Generating K-Anonymous Logs of People-Tracing Systems in Surveilled Environments

Francesco Buccafurri, Gianluca Lax, Serena Nicolazzo, and Antonino Nocera

DIIES, University of Reggio Calabria, Loc. Feo di Vito, 89122 Reggio Calabria, Italy
{bucca,lax,s.nicolazzo,a.nocera}@unirc.it

Discussion Paper

Abstract. In surveilled environments, physical access of individuals can be achieved by a human through mechanical means such as locks and keys, or through technological means such as access control systems based on magnetic stripe, barcode, smart cards, biometric devices, RFID, cameras, and so on. Besides the importance of monitoring people accessing these places, another relevant issue concerns the possibility of tracking them inside the environment. Indeed, in this way, we can have information about the movements of people at any time and, in case of an incident, the analysis of these logs can be decisive to have a complete and fast reconstruction of this event. However, privacy right typically makes this solution unrealizable. In this paper, we discuss this topic and propose a technique to generate logs that allows us to trace people with a certain degree of uncertainty, in such a way that privacy is fully preserved. From this point of view, logs are generated according to a new \( k \)-anonymity property, for which we are able to guess the location of an individual, at a given time, with probability \( k^{-1} \). A number of experiments show that the proposed method reaches the target in a good way, thus validating the approach. An important aspect of our technique is that it is implementable via very cheap devices, which is a relevant issue in pervasive environments where wireless devices with limited processing capability and power have to be utilized.

1 Introduction

In surveilled environments, where the physical access of individuals is controlled, a high level of security is reached if people can be traced everywhere, using for example RFID-based technology, in such a way that we have logs reporting at any time people localization. Consider, for example, the case of a museum, an airport, a railways station, etc. The (a-posteriori) analysis of logs, can provide decisive information in case of a security incident. There are realistic possibilities to apply a similar approach because, usually, physical access to surveilled environments requires people registration. Unfortunately, in most cases, a similar solution is intolerable for privacy reasons, often not compliant with law requirements. In
other words, we cannot assume in general that the importance of having precise information about people’ location in critical environments is stronger than the right of keeping private the access to some places and, more in general, the exact movements that people do in an environment. Therefore, there exists a trade-off between surveillance requirements and privacy rights whose solution is still an open issue in the scientific community. This paper proposes a technique to generate logs that allows us to trace people, as described earlier, but introducing a certain degree of uncertainty, in such a way that privacy is fully preserved. Logs fulfill a non-classical $k$-anonymity property, for which we are able to guess the location of an individual, at a given time, with probability $k^{-1}$. This form of $k$-anonymity$^1$ represents a novel privacy requirement which acts in a specular way w.r.t. classical $k$-anonymity. The concept of $k$-anonymity was originally designed by Samarati and Sweeney in the field of database privacy [11, 12]. The general idea of $k$-anonymity location [5, 13, 4] is that the position of a user is given provided that the probability of identifying her is less than $k^{-1}$. A similar approach to protect location data consists in creating areas of confusion where the traces from several users converge [10, 9]. Location cloaking [6–8] aims to perturb location data by introducing random noise in order to guarantee user’s privacy. An extension of $k$-anonymity for specific types of data, like spatio-temporal data, is the $(k, \delta)$-anonymity [1, 2], which is specifically designed for uncertain trajectories defined as the movement of an object on the surface of the Earth.

Observe that, while the classical $k$-anonymity localization aims to satisfy privacy requirements by changing or extending the exact user positions in such a way that $k$ users are confused each other, our approach is such that, given a location log, it returns a number of $k$ possible users located in this place, with no detectable correlation. It is worth noting that this new concept of $k$-anonymity meets the security requirement of identifying the person(s) who was present in a given location at a given time, with uncertainty equal to $k$. This means for example that, in case of a security incident (theft, damage, an explosion, etc.) we can restrict the pool of suspected individuals from the whole (potentially huge) population to a few (i.e., $k$) people. An important aspect of our technique is that it is implementable via very cheap devices, which is a relevant issue in pervasive environments where wireless devices with limited processing capability and power have to be utilized. Finally, a number of experiments show that the proposed method reaches the target, thus validating the approach.

