gramas que luego son perturbadas con el mecanismo exponencial para garantizar ϵ-LDP; a continuación, la trayectoria es reconstruida minimizando una función de distancia definida sobre las tres dimensiones; y, por último, el mecanismo vuelve al dominio inicial eligiendo aleatoriamente una localización, tiempo y categoría concreta en cada región, asegurándose de que las localizaciones consecutivas de una trayectoria son alcanzables en el tiempo correspondiente.

V. Limitaciones

En esta sección exploramos las limitaciones presentes en el ámbito de la anonimización de trayectorias.

Como punto principal, debido a la gran complejidad de los datos de trayectoria, no existe ningún mecanismo de anonimización que funcione considerablemente mejor que todos los otros [41]. Además, hay una falta de consenso a la hora de evaluar la privacidad y utilidad de los métodos propuestos. Esto desemboca en dos grandes problemas: primero, dificultad en la comparación de mecanismos; y segundo, evaluaciones sesgadas, ya que diferentes propuestas evalúan sus mecanismos respecto a sus propias métricas, dando así una falsa intuición sobre la protección o utilidad que realmente proporcionan.

En los siguientes apartados exploramos las limitaciones en privacidad y utilidad que presentan los mecanismos. Como veremos, muchas de las limitaciones aparecen a consecuencia de las propiedades de las trayectorias, como son su alta dimensión y correlación, o su naturaleza dispersa.

V-A. *Limitaciones en las Garantías de Privacidad*

Aunque las nociones sintácticas pueden, en general, proporcionar mejores datos de utilidad que DP, son susceptibles de sufrir varios ataques bien conocidos (e.g., *ataques de contraste* o *ataques de vinculación de atributos*). Esto, junto con el hecho de que no son componibles [42], limita la aplicación de la tecnología sintáctica para proteger los datos de trayectoria en contextos dinámicos. Otro problema común en los métodos sintácticos es considerar solamente la dimensión espacial de las trayectorias, como hacen en [17], [39], [35]. Estos son susceptibles de sufrir ataques en las otras dimensiones, ya que estas contienen aún información sensible no protegida.

En el resto de la sección, revisamos las propuestas que se basan en la privacidad semántica. Aunque DP se presentó como una fuerte garantía de privacidad, la noción exacta que proporciona a veces no es clara, como se explica en [43]. Además, al igual que las nociones sintácticas [20], varios trabajos [44], [45], [46] han establecido la debilidad de esta noción cuando las correlaciones entre los atributos en la base de datos son notorias. Desgraciadamente, en los datos de trayectoria, existe un alto grado de correlación dado por su naturaleza (leyes físicas de los movimientos, restricciones de velocidad de las carreteras, comportamiento humano habitual, etc.) [8] y las relaciones sociales de los individuos. En consecuencia, se puede producir una violación de la privacidad, aunque se apliquen mecanismos de DP. Esto se debe a que la noción de DP asume una muestra aleatoria simple i.i.d. como base de datos de entrada. En el caso de trayectorias reales, esta hipótesis no se sostiene, por ello, las garantías totales de DP no pueden ser garantizadas.

Un análisis [12] de las nociones de DP adaptadas a los datos de las trayectorias sugiere que la privacidad *event-level* no es segura. La correlación entre las localizaciones cercanas en el tiempo pueden usarse para atacar fácilmente esta noción. Así mismo, cualquier localización visitada más de una vez estará desprotegida.

En [12] también afirman que la privacidad w-*event* falla porque las trayectorias de los usuarios son dispersas y no están periódicamente recogidas, por lo que podrían no caber en la ventana. Aunque, tanto la privacidad w-*event* como la ℓ-*trajectory* son más robustas que la *event-level* en términos de ataques de correlación, los atributos de un usuario que excedan el marco de la ventana quedarán desprotegidos, por ejemplo, los puntos inicial y final de una trayectoria.

