# Anonymity properties of stored or transmitted data taken from Bluetooth scans

David Evans
Computer Laboratory
University of Cambridge
15 JJ Thomson Avenue
Cambridge, UK CB3 0FD
Email: david.evans@cl.cam.ac.uk

Robert H. Warren
Department of Informatics
University of Zürich
Binzmühlestrasse 14
Zürich, Switzerland
Email: warren@ifi.uzh.ch

***Abstract*—Modern consumer wireless devices are increasingly powerful, making them attractive to use as wireless sensor nodes. At the same time, many use protocol suites such as Bluetooth which require devices to reveal data that may make for unique device identifiers. This paper explores this quantitatively through scans covering several thousand devices in different urban locations. In measuring the anonymity afforded by the elements of a Bluetooth device profile, we find that (i) attributes such as the device class are poor for linking sightings of the same device; (ii) the device name can provide a surprising amount of anonymity but when it does not it can be a very effective key to link devices with individuals; and (iii) frequently users exhibit privacy-adverse behavior, such as placing telephone numbers in device names or using nicknames that are statistically rare.**

## I. Introduction

Many wireless devices broadcast information about themselves and their owners. As these devices become pervasive it becomes increasingly effective to make use of this information for surveys, pricing, scheduling, security, or marketing. Many devices such as mobile phones, computers, headsets and headphones, cars, cameras, printers, keyboards, and mice use the Bluetooth set of protocols for short-range wireless communication. These protocols are used for transfer of files and address books, transmission of audio, and so on. A Bluetooth device will reveal information about itself as part of its normal operation, including its link-layer identity, the type of device it is, the services it offers, and a name that may be visible to and adjustable by the user.

Previous work has examined how the data transmitted by devices may be used to construct location services [1], inform opportunistic routing [2], infer social networks over large areas [3], and engineer transient networks [4]. Some projects have attempted large scale surveys of Bluetooth devices [5] with interesting implications for the use of Bluetooth as a wide area tracking tool.

Our focus here is different and is motivated by our work on transport monitoring. We use what we call *sensor nodes* to generate data describing physical phenomena and *gateway nodes* to relay these data to the middleware and ultimately to applications [6, §1]. Suppose that the sensor nodes use Bluetooth to communicate with the gateway nodes. The sensor nodes are not anonymous as far as the gateway nodes are concerned—the data sent between the two includes unique identifiers such as the link-layer address that can be used to link sightings of the same sensor node device. However, because all communication between the middleware and the sensor nodes is via the gateway nodes, applications can only know about the sensor nodes what the gateway nodes forward. The gateway nodes' choice in this determines the large-scale privacy properties of the system. Here we explore this by scanning Bluetooth devices to determine what can be inferred about the sensor nodes based on information that these nodes provide and gateway nodes could make available.

This paper is organised as follows. Section II defines what we mean by anonymity and discusses how it can be measured. Section III describes Bluetooth scanning and the data that devices make available. Section IV defines a notation for talking about the anonymity properties of device attributes while section V presents our scanning methodology and basic properties of our two data sets. Section VI describes results that have to do with anonymity and discusses what can be learned from them and section VII concludes and proposes future work.

## II. Measuring anonymity

Suppose that each of a set of actions $A$ is performed by some individual in the set $I$. The goal of an observer is to find for each action in $A$ the corresponding individual in $I$. The degree to which privacy is preserved—what we will call the degree of anonymity—relates to the difficulty of this task.

A great deal of work has been done developing quantitative measures of this difficulty, particularly by the designers of systems providing anonymous communication. (In that case, the activities are the sending or receiving of messages and the individuals are senders or receivers.) The most simple expression of anonymity is the anonymity set, "the set of all possible subjects who might cause an action" [7]. For a given action in $A$, the anonymity set is the set of individuals in $I$ that may have performed that action and the corresponding measure is the size of this set. Different actions may have different corresponding anonymity sets.

