Computers, Environment and Urban Systems 47 (2014) 58–67

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/compenvurbsys

Spatio-temporal patterns link your digital identities

Mikko Perttunen *, Vassilis Kostakos, Jukka Riekki, Timo Ojala

Department of Computer Science and Engineering, P.O. Box 4500, FI-90014 University of Oulu, Finland

ARTICLE INFO

Article history:
Available online 23 January 2014

Keywords:
Identity association
Mobility
WiFi
Bluetooth

ABSTRACT

An important challenge for mobility analysis is the development of techniques that can associate users' identities across multiple datasets. These can assist in developing hybrid sensing and tracking mechanisms across large urban spaces, inferring context by combining multiple datasets, but at the same time have important implications for privacy. In this paper we present a scheme to associate different identities of a person across two movement databases. Our two key contributions are the reformulation of this problem in terms of a two-class classification, and the development of efficient techniques for pruning the search space. We evaluate performance of the scheme on synthetic and real data from two co-located city-wide WiFi and Bluetooth networks, and show that the pruning has a remarkable effect on the performance of the scheme in identifying individuals across two distinct mobility datasets. Finally, we discuss the privacy implications of this scheme in the light of our findings.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

The increase of smart and mobile technologies over the last decade has resulted in the generation of large amounts of digital traces that reflect diverse aspects of our lives. For example mobile phones and vehicles can connect to various communication networks, leaving digital traces in doing so. Similarly, smart ticketing and wireless technologies also generate digital traces in the course of their use. An important challenge for mobility analysis is developing efficient techniques for comprehensively analyzing these discrete and often incomplete datasets. Our long-term research objective is to develop techniques to seamlessly combine and take advantage of such digital traces.

An important challenge is that digital traces often come in the form of distinct and possibly anonymized data. For instance, mobility traces collected via multiple networks, such as WiFi and Bluetooth (BT), are likely to use different identifiers for the same entities (devices or people) in each dataset. To overcome this challenge, and to take advantage of combined datasets, techniques to unite such datasets are required. One approach to achieve this is to develop ways to group the various identifiers used to reference the same entity or person across multiple datasets.

In this paper we present techniques that exploit spatio-temporal mobility patterns to achieve this exact type of stitching across datasets. The studied mobility traces were captured across a city using overlapping WiFi and BT scanners. As shown in Fig. 1, the movement of a device across the city (black line) results in two separate mobility traces, one for WiFi and another for BT. Hence, the formulation of the problem we study is: *Given the mobility traces captured using WiFi and BT scanners, identify the traces that were generated by the same entity (person or device).*

Our approach is based on the intuition that a WiFi identifier and a BT identifier often seen simultaneously by adjacent WiFi and BT access points are likely the same person and device. Grouping all identifiers (here: WiFi and BT) relating to the same device enables tracking entities even when only one of the radio interfaces is being used. An additional benefit is the expansion of the coverage of mobility data beyond the reach of each network, and taking advantage of both indoor and outdoor technologies.

In this paper we first present techniques for efficiently associating device identifiers across two distinct mobility trace datasets. Second we discuss the implications of our work in terms of developing smarter systems, and also in terms of privacy. In the following section we summarize related literature, and then we proceed to present our work.

2. Related work

A number of projects have previously attempted to accurately reconstruct mobility patterns by exploiting the digital traces that mobile devices generate. In the past, mobile phone tracking has been used as an approach to measure passengers dlows between parts of a city (Caceres, Wideberg, & Benitez, 2007; Gonzalez, Hidalgo, & Barabasi, 2008) and for estimating speeds and travel times (Bargera, 2007). In addition, Bluetooth and WiFi traces have been used to analyze people's mobility (Balazinska & Castro, 2003;

Corresponding author. Tel.: +358 44 3865204; fax: +358 8 553 2612.
E-mail addresses: mikko.perttunen@ee.oulu.fi (M. Perttunen), vassilis.kostakos@ee.oulu.fi (V. Kostakos), jukka.riekki@ee.oulu.fi (J. Riekki), timo.ojala@ee.oulu.fi (T. Ojala).

0198-9715/$ - see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compenvurbsys.2013.12.004

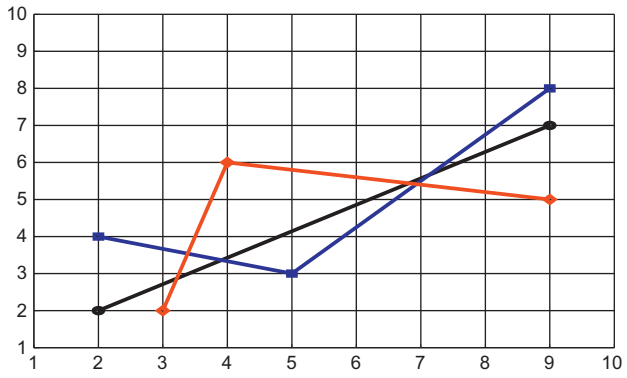

Fig. 1. Black line: the path a device follows across the city. Red line: the path as recorded by a network of WiFi scanners. Blue line: the path as recorded by a network of Bluetooth scanners. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Delafontaine, Versichele, Neutens, & de Weghe, 2012; Kostakos et al., 2009; Versichele, Neutens, Delafontaine, & de Weghe, 2012). Similarly, these technologies have also been used for traffic monitoring in the context of highways and major transport arteries, where the deployment of proximity-based scanners at strategic locations allows for the approximation of macro-travel behavior (Martchouk, Street, & Suite, 2010; Wasson, Sturdevant, & Bullock, 2008). These studies suggest that due to their popularity and widespread usage, proximity technologies are not just useful for capturing individual mobility traces, but can also be used to analyze the spatiotemporal behavior of masses.

Our own work on combining multiple mobility datasets is conceptually similar to a correlation attack, or association attack. Tan, Yan, Yeo, and Kotz (2011) define the correlation attack as follows: given the WiFi mobility traces for a number of known users, and a particular sequence of mobility traces from an unknown user, can the identity of the unknown user be matched to any of the known users? Tan et al. model the mobility of each known user by training a conditional random field (CRF) model using a sequence of features from the mobility traces as input. For any new sequence of features attributed to an unknown user, it is possible to estimate the similarity of features of the unknown user to each of the known users. This process generates a series of similarity scores, and the highest similarity score indicates the likely identity of the unknown user. It is important to note that in this case the system needs to decide between N possible answers, where N is the number of known users.

A closely related, but not identical, problem is that of mobility anomaly detection (Sun, Yu, Wu, Xiao, & Leung, 2006; Yan, Eidenbenz, & Sun, 2009): given the mobility history of a mobile user H, is a test mobility record R generated from the same user? Note that the mobility anomaly detection problem (MADP) considers the mobility of one user, whereas the correlation attack problem essentially classifies the input as being generated by one of N users.

A further attack relevant to our work has been considered theoretically. Specifically, the *constellation attack* amounts to associating all RFID identifiers of a person's belongings with the identity of this person (Garfinkel, Juels, & Pappu, 2005). In this sense, each person has a "digital shadow" of RFID identifiers consisting of potential identifiers in their clothes, shoes, jackets and coats, which moves around with them and can be used for tracking purposes.

Finally, some research has attempted to capture mobility using multiple communication networks, for instance by combining Bluetooth and WiFi (Anderson et al., 2009; Vu, Nahrstedt, Retika, & Gupta, 2010). These approaches rely on combining multiple such datasets to infer human contact patterns and the implicit social networks they reflect. However, these approaches entail end-users collecting such data while on the go, and as such rely on users' prior knowledge to combine the multiple datasets they collect. Instead, we are interested in settings where the environment itself makes these observations, not users and their own devices.

