k-Anonymous Crowd Flow Analytics

Valeriu - Daniel Stanciu
v.stanciu@utwente.nl
University of Twente
Enschede, The Netherlands

Maarten van Steen
m.r.vansteen@utwente.nl
University of Twente
Enschede, The Netherlands

Ciprian Dobre
ciprian.dobre@cs.pub.ro
University Politehnica of Bucharest
Bucharest, Romania

Andreas Peter
a.peter@utwente.nl
University of Twente
Enschede, The Netherlands

ABSTRACT

Measuring pedestrian dynamics using the signals sent from smartphones has become popular. Notably, Wi-Fi-based systems are currently widely deployed. However, many such systems have also become subject to serious debate due to privacy infringement. For some time, secure hashing of a smartphone's unique MAC address was considered to be sufficient, yet this method has been overruled by Europe's General Data Protection Regulation which states that an individual should not be identifiable from any dataset without explicit prior consent.

In this paper, we propose a novel anonymization technique that essentially anonymizes detected smartphones immediately at the sensor before any data on such a detection is stored for further analysis. Our solution borrows from the notion of k-anonymity, while avoiding its well-known drawbacks that lead to de-anonymization. Moreover, while ensuring what we coin detection k-anonymity, we also ensure high accuracy of counting measures when dealing with realistic pedestrian flows within crowds. We evaluate our solution both in a simulated environment and in a realistic environment reproducing real-life settings.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing**; • **Security and privacy** → **Privacy protections**.

KEYWORDS

crowd-monitoring, anonymization, pedestrian dynamics

ACM Reference Format:
Valeriu - Daniel Stanciu, Maarten van Steen, Ciprian Dobre, and Andreas Peter. 2020. k-Anonymous Crowd Flow Analytics. In *MobiQuitous 2020 - 17th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MobiQuitous '20), December 7–9, 2020, Darmstadt, Germany.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3448891.3448903

MobiQuitous '20, December 7–9, 2020, Darmstadt, Germany

ACM ISBN 978-1-4503-8840-5/20/11…$15.00
https://doi.org/10.1145/3448891.3448903

1 INTRODUCTION

Understanding pedestrian behavior in crowded public spaces has been a matter of interest for many years. Research within the crowd-dynamics field thoroughly explored movement patterns and different behaviors that can occur inside a crowd at different points in time [21, 25, 28]. It has already been shown that insights can be extremely valuable for urban planning [23], traffic optimization [20, 31], events organization [10, 12, 39, 41], footfall estimation [19, 34] or even public safety [16, 22, 40]. Various technologies have been employed, including video cameras, mechanical counters, RFID beacons and Infrared devices. With the advent of smartphones as personal devices constantly carried by people, an enormous amount of high-accuracy information became available, foaming from inside the crowd and boasting an unprecedentedly intimate whiff, creating opportunities for automated tracking through interfaces such as Bluetooth and Wi-Fi.

While the developments in this field are of undeniable importance, along with them numerous concerns regarding the privacy of individuals emerged. Regardless of the sensing technology, crowd-monitoring solutions have been built without taking privacy into account. More often than not, they rely on individuals having attached unique identifiers to them, leaving the door open for privacy-infringing situations such as malicious tactics of user tracking and profiling, unconsented surveillance or the more frequent nowadays but not less serious situation of unintentional personal data leakage. Moreover, the coming into effect of the EU General Data Protection Regulation [18] (GDPR) steered organizations towards taking personal data seriously; according to this regulation, the kind of information processed and stored with the intent of profiling a natural person, information which, combined with other external knowledge, could lead to uniquely identifying individuals, qualifies as personal data.

To address these concerns, attempts have been made to retrofit the existing systems with privacy-preserving capabilities [30]. The de-facto standard is based on pseudonymization, i.e. the process of replacing the personally identifiable information with a computed artificial identifier calculated by using a powerful one-way cryptographic function. However, it has been shown that this scheme could be broken in a matter of minutes due to the low entropy of the original identifiers [15], an attacker being able to brute-force the entire identifier space. So even by limiting the lifetime of an identifier, the method is still vulnerable. As a result, several crowd-monitoring initiatives have been halted [1, 4, 5], mentioning the privacy of the individuals as main reason. To the best of our knowledge, there is no available solution catering to these privacy needs and fully-adhering to the GDPR, so we question ourselves whether it is possible to come up with a mechanism that can preserve privacy by design while being able to fulfill crowd-monitoring needs.

In this paper, we propose a novel architecture for crowd-monitoring that preserves the privacy of all monitored individuals under anonymity guarantees while maintaining high accuracy of measurements. Our mechanism leverages k-anonymity principles on top of truncated identifiers, dropping the usage of unique identifiers and ensuring, for any formation of crowd-monitoring scenarios, that there is no individual having her privacy compromised. Moreover, the mechanism is computationally lightweight, running in linear time, and can be applied in a live manner right at the collection point even before the sensing data reaches the crowd-monitoring database, thus complying with requirements of anonymization on the fly. We evaluate our construction both in a simulated environment, to test edge cases and behavior when ranging different parameters, and in a realistic environment reproducing real-life settings.

The rest of the paper is organized as follows. Section 2 presents the system model, together with the theoretical grounds supporting our construction. Section III introduces the experimental setup, the metrics used and the employed mechanisms, while in Section IV a thorough evaluation is performed. In Section V a review of related literature is provided and then, finally, Section VI concludes the paper.

2 SYSTEM MODEL

2.1 Overview

Crowd-monitoring is the process of understanding the movement patterns of crowds of people inside a certain public or private environment. Regardless of the technology used for sensing (e.g., Wi-Fi or Bluetooth scanners, video cameras, and so on), it relies on detecting people passing by several collection points at different time intervals. For example, in the case of Wi-Fi, a mobile device regularly broadcasts probe requests containing its MAC address as a unique identifier, which can be subsequently picked up at a Wi-Fi scanner. In a naïve setting, a device detection is constructed at the scanner as a triplet containing a device's MAC address, a timestamp, and the scanner's identifier. Such triplets are stored in a central database for further analysis. Clearly, without taking further measures, privacy infringement is at stake. As an advancement of state-of-the-art methods, we propose to perform a novel anonymization process on the fly, directly at the scanner, or more general at collection points, before detections reach the server, a process that we will introduce later on in this paper. For clarity and without loss of generality, we will assume throughout this paper that Wi-Fi sensors are used.

In our construction, when we talk about movement patterns of crowds, we specifically refer to being able to understand pedestrian crowd flows, i.e. how people constituted in a crowd circulate through public spaces. To achieve this, we need to build our system in such a way that it offers high accuracy of measurements for this kind of scenarios while offering anonymity guarantees for all the data being stored.

2.2 Formalities

A Wi-Fi crowd-monitoring environment, as we define it in our construction, consists of:

- A set S of N scanners, which could be either access points, Wi-Fi sniffers, or any other device able to gather Wi-Fi messages. We make the assumption that scanners have nonoverlapping ranges and they run the protocol as expected.
- A set E of K epochs during which the system runs; the duration of the epochs is established according to the specificity of the environment. We assume that each epoch lasts τ time units. $T = K \cdot \tau$ is the total time span during which we perform crowd-monitoring activities.
- A set IDS of M people being detected throughout our system during T; each person is represented by a unique 48-bit identifier, be it a MAC address or other pseudonym.

