Estudi de la privacitat en dades de mobilitat: cas d'ús Swap Mobility Location

David Matos Xancó

Resum

Resum– En un món on l'ús de la tecnologia és gairebé imprescindible, els usuaris es veuen forçats a cedir dades de caràcter personal a tercers per a poder fer-ne ús. Aquestes, es publiquen a bases de dades públiques, generant així la necessitat de protegir la intimitat dels usuaris davant possibles atacs, mantenint, a l'hora, la integritat de les dades que seran aprofitades en estudis posteriors. Així mateix, l'objectiu principal d'aquest projecte és realitzar un estudi de l'actual estat de l'art en els mètodes de preservació de la privacitat en la publicació de les dades, fent èmfasi en les dades de les localitzacions dels usuaris, i implementar un dels mètodes estudiats posant solució al problema plantejat i estudiant els resultats.

Paraules clau– Dades, privacitat, trajectòries, emmascarament, anonimat, protecció, localització.

Abstract– In a world where the use of technology has become almost essential, users has been forced to relinquish / cede personal data to others so that they can make use of it. This data is published in public database, generating the necessity of protecting the intimity of the users in front of posstible attacks, mantaining, at the same time, the integrity of the data that will be resourceful in later studies. In this way, the main objective of this project is to carry out a study of the current state of Art in the preservation methods of the preservation of the privacy in the publication of data, emphasising in the data of users location and implementing one of the studied methods, drawing a solution to the considered problem and studying the results.

Keywords– Data, privacy, trajectories, mask, anonymity, protection, location.

1 Introducció

En l'actualitat, per a poder fer ús de les tecnologies, ja sigui en estudis mèdics, en investigacions demogràfiques, etcètera, necessitem acceptar obligatòriament l'entrega de les nostres dades personals a diferents empreses o organitzacions. Aquestes, ja sigui per interès mutu o per regulacions que ho exigeixen, es veuen en la obligació de publicar i compartir les dades que recullen. Això pot tenir efectes positius sobre els mateixos usuaris que ofereixen les dades personals, ja que molts cops se'n treu profit, però en ocasions pot arribar a posar en risc la privacitat dels mateixos. Recentment s'han vist casos relacionats amb el tractament incorrecte d'aquestes bases de dades, les quals han estat publicades sense un correcta procés d'anonimització i ha estat possible el reconeixement directe de diferents usuaris trencant, per tant, la privacitat d'aquestes persones. Arran d'aquest problema i possibles atacs a la intimitat dels usuaris, sorgeix la necessitat de trobar un mètode d'anonimitzar aquestes dades sensibles per tal que els usuaris no puguin ser reconeguts. Així doncs, les empreses que tracten aquest tipus de dades es veuen en la obligació de protegir els usuaris davant possibles atacs dirigits a l'extracció d'informació útil a partir de les dades personals, fent ús de diferents tècniques i mètodes.

• E-mail de contacte: davidignacio.matos@e-campus.uab.cat
• Menció realitzada: Tecnologies de la Informació
• Treball tutoritzat per: Jordi Casas Roma
• Curs 2019/20

2 Objectius

Tenint en compte la importància de l'anonimització d'aquestes dades de cara a la preservació de la privacitat dels usuaris, es plantegen els següents objectius:

- En primer lloc, estudiar l'estat de l'art en els mètodes de preservació de la privacitat en la publicació de les dades emfatitzant en les dades de localitzacions dels usuaris.
- Triar un mètode estudiat a l'estat de l'art i implementar-lo en llenguatge Python.
- Seleccionar un conjunt de dades públiques que contingui dades referents a la localització d'usuaris.
- Estudiar l'alteració de les dades en el procés d'anonimització, donant especial atenció al nivell de privacitat i de utilitat de les dades anonimitzades.

3 METODOLOGIA

Aquest projecte es pot dividir clarament en dues parts diferenciades. Un primer pas d'estudi de l'estat de l'art, i abast del projecte i una segona fase en la que s'escull un mètode a i el conjunt de dades a tractar, i s'implementa conjuntament amb un conjunt de proves per tal de determinar la privacitat, i utilitat del conjunt de dades que s'ha anonimitzat. Tenint aquesta estructura en compte, la metodologia en cascada és la més adequada.

4 PLANIFICACIÓ

Per tal de poder dur a terme el projecte dins els terminis que s'estableixen, s'haurà de fer una prèvia planificació que es seguirà exhaustivament 6.

- Primera fase - Inici: En la primera fase es farà una planificació del desenvolupament del treball i una descripció del problema juntament amb la descripció dels objectiu i metodologia.

- Segona fase - Documentació: Es fa un estudi de l'estat de l'art per a conèixer els diferents mètodes existents que posen en dubte la privacitat de les dades de trajectòria, així com les tècniques per a protegir els usuaris.

- Tercera fase - Implementació del model: S'escull un model dels anteriors i s'implementa per tal de veure el seu funcionament amb un conjunt de dades simulades.

- Última fase - Tancament: Realització de la memòria escrita i de la presentació final.

5 DEFINICIÓ DEL PROBLEMA

Les nostres vides cada cop es veuen més envoltades amb components on la tecnologia de la comunicació és omnipresent. Des trucar a algú per telèfon o enunciar qualsevol esdeveniment, fins navegar per la xarxa, són exemples de situacions en les que sense ser conscients, els usuaris deixen enregistrades dades que poden ser extretes més endavant per a diferents propòsits i beneficis de diferents persones. Una de les característiques que es destaca de les noves tecnologies, i que s'estudiarà en profunditat, és que sovint depenen d'una base de dades que inclou microdades sobre trajectòries. Aquestes microdades contenen informació personal sobre els individus i els seus moviments, les quals descriuen trajectòries en un espai i temps, és a dir, posicions geogràfiques dels usuaris. A continuació s'exposen 5 exemples de microdades basades de trajectòries per tal d'exemplificar la situació:

- **Serveis basats en ubicació:** Són aplicacions executades en dispositius mòbils que carreguen les dades sobre la posició d'un usuari, segons sigui necessari per a complir amb el servei. Exemples d'aplicacions serien Google Maps o Instagram.

- **Operadors de xarxes de telèfon:** Fan un monitoratge passiu a les seves xarxes per a recopilar dades sobre l'activitat dels seus usuaris amb fins que inclouen facturació, trànsit de dades o desarrelament de serveis en els que s'hagi d'afegir valor. Podem trobar rastres de la ubicació en l'antena del telèfon, el registre de trucades o el registre del servei a la xarxa del propi usuari.