Si nos centramos en propuestas específicas, surge un problema en métodos basados en *clustering* cuando consideramos el modelo de autocorrelación de las trayectorias. Un atacante con un buen modelo de correlación puede descartar la mayoría de uniones falsas de centroides, recuperando así las conexiones originales.

V-B. *Limitaciones en las Garantías de Utilidad*

Esta sección describe varias limitaciones que afectan a la utilidad de los datos. A continuación se exponen las limitaciones generales y, después, las más específicas relacionadas con las métricas y las metodologías.

Problemas generales. En primer lugar, destacamos algunas cuestiones generales que son inherentes a la naturaleza de los datos de las trayectorias. Los mayores desafíos en términos de utilidad aparecen debido al carácter único de las trayectorias, como la diversidad o alta dimensión de este tipo de datos.

Las trayectorias pueden ser muy dispares. Este suceso afecta a las nociones sintácticas como el k-anonimato. Los conjuntos de datos con trayectorias dispersas o cortas (alta unicidad) suponen un gran reto, ya que las trayectorias pueden tener poco solapamiento, lo que conlleva una inevitable pérdida utilidad, pues se necesita una mayor modificación de los datos para conseguir un grupo indistinguible. Del mismo modo, para las nociones semánticas (DP), la diversidad produce sensibilidades elevadas y la necesidad de añadir más ruido para lograr el mismo nivel de protección.

El factor de realismo también toma un papel importante para medir la utilidad. Los mecanismos perturbadores puede crear trayectorias imposibles, con localizaciones inalcanzables o incoherentes. Por ejemplo, los métodos basados en *clustering* [11], [26], [20] pueden dar lugar a nuevas localizaciones que podrían ser ilógicas, como coordenadas sobre edificios o ríos. Más generalmente, Gramaglia *et al.* [19] afirma que uno no puede basarse en datos aleatorios, perturbados o sintéticos para preservar la veracidad de los datos, ya que la adición de datos ficticios introduce sesgos imprevisibles en los datos saneados.

Además, los métodos como los de generalización podrían ser ineficientes para bases de datos de alta dimensión, debido a la *maldición de la dimensionalidad* [47].

Métricas. Tenemos unos cuantos problemas relacionados con la elección de las métricas de utilidad. Medidas de similitud como la distancia euclidiana o de Hausdorff, usadas en [17], [11], [25], [26], no tienen en consideración la coordenada temporal. Por lo tanto, dos trayectorias recorriendo la misma ruta, pero en tiempos diferentes, serán consideradas iguales según estas medidas, lo que claramente escode información y puede limitar su uso. Por ejemplo, no podrían usarse en la predicción de atascos, ya que ignoran factores como el flujo de tráfico.

Adicionalmente, métricas menores, como la *preservación de la longitud* o *de los lugares más visitados*, no deberían usarse en solitario, puesto que pueden devolver buenos resultados para ciertos mecanismos que no preserven los demás aspectos de las trayectorias, como las localizaciones, la forma o el tiempo.

Por último, se debe tener cuidado a la hora de elegir parámetros en algunas métricas, como por ejemplo en la *preservación de secuencias frecuentes* o en la fórmula de error de *queries* de conteos. Si se toma la preservación de secuencias de longitud K, con K grande, su evaluación de la utilidad no va a ser representativa.

Metodologías. A continuación, informamos de las deficiencias en las metodologías del estado del arte en términos de utilidad.

Ni los métodos de *clustering* [11], [26], [20] ni los de *conteos ruidosos* [32], [31] manejan, analizan o protegen la dimensión temporal de las trayectorias. Perdiendo por ende, utilidad en numerosas aplicaciones como es la predicción de atascos y generando patrones extraños como la eliminación de paradas. Asimismo, ambas metodologías generan trayectorias irreales debido a la substitución por el centroide, en el caso de *clustering*, y a la generación de conteos positivos de secuencias que inicialmente no existían, en los *conteos ruidosos*.

