De-anonymising individuals through unique patterns in movement data

Anonymised for Review

Anoymised for Review

Abstract. Recent developments have accelerated the need to track individual contact behaviours to counter the pandemic of COVID-19. Various solutions have been implemented to support contact tracing activities with respect to disease propagation. In particular, digital contact tracing approaches range from manual QR scanning to automated Bluetooth handshakes. However, a problem that all the contact tracing methods face is the issue of how to protect individuals' privacy while also dealing with the threat of a global disease. Approaches such as IMSI-catching address the data privacy issue by adopting rotating keys, which are aimed at minimising the potential risk of de-anonymisation. Rotating keys, however do not protect individual movement patterns or behaviour routines that can be inferred from contact tracing data. In this paper, we show that individual movement patterns in contact tracing data can be used to discover behaviour routines by finding routines and patterns that are representable as quasi-identifiers, and that such quasi-identifiers can enable re-identification. We demonstrate the practicality of our proposed de-anonymisation approach via a series of three (3) experiments based on both semi-synthetic and real-world datasets, to search for quasi-identifiers that reveal unique observation patterns. Since widely popular contact tracing apps employ similar movement and contact patterns, we discuss the risk of location based behaviour pattern tracking in spite protection mechanisms such as rotating keys that are currently used to protect handshake identifiers.

1 Introduction

The outbreak of the COVID-19 pandemic has resulted in the deployment of several digital contact tracing applications, that are aimed at providing healthcare practitioners with feedback on the spread of the disease. However, to enable tracing, users typically have to agree to part with some personal (sensitive) data. For instance, in the UK NHS COVID19 app, during the setup process, the app records the phone's make and model and asks the user for their postcode area. Following this, a unique installation identification number is generated along with a daily rotating identifier[1]. Bluetooth Low Energy (BLE) technology is used to monitor and record the daily identification number of other users nearby and broadcasts the identifier associated with the source phone. This identifier exchange works natively with Android and iOS and for basic operations do not require additional artefacts.

However, as has been shown in previous works, while unique identifiers, usually classified as pseudonyms, are good mechanisms to obscure the actual user personal

[1] The daily identifier is a random series of digits that is uniquely re-generated each day.

data cross-linking resulting in re-identification is still possible [4,41,37,8,43,39,36]. As Sweeney and subsequent works have pointed out, the primary cause of this issue is the presence of quasi-identifiers within the dataset [4,41,37,8,43,39,36,40]. While applications such as the UK NHS COVID19 app address this issue by employing rotating keys, rotating keys are not sufficient to guard against potential de-anonymisations since in effect they do not protect fine-grained data such as individual movement patterns or behaviour routines that can be inferred from contact tracing data. Such behaviour patterns can serve as quasi-identifiers, in the sense that the spatio-temporal data points comprising the data reveal information about a user. For instance, fine-grained data such as the date and time at which the app was consulted might result in inferences about the user's age, gender, and race, which might not be information the user intended to share with the app (or the associated service). The potential for such quasi-identifiers to enable de-anonymisations are often neglected in the context of suboptimal privacy. Yet, publishing or using data with untreated quasi-identifiers raises a high risk for de-anonymisation [30].

In this paper, we show that individual movement patterns in contact tracing data can be used to discover behaviour routines by finding routines and patterns that are representable as quasi-identifiers, and that such quasi-identifiers can enable re-identification. In particular, we assess the privacy risk posed by daily rotating unique identifier numbers that are being broadcasted frequently. To do so, we conduct an empirical study based on real-world and semi-synthetic datasets to re-identify individuals based on patterns or historical data. Based on daily rotating numbers, we recreate behaviour and movement patterns with the goal of illustrating how re-identifying individuals over multiple anonymous ID sessions occurs. As an example, we show that such quasi-identifiers can re-identify "Mayumi W.", a Portuguese speaking individual in Beijing China, through a cross-linkage of behaviour and movement patterns in real-world datasets. Since widely popular contact tracing apps employ similar movement and contact patterns, we discuss the risk of location based behaviour pattern tracking in spite protection mechanisms such as rotating keys that are currently used to protect handshake identifiers.

The rest of the paper is structured in the following manner: Related work is summarised in Section 2. A quick recap of the mechanisms of contact tracing through individual movement and contact patterns is offered in Section 3. The formalisation of discovering repeating movement patterns will be presented in Section 4. A series of experiments will illustrate the problem in different settings, both on real-world and semi-synthetic data, in Section 5. Section 6 finally concludes our results and suggests avenues for future work.