A detection is a triplet $(id, s, e), id \in IDS, s \in S, e \in E$, representing a person uniquely identified by id, sensed by scanner s during epoch e. Let $D(s, e) \subset IDS$ be the set of identifiers detected at scanner $s \in S$ during epoch $e \in E$. In our system we assume that, at the end of each epoch, the detections collected by the scanners undergo an *anonymization process* P:

Definition 1. Let 2^{IDS} denote the powerset of the set IDS. We define an **anonymization process *P*** as an algorithm $P : 2^{IDS} \times IDS \longrightarrow PIDS$ that takes a set of identifiers $A \in 2^{IDS}$ and an identifier $id \in A$ as input and outputs an anonymized identifier $pid \in PIDS$, where $PIDS$ denotes the set of all possible identifiers that are anonymized with respect to P, including the special symbol $\bot$ (which captures the "removal" of identifiers for anonymization).

By modelling P to take as input both an identifier id as well as an identifier set A in which id resides, we enable P to anonymize id depending on its "environment" A. To ease readability, for $id \in A \in 2^{IDS}$ we write $P(A, id)$ simply as $P_A(id)$ or even as $P(id)$ if there is no ambiguity about the underlying set A. For any $B \subseteq A$, we interpret $P(A, B)$ as $\bigcup_{b \in B} P_A(b)$. We note that for a set $A \in 2^{IDS}$, $P_A(A)$ defines a multiset for which $m(pid) = |\{j \in A \mid P_A(j) = pid\}|$ denotes the multiplicity of $pid \in P_A(A) \setminus \{\bot\}$[1]. The multiplicity $m(\bot)$ of $\bot$ in $P(A)$ is always set to 1 (as removed identifiers are assumed to be nonreconstructable).

Notation. For a detection set $D(s, e) \subseteq IDS$ and anonymization process P, we denote the multiset $P_{D(s,e)}(D(s, e))$ as $PD(s, e)$.

A simple example of such an anonymization process P is the truncation operation $trunc(id, nb)$ which removes all but the last nb bits from the binary number id, i.e., $trunc(id, nb) = id \mod 2^{nb}$, for all $id \in IDS$. In this example, $IDS \subseteq \{0, 1\}^{48}$ while $PIDS = \{0, 1\}^{nb}$. We will see more examples of anonymization processes P later in the paper.

After undergoing the process P, detections are stored as multisets in a database. The purpose of this *crowd-monitoring database* (*CMD*) is to provide meaningful answers to *crowd-monitoring queries*. Those queries are modelled again as multisets.

Definition 2. For scanners from S and epochs from E we define a **crowd-monitoring query** as a multiset $PD(s, e)$ and any AND-combinations of such multisets. In particular, a **simple query** is

[1] We write $m_A(pid)$ instead of $m(pid)$ when the context is ambiguous.

a single multiset $PD(s, e)$ for some $s \in S$ and $e \in E$. A **composite query** CD is an AND-combination of multisets over a collection $\mathcal{D} = \{PD(s, e)\}$ and is defined as the multiset

$$\{pid^m | pid \in \bigcup_{d \in \mathcal{D}} d; m = \min_d \{m_d(pid)\}\}.$$

Composite queries, as they are defined above, cover a broad spectrum of situations; many of these situations are not relevant for crowd analytics, while some are even impossible (such as detecting the same device at different locations at the same time). As we mentioned, we are interested in composite queries regarding *crowd flows*. Envisioning crowd flows, we expect to encounter people detected as moving together in the form of a crowd between different scanners over time. Under ideal circumstances, a crowd identified as being together at a certain point should be also detected in its entirety as it travels. However, in reality there are people leaving as well as joining a crowd flow, thus creating variations of *ideal crowd flows*. Ideal crowd flows and their variations form the focus of our investigations.

Definition 3. An **ideal crowd flow** (of size n) is a collection of detection sets $\mathcal{D} = \{D(s_1, e_1), \ldots, D(s_n, e_n)\}$ where $e_i < e_j$ for $i < j$, such that $\bigcap D(s_j, e_j) \in \mathcal{D}$. We call this situation "ideal" because in one of its detection sets we capture a crowd which is also *fully encountered* across all the other detection sets. Such an ideal crowd flow is depicted in Fig. 1.

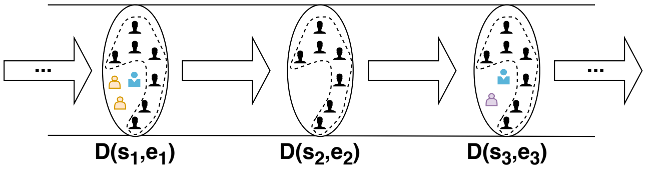

Figure 1: Ideal crowd flow

Definition 4. Let *CF* denote an *ideal crowd flow* of size n. Let $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_{n-1}\}$ be a set of percentages, where λ_i represents the percentage of devices that have left *CF* during e_i (i.e., these devices were detected during e_i, but no longer during e_{i+1}). Likewise, let $\Gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_{n-1}\}$ be a set of percentages, where γ_i represents the percentage of devices that joined *CF* during e_i, meaning that these devices were detected during e_i, but not during e_{i-1}. We define such a flow as a ***(Λ, Γ)-crowd flow***. We display an example in Fig. 2.

3 K-ANONYMOUS CROWD FLOW ANALYTICS

3.1 Metrics

We have shown in the previous section how *CMD* is built and what kind of crowd-monitoring queries are to be performed onto it. Now we focus on how to assess the effectiveness of a given anonymization process P in protecting the anonymity of individuals, as well as to measure its impact on the quality of outcomes expected from the system.

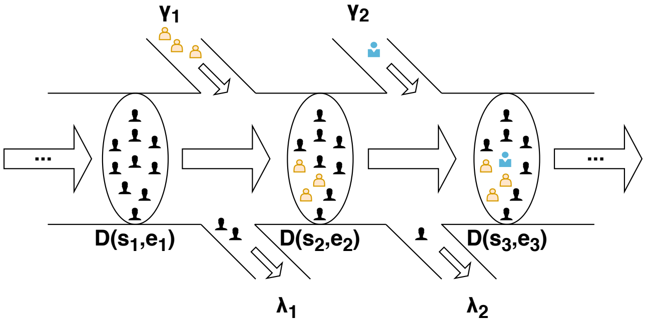

Figure 2: (Λ, Γ)-crowd flow

In terms of anonymity, we adapt the widely used notion of *k-anonymity* [32] to our setting of detection sets and introduce the notion *detection k-anonymity*.

Definition 5. We call an anonymization process P **detection k-anonymous** if $\forall A \subseteq IDS, id \in A : m(P_A(id)) \geq k$ or $P_A(id) = \perp$.

An anonymized identifier *pid* should correspond to at least k identifiers from the original set. Note that for sets smaller than k, the only option is to transform each identifier to $\perp$, since there are not enough identifiers in the original set to proceed differently.