Otro problema es la escasa utilidad que pueden ofrecer las aproximaciones de *conteos ruidosos* [32], [31], resultante de asumir implícitamente que las trayectorias contienen un gran número de prefijos y n-gramas comunes. Dado que el proceso añade ruidos a los conteos reales, si los conteos son pequeños, el ruido añadido a cada uno será más grande, con consecuencias fatales en términos de utilidad. Desgraciadamente, las bases de datos reales no siguen esta condición con mucha frecuencia (es decir, no podemos asumir que habrá muchos n-gramas comunes). Además, a causa del coste computacional, estos requieren bases de datos significativamente pequeñas, dificultando aún más su aplicación para bases de datos reales.

VI. Oportunidades

En esta sección, esbozamos posibles líneas de investigación futuras que pueden superar algunas de las deficiencias identificadas en las secciones anteriores. Dadas las limitaciones técnicas de las nociones sintácticas para proteger los datos dinámicos y su debilidad frente a ataques de contraste, consideramos conveniente centrarse en la tecnología de anonimización que ofrece garantías semánticas.

En términos de garantías de privacidad, el principal problema de la aplicación de DP en las trayectorias es que la correlación de datos puede violar el nivel de privacidad prometido. Una posibilidad para hacer frente a esto es adaptar nociones alternativas (basadas en la idea original de DP). Existen ciertas tentativas al respecto, como la propuesta en [46], pero estas no son completamente concluyentes. Se requiere más trabajo para el caso concreto del modelo de correlación presente en las trayectorias.

También es interesante elaborar un nivel de granularidad adecuado para este contexto. *User-level* es un nivel de protección robusto, aunque, no obstante, proteger una participación en la base de datos no siempre es necesario, ya que no suele ser muy sensible (conocer nuestra participación en la base de datos revela tan solo que tenemos un coche y vivimos en una ciudad o país), mientras que la posibilidad de conseguir una utilidad real de mis datos es escasa. Aunque tiene sentido relajar esta noción, ninguna de las granularidades presentadas consigue proteger robustamente los atributos y en ningún caso la identidad. Por ello, una granularidad que proteja todos mis atributos y tenga en cuenta las correlaciones de mi modelo sería deseable.

Por lo general, en el tratamiento de datos, los métodos de *clustering* son una buena opción. Respecto esto, estamos interesados en un algoritmo que tenga en cuenta el tiempo, y creemos que la agrupación de trayectorias enteras o subtrayectorias de las mismas, en lugar de en cada instante de tiempo, podría ser más prolífera en términos de utilidad. Esto también reduciría problemas como las incoherencias temporales de las trayectorias resultantes. Además, métodos de *clustering* que no solo dependen de la dimensión espacial, sino también de la temporal, podrían proporcionar una mayor utilidad, y ayudar a resolver ciertos inconvenientes como la eliminación de paradas.

Para terminar, nos gustaría mencionar ciertas métricas de utilidad que, dependiendo de las necesidades, podrían ser más eficaces para evaluar futuros algoritmos y resultados. Medir la similitud entre la base de datos original y anonimizada puede ser una buena opción si se usa medidas como EDR [48] o la propuesta en [28], que consideran el tiempo. Otras métricas como la *preservación de secuencias frecuentes* pueden ser útiles como métricas secundarias. Por otro lado, es conveniente implementar módulos de post procesado que aseguren un realismo de los datos. Cunningham *et al.* [28] evita la publicación de trayectorias imposibles, detectando cuando dos localizaciones consecutivas están demasiado apartadas para llegar de una a otra en el tiempo dado y corrigiéndolas. Claramente, se podrían incorporar algoritmos análogos a cualquier mecanismo de anonimización para garantizar que todas las trayectorias sean realistas.