2 Related Work

Works on digital contact tracing currently focuses on the efficiency of digital contract tracing solutions and their effect on disease control. Eames et al. acknowledge that contact tracing is an essential measurement in controlling infectious, if followed by treatment or isolation [9]. Further, in highly simplified versions of real infection processes, Eames et al. discuss implications and applications of contract tracing while also highlighting ineffective scenarios like with high-risk groups, with many possible transmission routes and a high incidence of infection. Klinkenberg et al. [16] concluded that tracing effectiveness need not be sensitive to the duration of the latent period and tracing delays; iterative tracing primarily improves effectiveness when single-step tracing is on the brink of being effective. The role of contact tracing has already been studied by Huerta et al. [15], who discussed that a major outbreak could be significantly reduced or even eliminated at a small additional cost through a mean-field model of contact tracing for the case of random graphs. Additionally, the influence of network topology on the contact tracing found that its effectiveness grows as the rewiring probability is reduced. Yasaka et al. developed a smartphone app that utilises self-created checkpoints for contact tracing, and promise anonymously self-reports of the user's health status [44]. Many national tracing apps follow a similar direction through either created checkpoints or dynamic users Bluetooth touchpoints that can be translated to checkpoints. Leith et al.'s work analysis the offered Google/Apple Exposure Notification (GAEN) service and highlights privacy concerns around data traffic to the apple and google servers [21]. Further, cookie hijacking and known privacy concerns of the advertising identifiers are elaborated. Further, Leith et al. analysis the properties of the Singapore OpenTrace app [20] and conclude that its usage of Google's Firebase Analytics service allows IP-based location tracking. Martinez-Martin et al. picks up a similar privacy discussion around national tracing apps [24], yet highlights that digital tracing efforts must employ ethical frameworks beyond anonymisation efforts. Rather Martinez-Martin et al. conclude that privacy contains a broader ethical context and should address the trade-offs between respecting individual liberties and protecting a society inherent in supporting public health. Rowe reflects on the general privacy paradox of digital tracing that encapsulates assurances of anonymity while still being able to trace individuals [38]. While acknowledging the greater public good, Rowe raises concerns about general acceptances of the loss of privacy. Zheng et al. have showed that routine behaviour pattern could be extracted from sparse mobile phone data using collaborative filtering [45]. This is already an important step, yet collaborative filtering relies on the similar-alike approach and requires a sufficiently large training dataset to refer to. Additionally, highly sparse datasets are extremely difficult with collaborative filtering known as sparsity problem [22,5,12].

Beyond these scopes, there is insufficient research on the privacy risk exposure in rotating identifiers in digital tracing apps that randomise individual, user-specific observations to the best of our knowledge.

3 Similarity of COVID19 tracing apps to movement data

The UK implementation of the contact tracing and mobile notification application ("NHS COVID19 App") promises strong privacy constraints. Same does the German version ("Corona-Warn App") that even published a full technical documentation[2], and its formal evaluation criteria for achieving privacy [3]. This application is running on individual phones, publicly available for everyone and endorsed by the respective national organisations. The NHS version inquiries the phone's make, model and the user postcode area as part of the initial setup process. Next, a unique installation identification number is being generated along with a rotating identification number. In the case of the NHS app,

[2] https://github.com/corona-warn-app/cwa-documentation/blob/master/overview-security.md

[3] https://github.com/corona-warn-app/cwa-documentation/blob/master/pruefsteine.md

the rotating identifier has a time-to-live (TTL) of one day, the German one just a few minutes. These rotating identifiers are locally broadcasted with Bluetooth Low Energy (BLE) technology. Simultaneously the COVID-19 tracing app monitors and records the daily identification number of other users in close proximity of the source phone. While the German version implements a fully distributed system from the beginning on, the NHS app initially stored identifiers in a centralised repository and after consideration moved to a decentralised architecture as well. In case of a compromised identifier, such centralised system is much more vulnerable to background knowledge attacks [23]. For the German version of the disease tracing solution known as ("Corona-Warn App"), for instance, the developer dedicated whole chapters in their publicly available documentation to security related questions and concepts[4]. Both solutions build on Google and Apple Exposure Notification (GAEN) service[5] that offloads the Bluetooth ID exchange to the mobile operating system. This GAEN service has been previously reviewed by Leith et al., who primarily raised concerns around potential IP-tracing opportunities for the TCP/IP package exchange in this decentralised setup [21,20].

In summary, the technical setup is quite similar but one of the essential differences between the German and the UK tracing solution is, that the rotating identifier number which exists in both UK and Germans version have a different time-to-live (TTL). The German approach incorporates a significantly smaller time-to-live and enforces more frequent ID rotation which is useful to reduce the time windows of recurring pattern appearance. However, rotating identifiers pose a privacy risk for users because they can be used to enable re-linking attacks and annul the users' privacy. The longer a temporary identifier is valid, the more likely we can observe behaviour points linked by the same identifier and therefore build better patterns for profiling. Decreasing the lifetime of an identifier, on the other hand, comes with the cost of increased processing needs as more identifiers are being collected, stored, transmitted, and later compared for proper tracing activities. Therefore, a trade-off based on the cost-benefit ratio of efficiency (infrastructure costs, mobile data ...) and privacy risk is needed.

We re-visit the issue of the trade-off between efficiency and privacy risk in subsequent sections, but first outline the approach of the behaviour pattern de-anonymisation with an analogue to IMSI catcher. Similar to IMSI catcher, each Bluetooth handshake can serve as a location indicator linked to a randomised identifier. In the following, we will exploit this to evaluate our challenges of re-identifying individual users.

4 Re-identification

Regardless of the dataset origin, we assume that the dataset contains location or contact tracing paired with a rotating identifier. We build upon the concept of IMSI catcher [11,28], which has been shown to be easily scaled even to the scope of an entire city [29]. In the following we will briefly explain the analogy of catching digital tracing app identifiers to IMSIs and how such IMSI catcher work. Afterwards, the slightly modified concept of IMSI catcher will be presented and applied to the world of disease tracing applications.

[4] https://github.com/corona-warn-app/cwa-documentation/blob/master/overview-security.md

[5] https://github.com/corona-warn-app/cwa-documentation/blob/master/solution_architecture.md#privacy-preserving-data-donation

4.1 IMSI catcher

In the telecommunication field, the International Mobile Subscriber Identity (IMSI) is a unique number bound to each cell phone that is used by the mobile network operator (MNO) to register a sim card to the infrastructure. For communication, these IMSIs are embedded in mobile cellular network and broadcast freely and frequently by network participants communicating over the mobile cellular network. The IMSI is used in the Signalling System #7 (SS7) for roaming billing purposes. However, SS7 has been shown to be vulnerable to global tracking incidents [14]. IMSIs, unlike the tracing IDs, are constant and are not rotated frequently.