Applying such a process on all the detection sets at collection points generates, by construction, multisets that can yield answers only to *detection k-anonymous simple queries*. We will show that this is sufficient in order to protect the anonymity of individuals under detection k-anonymity guarantees for any crowd-monitoring query, be it simple or composite, as it also exclusively leads to *detection k-anonymous composite queries*.

Definition 6. A (simple or composite) query CD is said to be **detection k-anonymous** if $\forall pid \in CD : m(pid) \geq k$.

Theorem 7. Consider a collection of detection k-anonymous simple queries $\mathcal{D} = \{PD(s, e)\}$ over a set of scanners S and epochs E. The composite query CD obtained by composition over these simple queries is also detection k-anonymous.

Proof. Consider an identifier $pid \in CD$. By definition of a composite query, we know that $m_{CD}(pid) = \min\{m_{PD}(pid)\}$ for any $PD \in \mathcal{D}$ for which $pid \in PD$. As each $PD \in \mathcal{D}$ is detection k-anonymous, we have that $m_{PD}(pid) \geq k$, and thus also $m_{CD}(pid) \geq k$. □

Anticipating further discussions on re-identification, apart from detection k-anonymity, an anonymization process is under scrutiny regarding its *l-surjectiveness*, as defined below.

Definition 8. We call an anonymization process P **l-surjective** if $\forall id \in IDS : m(P_{IDS}(id)) \geq l$.

When P is applied on the entire *IDS*, the resulting multiplicities represent the *actual* number of *physical* devices behind each anonymized identifier in the dataset. Therefore, in other words, an l-surjective anonymization process ensures that any resulting anonymized identifier is shared by at least l real devices in *CMD*, no matter the query. Imagine a trivial process which simply takes a query and makes each identifier in it occur k times. Despite detection k-anonymity being respected, an attacker can immediately trace back to individuals with a probability of 1. To avoid such a situation, l-surjectivity acts as a fallback solution, because the attacker can only guess correctly with a probability of $1/l$. Hence, it is highly important to choose the parameters of the crowd-monitoring system such that they lead to a satisfactory value of l, i.e. $l \gg k$.

Besides measuring the anonymity achieved by individuals, we are interested in the impact on the quality of outcomes expected from the system. Hence, we need to introduce an *accuracy* metric to express how far the answers to crowd-monitoring queries are from their original values after applying the anonymization process.

Definition 9. Let CD be a simple or composite crowd-monitoring query. Then, with $CD^* = CD \cup \{\perp\}$, let $IDS(CD^*)$ denote the identifiers in IDS detected by the scanners, as they were before applying the anonymization process P. We define the **query accuracy** as follows:

$$Acc(CD) = 1 - \frac{Abs(|CD| - |IDS(CD^*)|)}{|IDS(CD^*)|}$$

Abs denotes the absolute value. Special situation: if there was no identifier detected, respectively $|IDS(CD^*)| = 0$, then $Acc(CD) = 1$.

Anonymization could remove identifiers by transforming them to $\perp$, while manipulating the remaining ones together with their occurrences. Recalling this, what Definition 9 actually captures is the relation between the number of anonymized identifiers obtained as answer to the crowd-monitoring query and the number of original, nonanonymized identifiers, as they were before applying any anonymization process.

3.2 Mechanisms

We have as main goals achieving high accuracy for the kind of crowd-monitoring scenarios that we are interested in, as well as preserving the anonymity of all the individuals under detection k-anonymity guarantees. Let us then proceed on a quest addressing, as layers, different mechanisms needed for fulfilling these requirements.

In our system, we perform anonymization at scanner level on an epoch basis. We are willing to manipulate the detection sets in such a way that they deem detection k-anonymous simple queries. This is equivalent to saying that after applying anonymization, for each id from an input set $D(s, e)$, its associated pid should occur at least k times in an output multiset $PD(s, e)$. Following a layered approach, we chain several mechanisms, each of them representing an anonymization process by itself but inflicting changes only to the pid's that have not yet been manipulated to occur at least k times nor ending up to $\perp$.

Pseudonymization, a de-facto standard found both in literature and industry, represents a flavour of an anonymization process P, as it adheres to Definition 1. However, it is a weak mechanism since it simply does a one-to-one mapping of identifiers, leaving no way for k-anonymity aspirations. Nevertheless, it is important to apply it as a first step because it strips the identifiers from any connection with their original meaning.

Applied as a second layer on top of pseudonymization, the previously introduced truncation operation $trunc(id, nb)$ has the potential to achieve better results in terms of anonymization. It can lead to a many-to-one mapping of identifiers if the number of bits to truncate is intelligently chosen, in accordance with the size of the detection sets. While, regardless of the number of bits being truncated, the accuracy of simple queries cannot be affected (resulting multisets have the same sizes as the original sets), the situation is different for crowd flows as we discovered through experiments. To illustrate, we display in Fig. 3 the results of an example experiment concerning a $(\{30\}, \{50\})$-*crowd flow* containing 1000 identifiers and a desired anonymity of k=2. On the y-axis we show both the accuracy and the inherent k-anonymity achieved when ranging nb as displayed on the x-axis. By inherent k-anonymity we mean the fraction of people in a crowd flow that have their corresponding truncated identifier occurring at least k times after applying the truncation alone. When the parameters are chosen in such a way that the crowd-monitoring query gets close to being detection k-anonymous, the accuracy is the lowest. The accuracy gets higher when the inherent k-anonymity decreases and, as a consequence, the query gets farther from being detection k-anonymous. The inflection points of the curves can slide left- or rightwards when Λ, Γ, k or when the crowd size are changed, but the pattern remains the same. Besides that, the truncation operation, as an anonymization process, cannot be even by definition detection k-anonymous because it does not work for settings in which $A \subseteq IDS, |A| < k$.

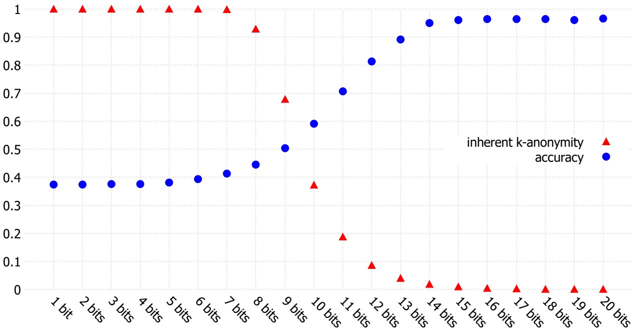

Figure 3: Detection k-anonymity versus accuracy when performing truncation on a $(\{30\}, \{50\})$-*crowd flow* with 1.000 identifiers, *k*=2, *nb* ranges from 1 to 20

Building on top of the detection k-anonymity inherently obtained through truncation, to preserve the high accuracy of queries and, at the same time, resolve the remaining nonanonymized individuals, we present, as a third layer, a *correction mechanism*. A simple method would be to drop the truncated identifiers corresponding to nonanonymized individuals, but this would dramatically lower the accuracies. The same holds for another method at hand, which is inserting copies until each truncated identifier reaches at least k multiplicity. We hypothesize that using a smart combination of consistently adding copies or removing identifiers has a minimum impact on the accuracy.