## 2 The Reference Scenario

In this section, we describe the scenario considered in our study, the addressed problem and the proposed solution. We consider a tracing system in surveilled environments where logs must be kept for security purposes (i.e., banks, museums, offices, tribunals, etc.). In particular, we require that it should be possible

--

$^1$ Observe that we overload the classical term $k$-anonymity (with different meaning but based on the same principle) in the specific localization context.
to guess, with a maximum approximation of $\delta_s$ meters, the location of a given person in $\delta_t$ seconds with a probability $k^{-1}$, where $\delta_s$, $\delta_t$, and $k$ are positive integer numbers. In words, for example by fixing $\delta_s = 10$, $\delta_t = 60$, and $k = 3$, we require that, in case of necessity, it must be possible to determine in at most 1 minute, three possible places where the person can stay with an approximation up to 10 meters. The approach we follow to address this problem is based on the use of active RFID tags and readers. First, the environment is partitioned into a number of cells by properly positioning the RFID readers in known locations. The position of each reader depends on the power transmission level and the read distance and has to guarantee that each portion of the place is covered by a reader. Moreover, each person is equipped with an active RFID tag that replies to reader’s requests by sending a (dynamic) integer number, say $ID$. Periodically, each tag sends the dynamic $ID$ to the nearest reader. The readers have a fixed and known position and are connected to a local network in such a way that all read information (i.e., the ID sent by the tag and the position of the reader that received this ID) can be processed by a server. The idea is that the ID generated by each RFID tag changes (in a pseudo random way) at each reading and only the server is able to calculate which ID is generated by a person. As a consequence, a malicious attempt to track a person movement by sniffing the generated number or by accessing the logs on the server fails if the attacker cannot associate the generated ID to the person. To allow the localization of a person with a probability $k^{-1}$, our approach implements a mechanism by means of which we can guarantee with a high probability that at any time there exist $k$ different people sending the same ID.

The protocol used to locate a person $P$ works as follows. The server computes the current ID of $P$ and a request for the tag with this ID is sent by all the readers. The $k$ tags with this ID ($P$ is among them) respond to the request, and by the knowledge of the location of the readers that have received a reply, it is possible to guess the location of $P$ with probability $k^{-1}$.

**Dynamic ID generation**

The notion of $k$-anonymity introduced above implies that the probability of identifying a person is $k^{-1}$, where $k$ is a given anonymity requirement. Our approach reaches this goal by suitably reducing the domain of IDs, that is, the number of possible values of an ID, to force that more people generate the same ID. Our method is $\epsilon$-approximate as the above requirement is guaranteed with a probability equal to $1 - \epsilon$, with $0 < \epsilon < 1$. Our solution is parametric w.r.t. two integer values $d$ and $a$: $d$ is an integer positive number and $a$ is a real number in the interval $[0, 1]$. The notations used throughout the paper are the following: $p$ is the number of people; $d$ is the number of possible IDs; $C$ is the number of colliding people (random variable); $k$ is the anonymity requirement; $\epsilon$ is the approximation degree; $d_R - 1$ is the maximum integer generated by the PRNG.

Each tag contains a clock, a pseudo-random generator PRNG generating integer numbers in the interval $[0, d_R - 1]$ with $d_R \gg d$, and a random permutation function RPF operating on the set of integers $[1, p]$. The seed of the PRNG of
each tag is known by the server and a different integer $\tilde{ID}(0)$ belonging to the interval $[1, p]$ is assigned to each of the $p$ people. As for the random permutation function, we use the following: $RPF : Z_p^* \rightarrow Z_p^*$, where $Z_p^*$ is the multiplicative group of $Z_p$, $Z_p$ is the set of (equivalent classes) of integers (mod $p$), and $RPF(i) = i \cdot g \pmod{p}$, for any $g \in \{1, \ldots, p\}$. Observe that $RPF$ works as a permutation, as $Z_p^*$ is an additive cyclic group, and every $i \in \{1, \ldots, p\}$ is a generator for $Z_p^*$ when $p$, as in our case, is prime. To avoid local small cycles of some IDs (recall that the whole permutation has a maximum cycle according to the definition of cyclic group) we can change $g$ at each application step of $RPF$.