In summary, there has been substantial work on relating digital identifiers to the identities of users. However, most previous work has only considered a single tracking modality, be it WiFi, BT, or RFID. When this has not been the case, the method relied on prior end-user knowledge, and on end-user personal devices for collecting such data, as opposed to systems embedded in the environment itself. Our own work extends previous work by showing how multiple sensing modalities, or datasets, can also be leveraged to associate a person or entity with multiple identifiers.

In the following section we describe the datasets that were used to develop and evaluate our concept.

3. Data

Collecting a large set of labeled data from real people was a challenge in this study. Therefore, while we used our real-world data to evaluate our pruning scheme, we evaluate the classification system through using synthetic data, the features of which could be easily controlled.

3.1. BT dataset

For the purposes of this study, we obtained data from a total of 47 Bluetooth access points acting as "scanners" deployed across the city center. These were hardware enclosures installed on top of traffic lights or lamp posts; typically they were placed at 3–4 m from the ground level. The enclosures contained a BlueGiga 2293 Bluetooth access server with three independent transceivers. Traffic light posts are convenient places to deploy such technology, because they are a source of constant power supply, they allow for protection from vandalism since the hardware can be placed out of reach of humans, and usually provide an unobstructed environment with clear lines of sight and optimal conditions for proximity technologies to operate.

The access points were programmed to conduct continuous discovery cycles of 10.24 s, as specified in the Bluetooth protocol standard. For each scanning cycle the time and the unique ID of the discovered devices were recorded. The data was sent in real-time to a central server that aggregated all data, and also contains accurate GPS coordinates of each BT access point. Ultimately, the data was processed to provide the following records: the identity of the scanner, the identity of the device detected near the scanner, the start time and end times during which the detected device was nearby the access point.

The aggregated data followed patterns that closely resemble the expected daily fluctuations. For instance, the plot in Fig. 2 shows the hourly number of detections for the BT access points in a lobby in the campus area over a two-month period. The peak of the number of devices seen occurs at around 12.30 pm.

3.2. WiFi dataset

We obtained access to a WiFi network consisting of 1140 access points across a city center. The city center area has dense coverage, while there are access points outside downtown as well. The WiFi access points record the unique ID of the connecting devices and the start and end time of the connection. The data is sent to a

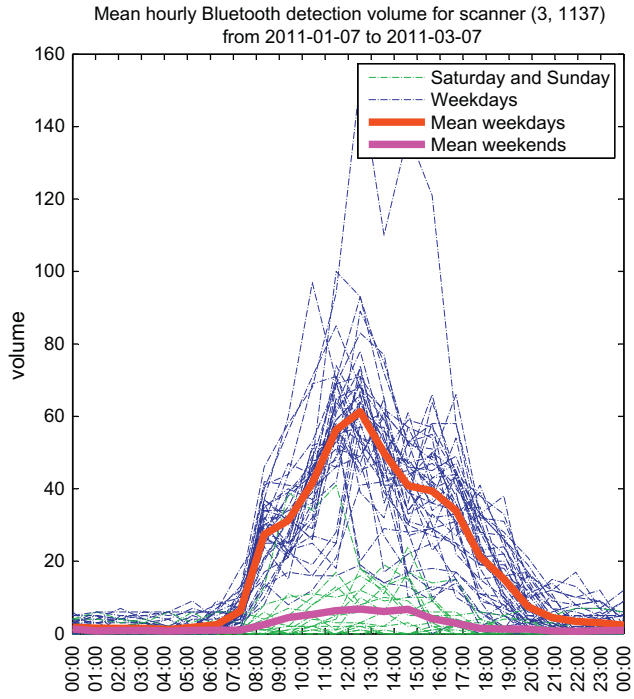

Fig. 2. Hourly volume of Bluetooth detections over a period of two months at a location.

central server where it is stored in a database. The database contains accurate GPS coordinates of the WiFi access points.

The plot in Fig. 3 shows the hourly number of detections for the WiFi access points in a lobby on a campus, over a two-month period. The peak of the number of devices seen occurs at around 12.30 pm, the same as what was seen in the BT access point data on left.

To give an approximate scale of our datasets, on average 4000 distinct WiFi devices and about the same volume of Bluetooth devices connect to the networks in one day.

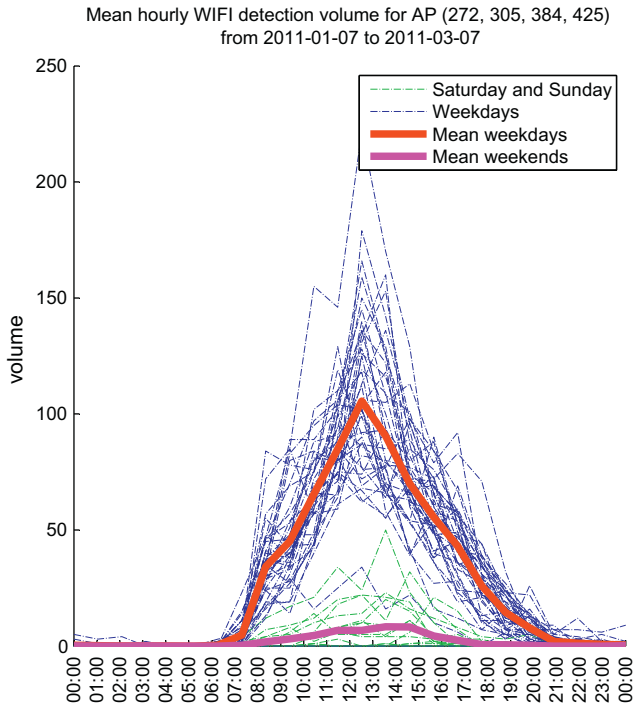

Fig. 3. Hourly volume of WiFi detections over a period of two months, for the same location as in Fig. 2.

3.3. Session durations

Figs. 4 and 5 depict histograms of session durations for a BT and a Wifi access point, respectively. Because the BT and WiFi data collection systems differ in the way they create sessions, we have filtered out sessions shorter than 0.1 s and longer than 1 h. We ran the analysis with varying length thresholds, and the mean WiFi session duration was consistently about half as long as the mean BT session duration. The reason for filtering out short sessions is that BT is far more likely to register short-lived sessions, because the hot spots operate in scanning mode, where as WiFi operates in communication mode, requiring the client device to connect for the WiFi network. Long sessions are filtered to avoid including

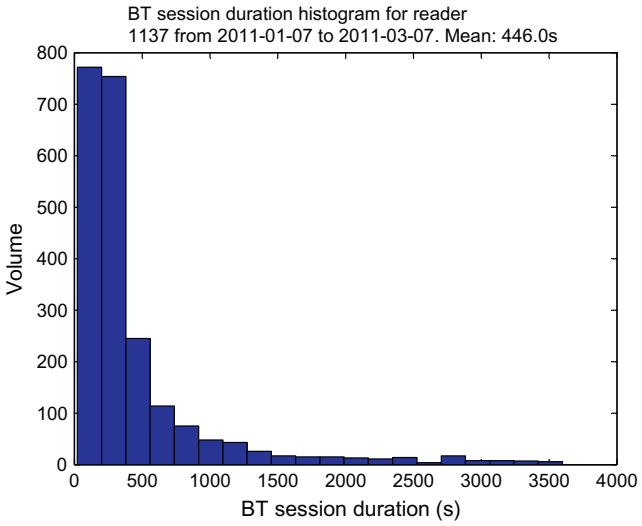

Fig. 4. BT session durations over a period of two months at a location.