VII. Conclusiones

En la primera parte de este artículo se ha analizado los avances actuales en la anonimización de los datos de trayectoria. Hemos examinado cómo se representan estos datos y qué aspectos pueden capturar; y hemos revisado las métricas y los mecanismos de anonimización más relevantes que proporcionan tanto protección sintáctica como semántica. Esta disección de la tecnología actual nos ha permitido profundizar en las limitaciones de las soluciones actuales, en cuanto a las garantías de privacidad prometidas, y la utilidad que queda tras la anonimización. Esta segunda parte de nuestro trabajo ha identificado más específicamente los impedimentos técnicos, por los cuales una parte importante de la tecnología examinada puede no proteger eficazmente la privacidad de los individuos y/o preservar la mayor parte de la utilidad de los datos de la trayectoria.

Agradecimientos

Javier Parra Arnau es beneficiario de una beca de investigación Alexander von Humboldt. Este trabajo también ha recibido el apoyo de la Fundación "la Caixa" (código de beca LCF/BQ/PR20/11770009), del programa H2020 de la Unión Europea (acuerdo de subvención Marie Skłodowska-Curie n.º 847648), del Gobierno de España en el marco del proyecto "COMPROMISE" (PID2020-113795RB-C31/AEI/10.13039/501100011033), y del proyecto BMBF "PROPOLIS" (16KIS1393K). Los autores del KIT cuentan con el apoyo de KASTEL Security Research Labs (Tema 46.23 de la Asociación Helmholtz) y de la Estrategia de Excelencia de Alemania (EXC 2050/1 'CeTI').

