Anonimización de trayectorias por medio de intercambios

Julián Salas
Internet Interdisciplinary Institute (IN3)
Universitat Oberta de Catalunya (UOC)
Center for Cybersecurity Research of Catalonia (CYBERCAT)
Barcelona, Spain
jsalaspi@uoc.edu

David Megías
Internet Interdisciplinary Institute (IN3)
Universitat Oberta de Catalunya (UOC)
Center for Cybersecurity Research of Catalonia (CYBERCAT)
Barcelona, Spain
dmegias@uoc.edu

Vicenç Torra
School of Informatics
University of Skövde
Skövde, Sweden
vtorra@his.se

***Resumen*—El estudio de datos de movilidad puede ayudar a mejorar la toma de decisiones, así como planear los transportes en áreas metropolitanas o localizar los servicios en una ciudad. Sin embargo, si los datos geo-localizados son asociados a un individuo específico, pueden poner en en riesgo su privacidad, al permitir conocer todas sus localizaciones, sus horarios y revelar sus hábitos. Por estos motivos, antes de publicar datos de movilidad (trayectorias) es importante anonimizarlas, es decir, garantizar que no puedan ser asociadas a individuos específicos. En este trabajo, presentamos un método de anonimización perturbativa, que funciona de un modo similar al *rank swapping*, preserva la información agregada de los datos de movilidad, tales como el número de vehículos que viajan de un punto de interés a otro, durante un periodo específico de tiempo, y su vez, mantiene las identidades de los individuos y sus puntos de interés protegidos.**

Index Terms*—anonimización de trayectorias, privacidad in GIS/ datos espaciales, intercambio de trayectorias, *swapping

I. Introducción

Los servicios basados en la localización (*Location-Based Services*, en lo sucesivo, LBS) han sido posibles y van en aumento gracias a las técnicas de localización como son GPS, GSM o RFID. Pueden ser utilizados también en sistemas de transporte inteligente, por ejemplo, usar los vehículos como sensores para recoger información sobre la temperatura, las condiciones viales o los atascos de tráfico.

Por otra parte, los registros de localización comportan ciertos riesgos para la privacidad de los usuarios. Si su identidad real puede ser inferida en base a dichos datos, las preferencias, los hábitos e incluso los horarios de los individuos podrían ser revelados. Por este motivo, se han desarrollado diversos métodos de anonimización para proteger la identidad de los usuarios y sus ubicaciones exactas. El *survey* [1] revisa algunos ataques, garantías y transformaciones de los datos para proteger la privacidad en los LBS.

La privacidad en la minería de datos sociales tiene que ser protegida para promover la participación y garantizar los derechos de los sujetos de los datos. Algunos usos de la minería de datos sociales son: comprender el comportamiento humano a través de descubrir perfiles sociales individuales y de analizar el comportamiento colectivo, cf. [2]. Así como estudiar la propagación de epidemias, el contagio social y la evolución del sentimiento y la opinión.

La minería y el análisis de datos espacio-temporales es útil para publicidad, o para hacer políticas públicas, de salud o de movilidad. Al mismo tiempo, este tipo de datos puede revelar localizaciones sensibles, tales como la direcciones de las casas, los trabajos, o los sitios que los individuos visitan y no quieren hacer públicos porque esto revelaría sus preferencias privadas (como pueden ser políticas, religiosas o sexuales).

I-A. Literatura relacionada

Diversas soluciones han sido propuestas para anonimizar trayectorias antes de ser publicadas. Abul et al. [3], propusieron el modelo de (k, δ)-anonimato, que consiste en publicar un volumen cilíndrico de radio δ que contenga la trayectoria de un objeto en movimiento. Esta idea es una extensión del conocido concepto de k-anonimato para bases de datos [4].

Terrovitis y Mamoulis [5] consideran un dominio espacial discreto. En este caso, las trayectorias de los usuarios son expresadas como secuencias de POIs. Usan como ejemplo las tarjetas RFID de la compañía Octopus en Hong Kong[1], que recoge las historias de las transacciones de sus clientes, y está interesada en publicar las secuencias de transacciones que ha hecho una misma persona para extraer sus patrones de movimiento y de conducta. Sin embargo, si un usuario usa su tarjeta para pagar en distintas tiendas que pertenecen a la misma cadena, dicha compañía podría reidentificarlo si su secuencia de compras es única en la base de datos de trayectorias que ha sido publicada. En este caso la trayectoria se puede representar como una trayectoria en un grafo en el que los nodos representan todas las direcciones de las tiendas que aceptan tarjetas Octopus.