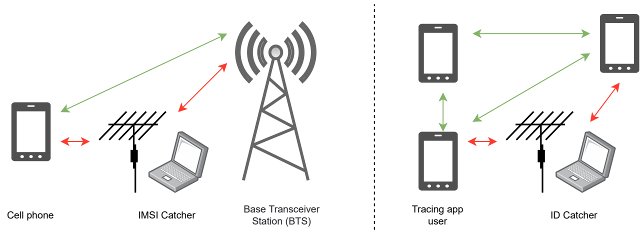


Fig. 1: Concept of IMSI Catcher

The IMSI catcher are an adversarial mechanism for stealing (or hijacking) the IMSI associated with a given cell phone. For this purpose, the concept of IMSI catcher utilises the behaviour where locally cell phones register with the Base Transceiver Station (BTS) to link wireless into the MNO network. Now, *active IMSI catcher* use the attack vector of mimicking or simply predominating the BTS reception in the way that the cell phone actually registers with the IMSI catcher as tower of the highest signal. The IMSI catcher then act as proxy to the actual BTS (see Figure 1). *Passive IMSI* catcher just read along with the broadcasted messages and parses the IMSI details from the communication feed. There have been initiatives to discover and quantify IMSI catcher as countermeasure [7], but we will not deepen this for now.

The objective of IMSI catchers is, to validate the location of the IMSI, and correspondingly the owner of the cell phone. Often, as an addition, selective message packages can be jammed, blocked, dropped, or attempted to be decrypted [26,6]. Since the IMSI is a super identifier that does not naturally change, digital COVID tracing apps, like the UK NHS one, try to avoid the re-identification risk by updating broadcasted identifier numbers. At the first sight, an attack is much more difficult to map together different chunks of observation points in absence of a constant identifier.

Conceptually, the setup for an adversarial mechanism for stealing (or hijacking) rotating identifiers of digital tracing apps is similar to passive IMSI catchers. Bluetooth-based COVID tracing apps willingly broadcast their identifiers. To collect these, no sophisticated BTS technology is needed rather a cheap phone or just Bluetooth chip and some lasting electricity power like a power bank. As the range is limited, several are needed, or their signalling capacity boosted.

4.2 Movement patterns as quasi-identifier problem

However, humans stick to repeating behaviour patterns as these routines disburden our active cognition [2,13]. We routinely use the same way to work, stop at our favourite coffee shop and use the same facilities for grocery shopping. Repeatedly. These routines are our entrance point, as they serve as connecting points to re-identify an individual behaviour [3,13]. Discovering routine behaviour patterns is not new, and has been proven before also in the field of telecommunication [45]. We use similar methodology, yet not rely on collaborative filtering as it suffers the sparsity problem [22,5,12]. Instead, existing techniques to link behaviour attributes to discover quasi-identifiers are being utilised [32,35]. Repeating behaviour patterns like movement can be observed and have been successfully modelled in the telecommunication sector [45]. Data samples can be derived from the official tracing app documentation[6]. Each time series event is basically acting as an check-in equivalent to the movement patterns (see Table 1).

Table 1: Timeseries dataset of logged bypass ID

Datetime	Long	Lat	location	identifier	weather	temperature
5/06/2021 07:41:56	-0.088	51.5184	Moorgate station	12345	rainy	25°C
5/06/2021 07:58:56	0.1211	51.4910	Abbey Wood station	12345	rainy	26°C
5/06/2021 08:13:41	0.1210	51.4901	Greggs coffee shop	12345	sunshine	27°C
..	..	..	..	..	..	

Discovering the movement patterns in this dataset is achieved by searching for quasi-identifiers (QIDs) [17]. Quasi-identifiers (QIDs) are attribute value combinations with any tuple length [31]. For instance, this can bypass three observation points in a certain time window (e.g., jogging in the park at 5am and coffee from Greggs at 7am). A quasi-identifier can represent a unique behaviour that is repeatedly observed (daily, weekly, monthly). Upon re-discovery, we can assume that the daily rotating identifier belongs to the same user and cross-link their activity session with a high probability. Cross-linking allows us to analyse the additional gained activity insights, for instance, ophthalmologist visit on Thursday and optician visit on Friday.

More formally we therefore formulate movement pattern as:

Definition 1. *Unique movement pattern*
Let $O = \{o_1,..,o_n\}$ be a set of all observation points where o_j consists out of multiple describing attributes $o_i = \{f_1,..,f_n\}$. Further, $P := \mathcal{P}(O) = \{P_1,..,P_k\}$ is the set of all possible observation combinations, thus its power set. A set of selected observations $P_i \in P$ is called a unique movement pattern, if P_i identifies at least one individual uniquely within 24 hours, repeats more than once outside the same 24 hours and all observations $o_j \in P_i$ are not standalone identifiers (unique in themselves).

As these patterns must be repeating over days but unique by individuals $Q \subseteq O$ and $P \subseteq Q$. Thus, the problem of finding a unique movement pattern is a superset of the *Find-QID* problem [34].