Definition 10. Let us suppose that we apply a truncation operation keeping nb bits and let $CIDS(nb, k) \subset PIDS$ be a set of identifiers such that $|CIDS(nb, k)| = 2^{nb}/k$. For a resulting simple crowd-monitoring scenario PD we define a **correction mechanism** as a transformation $T : PIDS \longrightarrow PIDS, T(PD) = PD^*$, such that $\forall pid \in PD$,

$$m_{PD^*}(pid) = \begin{cases} m_{PD}(pid), \text{ if } m_{PD}(pid) \geq k \\ k, \text{ if } m_{PD}(pid) < k \text{ } AND \text{ } pid \in CIDS(nb, k) \\ 0, \text{ if } m_{PD}(pid) < k \text{ } AND \text{ } pid \notin CIDS(nb, k) \end{cases}$$

The correction mechanism, as we can see, affects only part of the identifiers: the nonanonymized ones. If we assume a uniform distribution of identifiers at query level, the mathematical expectation (when given enough queries; cf. law of large numbers) is that the inserted identifiers will perfectly balance the removed ones. At the same time, the mathematical expectation is that when composing simple queries into ideal crowd flows, the count of identifiers originally present in the intersection and removed by the mechanism to be equal to the count of the ones present in the intersection after being inserted for detection k-anonymity purposes, thus not affecting the accuracy at all. In reality, though, there will be some limited changes in the accuracy, which we measure through experiments as discussed in Section 4. There are two reasons why accuracy is affected. First, although we can ensure uniform distribution of identifiers globally by, for example, using a uniformly distributed hash function as pseudonymization mechanism, there is no way we can guarantee such uniformity at query level. Second, in reality we encounter (Λ, Γ)-crowd flows rather than ideal crowd flows.

A detailed description of the actual implementation of our entire detection k-anonymous anonymization process is presented in Algorithm 1. The algorithm works with any pseudonymization mechanism of choice, be it hashing, tokenization or other method. As a correction mechanism, we use a best-effort adaptation of Definition 10, which, under the assumption of simple query-level uniformly distributed identifiers, is identical with the original one, but in practice, when the assumption does not hold, it takes care that the accuracy of simple queries is not severely affected. Essentially, instead of fixing $CIDS$, for example, to a uniform random sample of size $(1/k)$-th of the original pseudonyms space, we systematically look at the IDs that violate detection k-anonymity, order them, and keep only the first $(1/k)$-th part.

4 EVALUATION

The anonymization process that we've introduced as a composition of three different mechanisms is detection k-anonymous, protecting the anonymity of individuals for any crowd-monitoring query. In this section we analyze the impact of applying this process on the accuracy of the (Λ, Γ)-crowd flows.

4.1 Simulated environment

To get a clear understanding of the behavior of our anonymization process, in our evaluation we generate detections to emulate the scenarios that we are interested in. By doing this we can freely test our design in numerous settings and we can focus on the process itself as a theoretical construction. There are a number of parameters that shape the experiments, in particular those related to physical settings (size of the crowd, number of epochs, percentages of leavers - Λ, percentages of joiners - Γ) and those related to the anonymization process (values of k, truncation nb parameter).

```
Input: DSE[ ] //Detections made by a scanner during an epoch;
Input: nb //Number of bits to keep;
Input: k //Desired value for k;
Output: PDSE[ ] //Anonymized detections;
TIDS := [ ];
foreach DSE as currentId do
    /* Compute the pseudonym of each identifier */
    pseudoId := computePseudonym(currentId);
    /* Apply the truncation operation */
    truncId := trunc(pseudoId, nb);
    if containsKey(TIDS, truncId) then
        count := getValue(TIDS, truncId);
        updateValue(TIDS, truncId, count+1);
    else
        addKeyValue(TIDS, truncId, 1);
    end
end
/* Apply correction to reach detection k-anonymity */
PDSE := [ ];
breakingPids := [ ]; //pids disobeying detection k-anonymity
foreach TIDS as (pid, count) do
    if count ≥ k then
        addCountCopies(PDSE, pid);
    else
        totalBreaking += count;
        add(breakingPids, pid);
    end
end
sortAscending(breakingPids);
breakingToKeep := floor(totalBreaking/k);
for i := 0 to breakingToKeep do
    addKCopies(PDSE, breakingPids[i]);
end
return PDSE;
```

Algorithm 1: Our anonymization process.

For the simplicity of the exposition and easiness of interpretation, we start by looking at Λ and Γ of length 1, representing crowd flows traveling from one scanner to another between two epochs. We assume to initially have a crowd of 1000 people; part of it travels to the next scanner, part of it leaves the flow, while new people join the flow on the way. The number of bits to keep nb depends on k and on the expected sizes of the crowd. For example, in case of a crowd of 1000 uniformly distributed identifiers, 9 bits ensure almost full inherent 2-anonymity but deem low accuracies; 11 bits still ensure some degree of inherent anonymity even for k equals 3 or 4, but with much higher accuracies. Thus, we fix nb to 11. We do not fix k though, since an acceptable value can only be decided when building the crowd-monitoring system, as part of the design process; instead we run experiments for different values. First we want to see what happens when the percentage of leavers increases and the percentage of joiners remains constant, then we look at the case in which the leavers are fixed and the joiners fluctuate.

In the first experiment we assume that the flow starts with a crowd of 1000 people and, before reaching a second scanner, a fixed

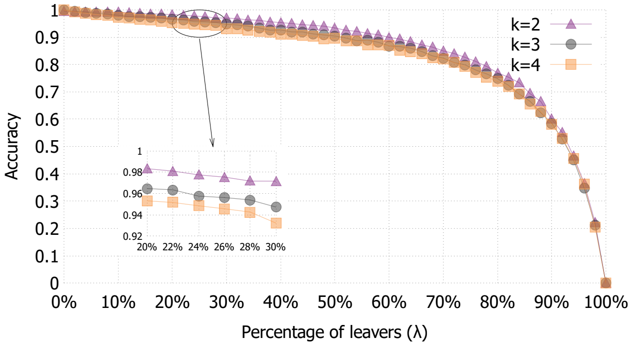

Figure 4: Crowd flows accuracy, γ=20, λ ranges from 0 to 100

number of 200 new people join the flow. In Fig. 4 we display the accuracy of the corresponding composite query having our detection k-anonymous process in action when the percentage of people leaving the flow ranges from 0 to 100%. For each pair (Λ, Γ) we perform 100 simulation runs and the mean values are displayed. Intuitively, the accuracy of queries decreases when the fraction of leavers increases. It slowly decreases as long as there are enough people remaining in the crowd flow; it abruptly decreases when there are more people joining the flow than remaining in it, but at that point we consider that we are not looking at a realistic crowd flow any more since we are already dealing with different crowds mixing together. For the configurations in which the percentage of leavers is lower than 70% the system achieves accuracies higher than 0.8 for all the tested values of k. Note, however, that for higher desired values of k the truncation operation should decrease the number of bits to be kept. The reason for this is to avoid ending up with each truncated identifier occurring k times only because, at that point, a privacy attacker can try guessing with a $1/l$ probability (as l-surjectiveness indicates) who is behind an anonymized identifier.