En [6] obtuvieron una aproximación similar que preserva los patrones secuenciales frecuentes (*frequent sequential patterns* [7]) por medio de transformaciones que agregan, eliminan o sustituyen algunos puntos de las trayectorias.

En [8] y [9], Hoh et al. se aborda el uso de datos de movilidad para planificación de transportes y para que aplicaciones que monitorizan el tránsito puedan proporcionar información sobre las condiciones del tránsito y del asfalto a los conductores. Para modelar las amenazas a la privacidad en dichos conjuntos de datos, asumen que un adversario no tiene información de cuales muestras pertenecen a un único usuario, aunque también asumen que usando algoritmos de

[1] http://www.octopuscards.com/

rastreo de objetos múltiples [10], las muestras posteriores se pueden vincular a un individuo que reporta periódicamente la información de su localización.

En [9] consideran el ataque de deducir las ubicaciones de las casas de los usuarios, por medio de heurísticas de clúster. Proponen técnicas de eliminación de datos como cambiar la frecuencia de muestreo (e.g., de 1 minuto a 2,4 y 10) para prevenir dichas inferencias. Mientras que en [8] proponen un algoritmo para perturbar las trayectorias de diversos individuos, haciéndolas parecer más cercanas para evitar que los atacantes descubran las trayectorias completas usando algoritmos de rastreo de objetos múltiples. Esto se lleva a cabo con una restricción de calidad, expresada como el error promedio entre la ubicación publicada y la ubicación real. Además, argumentan que se pueden garantizar niveles adecuados de privacidad si la densidad de usuarios es suficientemente alta.

Este trabajo está relacionado con las *mix-zones* que introdujeron en [11]. Las *mix-zones* son áreas en las que la localización de los usuarios no es accesible a las aplicaciones, de manera que cuando varios usuarios están presentes simultáneamente en una *mix-zone*, sus seudónimos pueden ser intercambiados para dificultar la asociación de una trayectoria que ha entrado a la zona con una que ha salido. El diseño que utilizan para proteger la privacidad de la ubicación intenta conservar las ventajas de los LBS y al mismo tiempo mantiene ocultas las identidades de los usuarios a las aplicaciones que reciben sus peticiones.

Para prevenir la reidentificación trivial, las comunicaciones entre los usuarios y las aplicaciones pasan por un intermediario de confianza y el seudónimo de un usuario cambia cuando cruza una *mix-zone*.

Para medir la privacidad en la ubicación, Beresford y Stajano [12] definen los conjuntos de anonimato como el conjunto de personas que visitan la *mix-zone* durante el mismo periodo de tiempo. Dado que la frontera y el tiempo en el que un usuario sale de una *mix-zone* está fuertemente correlacionado con la frontera y el tiempo en el que el usuario entra, esta información puede ser utilizada por un atacante, por este motivo, adaptan una métrica de teoría de la información que originalmente había sido propuesta por Serjantov y Danezis [13] para comunicaciones anónimas y considera las probabilidades de entrada y salida de mensajes a través de una red de *mix-nodes*.

Modelan estas probabilidades con una matriz de movimiento en la que se registran las frecuencias de entrada y salida de cada *mix-zone* en diversos momentos, y definen un grafo bipartito y con pesos, en el que los nodos representan los seudónimos de entrada y de salida y los pesos las probabilidades de que dos seudónimos representen a la misma persona. De esta manera un emparejamiento máximo en dicho grafo representa el mapeo más probable entre seudónimos de entrada y de salida.

Finalmente describen un método para llegar a soluciones parciales con la restricción de poder de computo, en un equilibrio entre la tratabilidad del problema y la precisión con la que se modela.

Otra aproximación que considera un sistema con un servidor no-confiable y clientes que se comunican en una red P2P para colectar trayectorias preservando la privacidad, fue propuesto en [14]. El objetivo es preservar el anonimato de los datos al guardarlos, transmitirlos o colectarlos. Esto se obtiene usando k-anonimato e intercambiando datos, sin embargo, aunque en cada fase los datos son anónimos, finalmente el servidor filtra los datos sintéticos y obtiene las trayectorias originales.

Una de las ventajas de que los usuarios sean quienes llevan a cabo la anonimización es que el proceso no es centralizado y los sujetos obtienen control, transparencia y más seguridad, como por ejemplo en [15].

II. Intercambio de trayectorias

En esta sección presentamos un método de intercambio de trayectorias para anonimizar datos de movilidad. Este método funciona de manera similar a las *mix-zones* pero en un espacio no delimitado, es decir, los usuarios que están cercanos pueden intercambiar sus trayectorias parciales sin necesidad de estar adentro de una *mix-zone*.