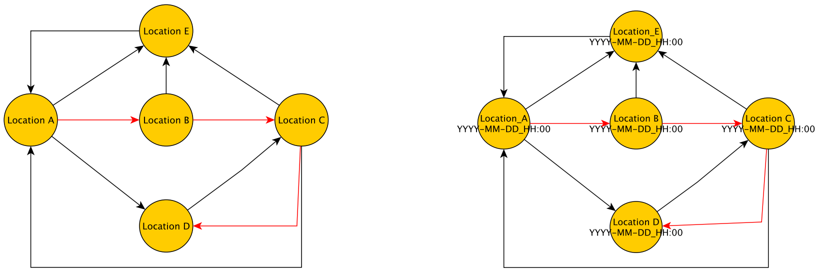
Example 1. Given the data described in Table 1, an observation point is the combination of a location and datetime. This observation point might be recorded multiple times when various devices pass by (for instance in the same train). Yet, multiple observation points can become a unique combination if the sequence of recording of the same observation points is unique. This happens, if device 12345 travels from Moorgate to Abbey

[6] https://github.com/corona-warn-app/cwa-documentation/blob/master/solution_architecture.md#privacy-preserving-data-donation

Wood, Greenwich around 8 am but then disembark as only passengers and check-in to a different observation point afterwards. Now, the unique combination of these observation points form a quasi-identifier that is unique for that observation period (e.g. 24 hours). But only if the unique sequence of observation points repeats outside of the observation period, the quasi-identifier becomes a repeating movement pattern (e.g. every morning an individual travels from Moorgate station to Abbey Wood station and buys coffee at Greggs coffee shop).

The procedure of evaluating describing attributes whether they form quasi-identifiers is prototypically implemented and publicly available on github.com[7]. In the first step the combination of all describing attributes are generated. Afterwards, these QID candidates are being evaluated if they qualify as standalone quasi-identifier through their uniqueness. Given these loops, the reader will briefly recognise the discovery of QIDs compute intensive given the iterative and incremental nature (in fact NP-hard and W[2]-complete problem [27]). Yet, for the present case existing solvers will be utilised to handle the exponential compute need for the QID search. Podlesny et al. showed a scalable approach for QID discovery in large, high-dimensional data with massive parallelisation through GPU compute [34]. The same will be applied to the present problem.

In the following section, we will demonstrate the practicability of such pattern recognition through a series of experiments.



(a) Visualisation of observation points

(b) Visualisation of observation points including datetime

Fig. 2: Graph visualisations

4.3 QID search as graph problem

Previously, we presented that movement patterns can be projected on the search of quasi-identifiers (*Find-QID problem* [34]). These patterns must be repeating over days but unique by individuals. Therefore, the problem of finding a unique movement pattern is a superset of the QID problem. As movement patterns consist out of individual observation points, each of these points assembles an observation track (see Figure 4). Subsets of these tracks might overlap, are fully identical or disjunct. This can be visualised well in a two-dimensional room, but adding dates and times becomes more complex. Various research fields have invested heavily in solving and optimising similar issues; therefore we translate the problem of finding overlaps in these time series to

[7] `https://github.com/jaSunny/qids-on-movement-data`

the field of graph theory. Figure 2a illustrates a simplified graph structure where nodes answer to location and edges to aggregated movements between these locations. The edge probability mapped to each edge is derived from the likelihood of the appearance of a movement exactly between the locations. In a more complex setting delineated in Figure 2b nodes are not only representing arbitrary locations but locations at a certain time in the day. Aggregating the exact timestamp to minutes or hours corresponds to allowance in the time windows that have been discussed previously (see Section 4.2). This trade-off may introduce false-positives similar to the discussion of the size of *k* in *k*-anonymity [27]. Quasi-identifiers (QIDs) remain a unique combination of describing attributes, yet in the context of a graph, a QID answers to unique concatenations of edges. Uniqueness can be approximated based on a low edge probability or exactly determined by comparing the graph of individual movement history to the consortium of all movement observations. Suppose the former graph is a true subset of the latter, and no full overlapping other graphs of any other individual movement history exist (see Figure 3). In that case, one can declare this a quasi-identifier.
Comparing graph structures is not completely new and known as "subgraph isomorphism problem" [10,42]. Various research has been conducted on solving this subgraph problem efficiently, represented by Ansari et al. [1] and McCreesh et al. [25] recent contributions. We will use this state-of-the-art to resolve the question of "subgraph isomorphism". Answering the question of "subgraph isomorphism", bridges us the solution to find overlapping movement tracks, that implements the discovery of quasi-identifiers.

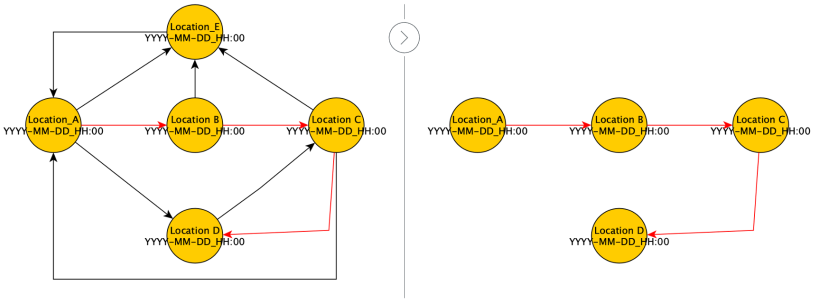


Fig. 3: Movement track as sub-graph problem

In the following section, we will conduct a series of experiments to implement the outline approach and demonstrate its practicality with a mix of real-world and semi-synthetic datasets.

5 Experiments

To assess our previously presented concept on feasibility, we follow a quite simple set up in the beginning. Because distributing various cheap cell phones connected to lasting power banks across major public infrastructure points might cause a bit of stress to emergency services, we will utilise existing real-world GPS datasets to validate the hypothesis of behaviour patterns in movement data through three separate experiments.