While the anonymity set captures the uncertainty inherent in observing actions in $A$, it says nothing about other knowledge

available to the observer. In particular, the observer may be able to assign different probabilities of having performed a particular action to different individuals. This idea is formalised in information-theoretic measures of anonymity, such as those of Díaz et al. [8] and Serjantov and Danezis [9]. In general we make no assumptions about the information available to the recipient of the data sent by the gateway nodes. For this reason we use the size of the anonymity set as our metric.

## III. Bluetooth scanning

A device equipped with a Bluetooth radio may initiate a scan for other nearby Bluetooth devices. Each device can configure itself to be either visible or invisible (what Bluetooth calls being "discoverable" and "undiscoverable") to such scans. (While there are techniques to scan for undiscoverable devices, for example that of Cross et al. [10], here we consider only devices that are discoverable.) The device doing the scanning will collect information from discoverable devices in its vicinity; the amount of information reported by a device in response is not normally adjustable by the user. Each time a device is seen by a scan it is called a *sighting*.

The device characteristics that can be identified include:

1) The **name** of the device. Sometimes this can be set by the user.
2) The medium access control (MAC) **address** of the device. This uniquely identifies the Bluetooth hardware and is almost always permanently assigned by the manufacturer and cannot be changed by the user. Since the MAC address identifies the Bluetooth hardware only, it may correspond to a Bluetooth interface "dongle" and not to the actual computing device that is hosting it. As address ranges are assigned to specific manufacturers, it is possible to identify a possible set of device models from a hardware address.
3) The **device major class** identifies the type of device. The valid devices classes are miscellaneous, computer, phone, network, audio/video, input device, imaging system, wearable (watch), toy (action figure) and unknown.
4) The **device minor class** further breaks down the class of device. These are defined in the context of a major device class and they are numerous, so we do not reproduce the values here. The minor class may indicate the specific type of device (a mobile telephone as opposed to a cordless telephone, say) or its current utilisation.
5) Each device offers a number of **services** to the outside world, such as presenting its serial ports, offering to route network packets, and so on. Each service is essentially a handle for a specific API.

We call each of these types of information about a device an *attribute* and summarise them in Table I. Note that the attribute `class` has a value that is a concatenation of 8 bits representing the device's major class and 8 bits representing the device's minor class.

Normally, an individual's degree of openness can be inferred from the level of security imposed on his or her device. However, the Bluetooth protocol separates privacy configuration of device visibility (discoverable versus undiscoverable) from the authorisation and access control needed to use the services offered by the device (whether the device requires a security PIN). This muddies the idea of informed consent to having one's information processed by others. It is not unreasonable to suppose that a user is not consenting to reveal information from an undiscoverable device, given the resources that must be expended to find the device. But if the device is discoverable, whether or not it requires a PIN to use its services, it will make the attributes described above available as part of its normal operation. Recipients do not need to use any clandestine behaviour or exploit device shortcomings. Also, the availability does not imply misconfiguration of the device akin to an open wireless access point. But because the information is part of the initial device identification process and is available before any authentication or access control takes place, what can we conclude about consent if the user has assigned a PIN? There is no other means to link devices and it is common for users to leave their devices discoverable for practical reasons, so exposure of information is highly likely in spite of the requirement of a PIN. Perhaps this is a lesson in that the design of future devices should require no permanent identifiers and use discovery and device linkage methods that are privacy aware.

TABLE I
THE DEVICE ATTRIBUTES THAT WE EXAMINE.

| Number | Name | Description |
|---|---|---|
| 1 | `class` | The class of the device |
| 2 | `address` | Device MAC address |
| 3 | `name` | Device name |
| 4 | `services` | The list of services offered by the device |

## IV. Notation: combinations of attributes

What follows defines a notation that we will use for subsequent discussion and to label graphs.

Let $D$ be the set of all devices. Suppose that $d \in D$ is a particular device and that its first sighting is at time $\tau(d)$. We define $D_t$ to be the set of devices whose first sighting is at or before $t$, *i.e.*, $D_t = \{d | \tau(d) \leq t\}$. This means that $|D_t|$ is the number of unique devices seen up to and including time $t$.