Aunque consideramos que el algoritmo se puede aplicar en tiempo real y P2P, en este trabajo definiremos una primera alternativa en la que los datos se anonimizan después de haber sido recogidos.

Al intercambiar sucesivamente las trayectorias parciales, estas se mezclan y cuando la trayectoria generada es recuperada por el servidor está formada por pequeñas fracciones de trayectorias de diversos individuos que se han encontrado durante el día, como en la Figura 1.

De este modo la relación entre los sujetos y sus datos se ofusca y al mismo tiempo se mantienen los datos agregados sin modificar, tales como el número de usuarios en un momento y una ubicación determinados.

En la siguiente sección presentamos definiciones y las suposiciones necesarias para formalizar nuestro método.

II-A. Definiciones

Suponemos que tenemos una tabla en la que la i-ésima observación es una tupla (ID_i, lat_i, long_i, t_i) que consiste en el identificador del individuo ID_i, la latitud (lat_i), la longitud (long_i) y el registro de tiempo (t_i).

La trayectoria T_x de un individuo x consistirá de todas las observaciones con identificador $ID_i = x$ ordenadas por sus registros de t_i. También se puede representar como $T_x = (x_1, x_2, \ldots, x_m)$ si hay m observaciones para el individuo x.

Diremos que dos individuos *se encuentran* o que sus trayectorias se cruzan en los puntos x_i y y_j, si han estado co-localizados, y lo denotamos con $x_i \approx y_j$. Estar *co-localizados* significa que han estado aproximadamente en el mismo lugar en el mismo momento, por lo tanto, estar co-localizados depende de los parámetros de proximidad χ y tiempo τ.

Definimos un *emparejamiento* como un subconjunto máximo de parejas de elementos de un conjunto

Denotamos por $Sw(T)$ a la trayectoria que resulta después de haber aplicado todos los intercambios a la trayectoria original T.

A continuación, definimos dos primitivas para nuestro algoritmo: *generar emparejamiento aleatorio* e *intercambiar*.

1. *Intercambiar (Swap):*
   Dadas dos trayectorias $T_x = (x_1, \ldots, x_i, x_{i+1}, \ldots)$ y $T_y = (y_1, \ldots, y_j, y_{j+1}, \ldots)$ que se encuentran en los puntos x_i y y_j, un swap de T_x

con T_y en los puntos x_i y y_j tiene como resultado $Sw(T_x) = (y_1, \ldots, y_j, x_{i+1}, \ldots)$ y $Sw(T_y) = (x_1, \ldots, x_i, y_{j+1}, \ldots)$.

2. *Generar emparejamiento aleatorio:* Dado un conjunto de elementos $S = s_1, s_2, \ldots, s_m$, generamos un emparejamiento aleatorio haciendo parejas entre los primeros $m/2$ y los siguientes $m/2$ números, después de haber hecho una permutación de todos los números m.

En caso de que el número de elementos m sea impar, para generar un emparejamiento es necesario dejar fuera un elemento al azar.

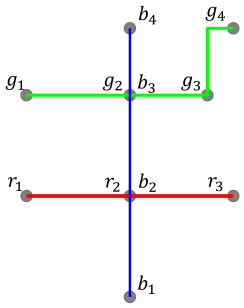
II-A1. Intercambio de trayectorias que se cruzan: Consideramos que las trayectorias de dos usuarios se cruzan si han estado colocalizados de acuerdo con los parámetros de proximidad χ y tiempo τ.

A continuación, simulamos el protocolo P2P en el que los usuarios intercambian sus IDs cuando están suficientemente cerca uno del otro.

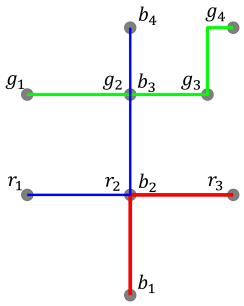
Calculamos los usuarios que están en contacto en un intervalo de tiempo determinado, y definimos un emparejamiento aleatorio entre ellos. El intercambio se lleva a cabo por parejas, sin embargo, podría haberse definido por una permutación como en [12].

Notemos que intercambiar seudónimos (IDs) es equivalente a intercambiar las trayectorias parciales.