- **Dispositius mòbils equipats amb interfícies Wi-Fi:** Aquests estan constantment enviant missatges per tal de descobrir punts d'accés propers. Aquests punts, coneguts com AC, registren l'adreça MAC dels dispositius que emeten aquestes ones. Alguns exemples els trobem als nostres dispositius mòbils els quals poden seguir en gran mesura els moviments dels usuaris.

- **Sistemes moderns de navegació:** Aquest servei de navegació proveeix als usuaris d'informació en temps reals sobre l'estat de les carreteres i les seves condicions, però també permet recopilar dades sobre el posicionament del vehicle. Aquestes dades són usades pels proveïdors de sistemes de navegació i per companyies de assegurances per tal de poder perfilar perfils de conducció i nivells de riscos associats.

- **Pagaments electrònics:** Aquest tipus de pagament permet a les empreses del sector bancari monitoritzar els moviments dels clients a mesura que usen les targetes de crèdit.

Aquests són exemples on es pot veure com les tecnologies permeten la recopilació de microdades de trajectòries a gran escala. Aquesta informació permet a la construcció de grans bases de dades que emmagatzemen aquestes trajectòries. És per això que s'obté un gran interès en explotar aquestes microdades per conèixer el perfil dels usuaris creixent en un mercat multimilionari emergent. Es crea així, una necessitat totalment nova de recopilar, emmagatzemar, i comercialitzar microdades de trajectòria. En aquest tràfic de dades cal aplicar processos d'anonimització per a preservar la intimitat dels usuaris i garantir la seguretat de la informació privada. Les dades són publicades per a extreure informació útil, però aquesta pot ser robada donats atacs, usant mètodes i algoritmes de la mineria de dades que busquen patrons en les dades que els permetin entendre l'estructura de les dades i fer prediccions futures sobre el comportament dels usuaris, per part de terceres persones El procés d'anonimització de les dades té implícit el concepte de modificació de les pròpies dades, aquesta farà que la seva utilitat en processos o estudis posteriors sigui menor o de menys qualitat, pel que s'ha de trobar un equilibri entre l'anonimització i la usabilitat de les dades.

S'ha d'entendre el risc que es presenta, ja que un atacant podria recopilar informació suficient per a re-identificar un usuari dins una base de dades que conté atributs sensibles, com per exemple la ubicació del lloc de treball, habitatge, o llocs freqüentats durant rangs determinats d'horaris, entre altres. Per tant, si l'atacant tingués èxit podria saber informació privada sobre els llocs que els usuaris freqüenten com llocs de culta que revelin les preferències religioses, locals de caràcter eròtic o sexual, etcètera.

6 MODELS TEÒRICS DE PRIVACITAT

Existeixen dos grans models per a limitar el risc de divulgació en els processos de publicació de les dades. Els podem

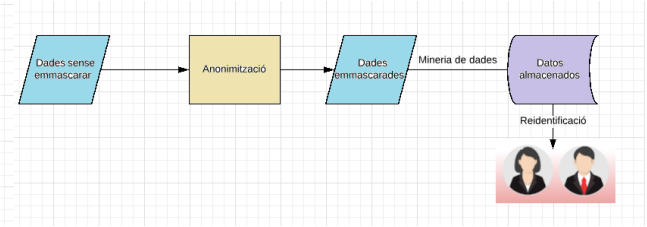


Fig. 1: Mineria de dades. [21]

classificar de la següent manera:

- **Protecció no interactiva:** Es genera i publica una versió protegida d'un conjunt de dades original. Es fa servir majoritàriament quan es vol fer un anàlisi de dades que són desconegudes en el moment de la publicació Figura2a.

- **Protecció interactiva:** Genera una versió protegida que es retorna en el moment que un usuari faci una consulta a una base de dades amb una finalitat analítica. Normalment es fa servir les dades prèviament conegudes Figura2b.

Dins aquests grans blocs, trobem els següents mètodes destacables:

1. **Pseudo-anonimització:** aquest model de protecció no interactiu va ser el primer model pensat per protegir, d'alguna manera, les dades dels usuaris. Consisteix principalment en eliminar els identificadors personals de la base de dades i substituir-los per algun identificador pseudoaleatori. Addicionalment es pot afegir soroll aleatòries en la resta d'atributs del conjunt, especialment en els quasi-identificadors. Un atacant no podrà saber de forma exacta si l'informació que obté al entrar a la base de dades és l'original o ha estat alterada. El principal problema d'aquest model és que el fet d'afegir massa soroll a les dades originals podria anular la seva pròpia utilitat, i per tant s'estableixen dos principis que el soroll ha de complir:

    - Que sigui suficient per a que un atacant no pugui saber si les dades han estat modificades o no.
    - L'informació general del conjunt s'ha de preservar amb l'objectiu que anàlisis futurs s'aproximin el màxim possible als fets amb dades originals.

2. k**-anonimitat:** aquest és un altre model teòric de protecció no interactiu. És una propietat de les dades que ens assegura que un individu no podrà ser identificat de la resta de k-1 individus existents a la base de dades tal i com s'explica a l'article [4]. La principal avantatge que experimenta aquest model és que un atacant no podrà identificar la seva víctima amb una probabilitat superior a $\frac{1}{k}$. És per això que s'ha de tenir en compte que com més alt sigui el valor de k més augmentarem la privacitat, i per contrapartida més reduirem la utilitat de les dades.

3. **Privacitat diferencial:** el model de privacitat diferencial està fet per a la protecció interactiva enfocada a bases de dades estadístiques, és a dir, consultes a bases de dades. En aquest context el mecanisme d'anonimització es troba entre l'usuari i la base de dades. Per tal d'assegurar que els usuaris estaran totalment protegits, aquest model defensa que el fet d'afegir o eliminar un individu d'un cert conjunt de dades no ha de suposar una alteració en els resultats d'una consulta a la base de dades. Aquesta afirmació és basa en el fet que si al afegir un conjunt de dades, d'un usuari en concret a una base de dades aquesta pateix una alteració suficients com per ser apreciada per un atacant, aquest usuari seria fàcilment identificable i per tant estaria en risc la seva privacitat. Per exemple, suposem que es té un conjunt de dades que es diferencia en només un element d'un altre conjunt de dades, i que per un altre banda, tenim un algoritme renderitzat del tipus ϵ-diferencial. Aquest element és el que regula directament la quantitat de diversitat que es pot trobar en el resultat d'una consulta a la base de dades quan s'elimina o afegeix un individu. Per tant, la privacitat diferencial assegura que el coneixement que un atacant pot adquirir estarà limitat segons aquest paràmetre. És per això que la privacitat diferencial és una condició que es troba en el mecanisme de publicació i no en el conjunt de dades en sí.