Let $A$ be the set of all device attributes and let $x$ be a particular selection of attributes. Clearly $x \subseteq 2^A$. We write $x$ as a number using the bits from column 1 of Table I that correspond to the attributes of interest; $\{\texttt{class}, \texttt{name}\}$ is thus combination $1010_2$ or 10. Let $x(d)$ be the values of attributes $x$ possessed by device $d$.

Suppose that we select a set of attributes $x$ and assign the identifier $x(d)$ to each device $d$. We refer to this as *considering* the attributes in $x$. Define $f_D(x, v)$ to be the size of the anonymity set when attributes $x$ are considered and they have values $v$, given that devices in $D$ have been seen. It is easy to be convinced that

$$1 \leq f_D(x, v) \leq |D|$$

for all $D$, $x$, and $v$. At one extreme are quantities like $f_D(\{\texttt{address}, v\})$, which we expect to be 1 for all $D$ and $v$

because each device's MAC address is unique. On the other hand, if every device in $D$ has the same value for all the attributes in $x$ then $f_D(x, v) = |D|$ because each device will contribute 1 element to the anonymity set. For convenience, let $f_D(x) = \max_v f_D(x, v)$. $f_D(x)$ is the maximum anonymity set size that a device could expect given that attributes $x$ are considered, assuming that it had the most commonplace values for those attributes (a particular device might not, of course—its values might make it unique).

For any set of attributes $x$ and their values $v$, we expect that

$$f_{D_{t_1}}(x, v) \leq f_{D_{t_2}}(x, v)$$

for any $t_1 \leq t_2$. In other words, as you accumulate sightings of more devices, the size of the anonymity sets of the attributes contained in those sightings can only go up. For example, when you have seen just one device, it is unique. When you see a second, if it is different from the first then it too is unique and the anonymity set has not increased in size. If however it is the same, then the anonymity set size is now 2.

## V. Scanning methodology and data set properties

We made use of the btsscan utility created by Tim Hurman [11] at Pentest Ltd. to make a continuous scan for Bluetooth devices. In our case, the data about a particular device are fixed at those found in the device's first sighting.

We have used two sets of scanning results, collected by us in two different cities. The first was collected over about three months within the computer science department of the University of Waterloo by placing a Bluetooth dongle on the back of an office door. The door faced one of the main thoroughfares between two large buildings on the university campus; large windows on the opposite side of the office had a view over a large car park. No effective range measurements were taken. The experiment spans just over 88 days from early January 2007 and consists of 429,925 sightings of 486 devices. Some initial conclusions were presented in poster format [12], including an anecdotal case where the weekly schedule of a phone owner was inferred from device observations.

The second dataset was collected from an apartment close to the ground, overlooking a large car park and a major thoroughfare in a large city in Canada. The survey was run for a period of just under 4 months; the dongle was placed on an exterior window and scanning for new devices was continuous. No effective range measurements were taken during the experimental period, but anecdotal evidence suggests the range was over 40 meters. The data cover just over 101 days from late November 2008 and consists of 136,819 sightings of 2,051 devices.

### A. Rate of device discovery

Fig. 1 shows plots of $|D_t|$ versus $t$ for data sets 1 and 2. (Remember that this is a count of *new* devices seen, not device sightings; it is possible that many devices are scanned but that none of them are new.) We can see that new devices continue to be found throughout the scans. The University of Waterloo held its winter "reading week" from February 17th