Table 2: Dataset 2 - Movements points

UUID	DateTime	Long	Lat
9996	2008-02-02 14:37:54	116.46125	39.9152
9996	2008-02-02 14:42:56	116.46118	39.91497
9996	2008-02-02 14:47:58	116.46118	39.91494
9996	2008-02-02 14:53:00	116.46119	39.91503
9996	2008-02-02 14:53:00	116.46119	39.91503
9996	2008-02-02 14:58:02	116.46124	39.91491
9996	2008-02-02 15:03:04	116.46124	39.91494
..	..	..	..

Hardware. Our examination runs on a GPU-accelerated high-performance compute cluster, housing 160x CPU cores (E5-2698 v4), 760 GB RAM, and 6x Nvidia GeForce RTX 3060[8] with combined 21,504 CUDA cores and 72 GB dedicated GPU memory.

Dataset. In a first step, we demonstrate that behaviour routines and movement patterns can be discovered on a smaller scale. To do so, instead of synthetic data we utilise real data from one-week trajectories of 10,357 taxis with about 15 million sensor events and the total distance of the trajectories reaches 9 million km. Following the same evaluation criteria formalised by the Chaos Computer Club (CCC), a well-reputed European hacker collective, a system should ensure data economy/minimisation, anonymity, no central entity to trust, unlinkability, unobservability of communication, no creation of central movement or contact profiles [9]. With the argumentum e contrario, the successful discovery of a quasi-identifier that exposures personally identifiable information (PII) should not be possible. The dataset represents a time series expressed in Table 2, original real-world data from Zheng et al. [46] and is publicly available [46] to ensure the reproducibility of the experimental results.

The first experiment demonstrates that behaviour patterns in movement data are omnipresent. With the confidence in mind that these patterns exist, we project the same characteristics of this real-world dataset on a semi-synthetic COVID tracing dataset. This is similar to if we would have collected our phone collecting gadgets from the public infrastructure points. The tracing dataset has the same semantics as the previous GPS dataset, instead of timely synchronised data gatherings acting as checkpoints we have Bluetooth touchpoints as checkpoints. This compiled large dataset will then be used in the second experiment to discover behaviour movement patterns on a large scale. Each recognition among the monitored infrastructure points is a puzzle piece, or a describing attribute can form quasi-identifiers in combination. Finding these quasi-identifiers is not easy and NP-hard, but we utilise previous work for acceleration [35,31]. The third experiment will transpose the same methodology to social media data, check if these can be cross-linked to previous patterns in movement data and utilise Podlesny et al. previous work [34] to discover quasi-identifiers on a large scale in Twitter data. For transparency, we describe the experimental hardware and dataset in the following.

[8] https://www.nvidia.com/de-de/geforce/graphics-cards/30-series/rtx-3060-3060ti/

[9] https://github.com/corona-warn-app/cwa-documentation/blob/master/pruefsteine.md

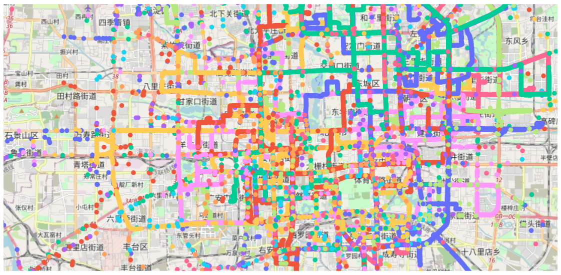


Fig. 4: Visualisation of movement observation points

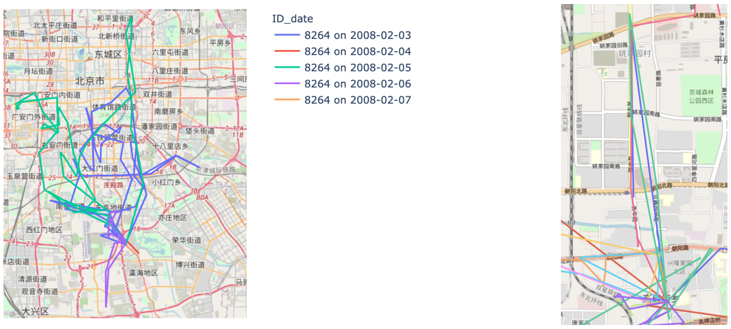
In the second step, we scale this dataset to a semi-synthetic one compiled to a typical entity-attribute-value (EAV) model that depicts a time-series log similar to previous IoT sensors. Yet, timestamps, location and broadcasted Bluetooth identifier have been enriched and aggregated as describing attributes following the same distribution of the cap dataset. A few environmental information like weather, day, temperature and special events in the location that might differ from typical behaviour have been added (see Table 1). Finally, we utilise COVID tweet data to evaluate cross-linkage and quasi-identifier discovery opportunities in social media data. The *TweetsCOV19* dataset originates from Lamsal et al. work [18,19] and consists out of roughly 858643 tweets posted with 11 hydrated describing attributes[10] (originally 3x with tweet ID, long, lat) based on 14x COVID19 related hashtags. The dataset is available for download via IEEE dataport[11]. We filtered and updated this dataset for the same region as the taxi data, Beijing China. Finally, the dataset includes an encrypted username, timestamp, URL, entities, and meta data like user followers, friends, and retweets. We exclude the tweet ID for cross-validation of successful de-anonymisation and received in total 120000 tweets with geographical information from years 2019 to 2021.

Evaluation. In the first step, the real-world dataset is evaluated for behaviour patterns and whether these can be found through analysing movement characteristics. A unique behaviour pattern is a combination of describing attributes that form a unique quasi-identifier. Therefore, practically, one searches for the existence of quasi-identifiers (QID) using the implemented Algorithms from Section 4. As illustration of the raw dataset, Figure 5a shows a selection of individual tracks done projected on a map. Figure 4 shows the individual observation points. For a trained eye, some overlapping patterns can be observed. Next, we add the time to our considerations as this is an important factor for routines under observation [17].