7 Anonimització de localitzacions i dades temporals

S'entén per dades temporals el conjunt de dades formades principalment per parelles, referents al temps i referents a la ubicació (coordenades). Donant informació sobre les dimensions espai i temps d'un usuari. La unió del conjunt de ubicacions en les dades temporals d'un usuari forma la trajectòria del mateix. Per a preservar la privacitat dels usuaris es recullen diferents tècniques en dos grans grups:

- **Tècniques d'anonimització de punts espai-temporals**, que tenen en compte únicament la localització del moment actual.

- **Tècniques d'anonimització de trajectòries:** Que tenen en compte una successió de punts al llarg del temps

Les tècniques d'anonimització que s'han estudiat són proposades per tal de minimitzar l'impacte o reduir en cert nombre els atacs que es poden donar i que posen en risc la privacitat en relació a la localització d'un usuari. Es diferencien dos grans principis dins l'estudi:

- Indistingibilitat

- Desinformació

A més d'aquests dos principis es troben diferents treballs que adopten nocions menys rigoroses de la privacitat, els quals fan una menció més general d'aquest terme i que s'agrupen en un tercer:

- Mitigació

En les següents seccions es fa una descripció dels principis de privacitat més destacables i es destaquen les tècniques

(a) Protecció interactiva: Les dades no es publiquen, s'ofereixen als usuaris després de la seva petició.

(b) Protecció no interactiva: Es genera una versió protegida de les dades que s'entrega a tots els usuaris.

Fig. 2: Protecció interactiva vs protecció interactiva

emprades més importants per a complir amb aquests principis. Tota base de dades té uns criteris de privacitat que vol protegir, i cada principi s'especialitza en algun d'aquests criteris. En aquest treball s'estudia els principis que busquen protegir la privacitat de les dades espai-temporals dels usuaris.

7.1 INDISTINGIBILITAT

Recomana que cada registre d'una base de dades sigui indistingible d'un subconjunt de registres existents a la mateixa base de dades, eliminant així la unicitat [1]. Aquest objectiu s'aconsegueix implementant k-anonimitat o alguns dels seus mètodes derivats com ℓ-diversity a o t-closeness a [18] i [20]. La idea és que un grup de punts espai-temporals de cada usuari en les microdades de trajectòria no sigui distingible per almenys k-1 altres usuaris en la mateixa base de dades.

Les solucions proposades es basen en les diferents variacions del model de k-anonimitat, per tant, s'ha de tenir en compte que encara no s'ha determinat quin és el millor valor de k per a preservar la utilitat de les microdades de trajectòria. En segon lloc, s'ha de tindre en compte que aquest model ofereix solucions únicament pels atacs del tipus enllaç de registres.

7.1.1 K-ANONIMITAT VIA GENERALITZACIONS ESPAI-TEMPORALS

És la tècnica base usada per aconseguir la k-anonimitat en les microdades de trajectòria. Tracta de reduir l'exactitud espacial, així com la granularitat temporal dels punts de les trajectòries que es troben a la base de dades, ocultant els punts d'una trajectòria amb altres punts d'altres trajectòries. Seguint aquesta filosofia, com s'ha explicat anteriorment, es pot arribar a l'extrem de perdre suficient exactitud en els registres de les dades i que aquestes quedin inservibles. Podem veure un exemple d'actuació a la Figura 3 .

Els investigadors Zang i Bolot proposen a [23] una primera opció basada en el decrement de la unicitat mentre que la granularitat espacial de la trajectòria també ho fa, però apel·la a la dificultat de fer-ho donat que si un atacant sabés les 3 localitzacions més visitades per part d'un usuari seria inútil intentar trobar la 2-anonimitat ja que s'hauria de publicat una quantitat de dades realment petita. En un posterior estudi [12] , a més, discuteixen sobre la dificultat d'aplicar k-anonimitat, explicant que en una base de dades prou gran es té moltes localitzacions comunes entre tots els usuaris però que sempre hi haurà localitzacions úniques que serà molt difícil d'ocultar. A més, demostren en el mateix estudi com introduint una $k > 2$ es perdria utilitat en les dades anonimitzades. A partir d'aquestes observacions, els autors proposen una tècnica anomenada GLOVE, un algoritme que aconsegueix finalment la k-anonimitat de les microdades de les trajectòries via generalitzacions d'espai-temporals. Explica com en el seu mètode s'aplica la reducció de la granularitat a cada punt de forma individual, a diferència de com s'havia estat fent fins el moment, on s'aplicava al conjunt de punts de la trajectòria. Basant-se en aquests fets, els mateixos autors defineixen una mètrica creada per ells mateixos, *fingerprint stretch effort* que quantifica la pèrdua necessària de granularitat per tal d'ocultar cada mostra d'una trajectòria amb la mostra més propera d'una altre trajectòria. La millora d'aquesta tècnica, permet anonimitzar una base de dades de desenes de milers de registres, conservant registres fins a 1 kilòmetre de la trajectòria inicial en la dimensió espacial i 1 hora en la temporal, demostrant que es perden menys dades a mesura que s'incrementa la base de dades.

7.1.2 K-ANONIMITAT VIA SUPRESSIÓ

Aquesta tècnica, es basa en esborrar els punts espai-temporals de la trajectòria original. Diferents estratègies han estat presentades, les més destacables com la proposa un algoritme que elimina iterativament alguns punts de les trajectòries fins a satisfer la k-anonimitat tal i com es diu a [19]. El funcionament és el següent: per cada iteració, els punts que trenquen la k-anonimitat són detectats, i el que comporta una distorsió Euclediana mínima és seleccionat per a ser esborrat. Aquest fet comporta diverses hipòtesis sobre atacs i format de dades:

- Les trajectòries són purament espacials, no tenen la dimensió del temps.
- La dimensió de l'espai es dividirà en un número definit de localitzacions.
- Els possibles atacants seran menys, i tota la possible informació que poden tindre és anonimitzable.