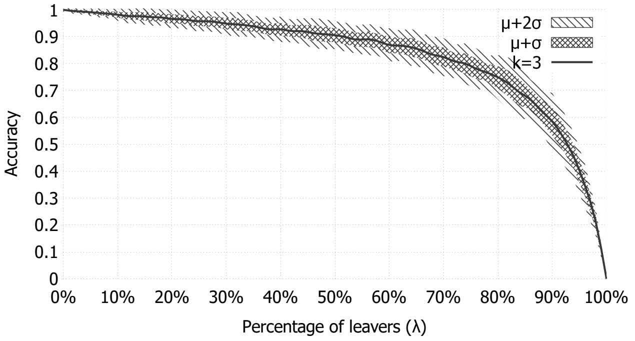

Figure 5: Standard deviations of crowd flows accuracy, γ=20, λ ranges from 0 to 100

For the same settings as above, we plot, for each configuration, the standard deviations within the 100 simulated runs. For clarity reasons, we choose to show this graph separately for one specific value of k; the graph looks similar for other values of k as well. 68% of the accuracy values are within the dotted surface while the striped surface covers 95% of them. Thus, as we can see, they are close together, the standard deviation ranging between 0.003 (0% leavers) and 0.052 (90% leavers).

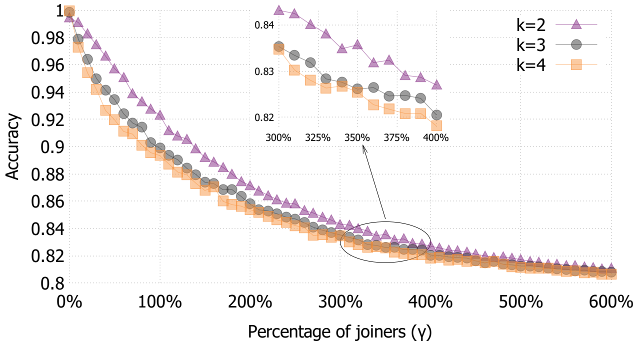

Figure 6: Crowd flows accuracy, λ=20, γ ranges from 0 to 600

A second experiment concerns situations in which the percentage of people leaving the flow remains constant, while the number of people joining the flow increases. Again we start with a crowd of 1000 people, a fixed percentage of 20% of them leave the crowd before reaching a second scanner and between 0 and 600% new people join; we perform 100 runs for each (Λ, Γ) pair and we display the mean value. The results are displayed in Fig. 6. We interpret them as follows: the accuracy stays above 0.9 as long as the remaining crowd is larger than the number of new people joining the flow and it can only go as low as 0.8 (note the y-axis) when there are 6 times more people joining the crowd than they were originally in it (in our case, 6000 new people joining). This lower bound has a theoretical explanation, being dictated by the value of Λ. In a worst-case scenario, the number of actual leavers may go completely undetected. This can happen when among the increasing number of joiners there are, after anonymization, enough identifiers to match all of the 200 leaving persons. Calculated according to the accuracy formula, the theoretical lower bound in this case is 0.75.

Now that we understand how our detection k-anonymous process influences the accuracies of crowd-monitoring queries when applied to simple (Λ, Γ)-crowd flows, let us move forward and investigate realistic scenarios from a well-known real-life deployment.

4.2 Reproducing real-life deployment settings

To get insights on how people use public underground transport and to explore potential improvements, in 2016 Transport for London conducted a pilot Wi-Fi data collection experiment across 54 London Underground stations [2], publishing their findings in a detailed review [3]. Salted hashing of MAC addresses was used as a privacy-preservation mechanism, a typical example of pseudonymization bearing all the pitfalls mentioned in the introduction. Starting with July 2019, data collection is performed across the whole London Underground network, using tokenization (i.e., the assignment of a unique random value to each MAC address) instead of hashing. Considering that tokenization does not solve the previously presented issues of pseudonymization, we investigate what impact our anonymization process has on their results, arguing that our solution can be successfully applied in such settings, evolving from pseudonymization to anonymization.

The experimental Wi-Fi crowd-monitoring environment, in this case, consists of a set S of 1070 scanners distributed across the 54 stations, a set E of epochs covering the total timespan T of the experiment (from 21 November to 19 December 2016) and a set IDS of 5.6 million devices detected by any of the 1070 scanners during the experiment. Choosing, for example, the epoch length τ as 1 minute would mean that the total number of epochs in this experiment is 41760. Then, fixing the number nb of bits to be kept to 11 and running 100 experiments concerning 5.6 million uniformly distributed identifiers, we could see that, on average, at least l identifiers map to each anonymized identifier, with $l = 2559$. The main concern of the study, besides looking at statistical values and measurements, was to visualize the real flows of people inside the Underground network, to see the specific routes that are chosen between a source and a destination station, to measure train-level congestion and crowdedness. This is equivalent with having a look at the devices detected by a number of scanners in a sequence of epochs representing a journey, successfully mapping to the *crowd flows* introduced in this paper.

As building blocks, we need to correctly identify scanner-epoch combinations in order to be able to spot the devices carried by persons taking a specific train, as well as the devices carried by persons who get off a train. By doing this, we are able to model the whole range of situations, i.e. the start of a journey, intermediate connections, and the end of a journey. Assuming that the train schedules are known, the solution for identifying the size of the crowd making a journey on a specific route should take into account the detections made in the following settings:

- Scanner s_s on the platform of the origin station, epoch e_s before a train arrives
- Scanners on the platforms of the connecting stations, epochs before the intermediate trains arrive
- Scanner s_d on the platform of the destination station, epoch e_d after the train of that journey has completely departed from the destination station

Some could argue that the assumption that the devices are indeed detected within the specified epochs is unrealistic due to heterogeneous crowd dynamics or sensing technology limitations. That does not affect our argumentation though since it is not related to our anonymization process; this has to do with the baseline functioning of the crowd-monitoring system itself, which is a distinct problem.

The London Underground Network has some particularities making us claim that most of the journeys can be uniquely modelled through two-epoch crowd flows, regardless of the source, destination or number of connecting stations. Looking at, for example, all the 17 routes between King's Cross St. Pancras and Waterloo, as they are presented in the published review [3], one can immediately see that 12 of them have unique combinations of source and destination platforms. This means that in this case, for each source platform s and destination platform d, it is enough to simply look at the detections made by scanners s_s and s_d during the epochs matching the beginning and the end of the analyzed journey. Detections made at intermediary stations are not needed for shaping the crowd flow since the routes are already unique by source and destination. The remaining five routes have the same combinations of source and destination platforms, but contain different connecting stations. Even if it seems rare, we could encounter the following situation: some people begin a journey on the same train, they then take different paths at some point, and then they end up, again, on the same train, arriving together at the destination. To model these alternative paths, three-epoch crowd flows are needed. Please note, however, that if, according to the circulation schedule, the alternative paths cannot lead to boarding on the same final train, a two-epoch crowd flow is still sufficient for modelling even these granular routes.