Locations and times describe attributes in the present dataset; therefore, as explained earlier, searching and finding unique combinations of these describing attributes helps identify repeating behaviour routines. The presence of these unique combinations, also known as quasi-identifiers is delineated in Figure 5b where the blue and the green line illustrate the same movement pattern. Having confirmed that quasi-identifiers are present in the real-world dataset, we know that movement patterns exist. Therefore, we pivot

[10] `https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet`

[11] `https://ieee-dataport.org/open-access/coronavirus-covid-19-geo-tagged-tweets-dataset`



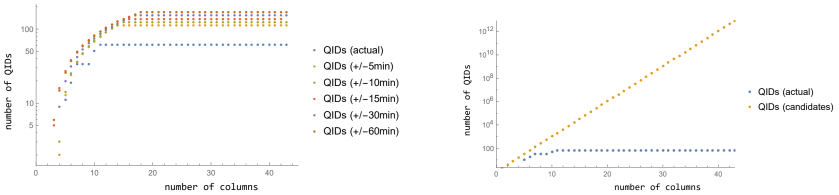
(a) Visualisation of selected movement tracks

(b) A unique quasi-identifier as movement pattern visualised

Fig. 5: Projected growth of discovered QID

our attention towards the second dataset. This semi-synthetic dataset is much larger and reflects the complexity challenge more evidently. The same QID search scheme is applied on this data and Figure 6b delineates the results. The number of observation points corresponds to a unique location over the quantity of quasi-identifier candidates (see Figure 6b). The exponential growth becomes apparent when carefully considering the increase on the log scale. In total, 1,330,995 unique GPS coordinates appear more than once. When combining more than 3 locations 62 quasi-identifiers with tuple length of 5 can be found (3x locations, 1x tracing ID, 1x datetime attribute). Interestingly, a few observation points are already sufficient to start forming actual quasi-identifiers (QID), as we see their discovery starting with five unique locations. This increase of actual QIDs and their candidates are depicted in Figure 6. That implies, in turn, that already with low efforts and handful of receivers' unique movement patterns can be found. The interesting point is whether we can re-identify the same unique patterns over a longer period of time. The challenging part is the deviation within the time-span of the observed repeating patterns. Missing a train or bus might happen and therefore a tolerance window is desired. The lower the accuracy, the higher the more false positives generated. Figure 6a illustrates the increase and decrease of quasi-identifiers through widening the time tolerance between 0 and 15 minutes. It becomes evident that the larger the time window, the more generous the pattern recognition is and therefore a larger number of quasi-identifiers be found. At the same time, more false-positives will be triggered.

In the third experiment, we extend previous work from Podlesny et al. [33] that successfully de-anonymised individuals in a publicly available Twitter dataset on the US Presidential election 2020 through the discovery of quasi-identifiers. In our case, we are using a newly generated Twitter dataset on recent COVID content which was presented earlier. As part of the experiment, the objective is to search for quasi-identifiers that unique identify individual users. Figure 7a delineates the runtime growth over the increasing number of tweet records while Figure 7b depicts the discovered quasi-identifiers. These quasi-identifiers could be used to re-identify individual tweet data. We discovered some convenient entrance points to cross-link these QIDs and their patterns



(a) Growth of discovered QID with different time windows applied

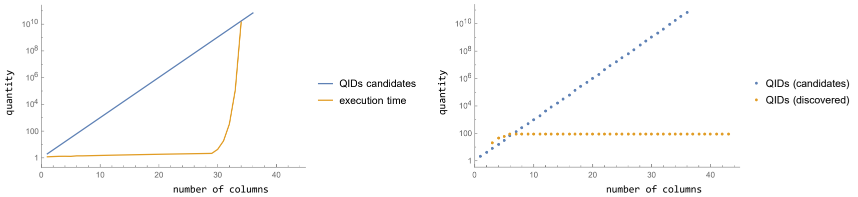
(b) Growth of QID candidates over available observation points

Fig. 6: Projected growth of discovered QID

to auxiliary datasets like the taxi one. Individuals tend to start tweeting around main transportation hubs like a train station or airport, addressing these geographic areas. This way, tweet content and location as well as date time can be linked. If the same individual now takes a cap, drive to their destination, hop off and tweet again, movement pattern and social media content can be linked. An example is a user named "Taiwan number one" who landed in Beijing China, tweeted on Aug 25, 2021 6:46 PM in Portuguese about taxis. This user repeatedly posted references to the same aliexpress live stream, which could have been crossed linked to a user identified as "Mayumi W." (full name has been removed for privacy reasons). Now, given the tweet activity, knowledge about behaviour preferences and location, we can re-identify the user and their movement pattern. Given ethical adherence, we cannot finally confirm that this user did actually sit in the cap that drove between the two tweet locations on the matching time. Still, we appreciate the high probability of these events occurring.

Discussion. To summarise, the experiments have shown three key insights. First, we were able to validate the existence of unique behaviour patterns in ordinary movement data. A prerequisite is that they offer fine-grained observations already based on a few data points. Second, even on larger scale, like in the case of digital digital tracing solutions where millions and billions of data points are being collected, these patterns can be successfully searched and identified. Third, smaller variations on movement timestamps can be smoothed potentially generating more false positives as side effects.
However, this only works with the possibility of observing data points over a longer period of time. Discovering these patterns in just 15-minute time windows seems not to be practical, at least our experiments have not been successful in discovering meaningful insights yet. As the UK NHS COVID app does not rotate the identifier every 15-minutes, rather once a day, this opens a much longer observation period. Therefore, more behaviour routines fall into one cycle and linkable data points increase the value of cross-link identifier sessions. In sum, this already presents an easy solution to the outlined privacy concerns. Simply decreasing the time to live for rotating identifiers is sufficient to counter the presented attack vector. For repeatability, we have made our source code publicly available on github.com[12].