7.1.3 K-ANONIMITAT VIA GENERALIZACIÓ I SUPRESSIÓ

Els investigadors Gramaglia i Fiore [9] destaquen que la supressió serà beneficiosa per a la k-anonimització donat que, descartant alguna petita fracció de punts únics determinats, s'eliminarà un conjunt de trajectòries on el nivell de precisió era prou alt. Fan referència també, a la impossibilitat de generalitzar trajectòries que tenen un nombre diferent de

[1] Mesura que caracteritza la diversitat de moviments d'un mateix individu. A major unicitat de les microdades de trajectòria major probabilitat que un atacant pugui relacionar informació d'una víctima

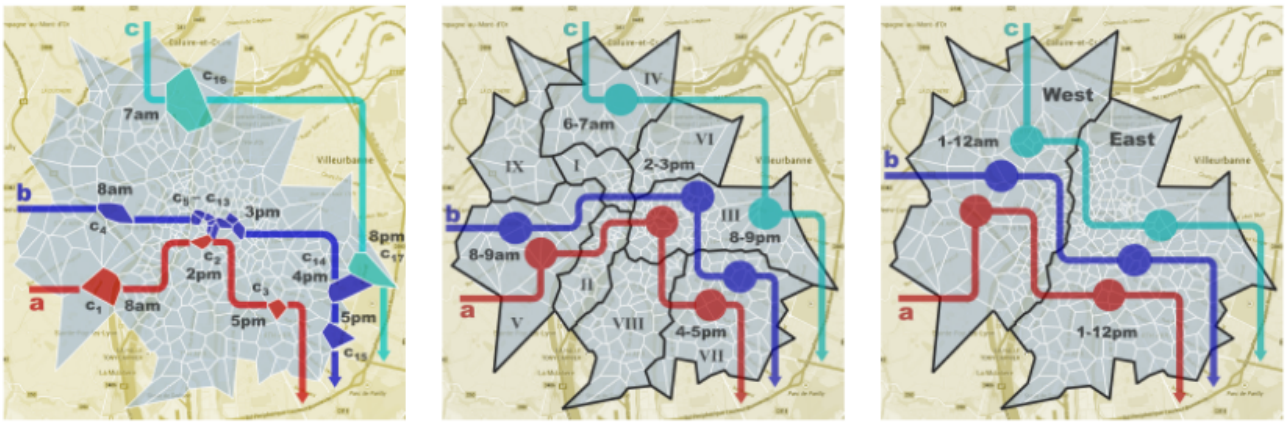


Fig. 3: k-anonimitat via generalització espai-temporal. Les rutes a la imatge de l'esquerra són totalment exactes, per tant, per cada imatge es van generalitzant amb els punts d'altres rutes fins aconseguir trajectòries que passin per punts similars, eliminant també altres microdades sensibles com l'hora. Imatge extreta de [1]

punts, és a dir, no es pot realitzar aquest mètode ja que comportaria generalitzar dos o més punts d'una trajectòria amb tant sols un punt d'una altre. Per altre banda, es basen en una mètrica de similitud de parelles de trajectòries diferent [2]. Aquesta mètrica de cost de registre i mostra com pot aconseguir 2-anonimització eliminant tant sols el 2 o 3% de les dades, mentre que anteriorment hauria comportat una eliminació d'un 25%, com a mínim. No ofereixen resultats concrets, però mostren com els resultats en *clustering* conserven una precisió del 50-90%. KAM_CUT i KAM_REC són tècniques proposades posteriorment. La primera és usada per a grans conjunts de dades i crea una estructura d'arbres de trajectòries on els nodes pare representen les més comunes i els fill les menys comunes, però més completes, dels mateixos usuaris. A partir d'aquest arbre, i triant una determinada k, s'eliminen les branques amb menys de k trajectòries compartides. Kam_rec és una extensió de la primera per a conjunts de dades més petits, i intenta reinserir punts de les sub trajectòries que s'han eliminat buscant la seva subseqüència de punts més llarga amb la condició que apunti a alguna trajectòria que es trobi encara en l'arbre , o que sigui compartida per almenys una altre sub trajectòria que ha estat eliminada.

7.1.4 K-ANONIMITAT VIA MICRO AGREGACIÓ I SUPRESSIÓ

Per explicar la micro agregació cal mencionar que consta de dos passos principals:

- **Partició:** En aquest primer pas s'agrupen les trajectòries inicials que són més semblants, de forma que els clústers que es generen al final tindran una cardinalitat igual a K.
- **Agregació:** Les trajectòries d'un clúster seran substituïdes per un prototip de clúster, calculat per un operador sobre els punts espai-temporals del mateix clúster. Fent això s'aconsegueix k-anonimitzar el conjunt de dades fent que les trajectòries en el clúster siguin igual al prototip.

La introducció de noves mètriques de similitud entre parelles, com per exemple la mètrica de distancia sincronitzada[6], entre trajectòries són introduïdes com a millores de l'anterior. "SwapLocation" [15] és una tècnica coneguda per l'ús d'aquesta similitud entre parelles, Per a cada trajectòria en un clúster canvia tots els punts espai-temporals per punts d'altres trajectòries en el mateix clúster. Aquest canvi ha de respectar els llindars de temps i espai definits en un principi. A més, si un punt no té possibilitat de realitzar cap canvi amb un altre degut als llindars, es suprimeix. En els resultats que mostren els autors, es fan servir dades reals i sintètiques per una k=10 hi ha una important supressió : amb un llindar espacial de 1 km s'obté una eliminació = 50% de les trajectòries i el 80% dels punts. Si s'augmenta el llindar a 3km es redueix la supressió a un 5% i els punts es mantenen. En el cas de dades reals i una K=2, el 29% dels punts són eliminats i la distorsió espacial se situa en 2.4 km.

7.2 DESINFORMACIÓ

La desinformació indica que el guany en termes de quantitat que un atacant obtindrà després de realitzar un atac ha de ser molt petit. Aquest principi té com a objectiu la protecció contra atacs probabilístics, i usa la privacitat diferencial com a criteri estàndard per tal de poder complir-lo. El primer cas proposat fa referència a l'ús d'un model Laplacià [8] per a afegir soroll a una sortida en forma de vector escalar. Aquest, proposa un segon cas en el que la sortida es dóna en forma de distribució de probabilitats a través d'un conjunt de resultats, aconseguint privacitat diferencial a través de renderitzar les probabilitats fent ús d'un mecanisme exponencial. En les dues situacions exposades s'obtindrà com a resultat un vector amb soroll, el qual vindrà donat per un element ϵ així com la diferència màxima entre totes les sortides possibles quan es suprimeix un registre únic.