The accuracy of the crowd-monitoring queries related to (Λ, Γ)-crowd flows is highly influenced by the percentages of leavers and joiners. In the current environment, the leavers are those detected during e_s but remaining on the platform after the source train departs (we can assume that they are waiting for another train), plus the ones taking the source train and going to a different destination than the one that we are looking at. The joiners are the persons detected on the platform after the destination train has left and were either there before the train arrived (we can assume, again, that they are waiting for another train) or came by train but from another source station than the one that we are looking at. We already know from previous experiments that our anonymization process performs well in terms of accuracy when the leavers and joiners are not overwhelmingly high relative to the size of the crowd, indicating popular routes as candidates for high accuracies. We can then immediately see that crowd flows originating or ending on platforms which have a unique line passing through them have a higher chance of achieving high accuracy. Since there are no trains going somewhere else to be waited for, the leavers and joiners would be at a minimum.

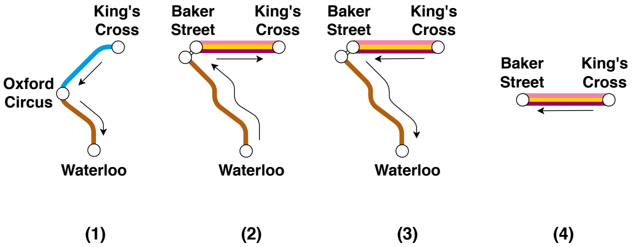

Figure 7: London Underground route types

With respect to the layout of the source and destination platforms, we identify four categories of routes, which we also depict in Fig. 7:

1. Unique lines going through both source and destination platforms, e.g. King's Cross St. Pancras (light blue) - Oxford Circus - Waterloo (brown)
2. One line going through the source platform and multiple lines through the destination platform, e.g. Waterloo (brown) - Baker Street - King's Cross St. Pancras (yellow/violet/pink)
3. Multiple lines going through the source platform and one line through the destination platform, e.g. King's Cross St. Pancras (yellow/violet/pink) - Baker Street - Waterloo (brown)

(4) Multiple lines going through both source and destination platforms, e.g. King's Cross St. Pancras (yellow/violet/pink) - Baker Street (yellow/violet/pink)

We perform example experiments regarding the four different categories of routes, running 100 rounds and computing the mean accuracy for each. For comparison reasons, we fix the following settings: people that enter and exit a train (200), people following the analyzed route (100), people on the train to destination coming from other directions (200-100=100), total number of people on a platform having 3 lines going through it (500). Recalling that within (Λ, Γ)-crowd flows the values of λ and γ represent percentages, the routes can be mapped to crowd flows like this:

(1) ({50},{50})-crowd flow
(2) ({50},{200})-crowd flow
(3) ({80},{20})-crowd flow
(4) ({80},{80})-crowd flow

Table 1: London Underground routes accuracies

Route type	Accuracy (k=2)	Accuracy (k=3)	Accuracy (k=4)
(1)	0.9502	0.94	0.9195
(2)	0.8742	0.8589	0.8493
(3)	0.8651	0.8443	0.8378
(4)	0.6194	0.5788	0.5774

The results of the experiments can be seen in Table 1. These results would be achieved using the already existing sensing infrastructure, without any modifications, as it is currently deployed in the London Underground Network. The impact that our anonymization process has on the accuracy of crowd monitoring queries concerning the first three types of routes is low. For the fourth type of route, we cannot accurately capture the crowd flow by using the existing sensing infrastructure alone. The reason is that we are trying to identify a relatively small crowd in relation to an overwhelming number of leavers and joiners. A solution at hand for accurately capturing this situation would be augmenting the sensing infrastructure with scanners placed directly on the trains. Otherwise, measurements shall be done only for situations where either the source or the destination allows us to do accurate counting.

5 RELATED WORK

Performing crowd-monitoring by leveraging the communication interfaces of the widely-available modern smartphones is currently done at large scale. Numerous ways of doing it are already out there, having different approaches on individuals' privacy or anonymization issues. To have a clear understanding of the domain, in this section we are going to first look into works related to crowds being monitored, focusing more on pedestrian tracking, flow identification and privacy-preservation approaches. Then we will dig deeper into anonymity matters, k-anonymity and state of the art developments in this field influencing our work.

5.1 Crowd-monitoring and privacy

Mobile devices have communication interfaces that allow us to detect them when they are in the vicinity of a sensing infrastructure. Research has shown that Wi-Fi and Bluetooth interfaces [6, 12, 33] are highly appropriate for unobtrusively detecting the behavior of crowds of people. In Wi-Fi setups, the MAC address of the devices carried by people is detected by fixed scanners whenever they transmit probe requests meant to discover available networks. Bluetooth sensing is performed by fixed scanners which send, periodically, inquiry requests to nearby devices and then receive responses containing the MAC addresses of the devices. For pedestrian monitoring however, Wi-Fi proved to be the better choice due to higher range, more discoverable devices and lower deployment costs [33]. For an extensive study of the matter, we refer the reader to [17].

Being able to collect precise information on the whereabouts of individuals without any explicit consent is a major privacy problem. This has been investigated by both hardware manufacturers and for crowd-monitoring deployments.

Hardware manufacturers tried to address this issue by implementing MAC address randomization, so that each time a device sends out probe requests, a random pseudonym address is used instead of the real address. In [27], however, building on work presented in [37], the authors show that there is a wide range of techniques which can effectively derandomize most of the implementations on the market. Besides that, devices use their real MAC addresses when they are connected to a network, a situation in which randomization is of no use.

Regarding crowd-monitoring, too few measures have been taken to address privacy concerns. The prevalent approach relies on using pseudonyms instead of the real identifiers, computed by the collecting side either by using one-way hash functions, encryption or randomization. Demir et al. [15] investigated the various employed techniques, concluding that many of the schemes in use can be broken in a matter of minutes and, anyhow, none of them is safe in the long term due to computational power constantly increasing. Furthermore, a survey by Draghici and van Steen [17] discovered that not using privacy-preserving methods is far from being a rare event.

A recent work by Allagan et al. [8] tries to solve the privacy problem in Wi-Fi crowd-monitoring by introducing differentially pan-private Bloom filters (BLIPs). Their method works well for epochs containing large crowds (e.g. tens of thousands), but, due to its differential privacy nature, it falls short when it comes to smaller crowds; our proposed solution can deal with all kinds of crowd sizes.

5.2 Anonymity

Anonymity, as a means of achieving privacy, is defined in [38] as noncoordinatability of traits such that a person is nonidentifiable. Privacy regulations regarding data processing carefully consider this aspect. For example, GDPR recital 26 [18] states that if personal data is rendered anonymous in such a manner that the data subject is no longer identifiable, then data protection principles do not apply any more, thus indicating anonymization as a very powerful mechanism for achieving privacy. Along with this, it explicitly mentions pseudonymization as a counterexample. Our anonymization process is tailored in such a way that every individual present in *CMD* is proven to be protected under detection k-anonymity guarantees, no matter what crowd-monitoring query is performed.

First introduced in [32] by Samarati and Sweeney as a privacy-preserving policy for data releasing and then extended in [35], k-anonymity is defined as the kind of protection achieved when the information about a person contained in a release is indistinguishable from k-1 other persons. Opportunely, k-anonymity is indicated as an acceptable anonymization technique by the European Data Protection Board[11], making it a strong starting point when designing a mechanism to protect information about individuals under GDPR. In our case, the information to protect would be the mere presence of a person near a scanner during an epoch or in a crowd flow. This presence is indicated by the person having a unique identifier which is displayed among the results of a crowd-monitoring query. Hence, pursuing detection k-anonymity comes naturally. However, to the best of our knowledge, there is no investigation performed regarding this particular setting.