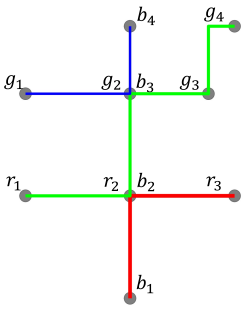
En la Figura 1, presentamos un sencillo ejemplo de tres trayectorias que se cruzan T_r, T_g, T_b. Suponiendo que se mueven de izquierda a derecha y hacia arriba, $T_r = (r_1, r_2, r_3)$, $T_g = (g_1, g_2, g_3, g_4)$ y $T_b = (b_1, b_2, b_3, b_4)$. También estamos suponiendo que la trayectoria azul se encuentra con la trayectoria roja primero ($b_2 \approx r_2$), y después con la verde ($b_3 \approx g_2$). En este pequeño ejemplo, podemos ver como los múltiples intercambios preservan partes de las trayectorias intactas, aunque al final cada trayectoria contiene partes de las otras con las que se ha cruzado, como por ejemplo la trayectoria verde que termina con un segmento de la azul, un segmento de la roja y un segmento de la trayectoria verde original, es decir, $Sw(T_g) = (r_1, r_2, b_3, g_3, g_4)$.



(a) Trayectorias originales



(b) Después del primer intercambio



(c) Después del segundo intercambio

Figura 1: Tres trayectorias antes y después del intercambio

II-B. Protección contra reidentificación

Es bien sabido que la de-identificación no necesariamente significa anonimización. Los mismos atributos que se utilizan para extraer conocimiento, pueden ser usados para encontrar a un individuo específico y distinguirlo únicamente, relacionando sus datos con su identidad.

En el contexto de datos geo-localizados, en [16], se mencionan los siguientes ataques a la privacidad de los individuos. Al identificar los POIs de un individuo, es posible inferir sus hábitos (e.g., hace deporte cada mañana, hace viajes largos en verano), las ubicaciones que visita frecuentemente y pueden estar relacionadas con sus creencias políticas o religiosas, o incluso con problemas de salud (como pueden ser clínicas y hospitales). También es posible deducir sus horarios, predecir sus ubicaciones y conocer sus ubicaciones anteriores, además de deducir sus relaciones personales observando con quien comparten frecuentemente sus ubicaciones. Asimismo, dichos hábitos y localizaciones pueden ser fácilmente utilizado para reidentificar a los individuos detrás de los datos, por ejemplo, los pares de ubicaciones casa/trabajo.

Golle y Partridge estudiaron en [17] individuos que revelaron las ubicaciones de sus casas y trabajos, con un ruido o una granularidad del tamaño de una manzana, un kilómetro o decenas de kilómetros (bloque censal, área censal o condado) y mostraron que los tamaños de los conjuntos de anonimato son respectivamente 1, 21 y 34980. Es decir, cuando la granularidad de los datos es del orden de un bloque censal, los individuos son únicamente identificables por sus ubicaciones de casa/trabajo, y para granularidades de área censal o condado, los individuos comparten el par de

Algorithm 1: algoritmo offline para intercambiar trayectorias

Input: Base de datos de trayectorias. Parámetros de tiempo τ y proximidad χ.
Output: identificadores de las trayectorias después de los swaps $Sw(T_i)$.
Partición de los registros de tiempo $t = \bigcup \tau_j$ en intervalos de longitud τ
for *Cada par de registros* i, j *en el intervalo* τ_j **do**
  if $dist(l_i, l_j) < \chi$ **then**
    agregar i, j a la lista de registros cercanos S_{τ_j} en el intervalo de tiempo τ_j
  end
end
generar un emparejamiento aleatorio con los registros en S_{τ_j}
ordenar todos los swaps en $\bigcup S_{\tau_j}$ por tiempo
for *para cada par* $i \approx j$ *en* $\bigcup S_{\tau_j}$ **do**
  intercambiar T_i con T_j
end
return *trayectorias intercambiadas* $Sw(T_i)$

ubicaciones casa/trabajo con 20 y 34979 individuos.

Zang and Bolot en [18], infieren las primeras N localizaciones de un usuario usando un registro de llamadas y las correlacionan con información accesible públicamente como datos del censo. Muestran que las primeras 2 localizaciones probablemente corresponden con la ubicación de casa y del trabajo, y posteriormente, que los conjuntos de anonimato se reducen drásticamente usando dichas localizaciones.

Por tanto, es crucial proteger las direcciones de casa y del trabajo de los usuarios para protegerlos contra la reidentificación y ofrecerles mínimas garantías de permanecer anónimos.

En nuestro caso, las trayectorias intercambiadas no permitirán seguir a un individuo específico y los sitios que visita, de manera que no permitirán ofrecerle información personalizada o clasificarlo, sin embargo, esto también es una forma de proteger su privacidad.