7.2.1 PRIVACITAT DIFERENCIAL SINTÈTICA

El primer treball que representa aquesta idea de privacitat, presentat per l'investigador Chen [5], explica que es basa en la construcció d'un arbre jeràrquic en el que les trajectòries s'agruparan en base a sub seqüències de localitzacions coincidents. En primer lloc, es creen els nodes fills de la iteració

[2] Fingerprint Stretch Efoort

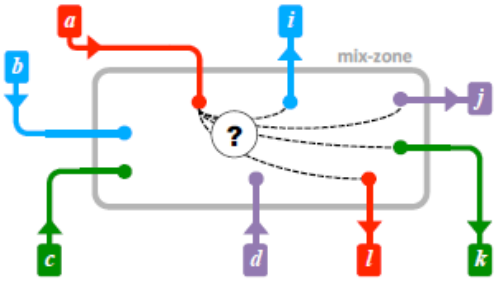


Fig. 4: Quan els usuaris entren a la zona mixta perden el pseudo identificador que tenien assignat, de forma que al sortir es reassigna a un altre usuari. Imatge extreta de [1]

base (root) i es fa el mateix per totes les iteracions. En segon lloc, un cop es té un primer arbre, s'afegeix soroll a cada node generalitzat fent ús del model Laplacià. A continuació, els nodes que tinguin un soroll per sota d'un llindar establert en un principi no s'expandiran més, i només ho faran els que superin el llindar i fins a un nivell d'expansió establert per l'usuari. Per acabar, es finalitza l'arbre amb els sorolls de Laplace dividits de forma que la suma dels valors dels nodes fills no puguin tindre un valor més alt que els nodes pares.

7.2.2 MITIGACIÓ

L'objectiu principal és reduir el risc de privacitat associat a les dades sense definir un principi clar de privacitat. Per tal de mitigar els riscos existents de la privacitat de les microdades de trajectòria, es proposa un model que afegeix soroll de forma aleatòria als punts espai-temporals, reduint així la resolució espacial o temporal de les dades, o com a mínim retallar les trajectòries. Aquesta estratègia, però, no assegura que es preservarà la privacitat de les dades. És aleshores quan es proposa un nou model presentat per a preservar la privacitat en els recorreguts que fan servir GPS, basat en la mitigació de les zones mixtes. Una zona mixta, és una regió espacial en la que els punts espai-temporals dels usuaris no són enregistrats i, a més, els pseudo identificadors dels usuaris seran modificats cada cop que entrin a una nova zona mixta. Si s'aconsegueix afegir quantitats suficients de trajectòries en una zona mixta serà molt difícil per a un atacant reconèixer un usuari en el moment que deixi aquesta zona. Com es pot apreciar, l'efectivitat de l'heurística dependrà completament de la quantitat de trajectòries espai-temporals que trevessen la zona mixta durant un interval particular de temps.

7.2.3 OFUSCACIÓ

Aquesta tècnica consisteix en afegir soroll per tal de distorsionar les dades de les localitzacions. És introduïda, per primer vegada, per evitar possibles atacs a través de la mineria de dades de localització per Agrawl i Srikant [4] i formalitzada més tard, amb el nom d'ofuscació, per Duckham i Kulik [7]. Cal destacar que uns anys més tard s'afegeixen diferents models a aquesta tècnica basats en l'addició d'una quantitat elevada de soroll aleatori a representacions de grafs socials, amb els que demostren la capacitat per a reduir considerablement l'èxit en atacs d'enllaç de registre. Els atacs d'enllaç de registre pertanyen a la categoria més investigada d'atacs, i defineixen el seu objectiu com la relació dels registres de les microdades de la trajectòria d'un usuari amb informació privada de la víctima. Aquesta informació inclou identificadors personals i dades privades sobre la seva mobilitat. Per tant, un atacant podria aconseguir la relació d'aquests registres si la base de dades objectiu conté atributs sensibles.

7.2.4 ENCOBRIMENT

Es recolza en reduir la granularitat[3] de les dades de trajectòria en dimensions d'espai o temps. Considerant la dimensió temporal, Hoh [2] defineix una tècnica que demostra la capacitat de reduir la re identificació d'un 85% a un 40% incrementant d'un a quatre minuts l'interval de mostreig de les microdades. En relació a la dimensió espacial, Murakami defensa a [13] la idea de suprimir uns punts determinats de cada trajectòria per tal de reduir les oportunitats per un atac d'enllaç de registre. El mateix autor afirma que amb la eliminació de 5 punts de cada trajectòria a cada base de dades es pot arribar a suprimir l'èxit d'atacs d'enllaç de registres fins a la meitat, tot i que continuaria sent un número molt alt. Posteriorment es presenten diferents propostes que continuen treballant amb la dimensió espacial, on afirmen que han treballat sobre una base de dades pròpia i que reduir la precisió geogràfica no té un clar efecte positiu sobre la unicitat. Posen en evidència aquests aclariments amb un experiment fet sobre aquesta base de dades, en el que si un atacant sap com a mínim 8 punts de la ruta de l'usuari objectiu la probabilitat d'èxit s'eleva fins el 50%.

7.2.5 SEGMENTACIÓ

Tècnica presentada per Song a [17] on proposa la segmentació de cada trajectòria i l'ús de diferents pseudo identificadors per a cada segment resultant. Aquesta segmentació és proposada donat que la unicitat augmenta paral·lelament amb la longitud de de cada trajectòria, de forma que pretén reduir-la al màxim per tal d'afavorir el resultat de l'anonimització, i que aquesta sigui en conclusió, menys única. Aquesta és una proposta teòrica, ja que com demostra el mateix autor en un experiment realitzat a una base de dades pròpia el 80% de les particions que es realitzaran seguiran sent úniques. Aquesta reduiria la utilitat de les mateixes, donat que evitarien molts anàlisis que requereixen de la informació completa sobre els moviments dels usuaris.

7.2.6 INTERCANVI

Salas a [15] en l'intercanvi de porcions de trajectòries de forma iterativa entre els diferents usuaris, fent així de les trajectòries resultants segments composats de trajectòries de múltiples usuaris. Aquesta tècnica és coneguda com a SwapMob. Resultats de tests demostren la reducció real en l'efectivitat d'atacs d'enllaç de registre, tot i que afirmen que si un atacant sabés 10 punts podria arribar a relacionar fins el 42% dels usuaris i aprendre el 50% de les trajectòries originals en el 5% dels casos.

[3] Representa el nivell de detall amb el que desitja emmagatzemar les dades tractades.