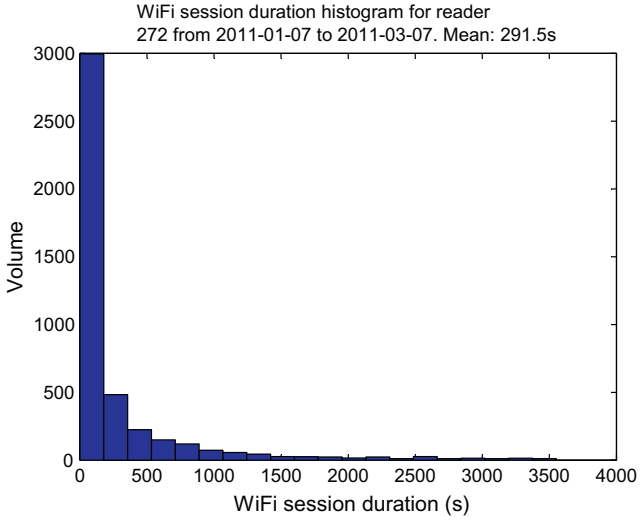

Fig. 5. WiFi session durations over a period of two months at a location.

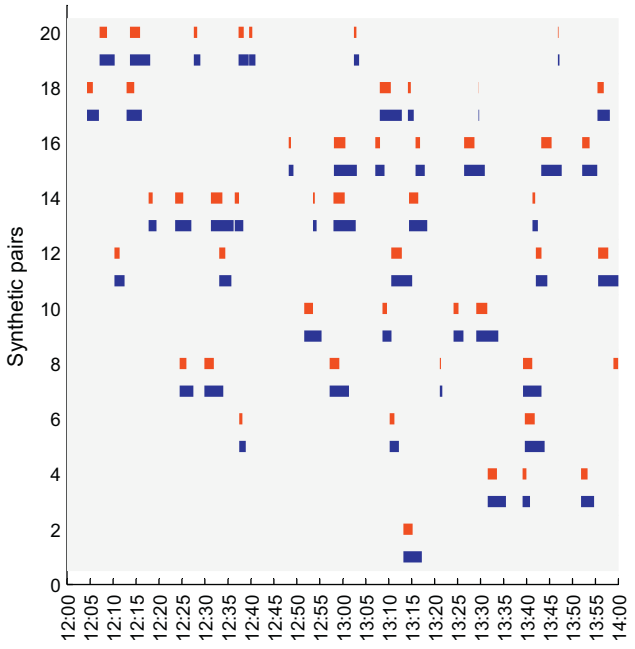

Fig. 6. Synthetic traces generated with Listing 1. Each simulated device has one WiFi (red) and one Bluetooth (blue) identifier. The *y*-axis is the ID of the device, and each device has a pair of WiFi and Bluetooth identifiers which are grouped together for visual clarity. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

devices that are always near a hotspot or are connected for a very long time for some other reason. By removing these outliers, the cleaned data better describes the data we use in this paper to associate BT and WiFi identiers, that is, devices that move past an access point pair creating a relatively short-lived session in both networks.

3.4. Synthetic data

An important challenge we faced in validating our work was obtaining substantial amounts of ground truth. In our case ground truth refers to collecting sets of WiFi and BT identities that we know belong to the same entity. For this reason we mostly rely on synthetic data to establish a comprehensive dataset for testing and developing our analysis. To synthesize a dataset with N distinct devices (i.e. people), the algorithm in Listing 1 was used.

The algorithm simulates data that would be collected at a single location where one WiFi and one BT access point would be present (as in Fig. 7), and for a particular period of time, say one day or one week. Each simulated device is given one WiFi and one BT ID. Each is then simulated as visiting the location a random number of times, each time for a random duration between *mindur* and *maxdur*, thus generating records in the WiFi and BT synthetic datasets. We believe that the relative session durations from WiFi and BT is a factor that most strongly binds together the real data and the synthetic data. In the experiments, we set the WiFi session duration to half of the BT session duration on average.

Hence, pretending that we do not know the original association between WiFi and BT IDs, our problem is to detect those in the final data records. A visual representation of the synthetic data is shown in Fig. 6. The more visits within a time frame, the denser the dataset becomes. Thus we are able to control the complexity of the dataset, and the subsequent difficulty in detecting the matching identities.

Algorithm 1. Synthetic data generation

1: $btdevs \leftarrow (btID_1 \dots btID_n)$
2: $wifidevs \leftarrow (wifiID_1 \dots wifiID_n)$
3: $truepairs \leftarrow btdevs \cdot wifidevs$
4: **for** p_i *in* $truepairs$ **do**
5: $num\_occurrences \leftarrow random\ in\ [1, maxoccur]$
6: **for** $eacho_j$ *in* $num\_occurrences$ **do**
7: $p_i.bt\_ses_{start}(j) \leftarrow random\ in\ [minT_1, maxT_1]$
8: $p_i.wifi\_ses_{start}(j) \leftarrow p_i.bt\_ses_{start}(j)$
9: $p_i\_bt.dur(j) \leftarrow random\ in\ [mindur, maxdur]$
10: $p_i\_wifi.dur(j) \leftarrow random\ in\ 0.5 \cdot [mindur, maxdur]$
11: $p_i.bt\_ses_{end}(j) \leftarrow p_i.bt\_ses_{start}(j) + p_i\_bt.dur(j)$
12: $p_i.wifi\_ses_{end}(j) \leftarrow p_i.wifi\_ses_{start}(j) + p_i\_wifi.dur(j)$
13: **end for**
14: **end for**
15: $allpairs \leftarrow btdevs \times wifidevs$
16: **for** p_i *in* $allpairs$ **do**
17: **if** $p_i \in truepairs == false$ **then**
18: $p_i.truepair \leftarrow false$
19: **end if**
20: **end for**
21: $output : allpairs$

4. Approach

We first introduce some concepts that we use throughout the rest of the paper. Specifically, we define:

- *Candidate pair* as a pair of WiFi and BT identities that may or may not belong to the same entity.
- *True pair* as a pair of WiFi and BT identities that belong to the same entity.
- *Non-pair* as a pair of WiFi and BT identities that do not belong to the same entity.

We cast the problem of identity association into a *two-class classification problem*, where a candidate pair can be classified as either class 1 (non-pair) or class 2 (true pair). Using a large set of examples from both classes, a model is trained to predict the class of a candidate pair in a test set. Our intuition is that true pairs exhibit higher similarity of spatio-temporal patterns than non-pairs. The result of the classification process can then be checked against the ground truth. In summary, the identity association process we follow is:

1. Generating candidate pairs.
2. Pruning the candidate pairs.
3. Extracting features and classifying.

4.1. Step 1. Generate candidate pairs

The first step is to identify candidate pairs by locating a suitable setting to conduct our analysis. In our case, a suitable setting is a location where a WiFi and a BT access point are within close range of each other. Such a setting facilitates narrowing down our analysis and reducing the complexity of the data we have to deal with. Also, the synthetic data is a good approximation of such a setting.

As mentioned, in the present work we use real data to evaluate the pruning concept and synthetic data to evaluate the classifier. To generate synthetic *candidate pairs* the algorithm in Listing 1 is used. For generating candidate pairs from real data, we use the identified adjacent WiFi and BT access points. Fig. 7 illustrates

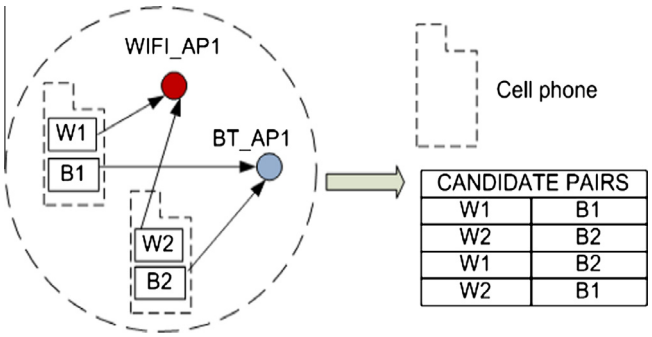

Fig. 7. Creating candidate pairs requires two adjacent access points, one WiFi (red) and one Bluetooth (blue). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

how we generate candidate pairs in such a setting. As the WiFi access point sees two WiFi IDs and the BT access point sees two BT IDs, four candidate pairs are generated. That is, for N Bluetooth IDs and M WiFi IDs, $(N \cdot M)$ *candidate pairs* are produced. As the number of devices in the environment increases linearly, the number of candidate pairs $(N \cdot M)$ grows exponentially, and hence, the search space possibly becomes very large.