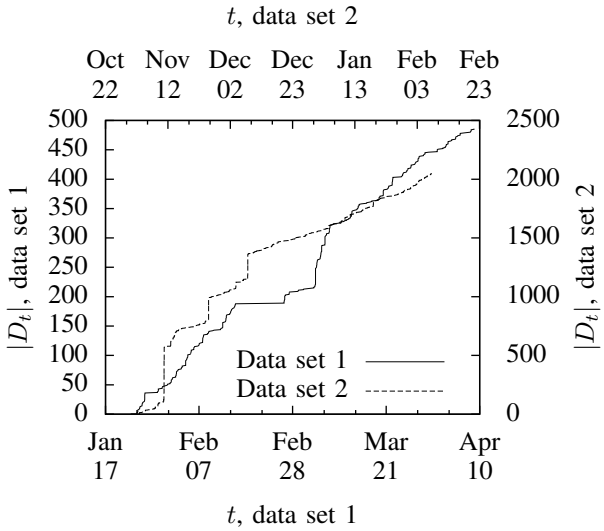



Fig. 1. $|D_t|$ versus $t$

to February 23rd; we suspect that this led to a change in the set of people walking past the scanner and a corresponding rarity of observation of new devices. For the second data set, a dramatic rise in the number of new devices seen occurred on November 11th, Remembrance Day. We are unsure of the reason for this as there is no parade or memorial service close to the survey site.

### B. Device popularity

Table II shows, for each data set, the top five device classes. For each of these the table shows the name of the class as assigned by the scanning software, the major and minor classes (note that different major classes may have the same name), the number of devices seen that belong to this class, and the percentage of devices that this makes up.

## VI. Results and discussion

### A. Potential anonymity

We now answer the question of how anonymous it is possible to be, given that a given combination of attributes about a device is considered. Fig. 2 shows the maximum anonymity set size for the indicated attribute combinations versus the number of devices seen. The attribute combinations that include the device address (namely 4, 5, 6, 7, 12, 13, 14, and 15) have, as expected, an anonymity set of size 1 no matter how many devices are seen and so are not shown.

The anonymity available is non-trivial. Even considering the device name, potentially anonymity set sizes of several dozen are possible. Interestingly, there are times when the anonymity available does not increase even though new devices are found; these correspond to the horizontal parts of the plots. At such times, the new devices that appear do not fit into one of the existing anonymity sets—they are unique. One can think of them as adding new "islands" of crowds of devices. Subsequent devices may form a part of these crowds, initially small as they may be.

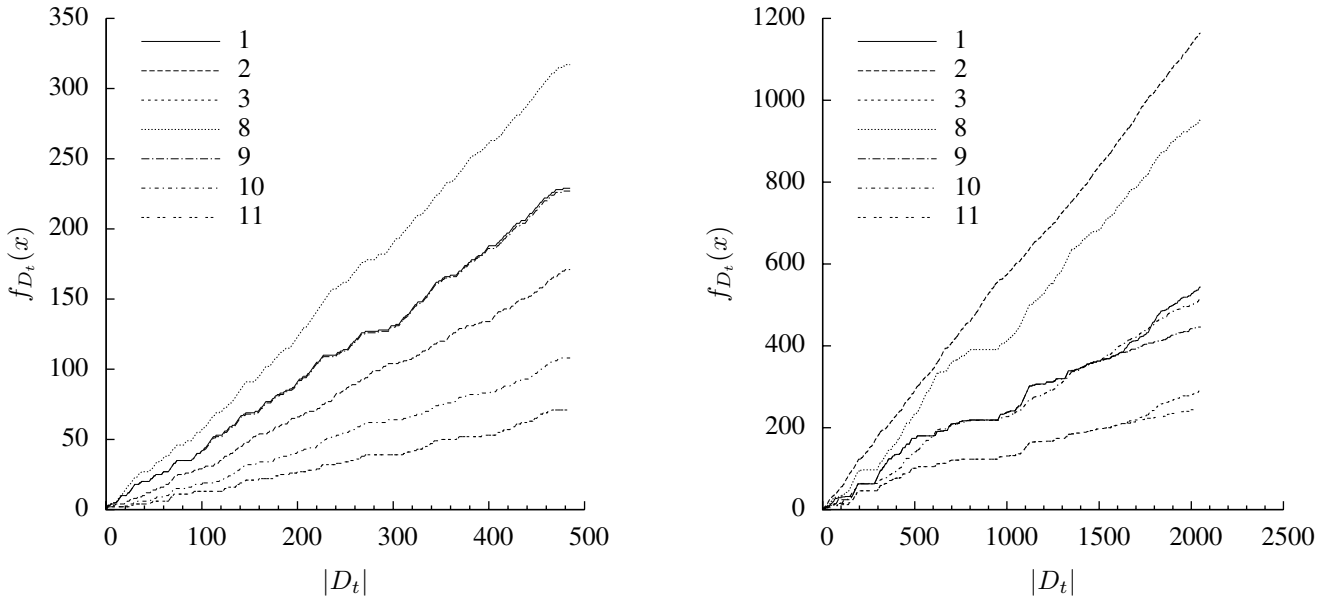
TABLE II
THE FIVE MOST SEEN CLASSES OF DEVICES