Algorithm 1: Swap Mobility Location

```
for element in coordenades do
    Creem un diccionari que guardi les coordenades per cada usuari
end
for element in len(diccionari.items()) do
    for actual in diccionari.values()[element] do
        for seguent in diccionari.values()[element+1] do
            if actual==seguent then
                afegir a un array que omplirem amb el valor de les interseccions
            else
                No fer res
            end
        end
    end
end
if nombre de vegades que es repeteix una localització més gran que lambda then
    Afegir la localització al array
else
    No afegir la localització
end
for element in interseccions do
    for elementValorsInicials in coordenades do
        while contador ¡ valors que hi hagi a l'element do
            if valor d'aquesta intersecció == element then
                guardem el identificador de l'usuari d'aquesta posició
            else
                Següent element
            end
        end
    end
end
while recorrem tots els ids repetits do
    trobem aquest identificador a les coordenades incials for x in identificador.Values() do
        busquem el valor que tenim emmagatzemat d'aquest identificador que és repetit. Afegim les localitzacions següents a un array
    end
end
for x in la resta de identificadors repetits do
    for y in les posicions d'aques id en les coordenades inicials do
        if si trobem la coordenada que busquem then
            Afegim les localitzacions següents a un array
        else
        end
    end
    fem el swap dels dos arrays
end
```

7.2.7 ZONES MIXTES

Les zones mixtes són considerades el model més popular de tècniques de mitigació de risc de privacitat, és considerat també, molt important per a preservar la privacitat en models que es basen en la ubicació. Assumeix que un atacant podria rastrejar el seu objectiu en el pas del temps, obligant a protegir la seqüència completa dels punts espai-temporals i no individualment, punt per punt, tal i com farien els models de serveis basats en localització. El primer model proposat per Beresford a [3] garanteix que dins la zona mixta els atacs per enllaç de registre no es podran realitzar, sempre i quan el nombre de trajectòries que surten d'una zona mixta durant un mateix interval de temps sigui suficientment gran i si la seva entropia de mobilitat [4] és suficientment gran. Es destaca, també, la tècnica introduïda per Hoh i Gruteser [10] anomenada confusió de ruta, la qual intenta protegir els usuaris d'un possible enllaç de punts de la trajectòria al llarg d'un període de temps evitant que un atacant pugui reconstruir la trajectòria inicial. El funcionament és el següent: en lloc de basar-se en una zona en la que es barregin les trajectòries, es produeix una confusió de ruta, és a dir, sempre que dos o més rutes d'usuaris qualsevol siguin prou semblants entre elles la informació sobre la ruta d'un usuari es pertorbarà de forma que l'atacant confongui les seves rutes. La Figura 4 mostra el seu funcionament. Una alternativa presentada per Hoh [11], es basa en l'introducció d'un model que, a més d'operar en la dimensió de l'espai, opera també en la del temps. De nou, es continua refinant aquesta tècnica, i sorgeix la necessitat de crear zones mixtes al voltant del conjunt de localitzacions sensibles que els usuaris visiten com a mesura de major protecció. Com a conclusió, es pot dir que el problema roman en trobar el nombre de zones mixtes òptim a desplegar. Aquest problema, és tractat per Liu [22] que proposa heurístiques que tenen en compte la influència de la densitat de la trajectòria. També es pot afegir, que en comparació amb les tècniques de segmentació o basades en intercanvi, aquesta limita la utilitat de les microdades de trajectòria de cara a anàlisis que precisen d'un seguiment exhaustiu dels usuaris al llarg del temps ja que no es podrien re identificar un cop canviat els pseudo identificadors.

8 CAS PRÀCTIC

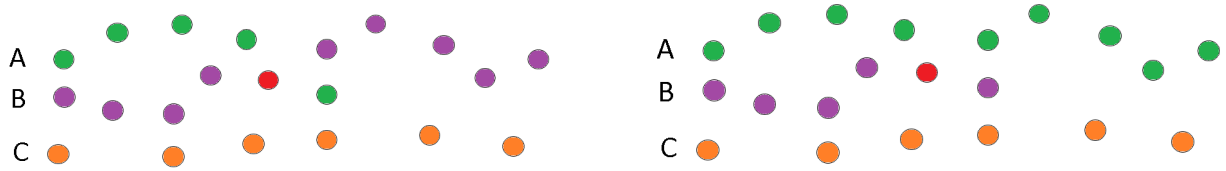
En aquesta secció veurem la implementació d'un mètode estudiat en l'estat de l'art per tal de preservar la privacitat de les trajectòries dels usuaris, fent una posterior avaluació del nivell de privacitat i utilitat de les dades anonimitzades.

Donada l'aproximació que aquest mètode proporciona, les fonts d'informació consultades i la complaxitat dels mètodes el mètode escollit ha estat "Swap mobility location". L'experiment consta d'una part dedicada a la construcció del codi 1 i la cerca de les dades, s'ha de tindre en compte que aquest no s'ha trobat de forma oberta ("open source"), sinó que s'han trobat petites explicacions en forma de pseudocodi com a articles com [16]que s'han usat per a extreure una lleugera idea per a la posterior programació del codi.

[4] L'entropia de mobilitat és la diversitat de direccions que prenen un cop deixen la zona mixta.

TAULA 1: Execucions pels diferents mètodes.

Execució	Temps		Canvis		Usuaris no modificats		Aprox.
(a) i (b)	(a)	(b)	(a)	(b)	(a)	(b)	(a) i (b)
Execució sense retalls. 100 usuaris	30'	30'	8	4	91	92	0 m
Execució retallant un decimal. 100 usuaris	45'	42'	24	15	67	70	10 m
Execució retallant dos decimals. 100 usuaris	6"	6"30 '	180	32	5	36	500 m
Execució sense retalls. 1000 usuaris	52"	32"	68	40	916	920	0 m
Execució retallant un decimal. 1000 usuaris	3h 15"	1h 30"	259	88	677	824	10 m
Execució retallant dos decimals. 1000 usuaris	4h 45"	4h	1040	315	179	370	500 m



(a) Trajectòria prèvia.

(b) Trajectòria posterior a l'aplicació de l'algoritme.

Fig. 5: Resultat de l'aplicació del mètode

8.1 SWAP MOBILITY LOCATION

A la Figura 5, es pot apreciar un resum del seu funcionament, l'algoritme busca punts coincidents en les rutes dels usuaris, i un cop detectats s'usen per a intercanviar les rutes, de tal forma que es mantenen els punts inalterables però associats a un altre usuari. S'han plantejat dos models d'algoritme:

- Mètode de intercanvis múltiple: En aquest mètode un usuari patirà tants intercanvis com punts d'encreuament es trobin. És a dir, si una ruta concreta té *n* punts d'encreuament amb altres *m* rutes, aleshores es produiran n/m intercanvis de rutes.