[12] https://github.com/jaSunny/gpu_qid_discovery



(a) Execution time for QID discovery

(b) Growth of Tweet QIDs

Fig. 7: Observed quasi-identifier (QID) characteristics

6 Conclusions and Future Work

In this paper, the issue of de-anonymisations of individuals based on behaviour patterns linked through arbitrary rotating identifiers has been discussed and their implication on privacy exposure through experiments explored. For that purpose, an analogy to IMSI catcher has been offered and an approach discussed to find users' behaviour routines by discovering quasi-identifiers (QID) in their movement patterns. By transposing the issue of identifying movement patterns to the search of quasi-identifiers (QID), we can leverage existing research and techniques to evaluate scalable experiments.

In our experiments, we demonstrate that, already with a small sample of GPS based movement observation, one is able to re-identify individuals. Further, the shorter the tracing ID rotation is conducted, the more information can be restored and the easier is a de-anonymisation of individuals. These individuals could be digital tracing app users. While for the UK version of the COVID19 tracing app such re-identification seems realistic even on a large scale, the same does not hold true for the German version that implements a 15-min rotation of tracing identifiers. As part of our evaluation, we highlighted the importance of this sliding time-window for anonymisation guarantees. Further, we also showed that the combination of location and time data can form quasi-identifiers that helped to cross link movement and social data to re-identify "Mayumi W.", a Portuguese speaking individual in Beijing China.

Future work may address the scalability of Bluetooth tracing ranges to minimise the need for implemented observation points or conduct a large field test on the practical feasibility even further. In addition, smoothing effects on GPS data and their implication on the quasi-identifier discovery could be extended through further experiments.

Ethical considerations. As part of our work, we used existing datasets to mimic user behaviour and movement patterns to evaluate our hypothesis and prove our concepts in near-real experimental situations. The datasets have been collected initially in other research studies, with user consent given and made public for other research venues. For privacy discussions, the authors encourage public discussions around the balance between individuals' privacy and the greater public good and appreciate that the conclusions might differ based on culture, regional, historical, and legal backgrounds.

References

1. Ansari, Z.A., Abulaish, M., et al.: An efficient subgraph isomorphism solver for large graphs. IEEE Access 9, 61697–61709 (2021)

2. Banovic, N., Buzali, T., Chevalier, F., Mankoff, J., Dey, A.K.: Modeling and understanding human routine behavior. In: CHI Conference on Human Factors in Computing Systems. pp. 248–260 (2016)
3. Banovic, N., Wang, A., Jin, Y., Chang, C., Ramos, J., Dey, A., Mankoff, J.: Leveraging human routine models to detect and generate human behaviors. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. pp. 6683–6694 (2017)
4. Barth-Jones, D.: The're-identification'of governor william weld's medical information: a critical re-examination of health data identification risks and privacy protections, then and now. Then and Now (July 2012) (2012)
5. Bobadilla, J., Serradilla, F.: The effect of sparsity on collaborative filtering metrics. In: Proceedings of the Twentieth Australasian Conference on Australasian Database-Volume 92. pp. 9–18. Citeseer (2009)
6. Cattaneo, G., De Maio, G., Petrillo, U.F.: Security issues and attacks on the gsm standard: a review. J. Univers. Comput. Sci. 19(16), 2437–2452 (2013)
7. Dabrowski, A., Pianta, N., Klepp, T., Mulazzani, M., Weippl, E.: Imsi-catch me if you can: Imsi-catcher-catchers. In: Proceedings of the 30th ACSAC Conference. pp. 246–255 (2014)
8. De Montjoye, Y.A., Radaelli, L., Singh, V.K., et al.: Unique in the shopping mall: On the reidentifiability of credit card metadata. Science 347(6221), 536–539 (2015)
9. Eames, K.T., Keeling, M.J.: Contact tracing and disease control. Proceedings of the Royal Society of London. Series B: Biological Sciences 270(1533), 2565–2571 (2003)
10. Eppstein, D.: Subgraph isomorphism in planar graphs and related problems. In: Graph Algorithms and Applications I, pp. 283–309. World Scientific (2002)
11. Fox, D.: Der imsi-catcher. Datenschutz und Datensicherheit 26(4), 212–215 (2002)
12. Grčar, M., Mladenič, D., Fortuna, B., Grobelnik, M.: Data sparsity issues in the collaborative filtering framework. In: International workshop on knowledge discovery on the web. pp. 58–76. Springer (2005)
13. Hamermesh, D.S.: Routine. European Economic Review 49(1), 29–53 (2005)
14. Holtmanns, S., Rao, S.P., Oliver, I.: User location tracking attacks for lte networks using the interworking functionality. In: IFIP Networking conference. pp. 315–322. IEEE (2016)
15. Huerta, R., Tsimring, L.S.: Contact tracing and epidemics control in social networks. Physical Review E 66(5), 056115 (2002)
16. Klinkenberg, D., Fraser, C., Heesterbeek, H.: The effectiveness of contact tracing in emerging epidemics. PloS one 1(1), e12 (2006)
17. Koot, M.R., Mandjes, M., van't Noordende, G., de Laat, C., et al.: Efficient probabilistic estimation of quasi-identifier uniqueness. Proceedings of NWO ICT. Open (2011)
18. Lamsal, R.: Coronavirus (covid-19) geo-tagged tweets dataset (2020), https://dx.doi.org/10.21227/fpsb-jz61
19. Lamsal, R.: Design and analysis of a large-scale covid-19 tweets dataset. Applied Intelligence 51(5), 2790–2804 (2021)
20. Leith, D.J., Farrell, S.: Coronavirus contact tracing app privacy: What data is shared by the singapore opentrace app? In: International Conference on Security and Privacy in Communication Systems. pp. 80–96. Springer (2020)
21. Leith, D.J., Farrell, S.: Contact tracing app privacy: what data is shared by europe's gaen contact tracing apps. In: IEEE Conference on Computer Coms. pp. 1–10. IEEE (2021)
22. LI, C., LIANG, C.y., YANG, S.l.: Sparsity problem in collaborative filtering: a classification. Journal of Industrial Engineering and Engineering Management 1 (2011)
23. Martin, D.J., Kifer, D., Machanavajjhala, A., Gehrke, J., Halpern, J.Y.: Worst-case background knowledge for privacy-preserving data publishing. In: 2007 IEEE 23rd International Conference on Data Engineering. pp. 126–135. IEEE (2007)
24. Martinez-Martin, N., Wieten, S., Magnus, D., Cho, M.K.: Digital contact tracing, privacy, and public health. Hastings Center Report 50(3), 43–46 (2020)