| Name | Major class | Minor class | Number | % devices |
|---|---|---|---|---|
| Phone/Mobile | 2 | 4 | 317 | 65.23 |
| Phone/Smart phone | 2 | 12 | 98 | 20.16 |
| Computer/Laptop | 33 | 12 | 16 | 3.29 |
| Computer/Laptop | 1 | 12 | 15 | 3.09 |
| Computer/Palm sized PC-PDA | 1 | 20 | 8 | 1.65 |

(a) Data set 1

| Name | Major class | Minor class | Number | % devices |
|---|---|---|---|---|
| Phone/Mobile | 2 | 4 | 953 | 46.47 |
| Audio-Video/Hands free | 4 | 8 | 390 | 19.02 |
| Phone/Smart phone | 2 | 12 | 371 | 18.09 |
| Phone/Mobile | 34 | 4 | 122 | 5.95 |
| Uncategorised | 0 | 1 | 56 | 2.73 |

(b) Data set 2




(a) Data set 1

(b) Data set 2

Fig. 2. $f$ versus $|D_t|$

### *B. The anonymity one can expect*

Fig. 2 tells us that the anonymity afforded a given device depends on the attributes of that device. For example, suppose that a manufacturer sets each of its new devices' names to "New." One of these devices having the name "New" will be indistinguishable from the others that have not had their names changed. On the other hand, a device having a very unusual name stands a greater chance of being easily distinguished.

Fig. 3 shows a histogram of $f(x, v)$ for $x = \{\texttt{service}\}$ as $v$ varies over all values in each data set. (Obviously similar histograms could be plotted for other attribute combinations.) The degree of anonymity that a device can expect is quite varied. Within data set 1, while there is an anonymity set of size 229 and any device in this set will be indistinguishable from any other, there are 6 that are each of size 1. These 6 devices are afforded no anonymity at all: the list of services that each offer is a unique identifier.

This leads to the question, what combinations of attributes identify devices? For example, as we have discussed, a device's MAC address is unique, so we would expect any combination containing the address to have $n$ anonymity sets of size 1 where $n$ is the number of devices. Table III shows, for each attribute combination, the number of sets of size 1 and this as a percentage of all devices, indicating the fraction of devices that are identifiable. Combination 4 is the device address and can be used to identify 100% of the devices; other combinations that include the address have the same property and are not shown.

The device name can be used to identify about 41% and 22% of the devices in the two data sets, respectively. We