- Mètode de intercanvi únic: En aquest cas, un usuari amb *n* punts d'encreuament tindrà només un intercanvi de ruta amb un altre usuari. De forma que el número màxims de modificacions de rutes que s'obtindran pel conjunt de usuaris serà de 1 canvi.

Cal aclarir que un cop un usuari ha patit un canvi de ruta, i en cas de contindre més punts coincidents dins la seva trajectòria, si aquests, degut al canvi passen a formar part d'un altre usuari s'obviaran. Això es fa per a conservar en la mesura del possible la integritat de les dades per a que posteriorment no quedin inservibles.

8.1.1 INCONVENIENTS DEL MÈTODE

Com es pot veure a la Figura 5 , aquest mètode no aplicarà per a usuaris que no tinguin cap punt en comú amb altres, és a dir, si un usuari està registrat a la base de dades amb una ruta única i no coincident, l'algoritme anterior no aplicarà cap canvi en la seva trajectòria. Aquest podria ser un problema en el cas que en la base de dades objectiu tots els usuaris tinguessin una ruta única, de forma que un atacant podria realitzar una re identificació del conjunt d'usuaris.

8.2 DADES

El conjunt de dades és extret d'una base de dades [14] que conté les diferents trajectòries a partir del GPS d'usuaris que han fet servir el servei de taxis de Pequín des del 2 de Febrer del 2008 fins el dia 8 de Febrer del mateix any, a Pequín. L'experiment es realitza amb un conjunt de 100 usuaris i posteriorment amb 1.000, on el nombre de localitzacions és variant per cada un. Les dades s'emmagatzemen en fitxers diferenciats per cada usuari, en cada fitxer el format apareix com un cúmul de línies, on cada línia representa una localització en un moment determinat per usuari tractat. Dins la línia trobem les següents dades, separades per comes: `ID usuari, Data (format YYYY-MM-DD), Hora (format HH:MM:SS), Longitud.`

En la realització d'algun dels experiments s'ha fet un retall en els decimals de les dades, ja que usant les dades originals, amb una precisió de 5 decimals, s'obté una precisió en la localització de més o menys 3 metres fent que hi hagi molts pocs punts d'intercanvi. Per a solucionar aquest problema, s'ha reduït la precisió de les localitzacions, eliminant fins a 2 decimals, proporcionant així un rang de 20 metres en la precisió de la localització.

8.3 RESULTATS

En aquesta secció, s'avalua el mètode implementat i es comprova el seu ús en un cas real, centrant-nos en el nivell de privacitat i utilitat aconseguit per les dades anonimitzades. S'avaluen diferents paràmetres d'anonimització, així com diferents configuracions de dades. Es realitzaran dos experiments per cada model presentat (intercanvis múltiples i intercanvi únic), el primer amb 100 usuaris i posteriorment amb 1000. Cada experiment consta de 3 execucions, la primera sense retallar decimals de les dades, el segon retallant un decimal, i l'últim retallant-ne dos.

A la Taula 2 es mostren els resultats referents a la desviació dels punts finals obtinguda degut a l'anonimització. S'adjunta una comparació dels punts inicials i finals d'un

TAULA 2: Trajectories

Execució	ID	Longitud inicial	Latitud inicial	Longitud final	Latitud final
Trajectòries inicials	1	116.51172	39.92123	116.54723	39.90841
	2	116.36422	39.88781	116.26975	39.92127
	3	116.35743	39.88957	116.43003	39.93209
	4	116.47002	39.90660	116.52252	39.92111
	5	116.62934	39.82726	116.59029	39.83682
execució sense retalls	1	116.51172	39.92123	116.54723	39.90841
	2	116.36422	39.88781	116.26975	39.92127
	3	116.35743	39.88957	116.43003	39.93209
	4	116.47002	39.90666	116.52252	39.921111
	5	116.62934	39.82726	116.59029	39.83682
Execució retallant un decimal	1	116.5117	39.9212	116.5472	39.9084
	2	116.3642	39.8878	116.2697	39.9212
	3	116.3574	39.8895	116.4300	39.9320
	4	116.4700	39.9066	116.5225	39.9211
	5	116.6293	39.8272	116.5902	39.8368
Execució retallant dos decimals	1	116.511	39.921	116.590	39.836
	2	116.364	39.887	116.465	39.92
	3	116.357	39.889	116.318	39.951
	4	116.470	39.906	117.107	40.114
	5	116.629	39.827	116.289	39.957

conjunt d'usuaris, sense haver aplicat l'algoritme, i per les tres execucions que s'han esmentat.

Els individus agafats en cada una de les execucions són els mateixos, els 5 primers del conjunt total d'usuaris de la base de dades.

A la Taula 1, en canvi, es fa una comparació dels resultats obtinguts per a les tres execucions aplicades a cada mètode, sent (a) el mètode d'intercanvis múltiples i (b) d'intercanvi únic, amb la variació corresponent del nombre d'usuaris i per cada un dels escenaris plantejats, mostrant els resultats de temps d'execució, nombre de canvis realitzats per al conjunt d'usuaris i desviació en metres dels punts inicials degut al retall de decimals.

- Execució sense retallar decimals: Degut a l'alta precisió en la localització no s'observen variacions en les localitzacions finals i per tant el temps d'execució, és també menor.

- Execució amb un decimal retallat: El nombre de canvis efectuats al conjunt d'usuaris de la base de dades és major, però no s'aconsegueix reflectir cap canvi en els usuaris escollits com a mostra, com es pot apreciar a la Taula 2.

- Execució amb dos decimals retallats: En aquesta execució, i agafant d'exemple l'usuari 1, s'aprecia la variació en el punt final de la trajectòria, producte de l'aplicació del mètode. Es troben molts més intercanvis donada una menor exactitud en la localització, experimentant també un major temps d'execució i més usuaris protegits.

8.3.1 UTILITAT DE LES DADES

Retallar decimals fa que es perdi exactitud en quant a les trajectòries dels usuaris, però com s'ha comprovat, dins un escenari on l'exactitud de les localitzacions és tan concreta, es converteix en un recurs necessari per tal d'aconseguir un nombre més alt de coincidències i poder protegir la intimitat de més persones. Un major nombre de canvis implica també una major pèrdua d'integritat de les trajectòries reals dels usuaris, les quals per a poder ser aprofitades en posteriors estudis, necessiten ser semblants a les originals. Un cop realitzats els experiments, es conclou, prenent de referència la Taula 1 i el nombre de canvis efectuats per cada execució, que amb un major nombre de decimals retallats es trobaran més punts coincidents entre usuaris, i per tant, hi haurà més canvis a les trajectòries i una major protecció de la privacitat dels usuaris. A l'hora, però, segons la Taula 2 la integritat de les dades serà menor, fet que perjudica els interessos de tercers cara a posteriors estudis. L'augment notable en el temps d'execució davant un major nombre d'usuaris, i un major retall de decimals, és un detall important a tindre en compte davant l'escalabilitat a bases de dades més grans.