At the iteration $t > 0^2$, the tag computes $\tilde{ID}(t) = RPF(\tilde{ID}(t) - 1)$. In words, thanks to this function, we obtain that, at each iteration, each tag is associated with a new integer in $[1, p]$ and that two different tags do not collide on the same integer. Then, the tag computes the ID to generate at the iteration $t$ as:

$$ID(t) = \begin{cases} \tilde{ID}(t) \mod d & \text{if } PRNG(t) \leq a \cdot d_R, \\ PRNG(t) \mod d & \text{otherwise}. \end{cases}$$

**Setting of the parameters**

Now we discuss about the setting of the parameters $d$ and $a$ in such a way to obtain the desired privacy requirement. We want that, at each time, the probability that the set of the people with a given ID has cardinality greater than or equal to $k$ is $1 - \epsilon$, for a given (small) $\epsilon$. The parameter $d$ represents the number of different IDs that a tag can generate. Observe that the trivial solution $d = 1$ solves the above problem because all people send the same ID. However, in this case, no meaningful information about the location of any person is provided. As a consequence, we need to find the greatest $d$ satisfying the above property.

For this purpose, we introduce the random variable $C$ defined as the total number of people who, at a given time $t$, generate the same ID as a given person, thus, causing collisions. We study the probability $P(C = c)$, which is the probability to have exactly $c$ collisions, with $0 \leq c \leq p$.

First, consider the case $a = 1$. In this case, we obtain that $P(C = c) = 1$ if $c = \frac{d}{p}$, $P(C = c) = 0$ otherwise. Indeed, when $a = 1$, $ID(t)$ is equal to the deterministic value $\tilde{ID}(t) \mod d$ and the function $RPF$ guarantees that the possible IDs are uniformly distributed among people.

Then, consider the case $a = 0$. Now, $P(C = c) = \frac{(d-1)^{c-\binom{p}{c}}}{d^p}$. Indeed, $d^p$ is the number of possible distributions of IDs and $(d - 1)^{c-\binom{p}{c}}$ is the number of cases producing $c$ collisions. Moreover, $P(C = c) = \frac{1}{p-1} \frac{1}{d^p} (d - 1)^{p-c} \binom{p}{c} = \binom{p}{c} \left(\frac{1}{d}\right)^{c-1} \left(1 - \frac{1}{d}\right)^{p-c}$.

$^2$ Once the reading frequency time $x$ (per minute) is fixed, the $t$-th iteration occurs after $t \cdot x$ minutes.
Combining the two cases above, we obtain: \( P(C = c) = a + (1 - a) \left( \frac{p}{d} \right) \left( \frac{1}{2} \right)^c(1 - \frac{1}{d})^p \), if \( c = \frac{p}{d} \), \( P(C = c) = (1 - a) \cdot \left( \frac{p}{d} \right)^c(1 - \frac{1}{d})^p \), otherwise.

The problem we have to solve is to find \( d \) and \( a \) such that, for a given (small) \( \epsilon \), the following inequality holds: \( \sum_{c=k}^{p} p^C \geq 1 - \epsilon \), because we consider all cases in which we have at least \( k \) collisions. Observe that, setting \( a = 0 \) means that the IDs generation of each tag depends only on the random permutation function that guarantees a uniform distribution of IDs and a number of ID collisions deterministic fixed (the last aspect will be detailed in the first experiment of Section 3). However, a side effect of setting \( a = 0 \) is that the sequence of IDs generated by the same tag is periodic and the period is equal to the number of people. An attacker may exploit this periodicity to know the next IDs that a person will generate and, thus, to track person movements. The parameter \( a \) is used to prevent such attacks by randomly (actually, depending on the PRNG function) changing some elements of the deterministic ID sequence. The parameter \( a \) measures the probability that an element of the deterministic ID sequence is changed. Assuming that changing one every \( x \) elements of the deterministic sequence is enough to avoid tracking attacks, then we will set \( a = x^{-1} \).