Using k-anonymity alone is prone to several attacks, such as homogeneity attack and background knowledge attack, as suggested by Machanavajjhala et al. [26]. To avoid these, the authors propose another technique, l-diversity, to ensure that for every equivalence class of size greater or equal to k there are at least l well-represented values for the sensitive attribute. Considering the scanner and the epoch as nonsensitive attributes and the identifier as a sensitive attribute, we can clearly see that these attacks are not possible in our setting and l-diversity suddenly becomes a nonproblem. The reason is that detection k-anonymity is achieved by manipulating the original identifiers into anonymized identifiers occurring multiple times, so even if the anonymized identifiers in a query are all identical, in fact they correspond to distinct values. We do protect against another kind of diversity attack though, through l-surjectiveness, as previously described in the paper.

Anonymization techniques based on k-anonymity are present in numerous domains; relevant to our work are approaches regarding location-based services (LBS), moving-objects databases (MOD), as well as trajectory databases. All have in common the spatio-temporal dimension of the data being protected. In [9], Bettini et al. look into k-anonymity for location-based services, where a geo-localized history of user requests to a service provider can reveal sensitive information about individuals. Indirectly, such history is, in fact, a trajectory, and can be related to persons being sensed in a crowd-monitoring environment. However, their solution, which is based on historical k-anonymity, does not work for our settings since it only ensures that there are at least k people launching requests across the same spatio-temporal history, thus not protecting an individual's presence per se. In [7], Abul et al. propose (k,δ)-anonymity for trajectories, such that there are at least k trajectories found within a cylinder of uncertainty having the radius δ. Trajectory translations should be performed until such conditions are met. This concept is very closely related to ours and, if we consider the scanners as central points and their ranges as δ, translations are not even required because our system does not store localization information other than the position of scanners. In other words, any trajectory would already be within such a cylinder. Even so, we do not have trajectories in our system, but detections that deem trajectories only after they have already reached *CMD*. This is why this solution cannot be applied on the fly, at the collection point, as we demand. For the same reasons, similar solutions presented in [29], [36] or [14] do not suit our problem.

Finally, there are several studies about ensuring k-anonymity on the fly for data streams, such as [42], [13] or [24]. In essence, these methods ensure that streaming data is made k-anonymous before publishing, just like we do. In contrast to our work, all the existing works consider a setting in which a single trusted server collects and stores the *raw (nonanonymized)* stream of data (typically from one source) which it turns into a k-anonymous form before final publishing; this works by taking the complete history of the data stream into account for the anonymization procedure. Unfortunately, this approach does not work in our setting where the anonymization must happen at each data source in isolation (i.e., at each scanner in our case) before it reaches the server *and* without access to the history of the complete data stream that ultimately consists of the data from multiple sensors. This difference, i.e. the anonymization of data at each sensor in isolation as opposed to the anonymization at the central collection point, defines the major challenge that we tackle in our paper.

6 CONCLUSION

In this paper, we addressed the problem of privacy-preservation through anonymity in crowd-monitoring systems. Our aim was to ensure that the privacy of each monitored individual is preserved while the system can still offer meaningful insights regarding pedestrian dynamics. Having privacy-by-design principles in mind, we designed a lightweight anonymization process to be executed right on the crowd-monitoring sensors, before forwarding the data to a central server. This process manipulates the detected identifiers of individuals through a series of pseudonymization, truncation and correction operations. After these operations are executed, every individual whose smartphone is being monitored ends up being protected with anonymity guarantees, making our solution GDPR compliant as well. In our construction, we introduce detection k-anonymity as anonymity metric, ensuring that there is no crowd-monitoring query in which there are anonymized identifiers occurring fewer than k times each. Besides that, we introduce l-surjectiveness as a metric indicating the number of real devices behind any anonymized identifier.

We evaluated our anonymization process on pedestrian crowd flows, first in a purely simulated environment, then in an environment reproducing the real-life deployment from the London Underground Network. Results show that our anonymization process has a low impact on the accuracies of queries related to crowd flows suffering small perturbations, i.e., relatively few people leaving or joining the flow. In the realistic environment reproduction, the accuracy of such queries stays above 0.8 for all the tested values of k, topping at 0.9502 for k=2 in the case of a ({50},{50})-crowd flow. The impact is higher for crowd flows suffering big perturbations and having a relatively small size in comparison with the number of leavers and joiners. However, this is a desirable result, as we designed our system to also protect the anonymity of people in small crowds, hence offering lower accuracy in those cases. Our experimental results confirm this behavior.

In future work, we plan to realize a practical implementation of our anonymization process to be installed on scanners. Then we are going to deploy it within an actual crowd-monitoring system, to assess its behavior under real-life conditions.

REFERENCES

[1] 2015. https://marketinglaw.osborneclarke.com/advertising-regulation/jc-decauxs-pedestrian-tracking-system-blocked-by-french-data-regulator/. [Online; accessed 04-June-2020].

[2] 2017. https://www.gizmodo.co.uk/2017/02/heres-what-tfl-learned-from-tracking-your-phone-on-the-tube/. [Online; accessed 04-June-2020].

[3] 2017. http://content.tfl.gov.uk/review-tfl-wifi-pilot.pdf. [Online; accessed 04-June-2020].

[4] 2018. https://www.rtvoost.nl/nieuws/303852/Enschede-stopt-tijdelijk-met-wifi-tellingen-na-publicatie-Autoriteit-Persoonsgegevens. [Online; accessed 04-June-2020].

[5] 2018. https://www.autoriteitpersoonsgegevens.nl/nl/nieuws/bedrijven-mogen-mensen-alleen-bij-hoge-uitzondering-met-wifitracking-volgen. [Online; accessed 04-June-2020].

[6] Naeim Abedi, Ashish Bhaskar, and Edward Chung. 2013. Bluetooth and Wi-Fi MAC address based crowd data collection and monitoring: benefits, challenges and enhancement. (2013).

[7] Osman Abul, Francesco Bonchi, Mirco Nanni, et al. 2008. Never Walk Alone: Uncertainty for Anonymity in Moving Objects Databases.. In *ICDE*, Vol. 8. 376–385.

[8] Mohammad Alaggan, Mathieu Cunche, and Sébastien Gambs. 2018. Privacy-preserving Wi-Fi Analytics. *Proceedings on Privacy Enhancing Technologies* 2018, 2 (2018), 4–26.

[9] Claudio Bettini, X Sean Wang, and Sushil Jajodia. 2005. Protecting privacy against location-based personal identification. In *Workshop on Secure Data Management*. Springer, 185–199.

[10] Ulf Blanke, Gerhard Tröster, Tobias Franke, and Paul Lukowicz. 2014. Capturing crowd dynamics at large scale events using participatory gps-localization. In *2014 IEEE Ninth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)*. IEEE, 1–7.