9 CONCLUSIONS I TREBALL FUTUR

La revisió de l'estat de l'art, davant un àmbit en el que no s'ha avançat gaire i el qual és molt important per a preservar la intimitat de la població, ha fet prendre consciència davant els grans perills que presenten les dades que es cedeixen diàriament. La necessitat, a més, de mantindre la originalitat de les dades de cara futurs estudis, i anàlisis, serveix per a motivar-nos davant l'ample camp de treball que existeix actualment.

De cara a un futur, es planteja l'adaptació del codi davant diferents casos amb un volum molt més gran de dades, el qual sigui capaç d'anonimitzar el major nombre d'usuaris possibles amb un temps de resposta òptim, així com el descobriment i implementació de nous algoritmes que actuïn davant les debilitats que presenta l'implementat.

REFERÈNCIES

[1] Ulrich Aïvodji. Privacy of trajectory micro-data: a survey, 2019.

[2] Baik Hoh, M. Gruteser, Hui Xiong, and A. Alrabady. Enhancing security and privacy in traffic-monitoring systems. *IEEE Pervasive Computing*, 5(4):38–46, Oct 2006.

[3] A. R. Beresford and F. Stajano. Location privacy in pervasive computing. *IEEE Pervasive Computing*, 2(1):46–55, Jan 2003.

[4] Jordi Casas Roma and Cristina Romero Tris. *Privacidad y anonimización de datos*. Editorial UOC, 2017.

[5] Rui Chen, Benjamin Fung, Bipin Desai, and Nériah Sossou. Differentially private transit data publication: A case study on the montreal transportation system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 08 2012.

[6] Josep Domingo-Ferrer and Rolando Trujillo-Rasua. Microaggregation- and permutation-based anonymization of movement data. *Information Sciences*, 208, 11 2012.

[7] Matt Duckham and Lars Kulik. A formal model of obfuscation and negotiation for location privacy. volume 3468, pages 152–170, 05 2005.

[8] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. volume Vol. 3876, pages 265–284, 01 2006.

[9] Marco Gramaglia and Marco Fiore. Hiding mobile traffic fingerprints with glove. In *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies*, New York, NY, USA, 2015. Association for Computing Machinery.

[10] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady. Achieving guaranteed anonymity in gps traces via uncertainty-aware path cloaking. *IEEE Transactions on Mobile Computing*, 9(8):1089–1107, Aug 2010.

[11] Baik Hoh, Marco Gruteser, Hui Xiong, and Ansaf Alrabady. Preserving privacy in gps traces via uncertainty-aware path cloaking. New York, NY, USA, 2007. Association for Computing Machinery.

[12] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 106–115, April 2007.

[13] T. Murakami, A. Kanemura, and H. Hino. Group sparsity tensor factorization for re-identification of open mobility traces. *IEEE Transactions on Information Forensics and Security*, 12(3):689–704, March 2017.

[14] Klara Nahrstedt and Long Vu. Crawdad, 2012.

[15] Julián Salas, David Megías, and Vicenç Torra. Swapmob: Swapping trajectories for mobility anonymization. In *Privacy in Statistical Databases*, pages 331–346, Cham, 2018. Springer International Publishing.

[16] Julián Salas, David Megias, and Vicenç Torra. *SwapMob: Swapping Trajectories for Mobility Anonymization: UNESCO Chair in Data Privacy, International Conference, PSD 2018, Valencia, Spain, September 26–28, 2018, Proceedings*, pages 331–346. 01 2018.

[17] Yi Song, Daniel Dahlmeier, and Stéphane Bressan. Not so unique in the crowd: a simple and effective algorithm for anonymizing location data. In *Proceeding of the 1st International Workshop on Privacy-Preserving IR: When Information Retrieval Meets Privacy and Security co-located with 37th Annual International ACM SIGIR conference, PIR@SIGIR 2014, Gold Coast, Australia, July 11, 2014*, CEUR Workshop Proceedings, pages 19–24. CEUR-WS.org, 2014.

[18] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and S. Martínez. t-closeness through microaggregation: Strict privacy with enhanced utility preservation. *IEEE Transactions on Knowledge and Data Engineering*, 27(11):3098–3110, Nov 2015.

[19] M. Terrovitis and N. Mamoulis. Privacy preservation in the publication of trajectories. In *The Ninth International Conference on Mobile Data Management (mdm 2008)*, April 2008.

[20] Z. Tu, K. Zhao, F. Xu, Y. Li, L. Su, and D. Jin. Protecting trajectory from semantic attack considering k -anonymity, l -diversity, and t -closeness. *IEEE Transactions on Network and Service Management*, 16(1):264–278, March 2019.

[21] Matos David Xancó. Mineria de dades, 2019.

[22] Xinxin Liu, Han Zhao, Miao Pan, Hao Yue, Xiaolin Li, and Yuguang Fang. Traffic-aware multiple mix zone placement for protecting location privacy. In *2012 Proceedings IEEE INFOCOM*, pages 972–980, March 2012.

[23] Hui Zang and Jean 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*, MobiCom '11, page 145–156, New York, NY, USA, 2011. Association for Computing Machinery.

Appendices

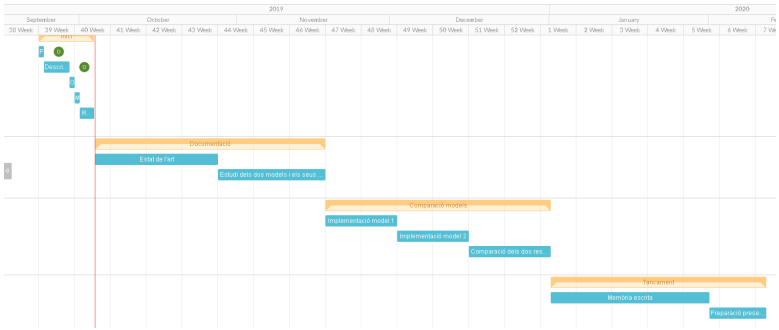


Fig. 6: Diagrama de Gantt. Planificació del treball