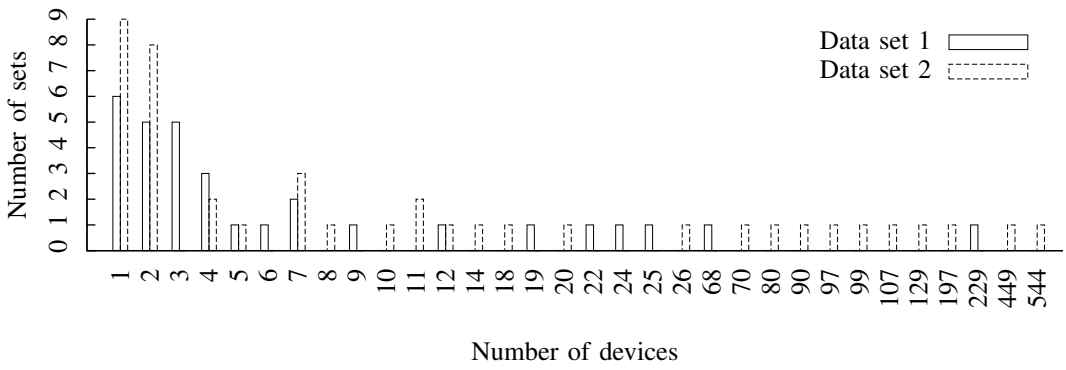



Fig. 3. Histogram of $f(x, v)$; $x = \{\texttt{service}\}$

TABLE III
NUMBER OF ANONYMITY SETS OF SIZE 1 FOR EACH ATTRIBUTE COMBINATION

| Combination | Number of sets of size 1 | | % devices that are unique | |
|---|---|---|---|---|
| | Data set 1 | Data set 2 | Data set 1 | Data set 2 |
| 1 | 6 | 9 | 1.23 | 0.44 |
| 2 | 201 | 454 | 41.36 | 22.14 |
| 3 | 218 | 483 | 44.86 | 23.55 |
| 4 | 486 | 2051 | 100.00 | 100.00 |
| 8 | 5 | 6 | 1.03 | 0.29 |
| 9 | 10 | 22 | 2.06 | 1.07 |
| 10 | 207 | 471 | 42.59 | 22.96 |
| 11 | 220 | 491 | 45.27 | 23.94 |

suspect that these numbers are as small as they are because a minority of users personalise their devices' names. Furthermore, referring to table II we see that devices in the "audio-video/hands free" class are far more prevalent in data set 2 than they are in set 1. It is not unreasonable to suppose that changing the name of such a device may be trickier than changing that of a phone—in fact, a user interface for doing this might not be provided—and there are 256 "audio-video/hands free" devices in data set 2 having an empty name. Furthermore, the user population for data set 1 is biased towards undergraduate students who may have the urge to link gadgets with themselves which leads to more unique device names.

The attributes that reveal little are the list of services offered and the device class; even considered together, as in combination 9, few devices can be uniquely identified.

Based on these results, a gateway node transmitting a device's class and/or the list of services that it offers compromises anonymity little. The device's name is less good and our data bear out the intuitive notion that the device's address contains no privacy whatsoever.

### C. Linking with individuals

So far we have made no assumptions about additional information that recipients of the data from gateway nodes might have. Now we shift focus somewhat and examine the efficacy of combining these data with others to link sensor nodes with individuals.

Our population of individuals is the online directory of members of the University of Waterloo.[1] Device users who are a part of the university will have their personal details in the online directory. We searched the directory using keyword queries formed from string tokens in devices' `name` attributes, excising device names that were blank, obviously product models, or brand names; the total number of queries was therefore less than 486. The results returned were manually

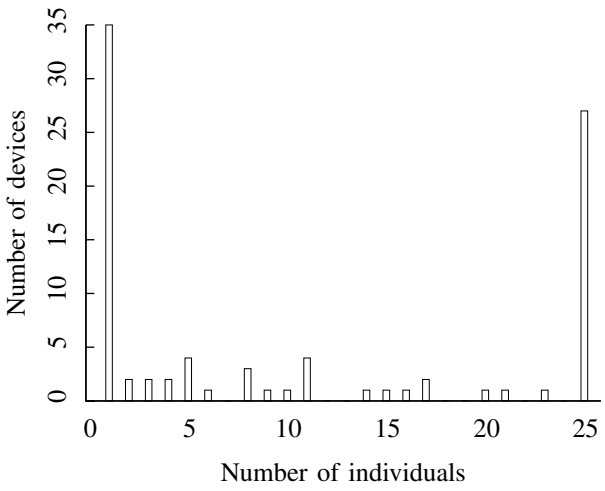



Fig. 4. Histogram of the number of individuals a device can be attributed to.