So far we have considered a single geographical location, and as locations are increasingly populated with devices it becomes challenging for a classifier to distinguish between real pairs and non-pairs. Considering multiple locations can provide even more data, but at the same time much of this data is noise, for two reasons. First, the growth of the volume of candidate pairs quickly becomes immense, as the devices seen at the same time grows. Second, the features of a random device pair seen during the observation period become increasingly similar on average. As on average about 4000 distinct BT devices and WiFi devices connect to the networks, a substantial amount of candidate pairs are created, in just one day. Therefore we require a way to reduce the search space through pruning, i.e. by removing candidate pairs from the search space. Next we describe this process.

4.2. Step 2. Pruning via counter-evidence

The sheer number of candidate pairs in our analysis requires us to identify ways to reduce the amount of data that we need to consider: we need an efficient way of pruning candidate pairs from further analysis. Because our approach relies on establishing enough evidence that the spatio-temporal patterns of two particular WiFi and BT identifiers belong to the same person or device, pruning candidate pairs reduces the cost at later stages in our analysis: feature extraction and classification.

Our approach to pruning the candidate pairs is based on the idea of *counter-evidence*, as illustrated in Fig. 8. Counter-evidence is effectively proof that a candidate pair cannot be a true pair. In our case, counter-evidence is obtained by considering pairwise WiFi and BT access points which are sufficiently far apart, such that one device cannot possibly connect to both of them simultaneously. Hence, a candidate pair seen at two different places simultaneously cannot belong to the same entity. The number of combinations of locations that can be considered for counter-evidence grows exponentially with the number of BT and WiFi access points. However, this is not a major constraint because the obtained counter-evidence is valid forever, and a lookup to this data is very fast.

As an example, in Fig. 8 the BT access point *BT_AP*1 sees $B1$ and the WiFi access point *WiFi_AP*2 sees the $W3$ at the same time instant, thus the candidate pair $(W3, B1)$ is pruned because it is physically impossible for $(W3, B1)$ to be a single device, since that implies it would have to be in two different places at the same time.

By pairwise analysis of all possible WiFi and BT access points with sufficient physical distance between them, a substantial amount of counter-evidence can be accumulated. In Fig. 9 we represent as set P all counter-evidence we can collect. Then, for each adjacent pair of WiFi, BT access points, we generate the sets of candidate pairs C_1, C_2, C_3. We then use our counter-evidence from set P to prune our candidate pairs, ultimately leaving us only with subsets Cf_1, Cf_2, Cf_3. These subsets can then be fed to the classification system. Thus, the scalability of the algorithm depends on the frac-

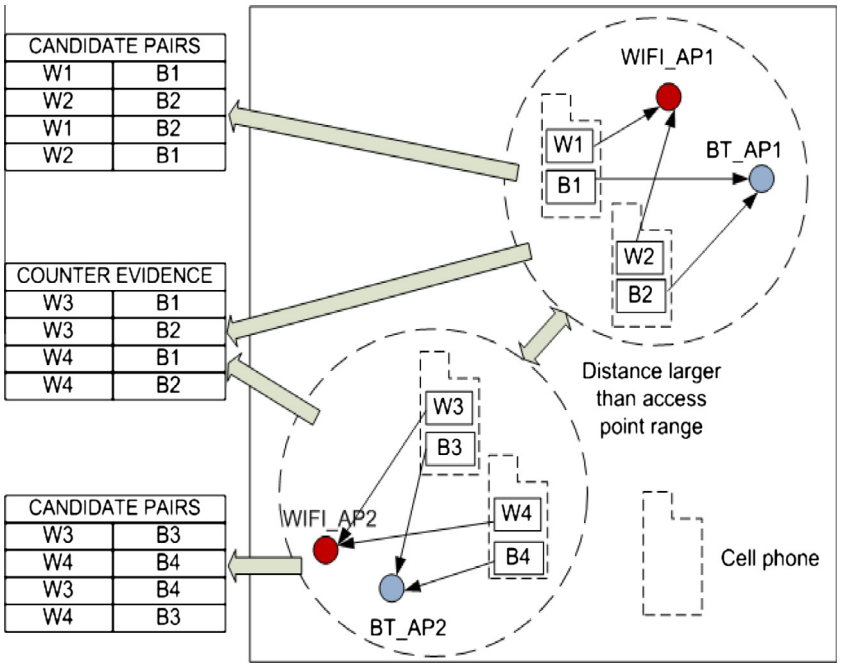

Fig. 8. Nearby access points generate candidate pairs (WIFI_AP1, BT_AP1) and (WIFI_AP2, BT_AP2). Considering pairwise access points that are far apart we can generate counter-evidence to prune the pairs (WIFI_AP1, BT_AP2) and (WIFI_AP2, BT_AP1).

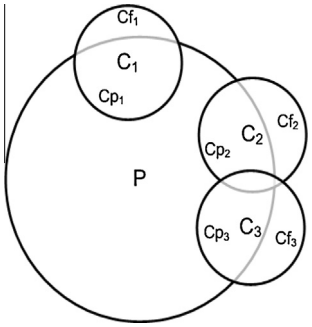

Fig. 9. P is the set of counter-evidence we have collected. Sets C_1, C_2, C_3 represent the candidate pairs we generated from adjacent pairs of WiFi, BT access points. Using counter-evidence we are able to remove the candidate pairs Cp_1, Cp_2, Cp_3. Thus, the candidate pairs we need to consider further are Cf_1, Cf_2, Cf_3.

tion of candidate pairs that can be pruned, because we need to extract similarity features and run the classifier only for the final sets of candidate pairs instead of all candidate pairs. Note also that in the figure C_2 and C_3 overlap, reflecting our assumption that several adjacent access point pairs may produce some of the same candidate pairs.

4.3. Step 3. Feature extraction and classification

At this point in our analysis we have identified a suitable location, and have fetched the data from that location to derive candidate pairs, and have pruned these candidate pairs. A key in our analysis is formulating the problem as a two-class classification problem. Thus, we need to identify features and characteristics that help us differentiate true pairs from non-pairs.

Nearest neighbor (NN) classifiers classify objects by determining the degree of similarity between the query object and the labeled objects. The query object's class is determined through majority count within the N nearest, that is, within most similar labeled objects. Thus, deploying a NN classifier does not involve a training phase, instead labeled data are used directly in the classification. Similarity of the objects is most often determined through calculating Euclidean distance between the feature vectors. It should also be noted that NN is parameter-free, simplifying experimentation. In other words, optimizing parameters for the classifier through running multiple passes of the experiments is not needed.

We chose to use a 3-NN classifier in the following experiments. We stress that in the present paper we focus on developing the concept and avoid delving into details such as finding the most efficient features and classifier for the problem. To simplify our analysis we consider just two features for a candidate pair:

- Total temporal overlap across the two datasets between the WiFi and BT. This measures the duration the two IDs were seen at the same place at the same time.
- Total temporal shift between WiFi and BT traces. This measures the aggregate difference in time between the two IDs appearing at the same place.