Nuestro algoritmo preserva el anonimato al disociar los segmentos de trayectorias del sujeto que los ha generado.

Aún en el caso que un atacante supiera algunos puntos que identifican a un individuo, no podría acceder a la trayectoria completa, puesto que la trayectoria publicada está formada por trayectorias de varios individuos

III. Evaluación

Hemos probado nuestro algoritmo en el conjunto T-drive de [19], [20], que contiene las trayectorias que han realizado 10,357 taxis del 2 al 8 de febrero de 2008 en Pekín. El número total de puntos es alrededor de 15 millones. Aunque, no todos los taxis aparecen cada día ni reportan sus posiciones con la misma regularidad. En promedio las reportan en un intervalo de alrededor de 177 segundos y 623 metros.

Asumimos que dos taxis han estado co-localizados si han estado a una distancia de menos de 111 metros ($\chi = 0{,}001$) en un periodo de 1-minuto ($\tau = 60$).

III-A. Reidentificación usando POIs y la ubicación de casa

En esta sección repasaremos métodos para inferir los POIs de un individuo, tales como la ubicación de su casa. Para probar que hemos protegido correctamente la ubicación de casa, compararemos la ubicación inferida antes y después de aplicar nuestro algoritmo.

Recordemos que uno de nuestros objetivos es proteger las ubicaciones relacionadas con los hábitos de los sujetos, y al mismo tiempo mantener los patrones agregados de movimiento.

Otro objetivo principal es proteger la asociación de las trayectorias a individuos específicos.

Como habíamos indicado anteriormente, si un atacante consiguiera asociar una trayectoria anonimizada con nuestro método a un individuo, no descubriría su trayectoria real, puesto que la trayectoria publicada estaría formada por segmentos de varias trayectorias. Por otra parte, aunque el atacante conozca algunos puntos reales y los intente utilizar para reidentificar a alguien, tendrá que encontrar una trayectoria anonimizada que también los contenga después del proceso de intercambio.

Para inferir los puntos de interés, en [21] consideran que, si un individuo permanece a menos de 200 metros de un punto determinado durante al menos 20 minutos, entonces es un punto de interés. En este caso, los parámetros para detectarlos serían distancia $\chi = 200$ metros y tiempo $\tau = 1200$.

En este trabajo, para obtener puntos de interés, discretizamos el espacio en celdas de 111 metros (0.001 grados decimales) y contamos las celdas más pobladas por cada individuo.

Definimos un punto de interés al centro una celda, si esta tiene al menos el 50 % de registros que la celda más poblada.

Analizando los tiempos, observamos que estos POIs también son POIs en el sentido de [21], aunque su definición podría considerar más puntos de los que hemos detectado.

Para inferir, la ubicación de la casa de un individuo, hemos seguido las ideas de [22] que discretizan el espacio en celdas de 25×25 km, y definen la ubicación de casa como la ubicación promedio de la celda más poblada. Ver también [23].

Igual que para los POIs, consideramos los puntos dentro de una celda de $t = 0{,}001$ grados de lado, es decir, aproximadamente 111 metros. Asignamos la ubicación de casa a la celda que tiene más puntos usando este parámetro.

Para validar que funciona nuestro método para deducir la ubicación de casa, generamos el histograma de las horas en las que cada individuo está cerca de la ubicación deducida de casa. Al analizar el histograma (Fig. 2), podemos observar que hay como mínimo alrededor de 5 % del total de taxis que están en (la ubicación deducida de) casa durante el día y este número incrementa consistentemente a partir de las 20 pm hasta las 6 am, que es cuando llega al máximo y decrece hasta alrededor de 5 % a las 10 am.

Este comportamiento es consistente con el comportamiento general de la población respecto a las horas que están en casa. Además, si asumimos que aquellas ubicaciones con más de 5 % de frecuencia relativa, son las ubicaciones correctas de casa, obtenemos que el porcentaje de ubicaciones correctas de casa es el 83,7 % del total, el cual es muy similar al que