Once \( a \) has been fixed, the parameter \( d \) is obtained as the greatest integer such that \( \sum_{c=k}^{p} P(C = c) \geq 1 - \epsilon \). This number can be found by a guess-and-check iterative method, which starts from the minimum value \( d = 2 \) and at each iteration increases by 1 the value of \( d \).

3 Experiments

In this section, for space limitations, we present only some of the experiments performed to validate the proposed approach. In the whole experimental campaign, we consider an area of 50k squared meters. Moreover, we set the number of possible people to 1000. The (adjustable) read range of such tags has been set to cover an area of about 40 meters (i.e., about a room). Each reader can detect hundreds of tags in few seconds and based on the signal strength received, the reader reports or ignores the received ID to avoid multiple reading of the same tag. Tags have small size and are attached to person’s wrist or ankle.

In our experiments, to simulate person shifts inside a given area, we built a prototype implementing one of the random-based mobility models described in [3].

In the first experiment, we test the performance of our technique for different values of the parameter \( a \) defined in Section 2. Specifically, we choose four different values for \( a \) (i.e., 0.0, 0.1, 0.3 and 1) and we measure the number of colliding people \( C \) against the percentage of IDs, called \( d_c \), experimenting this collision number in average. For this purpose, we fix the number of possible IDs to \( d = p/5 \) and we simulate 100 shifts for each person. In Figure 1, we report the results of this experiment.

From the analysis of this figure, we can observe that the peak of each curve decreases if \( a \) has high values. In particular, if \( a \) is equal to 1, our approach
assigns person IDs in a pseudo-random way, according to the generator PRNG described in Section 2. For this reason, the trend of the $d_c$ is smoother, thus involving that there is a high standard deviation in the number of collisions. By contrast, if $a$ is equal to 0, the IDs generation is deterministic and, hence, we observe a very peaked trend of $d_c$ for $C = 5$.

In the second experiment, we test the performance of our approach for different values of the parameter $d$ (i.e., $d = p/5$, $d = p/7$, $d = p/8$, $d = p/9$, and $d = p/10$), which defines the number of different IDs generated by people, setting $a = 0.1$. Also in this case, we measure the number of colliding people $C$ against the percentage of IDs ($d_c$) experimenting this collision number in average.

The result of this experiment is reported in Figure 2. The discrete-Gaussian-like curves associated with the different values of $d$ show decreasing height of the peak as $d$ decreases, whereas the width of the bell (and, hence, the standard deviation) behaves the opposite. This can be explained by considering the fact that a lower value of $d$ implies a greater number of people who generate the same ID. However, due to the presence of the $a$ parameter (representing the probability of assigning an ID different from that generated deterministically) the lower value of $d$ also implies that the number of people with a different ID is greater than the number of people who have been assigned a different ID when this assignment is done deterministically. These people will decrease the number of collisions for the ID they should have assigned deterministically and will increase the average number of collisions for other IDs. This causes the reduction of the height of the curve’s peak and the increment of its standard deviation.
Fig. 2. Number of person collisions versus the percentage of IDs experimenting it for different $k$ values.

4 Conclusions

In this paper, we propose a technique to generate logs that allows us to trace people in surveilled environments and introduces a certain degree of uncertainty, in such a way that privacy is fully preserved. The main feature of the approach concerns the cheapness of service implementation both from the computational point of view and from the economic perspective. We plan to extend our analysis studying the case in which having a number of subsequent readings and correlating them according to spatial and temporal constraints would lead to a breach of anonymity.

Acknowledgment

This work has been partially supported by the TENACE PRIN Project (n. 20103P34XC) funded by the Italian Ministry of Education, University and Research and by the Program “Programma Operativo Nazionale Ricerca e Competitività” 2007-2013, Distretto Tecnologico CyberSecurity funded by the Italian Ministry of Education, University and Research.

References