5. Results

5.1. Candidate pairs volume and complexity

Using the algorithm in Listing 1 we studied the effect of device crowding on the volume of candidate pairs, and subsequently the increase in complexity in classifying these. We generated synthetic data using a variety of conditions. Fig. 10 shows the total temporal overlap of each candidate pair in a dataset with 50 unique devices. Thus, this particular dataset has 50 true pairs (highlighted in red)

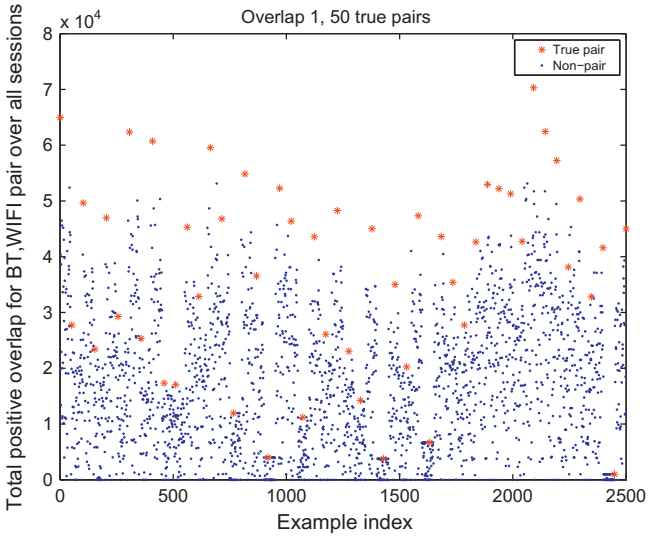

Fig. 10. Total temporal overlap (y-axis) for a synthetic dataset with 50 devices. There are consequently 50 true pairs (highlighted in red) out of about 2500 candidate pairs (x-axis). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

out of 2500 candidate pairs, created using the algorithm in Listing 1. Here the x-axis is a random order for each candidate pair, while the y-axis is the total amount of time (in seconds) that the particular candidate pair overlaps in the dataset. In other words, the higher the overlap, the more often this pair of WiFi and BT devices was "seen simultaneously" in our dataset.

Fig. 11 shows that already with 2500 candidate pairs, some of the non-pairs exhibit the same temporal overlap as some of the true pairs. This indicates that temporal overlap is not a sufficient feature to distinguish true pairs under all circumstances. This observation is confirmed when we consider the amount of "noise" in the data as the number of devices increases. Our analysis showed a non-linear increase in complexity and "noise" as more devices were added to the synthetic dataset. This is because as the number of devices increases, the true pairs and non-pairs overlap with increasing frequency in the dataset. Intuitively, adding more *non-pairs* makes it more likely that two devices overlap in time, simply because the time period becomes densely

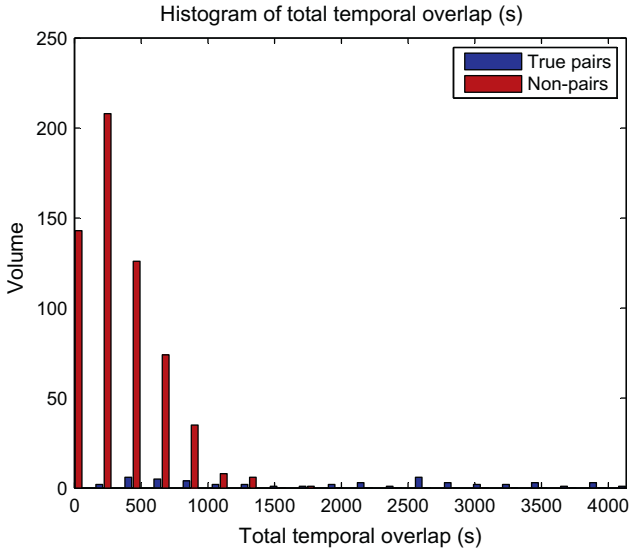

Fig. 11. Histogram of temporal overlap for true pairs and non-pairs for a synthetic dataset with 50 devices.

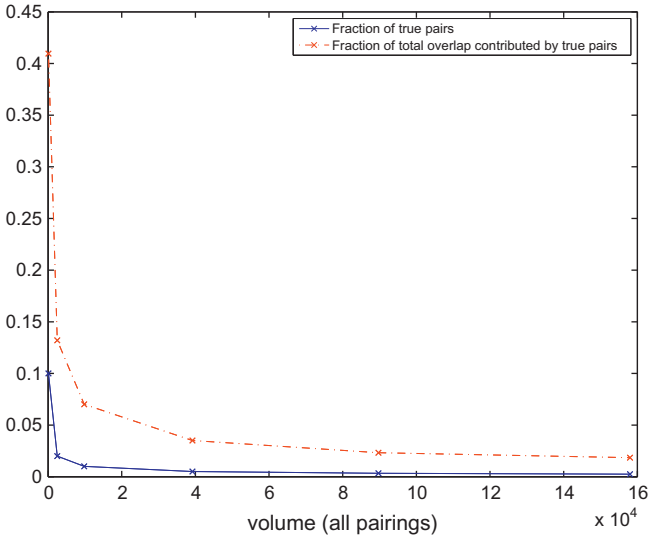

Fig. 12. The blue solid line shows the fraction of true pairs (y-axis) as the total number of candidate pairs increases (x-axis). The dashed red line shows the respective fraction of temporal overlap in the dataset that is attributed to true pairs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

populated. This trend is shown in Fig. 12, which demonstrates how increasing the number of candidate pairs (x-axis) results in a sharp decrease in the fraction of total true pairs (blue solid line) which becomes less than 1% in the long run. Subsequently, the percent of overlap accounted for by true pairs (red dashed line) also sharply decreases to less than 5% in the long run. These results show that beyond a "sweet spot" the relative frequency of true pairs, and the possible evidence that can be used to identify true pairs, is crowded out by large volumes of noise. This raises the issue of identifying a way to reduce this complexity and therefore reduce the volume of candidate pairs. One way to do this is via pruning as we describe next.

5.2. Pruning

We evaluated our approach to pruning via counter-evidence by considering data from real WiFi and BT networks. We wanted to establish whether this counter-evidence approach can be used to exclude a substantial part of the candidate pairs, or whether it results in only marginal gains. We tested our approach to collecting counter-evidence as follows. We consider two nearby access points in a particular location A, one WiFi and one BT, and for a single day of observations we derive a list of candidate pairs. These are pairs of (WiFi, BT) devices that were seen at the same place and at the same time in location A. Counter-evidence for a candidate pair in this set consists of either observing the same WiFi ID in another location concurrently, or correspondingly, observing the same BT ID in another place at the same time. Thus, we first considered the BT access point in Location A in conjunction with all other WiFi access points that were at least 200 m away from the BT in location A, and ran our analysis to search for counter-evidence. Subsequently we considered the WiFi access point of Location A in conjunction with all other BT access points at least 200 m away, thus again generating counter-evidence. We note that our access points are distributed across a dense urban environment with tall concrete buildings, and our measurements have shown that the effective range of both WiFi and BT access points is below our threshold.