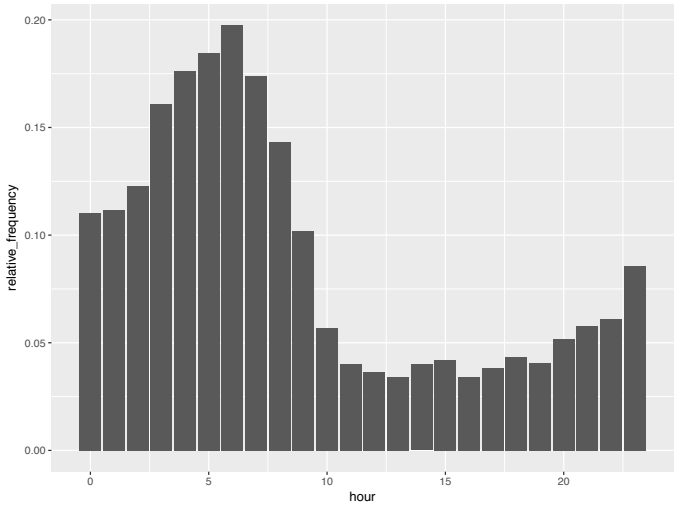


Figura 2: Histograma de las horas en que los taxis están en la localización deducida de sus casas, como porcentaje del total de mediciones para cada hora.

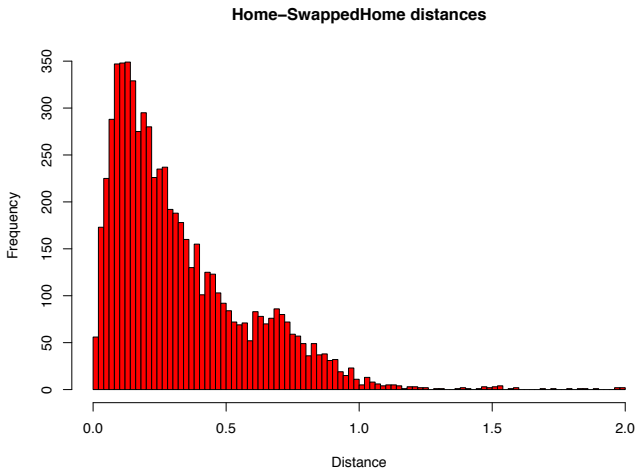


Figura 3: Histograma de distancias en grados decimales (i.e., $\times 111{,}32$ km) entre la ubicación de casa deducida antes y después de aplicar nuestro algoritmo.

obtuvieron para su conjunto de datos en [22] por inspección manual (85 %).

Por ejemplo, usando este método deducimos que el usuario 7828 estuvo en casa desde las 00:13:44 a las 09:11:44 en la fecha 2008-02-03 (33 registros consecutivos). Posteriormente, desde las 23:32:01 en la fecha 2008-02-03 hasta las 10:11:59 en la fecha 2008-02-04 (70 registros consecutivos), llegó a casa a las 02:30:57 el día siguiente (55 registros) y salió a trabajar a las 10:10:56, y así sucesivamente. Así, podemos obtener no solo la ubicación de casa, sino también los hábitos o el horario de diversas personas.

III-B. Ubicación de casa después de los swaps

Para comprobar los resultados de nuestro método, intentamos deducir la ubicación de casa de los usuarios antes y después de aplicar nuestro algoritmo.

Considerando que la ubicación de casa es el POI más relevante y usando la técnica mencionada en la sección anterior para inferir POIs como los sitios más frecuentados, podemos concluir que nuestro método sirve para proteger no solamente la ubicación de casa, sino también los otros POIs.

En la Figura 3 mostramos el histograma de las distancias entre la ubicación de casa deducida antes y después de aplicar nuestro algoritmo.

Hemos obtenido que los usuarios para los cuales la distancia entre la ubicación de su casa deducida antes y después de aplicar nuestro algoritmo es menos de 0.001 grados son el 7674, 9197, 1, 533, 8205, 6590, 9223, 4801 y el 8427, las distancias respectivas en metros son 1.37, 10.24, 10.72, 17.48, 22.77, 26.98, 65.76, 74.8 y 106.34.

Únicamente los usuarios 1, 7674 y 9197 no han cambiado la ubicación de casa antes y después de intercambiar las trayectorias. Estos usuarios son un $0{,}02\,\%$ del total de usuarios.

El usuario 1 no se ha movido de casa durante el 2008-02-03 hasta las 18:35:14, por lo tanto, no había intercambiado su trayectoria a las 18:05, que es la hora que definimos para enviar las trayectorias intercambiadas al servidor. La misma situación se puede observar para los usuarios 7674 y 9197. Sin

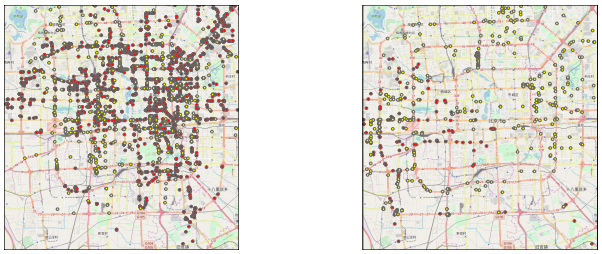


Figura 4: Puntos en las trayectorias originales y después de los swaps para los usuarios 7674 and 9197.