Referencias

[1] S. P. Gangadharan, "How can big data be used for social good?" *The Guardian*, 2013.
[2] B. Tarnoff, "Big data for the people: It's time to take it back from our tech overlords," *The Guardian*, 2018.
[3] S. Ovide, "Just collect less data, period." *New York Times*, 2020.
[4] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-Anonymity and its enforcement through generalization and suppression," SRI Int., Tech. Rep., 1998.
[5] C. Dwork, "Differential privacy," in *ICALP*, 2006.
[6] C. Dai *et al.*, "CenEEGs: Valid EEG selection for classification," *TKDD*, 2020.
[7] Y. Yang *et al.*, "TAD: A trajectory clustering algorithm based on spatial-temporal density analysis," *ESA*, 2020.
[8] Y.-A. De Montjoye *et al.*, "Unique in the crowd: The privacy bounds of human mobility," *Sci. Rep.*, 2013.
[9] J. Bambauer, K. Muralidhar, and R. Sarathy, "Fool's gold: An illustrated critique of differential privacy," UArizona, Tech. Rep., 2013.
[10] M. Fredrikson *et al.*, "Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing," in *USENIX Secur.*, 2014.
[11] J. Hua, Y. Gao, and S. Zhong, "Differentially private publication of general time-serial trajectory data," in *INFOCOM*, 2015.
[12] Y. Cao and M. Yoshikawa, "Differentially private real-time data release over infinite trajectory streams," in *MDM*, 2015.
[13] A. Hundepool *et al.*, *Statistical Disclosure Control*. Wiley, 2012.
[14] C. Clifton and T. Tassa, "On syntactic anonymity and differential privacy," *TDP*, 2013.
[15] A. Machanavajjhala *et al.*, "l-diversity: Privacy beyond k-anonymity," *TKDD*, 2007.
[16] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy beyond k-anonymity and l-diversity," in *ICDE*, 2007.
[17] O. Abul, F. Bonchi, and M. Nanni, "Never walk alone: Uncertainty for anonymity in moving objects databases," *ICDE*, pp. 376–385, 2008.
[18] G. Poulis *et al.*, "Apriori-based algorithms for k^m-anonymizing trajectory data," *TDP*, 2014.
[19] M. Gramaglia *et al.*, "Preserving mobile subscriber privacy in open datasets of spatiotemporal trajectories," *INFOCOM*, 2017.
[20] R. Chen *et al.*, "Privacy-preserving trajectory data publishing by local suppression," *Inf. Sci.*, 2013.
[21] C. Dwork, "Differential privacy: A survey of results," in *TAMC*, 2008.
[22] ——, "Differential privacy in new settings," in *SODA*, 2010.
[23] G. Kellaris *et al.*, "Differentially private event sequences over infinite streams," *VLDB Endow.*, 2014.
[24] J. Domingo-Ferrer and R. Trujillo-Rasua, "Microaggregation- and permutation-based anonymization of movement data," *Inf. Sci.*, 2012.
[25] S. Chen *et al.*, "RNN-DP: A new differential privacy scheme base on recurrent neural network for dynamic trajectory privacy protection," *JNCA*, 2020.
[26] M. Li *et al.*, "Achieving differential privacy of trajectory data publishing in participatory sensing," *Inf. Sci.*, 2017.
[27] O. Abul, F. Bonchi, and M. Nanni, "Anonymization of moving objects databases by clustering and perturbation," *IEEE IS*, 2010.
[28] T. Cunningham *et al.*, "Real-world trajectory sharing with local differential privacy," *arXiv preprint*, 2021.
[29] M. Luca *et al.*, "A survey on deep learning for human mobility," *arXiv preprint*, 2020.
[30] M. E. Gursoy, V. Rajasekar, and L. Liu, "Utility-optimized synthesis of differentially private location traces," in *TPS-ISA*, 2020.
[31] R. Chen, G. Acs, and C. Castelluccia, "Differentially private sequential data publication via variable-length n-grams," in *CCS*, 2012.
[32] R. Chen, B. Fung, and B. C. Desai, "Differentially private trajectory data publication," *arXiv preprint*, 2011.
[33] W. Wang *et al.*, "Travel trajectory frequent pattern mining based on differential privacy protection," *JWCMC*, 2021.
[34] T. T. Portela, F. Vicenzi, and V. Bogorny, "Trajectory data privacy: Research challenges and opportunities," in *GEOINFO*, 2019.
[35] Y. Dong and D. Pi, "Novel privacy-preserving algorithm based on frequent path for trajectory data publishing," *KBS*, 2018.
[36] M. Nergiz *et al.*, "Towards trajectory anonymization: A generalization-based approach," *TDP*, 2009.
[37] M. Gramaglia *et al.*, "GLOVE: Towards privacy-preserving publishing of record-level-truthful mobile phone trajectories," *TDS*, 2021.
[38] Z. Tu *et al.*, "Protecting trajectory from semantic attack considering k-anonymity, l-diversity, and t-closeness," *TNSM*, 2019.
[39] A. Monreale *et al.*, "C-safety: A framework for the anonymization of semantic trajectories," *TDP*, 2011.
[40] S. P. Kasiviswanathan *et al.*, "What can we learn privately?" *SICOMP*, 2011.
[41] M. Fiore *et al.*, "Privacy in trajectory micro-data publishing: A survey," *TDP*, 2020.
[42] J. Soria-Comas and J. Domingo-Ferrer, "Big data privacy: Challenges to privacy principles and models," *JDSE*, 2016.
[43] J. Lee and C. Clifton, "How much is enough? choosing ε for differential privacy," in *ISC*, 2011.
[44] Y. Cao *et al.*, "Quantifying differential privacy under temporal correlations," in *ICDE*, 2017.
[45] H. Wang *et al.*, "Why current differential privacy schemes are inapplicable for correlated data publishing?" *WWW*, 2021.
[46] B. Yang, I. Sato, and H. Nakagawa, "Bayesian differential privacy on correlated data," in *MOD*, 2015.
[47] C. C. Aggarwal, "On k-anonymity and the curse of dimensionality," in *VLDB Endow.*, 2005.
[48] L. Chen, M. T. Özsu, and V. Oria, "Robust and fast similarity search for moving object trajectories," in *SIGMOD*, 2005.