Fig. 13 shows the results of our search for counter-evidence. The figure shows the portion of candidate pairs for a particular location that can be eliminated on average by considering distant BT (BT

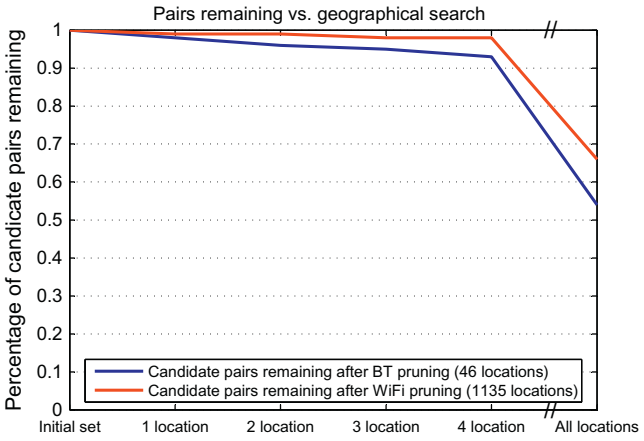

Fig. 13. The scope of searching for counter-evidence can be broadened by considering other locations to seek counter-evidence. Using all 46 Bluetooth locations with the nearby WiFi access points for creating counter-evidence results in approximately 45% reduction in the volume of candidate pairs. With WiFi, the reduction is approximately 35%.

pruning) and WiFi (WiFi pruning) access points to generate counter-evidence. Our results show that the amount of pruned candidate pairs follows a constant linear effect as a function of the number of locations, and also shows that, in general, looking for alternative BT access points to generate counter-evidence is more productive. Expressly, in our setup we can have up to 46 alternative BT access points to consider for counter-evidence, which can reduce our candidate pairs list by 45%. On the other hand, we have more than 1000 WiFi access points (depending on the geography) that can reduce the candidate pairs list by 35%.

In addition to deciding *where* to look for counter-evidence, one needs to also consider *when*. Hence, orthogonally to considering multiple locations, one can expand the temporal threshold when looking for counter-evidence. Fig. 14 shows that in fact the best time window within which to look for counter-evidence is the day the candidate pair was observed. In particular, we find that the candidate pairs can be reduced by 68% by limiting the scope to just the day the candidate pairs were observed. Broadening the scope to one extra day (e.g. the next day) reduces the candidate pairs list down a further 1%, while broadening even further to a week reduces a further 2%. Once again we find that BT pruning is more effective at generating counter-evidence than WiFi pruning.

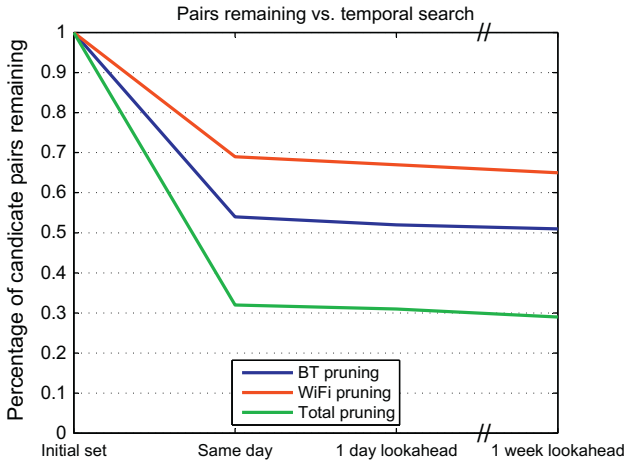

Fig. 14. The scope of searching for counter-evidence can be broadened by adapting the temporal domain of the search.

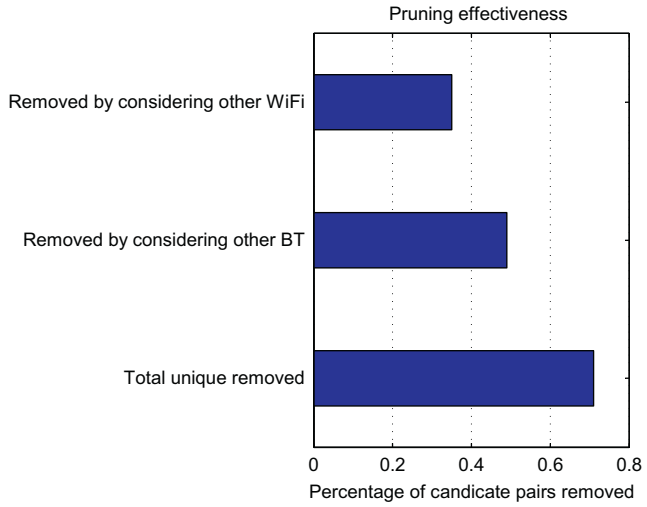

Fig. 15. Pruning effectiveness based on considering all available BT and WiFi access points for a 1-week look-ahead window. Total effectiveness is less than the sum of the two approaches because the two approaches generate overlapping counter-evidence.

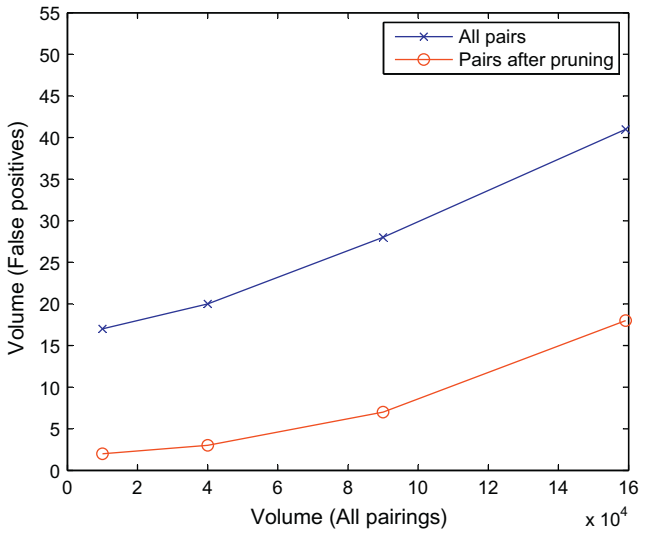

Fig. 16. False positives as a function of volume of all pairings for 3-NN classifier.

Finally, in Fig. 15 we summarize the overall gains that can be expected by using our pruning approach. We find that overall, pruning via counter-evidence can remove approximately 70% of the candidate pairs by exploiting the density and richness of the data to our advantage. We point out that searching for counter-evidence is a computationally intensive process, but it only has to be performed once for each candidate pair. After a pair has been recorded as a non-pair, simple lookup can be used instead of a counter-evidence search to prune it. The reason the search is intensive is the exponential number of combinations of locations that can be considered for counter-evidence. As the number of potential locations to consider increases linearly, the number of pairwise combinations of locations increases exponentially. Hence, it is important to prioritize the search space for counter-evidence, by considering both spatial and temporal thresholds.

5.3. Classifier performance

To evaluate the classifier, we used the synthetic data set. We applied the results of the pruning experiments in order to simulate real data as accurately as possible. That is, we analyzed classifier performance with and without pruning, removing the percentage of non-pairs that were successfully removed from real data.

In this experiment it is more meaningful to analyze the amount of false positives than the accuracy of the classifier, because there are far more non-pairs than true pairs, which means that classifying all examples as non-pairs would result in nearly 100% accuracy. Moreover, we are interested in analyzing how pruning affects classifier performance, thus the overall performance of the identity association scheme. Note that pruning directly affects the amount of non-pair examples in the test set. Non-pairs, in turn, are classified as either true negatives or false positives. We think that in the identity association problem false positives are more harmful than false negatives, that is, missing to associate a pair of identities from a set of evidence is not as bad as creating a false link between an identifier pair. A justification for this position is that in a real deployment more evidence would typically be obtained over time.

Fig. 16 shows the performance of the classifier in terms of the false positives. As expected, increasing the amount of candidate pairs results in more false positives. Pruning has an important effect on performance. By reducing up to 70% of non-pairs via the use of counter-evidence (see Fig. 15), the increase in false positives in the classifier output is more gradual than without pruning. The performance gain can be explained by the fact that removing a large percentage of non-pairs should remove some examples that are very similar to true pairs, that is, examples that the classifier might fail to classify correctly. Accuracy of the classifier is approximately 99% in all cases.