embargo, sus trayectorias han cambiado mucho después de los intercambios, como puede observarse en la Figura 4, en que la trayectoria original está representada en rojo y la modificada en amarillo. Por lo tanto, no hay una manera aparente en la que un atacante pudiera tener alguna certeza de que la ubicación de la casa que ha deducido sea correcta, además, hemos visto que acertaría con una probabilidad de a lo más $0{,}02\,\%$ en este ejemplo.

IV. Conclusiones

En este trabajo definimos un algoritmo para anonimizar trayectorias que no modifica los datos agregados de ubicación y tiempo. Preserva el número de individuos que hay en cada sitio, en cada momento, y modifica las trayectorias al intercambiar segmentos entre diversos usuarios. Para probar que las casas y los POIs son protegidos por nuestro algoritmo, utilizamos una heurística de clúster (basada en la frecuencia de visitas) para deducir las ubicaciones de las casas de los individuos en un conjunto con 15 millones de registros. Posteriormente, comprobamos que después de aplicar nuestro algoritmo de intercambio de trayectorias, no es posible deducir correctamente las ubicaciones de casa y otros POIs de los usuarios. Queda como trabajo futuro probar empíricamente otras medidas de utilidad y riesgo de los datos agregados.

Una posible herramienta para calcular la utilidad y el riesgo de reidentificación, será representar las trayectorias como un

"grafo de cruces", en el que los nodos representarán cuando se han cruzado dos o más individuos en una posición y un periodo de tiempo concretos y las aristas tendrán pesos que representen la cantidad de individuos que recorren el mismo trayecto entre dos cruces. Podemos adelantar que, usando esta herramienta, cualquier localización que no corresponda a un nodo en el grafo de cruces podría ser utilizada para identificar la trayectoria, puesto que será única. Además, sería interesante comparar la intersección de las trayectorias originales con las publicadas, para saber si aún en el caso de identificar una trayectoria, está podría no ser muy similar a la original según la distancia de Hamming o alguna otra métrica. El grafo de cruces que definiremos, también nos servirá para dar una aproximación al número de trayectorias que viajan de un sitio a otro.

Agradecimientos

Este trabajo está financiado parcialmente por el Ministerio de Economía y Competitividad a través del proyecto TIN2014-57364-C2-2-R "SMARTGLACIS". El tercer autor está financiado por el proyecto "Disclosure risk and transparency in big data privacy"(VR 2016-03346, 2017-2020). El primer autor agradece el apoyo de una beca postdoctoral UOC.

Referencias

[1] M. Terrovitis, "Privacy preservation in the dissemination of location data," *SIGKDD Explor. Newsl.*, vol. 13, no. 1, pp. 6–18, Aug. 2011. [Online]. Available: http://doi.acm.org/10.1145/2031331.2031334

[2] F. Giannotti, D. Pedreschi, A. Pentland, P. Lukowicz, D. Kossmann, J. Crowley, and D. Helbing, "A planetary nervous system for social mining and collective awareness," *The European Physical Journal Special Topics*, vol. 214, no. 1, pp. 49–75, Nov 2012. [Online]. Available: https://doi.org/10.1140/epjst/e2012-01688-9

[3] O. Abul, F. Bonchi, and M. Nanni, "Never walk alone: Uncertainty for anonymity in moving objects databases," in *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, ser. ICDE '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 376–385. [Online]. Available: http://dx.doi.org/10.1109/ICDE.2008.4497446

[4] P. Samarati and L. Sweeney, "Generalizing data to provide anonymity when disclosing information (abstract)," in *Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, ser. PODS '98. New York, NY, USA: ACM, 1998, pp. 188–. [Online]. Available: http://doi.acm.org/10.1145/275487.275508

[5] M. Terrovitis and N. Mamoulis, "Privacy preservation in the publication of trajectories," in *Proceedings of the The Ninth International Conference on Mobile Data Management*, ser. MDM '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 65–72. [Online]. Available: https://doi.org/10.1109/MDM.2008.29

[6] R. G. Pensa, A. Monreale, F. Pinelli, and D. Pedreschi, "Pattern-preserving k-anonymization of sequences and its application to mobility data mining," in *PiLBA*, 2008. [Online]. Available: https://air.unimi.it/retrieve/handle/2434/52786/106397/ProceedingsPiLBA08.pdf$\#$page=44