[11] European Data Protection Board. 2014. Opinion 05/2014 on Anonymisation Technique. https://www.pdpjournals.com/docs/88197.pdf. [Online; accessed 04-June-2020].

[12] Bram Bonné, Arno Barzan, Peter Quax, and Wim Lamotte. 2013. WiFiPi: Involuntary tracking of visitors at mass events. In *2013 IEEE 14th International Symposium on" A World of Wireless, Mobile and Multimedia Networks"(WoWMoM)*. IEEE, 1–6.

[13] Jianneng Cao, Barbara Carminati, Elena Ferrari, and Kian Lee Tan. 2008. Castle: A delay-constrained scheme for k s-anonymizing data streams. In *2008 IEEE 24th International Conference on Data Engineering*. IEEE, 1376–1378.

[14] Rui Chen, Benjamin CM Fung, Noman Mohammed, Bipin C Desai, and Ke Wang. 2013. Privacy-preserving trajectory data publishing by local suppression. *Information Sciences* 231 (2013), 83–97.

[15] Levent Demir, Mathieu Cunche, and Cédric Lauradoux. 2014. Analysing the privacy policies of Wi-Fi trackers. In *Proceedings of the 2014 workshop on physical analytics*. ACM, 39–44.

[16] Nikolaos Doulamis. 2009. Evacuation planning through cognitive crowd tracking. In *2009 16th International Conference on Systems, Signals and Image Processing*. IEEE, 1–4.

[17] Adriana Draghici and Maarten Van Steen. 2018. A survey of techniques for automatically sensing the behavior of a crowd. *ACM Computing Surveys (CSUR)* 51, 1 (2018), 21.

[18] European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

[19] Yuuki Fukuzaki, Masahiro Mochizuki, Kazuya Murao, and Nobuhiko Nishio. 2015. Statistical analysis of actual number of pedestrians for Wi-Fi packet-based pedestrian flow sensing. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*. ACM, 1519–1526.

[20] Antonio Guillén-Pérez and María Dolores Cano Baños. 2018. A WiFi-based method to count and locate pedestrians in urban traffic scenarios. In *2018 14th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*. IEEE, 123–130.

[21] Dirk Helbing, Péter Molnár, Illés J Farkas, and Kai Bolay. 2001. Self-organizing pedestrian movement. *Environment and planning B: planning and design* 28, 3 (2001), 361–383.

[22] Anders Johansson, Dirk Helbing, Habib Z Al-Abideen, and Salim Al-Bosta. 2008. From crowd dynamics to crowd safety: a video-based analysis. *Advances in Complex Systems* 11, 04 (2008), 497–527.

[23] Constantine E Kontokosta and Nicholas Johnson. 2017. Urban phenology: Toward a real-time census of the city using Wi-Fi data. *Computers, Environment and Urban Systems* 64 (2017), 144–153.

[24] Jianzhong Li, Beng Chin Ooi, and Weiping Wang. 2008. Anonymizing streaming data for privacy protection. In *2008 IEEE 24th International Conference on Data Engineering*. IEEE, 1367–1369.

[25] Jian Ma, WG Song, Siu Ming Lo, and ZM Fang. 2013. New insights into turbulent pedestrian movement pattern in crowd-quakes. *Journal of Statistical Mechanics: Theory and Experiment* 2013, 02 (2013), P02028.

[26] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. 2006. l-diversity: Privacy beyond k-anonymity. In *22nd International Conference on Data Engineering (ICDE'06)*. IEEE, 24–24.

[27] Jeremy Martin, Travis Mayberry, Collin Donahue, Lucas Foppe, Lamont Brown, Chadwick Riggins, Erik C Rye, and Dane Brown. 2017. A study of MAC address randomization in mobile devices and when it fails. *Proceedings on Privacy Enhancing Technologies* 2017, 4 (2017), 365–383.

[28] Mehdi Moussaïd, Dirk Helbing, Simon Garnier, Anders Johansson, Maud Combe, and Guy Theraulaz. 2009. Experimental study of the behavioural mechanisms underlying self-organization in human crowds. *Proceedings of the Royal Society B: Biological Sciences* 276, 1668 (2009), 2755–2762.

[29] Mehmet Ercan Nergiz, Maurizio Atzori, and Yucel Saygin. 2008. Towards trajectory anonymization: a generalization-based approach. In *Proceedings of the SIGSPATIAL ACM GIS 2008 International Workshop on Security and Privacy in GIS and LBS*. ACM, 52–61.

[30] Future of Privacy Forum. 2013. https://fpf.org/wp-content/uploads/10.22.13-FINAL-MLA-Code.pdf. [Online; accessed 04-June-2020].

[31] Mikko Perttunen, Vassilis Kostakos, Jukka Riekki, and Timo Ojala. 2015. Urban traffic analysis through multi-modal sensing. *Personal and Ubiquitous Computing* 19, 3-4 (2015), 709–721.

[32] Pierangela Samarati and Latanya Sweeney. 1998. *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression*. Technical Report. technical report, SRI International.

[33] Lorenz Schauer, Martin Werner, and Philipp Marcus. 2014. Estimating crowd densities and pedestrian flows using wi-fi and bluetooth. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*. ICST (Institute for Computer Sciences, Social-Informatics and …, 171–177.

[34] Balamurugan Soundararaj, James Cheshire, and Paul Longley. 2019. Estimating real-time high-street footfall from Wi-Fi probe requests. *International Journal of Geographical Information Science* (2019), 1–19.

[35] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (2002), 557–570.

[36] Manolis Terrovitis and Nikos Mamoulis. 2008. Privacy Preservation in the Publication of Trajectories.. In *MDM*, Vol. 8. 65–72.

[37] Mathy Vanhoef, Célestin Matte, Mathieu Cunche, Leonardo S Cardoso, and Frank Piessens. 2016. Why MAC address randomization is not enough: An analysis of Wi-Fi network discovery mechanisms. In *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security*. ACM, 413–424.

[38] Kathleen A Wallace. 1999. Anonymity. *Ethics and Information technology* 1, 1 (1999), 21–31.

[39] Jens Weppner, Benjamin Bischke, and Paul Lukowicz. 2016. Monitoring crowd condition in public spaces by tracking mobile consumer devices with wifi interface. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. ACM, 1363–1371.

[40] Martin Wirz, Tobias Franke, Daniel Roggen, Eve Mitleton-Kelly, Paul Lukowicz, and Gerhard Tröster. 2012. Inferring crowd conditions from pedestrians' location traces for real-time crowd monitoring during city-scale mass gatherings. In *2012 IEEE 21st International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises*. IEEE, 367–372.

[41] Martin Wirz, Eve Mitleton-Kelly, Tobias Franke, Vanessa Camilleri, Matthew Montebello, Daniel Roggen, Paul Lukowicz, and Gerhard Troster. 2013. Using mobile technology and a participatory sensing approach for crowd monitoring and management during large-scale mass gatherings. In *Co-evolution of Intelligent Socio-technical Systems*. Springer, 61–77.

[42] Bin Zhou, Yi Han, Jian Pei, Bin Jiang, Yufei Tao, and Yan Jia. 2009. Continuous privacy preserving publishing of data streams. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*. ACM, 648–659.