6. Discussion

In this paper we presented a scheme to associate the identities of a person in two distinct mobility traces datasets, and provide techniques for optimizing this process via the use of counter-evidence. A key contribution was the formulation of the problem as a two-class classification problem to distinguish true pairs from non-pairs. We note that the same problem can also be defined analogously to the MADP (Sun et al., 2006; Yan et al., 2009): "*given a BT association record, find the best matching WiFi association record*" or vice versa. Stated in this way, the problem becomes an N-class classification problem, where the goal is to select the correct WiFi identifier using the BT mobility traces as input. A difference is that in our own two-class formulation a single model is built to discriminate true pairs from non-pairs, that is, to model the difference of spatiotemporal relations of true pairs and non-pairs.

In contrast, in the MADP-type formulation, the device producing a mobility trace must be distinguished from N candidates, for example with a dedicated CRF model. Indicatively, the authors in Yan et al. (2009) reported accuracy rates from 50% to 95% and false alarm rates of 5% to 28%, while Sun et al. (2006) reported false alarm rates from 10% to 60% depending on conditions. However, it is hard to contrast these previous results with our own since the work is conducted under very different contexts. A more appropriate comparison is with the work reported in (Tan et al., 2011), where the authors carry out an association attack solely using WiFi traces, and report success rates from 60% to 73%. However, they also claim that their approach can successfully narrow down the identity of a person to a set of 20 possible identities for 98.51% of users. This approach of considering sets of possible solutions is possible due to the N-class classification approach, and in some cases may be preferable to our own.

6.1. Complexity and counter-evidence

One implication of our two-class formulation of the problem is that it leads to relatively large volumes of candidate pairs. Hence,

our second contribution was presenting an efficient way to prune the search space. The pruning scheme is based on the idea of counter-evidence, which reverses the problem: in addition to looking for evidence in favor of a candidate pair, we look for evidence against a candidate pair.

In terms of complexity, our analysis demonstrates an interesting phenomenon that we can use to our advantage. As the dataset become more dense and increases in complexity, more noise is added which makes classification harder. However, the sheer amount of noise added to the data results in a substantial increase in the counter-evidence that we can harvest by searching both across space and time. The results showed that the amount of pruned candidate pairs follows a constant linear effect as a function of the number of locations considered for counter-evidence, and that in general looking for alternative BT access points to generate counter-evidence is more effective. By considering all WiFi access points, the candidate pairs remaining was reduced by 35%, while by considering BT access points the reduction was 45%. Thus, adding a BT access point helps us prune more pairs than adding a WiFi access point.

One explanation for these results is that we expect users to manage their WiFi and BT transceivers differently. For example, prior work on users' practices with Bluetooth shows that when a person has a device with BT in discoverable mode, she is not likely to constantly manipulate its state, and few operations on the phone can trigger a change of BT's state (Kindberg & Jones, 2007). On the other hand, studies have shown that powering WiFi on and off is more frequent than BT, largely since many devices automatically activate the transceiver whenever connecting to the Internet, e.g. for browsing the Web or checking email (Ferreira, Dey, & Kostakos, 2011).

We also observed that candidate pairs can be reduced by 68% by limiting the scope to just the same day as when the candidate pairs were observed. Nevertheless, expanding the analysis a week ahead improved the result only by 3%. We believe this is due to an overwhelming number of people visiting a particular location once, or on a single day. This type of heavy-tailed distribution of visitor time spent at particular locations is characteristic of urban locations. In fact prior work suggests that for any given location only a small number of people account for "regulars", while most visitors at a particular location are ephemeral (Kostakos, O'Neill, Penn, Roussos, & Papadogkonas, 2010). For this reason we expect that the most effective time-window to look for counter-evidence is near the time when a candidate pair was observed.

Our pruning method makes the job of the classifier easier, that is, it not only improves efficiency, but also the accuracy of identity association. The effect of reducing the sheer number of candidate pairs is shown in Fig. 16: effectively, a reduction in the volume suggests a non-linear increase in the performance of our classifier. While we focused on pruning in order to reduce the search space of the pairing problem, looking for counter-evidence is itself computationally expensive. Note, however, that counter-evidence preserves well: recording counter-evidence for a pair once is valid forever. Considering a continuous data collection effort, it could be relatively efficient to monitor and record counter-evidence from the incoming data stream, even across large urban spaces.

6.2. Applications and privacy tensions

There are several applications of our work on matching user identities across multiple datasets. Primarily, this "stitching" approach makes it possible to develop hybrid sensing strategies. Various proximity-based technologies such as WiFi, Bluetooth and RFID/NFC can be combined based on their respective advantages. One strategy would be to rely on different technologies for indoor and outdoor sensing depending on the granularity and availability of each wireless technology. In urban spaces spanning large areas this can be a fruitful strategy, as would be the case for urban transport. For instance, WiFi networks can easily be combined with contactless ticketing systems (such as the Oyster card at transit stations or RFID onboard busses) to offer a comprehensive understanding of passengers' mobility in and out of public transit services.

Another benefit of a hybrid approach is that it substantially expands the coverage area of sensing by effectively merging multiple networks. A key requirement in this case is that the networks *must* have some physical overlap, as was the case in our own study. The existence of such overlapping regions is necessary to generate candidate pairs. We note that to generate counter-evidence we do not actually require overlapping segments. In fact, counter-evidence relies on the disjoint placement of sensors.

Of course, such hybrid sensing was already technically possible, but requires prior knowledge of the association between various IDs that belong to each person. This, in turn, depends on explicit feedback from every person being observed in the dataset, possibly in the form of a registration process. Our approach enables us to avoid such a process, and to analyze data on large numbers of people and crowds using such a hybrid strategy.

It is this exact capability of our approach, avoiding an explicit registration process, that raises a number of privacy concerns. Technically a BT or WiFi identifier relates to a device, not a person. But because we assume that devices move with people, one can infer the location of a person by knowing the location of the device. For this reason the unique device ID can be considered as personal data (van Lieshout et al., 2008). Collecting information about someone's traveling behavior, especially across multiple datasets, can be seen as breaching that person's informational privacy (information directly associated with that person), relational privacy (it may be used to relate him to other travelers who exhibit the same traveling pattern) and spatial privacy (as these travel patterns may reveal where that person has been).

While the focus of this paper is presenting an approach to link users' digital identities, we believe it is important to discuss how the related privacy concerns could be mitigated as much as possible. We argue that there are at least two instances where mitigation is important: in collecting the data, and in sharing the data. Prior work has already proposed a number of techniques that are applicable to mobility traces in general, including the adoption of digital pseudonyms (Chaum, 1985) and cryptographic techniques (Gabber, Gibbons, Kristol, Matias, & Mayer, 1999). However, in our case we propose developing techniques for thwarting or hampering the "stitching" of datasets as we have described here.

An important characteristic of our analysis is that the richness of the data due to associating multiple identities with a particular person increases hand-in-hand with privacy concerns. Since it is hard to break this relationship between data richness and privacy concerns, we envision a granular approach to managing this tension by manipulating the temporal validity of the identity associations we uncover. The obfuscating scheme we propose would ensure that the identities and pseudonyms associated with an entity across multiple datasets would be rotated every x hours. This approach would allow researchers to share their data by manipulating its richness. For instance, a dataset can be shared wherein the pseudonyms are rotated every 24 h, thus allowing for the analysis of daily patterns of behavior, but no more. The same data could instead use a 168 h (1 week) rotation, thus allowing for weekly patterns of behavior to emerge, but no more.