[7] R. Agrawal and R. Srikant, "Mining sequential patterns," in *Proceedings of the Eleventh International Conference on Data Engineering*, ser. ICDE '95. Washington, DC, USA: IEEE Computer Society, 1995, pp. 3–14. [Online]. Available: http://dl.acm.org/citation.cfm?id=645480.655281

[8] B. Hoh and M. Gruteser, "Protecting location privacy through path confusion," in *Proceedings of the First International Conference on Security and Privacy for Emerging Areas in Communications Networks*, ser. SECURECOMM '05. Washington, DC, USA: IEEE Computer Society, 2005, pp. 194–205. [Online]. Available: http://dx.doi.org/10.1109/SECURECOMM.2005.33

[9] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady, "Enhancing security and privacy in traffic-monitoring systems," *IEEE Pervasive Computing*, vol. 5, no. 4, pp. 38–46, Oct 2006.

[10] D. B. Reid, "An algorithm for tracking multiple targets," *IEEE Transactions on Automatic Control*, vol. 24, pp. 843–854, 1979.

[11] A. R. Beresford and F. Stajano, "Location privacy in pervasive computing," *IEEE Pervasive Computing*, vol. 2, no. 1, pp. 46–55, Jan. 2003. [Online]. Available: http://dx.doi.org/10.1109/MPRV.2003.1186725

[12] ——, "Mix zones: user privacy in location-aware services," in *IEEE Annual Conference on Pervasive Computing and Communications Workshops, 2004. Proceedings of the Second*, March 2004, pp. 127–131.

[13] A. Serjantov and G. Danezis, "Towards an information theoretic metric for anonymity," in *Proceedings of the 2nd International Conference on Privacy Enhancing Technologies*, ser. PET'02. Berlin, Heidelberg: Springer-Verlag, 2003, pp. 41–53. [Online]. Available: http://dl.acm.org/citation.cfm?id=1765299.1765303

[14] G. Gidófalvi, "Spatio-temporal data mining for location-based services," Ph.D. dissertation, Faculties of Engineering, Science and Medicine Aalborg University, Denmark, 2007.

[15] C. Romero-Tris and D. Megías, "User-centric privacy-preserving collection and analysis of trajectory data," in *Data Privacy Management, and Security Assurance: 10th International Workshop, DPM 2015, and 4th International Workshop, QASA 2015, Vienna, Austria, September 21–22, 2015. Revised Selected Papers*, J. Garcia-Alfaro, G. Navarro-Arribas, A. Aldini, F. Martinelli, and N. Suri, Eds. Cham: Springer International Publishing, 2016, pp. 245–253. [Online]. Available: https://doi.org/10.1007/978-3-319-29883-2_17

[16] S. Gambs, M.-O. Killijian, and M. Núñez del Prado Cortez, "Show me how you move and i will tell you who you are," *Trans. Data Privacy*, vol. 4, no. 2, pp. 103–126, Aug. 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=2019316.2019320

[17] P. Golle and K. Partridge, "On the anonymity of home/work location pairs," in *Proceedings of the 7th International Conference on Pervasive Computing*, ser. Pervasive '09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 390–397. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-01516-8_26

[18] H. Zang and J. Bolot, "Anonymization of location data does not work: A large-scale measurement study," in *Proceedings of the 17th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '11. New York, NY, USA: ACM, 2011, pp. 145–156. [Online]. Available: http://doi.acm.org/10.1145/2030613.2030630

[19] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang, "T-drive: Driving directions based on taxi trajectories," in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ser. GIS '10. New York, NY, USA: ACM, 2010, pp. 99–108. [Online]. Available: http://doi.acm.org/10.1145/1869790.1869807

[20] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "Driving with knowledge from the physical world," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '11. New York, NY, USA: ACM, 2011, pp. 316–324. [Online]. Available: http://doi.acm.org/10.1145/2020408.2020462

[21] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining interesting locations and travel sequences from gps trajectories," in *Proceedings of the 18th International Conference on World Wide Web*, ser. WWW '09. New York, NY, USA: ACM, 2009, pp. 791–800. [Online]. Available: http://doi.acm.org/10.1145/1526709.1526816

[22] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: User movement in location-based social networks," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '11. New York, NY, USA: ACM, 2011, pp. 1082–1090. [Online]. Available: http://doi.acm.org/10.1145/2020408.2020579

[23] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo, "Socio-spatial properties of online location-based social networks." *ICWSM*, vol. 11, pp. 329–336, 2011.