25. McCreesh, C., Prosser, P., Trimble, J.: The glasgow subgraph solver: using constraint programming to tackle hard subgraph isomorphism problem variants. In: International Conference on Graph Transformation. pp. 316–324. Springer (2020)
26. Meyer, U., Wetzel, S.: On the impact of gsm encryption and man-in-the-middle attacks on the security of interoperating gsm/umts networks. In: 15th International Symposium on Personal, Indoor and Mobile Radio Communications. vol. 4, pp. 2876–2883. IEEE (2004)
27. Meyerson, A., Williams, R.: On the complexity of optimal k-anonymity. In: Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. pp. 223–228 (2004)
28. Milke, V., Stroetmann, L.: Imsi-catcher (2014)
29. Ney, P., Smith, I., Cadamuro, G., Kohno, T.: Seaglass: Enabling city-wide imsi-catcher detection. Proc. Priv. Enhancing Technol. 2017(3), 39 (2017)
30. Ohm, P.: Broken promises of privacy: Responding to the surprising failure of anonymization. UCLA l. Rev. 57, 1701 (2009)
31. Podlesny, N.J., Kayem, A.V., Meinel, C.: Attribute compartmentation and greedy ucc discovery for high-dimensional data anonymization. In: Proceedings of the Ninth ACM Conference on Data and Application Security and Privacy. pp. 109–119 (2019)
32. Podlesny, N.J., Kayem, A.V., Meinel, C.: Identifying data exposure across distributed high-dimensional health data silos through bayesian networks optimised by multigrid and manifold. In: IEEE Intl Conf on Dependable, Autonomic and Secure Computing. pp. 556–563. IEEE (2019)
33. Podlesny, N.J., Kayem, A.V., Meinel, C.: Gpu accelerated bayesian inference for qid discovery in high-dimensional data. In: International conference on advanced information networking and applications (AINA). Springer (2021)
34. Podlesny, N.J., Kayem, A.V., Meinel, C.: A parallel quasi-identifier discovery scheme for dependable data anonymisation. In: Transactions on Large-Scale Data-and Knowledge-Centered Systems. Springer (2021)
35. Podlesny, N.J., Kayem, A.V., von Schorlemer, S., Uflacker, M.: Minimising information loss on anonymised high dimensional data with greedy in-memory processing. In: International Conference on Database and Expert Systems Applications. pp. 85–100. Springer (2018)
36. Polonetsky, J., Tene, O., Finch, K.: Shades of gray: Seeing the full spectrum of practical data de-intentification. Santa Clara L. Rev. 56, 593 (2016)
37. Price, W.N., Cohen, I.G.: Privacy in the age of medical big data. Nature medicine 25(1), 37–43 (2019)
38. Rowe, F.: Contact tracing apps and values dilemmas: A privacy paradox in a neo-liberal world. International Journal of Information Management 55, 102178 (2020)
39. Rubinstein, I.S., Hartzog, W.: Anonymization and risk. Wash. L. Rev. 91, 703 (2016)
40. Sly, L.: Us soldiers are revealing sensitive and dangerous information by jogging. The Washington Post 29 (2018)
41. Sweeney, L., Abu, A., Winn, J.: Identifying participants in the personal genome project by name (a re-identification experiment). arXiv preprint arXiv:1304.7605 (2013)
42. SysŁ, M.M., et al.: The subgraph isomorphism problem for outerplanar graphs. Theoretical Computer Science 17(1), 91–97 (1982)
43. Vessenes, P., Seidensticker, R.: System and method for analyzing transactions in a distributed ledger (Mar 29 2016), uS Patent 9,298,806
44. Yasaka, T.M., Lehrich, B.M., Sahyouni, R.: Peer-to-peer contact tracing: development of a privacy-preserving smartphone app. JMIR mHealth and uHealth 8(4), e18936 (2020)
45. Zheng, J., Liu, S., Ni, L.M.: Effective routine behavior pattern discovery from sparse mobile phone data via collaborative filtering. In: International Conference on PerCom. pp. 29–37. IEEE (2013)
46. Zheng, Y.: T-drive trajectory data sample (August 2011), t-Drive sample dataset