Another approach to protect against unwanted identity association across datasets is the introduction of fake counter-evidence strategically. We believe that fake counter-evidence can be planted across datasets with minimal qualitative side-effects. This is

possible because counter-evidence does not follow mobility patterns: counter evidence is collected across distant locations, and does not have any particular 'mobility' characteristics. Hence, if one wanted to hide a particular association between two particular device IDs, one would only have to plant a handful of records strategically across multiple datasets, such that they give rise to counter evidence. This would render the use of counter-evidence inappropriate for further analysis, and hence make the process of identity association much more challenging. However, mitigation of the privacy concerns offers a possible direction for dedicated research in the future.

6.3. Limitations of the study

A challenge for this study was collecting substantial amounts of true pairs from real people. This is potentially a breach of privacy, and for this reason we chose not to work directly with such data. Therefore, we used our real-world data only to generate counter-evidence, i.e. proof that some candidate pairs are definitely not true pairs. For the aspects of our work where ground truth knowledge of true pairs was required, such as in classification, we relied on synthetic data. While synthetic data can be well-controlled and manipulated, it is not clear how well it reflects the real world. Hence we acknowledge that synthetic traces may not fully capture the characteristics of actual data. For example, WiFi and Bluetooth have different spatial resolutions which is quite likely to affect the overlap times of the sessions at an access point pair. Nevertheless, we assume this phenomenom is handled by the feature extraction phase which should produce distinctive temporal features regardless of the likely biases in BT and WiFi session durations. With synthetic data, our objective was to clearly describe the circumstances and parameter settings under which our experiments were run and the accuracy at which true pairs can be identified from the synthetic data. Based on these results, we hypothesize that we can find discriminative features for real data once we fully understand its true characteristics.

It should be noted that applying an unsupervised method to categorize the real pairs and non-pairs might reduce the need for labeled data. However, labeled data would likely be required for validating the results of the unsupervised methods. Given our contribution towards developing the matching algorithms, future work can focus on making it easier to bootstrap this process by collecting such training data. Perhaps volunteers could be recruited using a game, or some other incentive mechanism could be used to initially collect such data.

In the present work we have considered linking identities across two networks. However, the scheme is also valid for multiple networks; In the feature extraction phase, several networks could be handled pairwise, thus generating multiples of the same descriptive features we used in 2-network case. The combination of these features constitute a feature vector which is used to train the classifier. In the pruning concept, the counter-evidence could also be handled pairwise between each two networks. For example, for three networks, this would mean two lookups to the counter-evidence for each candidate pair, instead of one lookup. Note that the classification would still be a two-class problem, namely classifying the real pairs and non-pairs.

6.4. Conclusion and ongoing work

We have described a scheme to associate multiple identities of an entity across two mobility datasets. Our main contribution is the formulation of this problem as a two-class classification, and the exploitation of spatio-temporal patterns and counter-evidence to optimize our analysis. We conclude that according to the obtained results our scheme is capable of associating identities from a large set of candidates. Pruning with counter-evidence improves accuracy and efficiency of the identity association remarkably. The paper opens several questions requiring further study. To mention a few: how to build an online identity association system on a city scale? What other networks and device identifiers could be associated to the set of identifiers? Which privacy guarding methods perform best in preventing identity association with our scheme?

In ongoing work we are interested in representing the movement of a device between access points as a trajectory in a 2D spatial domain and in considering identity association in the context of geometrical trajectory similarity assessment.

Acknowledgments

This work was supported by the Finnish Funding Agency for Technology and Innovation (TEKES) as part of the Urban Flows and Networks project and the Data to Intelligence program of DIGILE (Finnish Strategic Centre for Science, Technology and Innovation in the field of ICT and digital business).

References

Anderson, E., Phillips, C., Gonzales, H., Bauer, K., Sicker, D., & Grunwald, D. (2009). Sniffmob: Inferring human contact patterns using wireless devices. In *Proceedings of the 1st ACM international workshop on hot topics of planet-scale mobility measurements HotPlanet '09* (pp. 4:1–4:6). New York, NY, USA: ACM.

Balazinska, M., & Castro, P. (2003). Characterizing mobility and network usage in a corporate wireless local-area network. In *Proceedings of the 1st international conference on mobile systems applications and services MobiSys 03* (pp. 303–316). ACM Press.

Bargera, H. (2007). Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from israel. *Transportation Research Part C: Emerging Technologies, 15*, 380–391.

Caceres, N., Wideberg, J. P., & Benitez, F. G. (2007). Review of traffic data estimations extracted from cellular networks. *Engineering and Technology, 1*, 15–26.

Chaum, D. L. (1985). Security without identification: Transaction systems to make big brother obsolete. *Communication of the ACM*, 1030–1044.

Delafontaine, M., Versichele, M., Neutens, T., & de Weghe, N. V. (2012). Analysing spatiotemporal sequences in Bluetooth tracking data. *Applied Geography, 34*, 659–668.

Ferreira, D., Dey, A. K., & Kostakos, V. (2011). Understanding human-smartphone concerns: A study of battery life. In *Pervasive* (pp. 19–33).

Gabber, E., Gibbons, P., Kristol, D., Matias, Y., & Mayer, A. (1999). On secure and pseudonymous client-relationships with multiple servers. *ACM Transactions on Information and System Security, 2*, 390–415.

Garfinkel, S. L., Juels, A., & Pappu, R. (2005). Rfid privacy: An overview of problems and proposed solutions. *IEEE Security and Privacy Magazine, 3*, 34–43.

Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *Nature, 453*, 779–782.

Kindberg, T., & Jones, T. (2007). "merolyn the phone": A study of Bluetooth naming practices (nominated for the best paper award). In *Ubicomp* (pp. 318–335).

Kostakos, V., Nicolai, T., Yoneki, E., O'Neill, E., Kenn, H., & Crowcroft, J. (2009). Understanding and measuring the urban pervasive infrastructure. *Personal Ubiquitous Computing, 13*, 355–364.

Kostakos, V., O'Neill, E., Penn, A., Roussos, G., & Papadogkonas, D. (2010). Brief encounters: Sensing, modeling and visualizing urban mobility and copresence networks. *ACM Transactions on Computer–Human Interaction*, 17.

van Lieshout, M., & Kool, L. (2008). Privacy implication of rfid: An assessment of threats and opportunities. In S. Fischer-Hubner, P. Duqenoy, A. Zuccato, & L. Martucci (Eds.), *The future of identity in the information society* (pp. 129–141).

Martchouk, M., Street, N. C., & Suite, N. E. (2010). Analysis of freeway travel time variability using Bluetooth detection. *Journal of Transportation Engineering, 2051*, 1–30.

Sun, B., Yu, F., Wu, K., Xiao, Y., & Leung, V. C. M. (2006). Enhancing security using mobility-based anomaly detection in cellular mobile networks. *IEEE Transactions on Vehicular Technology, 55*, 1385–1396.

Tan, K., Yan, G., Yeo, J., & Kotz, D. (2011). Privacy analysis of user association logs in a large-scale wireless lan. *Information Sciences, 836*, 31–35.

Versichele, M., Neutens, T., Delafontaine, M., & de Weghe, N. V. (2012). The use of Bluetooth for analysing spatiotemporal dynamics of human movement at mass events: A case study of the Ghent Festivities. *Applied Geography, 32*, 208–220.

Vu, L., Nahrstedt, K., Retika, S., & Gupta, I. (2010). Joint Bluetooth/WiFi scanning framework for characterizing and leveraging people movement in university campus. In *MSWIM '10* (pp. 257–265). New York, NY, USA: ACM.

Wasson, J. S., Sturdevant, J. R., & Bullock, D. M. (2008). Real-time travel time estimates using media access control address matching. *ITE Journal, 78*, 20–23.

Yan, G., Eidenbenz, S., & Sun, B. (2009). Mobi-watchdog: You can steal, but you can't run! *Conference On Wireless Network Security*, 139–150.