[1]This means that data set 1 was used but the techniques described here are general; data set 2 could be linked with the telephone directory, for example.

reviewed for *prima facie* appropriateness. A histogram showing how many devices could be linked to a given number of individuals is shown in Fig. 4. (As the directory has a limit of 25 results per query, queries that matched more than 25 hits are lumped in with those that returned 25.)

Using this crude technique we were able to link each of 35 devices with a single individual. This suggests that while device names may have the potential for high anonymity, as discussed in VI-B, those that are unique can be very effective at identifying users. We observed the use of fully qualified phone numbers and personal names as the `name` attribute—clearly privacy-averse behaviour! We also discovered that while people may make use of nicknames, presumably for privacy or vanity reasons, they tend to make repeated use of the same nickname in different contexts. For example, where someone may have labelled a device "greatguy43", there may be an entry in the university directory with the nickname "greatguy43". In some cases, this information enabled us to link devices to individuals where no "real" name was included in the device's `name` attribute. The irony of this behaviour is that most of the nicknames used are "rare" and stand out against dictionary words. Not only does this make the nicknames easier to isolate within a dataset, it enhances linkability with records in other datasets because the terms are invariably rare in all of them and does not suffer from a large statistical normalisation penalty [13], [14], [15]. Inventing nicknames for privacy reasons is therefore counter productive.

## VII. Conclusions and future work

As the complexity of wireless and portable computing devices increases, it becomes more attractive to use them as wireless sensor nodes. As they communicate with gateways to make sensor data available to applications, they necessarily reveal to the gateway nodes information about themselves that could compromise the privacy of their users. This will be exacerbated by the volume of data exchanged and the ability to record data obtained on a massive scale. In this paper we have explored the anonymity properties of data from Bluetooth scans as an aid in understanding how the choices made by gateway nodes affect the privacy of sensor nodes that use Bluetooth. We have examined the anonymity that a device can expect, shown how easy it can be to link devices with individuals, and described some behaviour we have seen that suggests that users misunderstand the privacy implications of configuring their devices. Attributes such as the device class are poor for linking sightings of the same device, the device name can provide a surprising amount of anonymity but when it does not it can be a very effective key to link devices with individuals, and frequently users exhibit privacy-adverse behaviour, such as placing telephone numbers in device names or using statistically rare nicknames. Bluetooth's use of pre-set, lifetime hardware addresses (MAC addresses) is worrisome from a privacy perspective in that they allow trivially linking multiple sightings of the same device. It is not difficult to imagine an alternate approach where device addresses, even if required, would be ephemeral.

In the future we would like to make use of automated record linkage approaches and evaluate a device's settings to estimate the probability of its owner being identified using other databases such as phone books. The objective would be for every wireless device to perform these checks as part of their interactions, whereby nearby vulnerable devices could be pro-actively identified and their owners coerced to remedy the situation through social peer pressure.

## Acknowledgement

The authors would like to thank Greg Zaverucha for his suggestions and help on early versions of this work.

## References

[1] F.-L. Wong and F. Stajano, "Location privacy in bluetooth," *Security and Privacy in Ad-hoc and Sensor Networks*, vol. 3813, pp. 176–188, December 2005.

[2] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott, "Impact of human mobility on opportunistic forwarding algorithms," *IEEE Transactions on Mobile Computing*, vol. 6, no. 6, pp. 606–620, 2007.

[3] N. Eagle and A. (Sandy) Pentland, "Reality mining: sensing complex social systems," *Personal Ubiquitous Comput.*, vol. 10, no. 4, pp. 255–268, 2006.

[4] P. Hui, J. Crowcroft, and E. Yoneki, "BUBBLE rap: social-based forwarding in delay tolerant networks," in *MobiHoc '08: Proceedings of the 9th ACM international symposium on Mobile ad hoc networking and computing*. New York, NY, USA: ACM, 2008, pp. 241–250.

[5] "Bluetoothtracking," http://bluetoothtracking.org/.

[6] J. Bacon, A. R. Beresford, D. Evans, D. Ingram, N. Trigoni, A. Guitton, and A. Skordylis, "TIME: An open platform for capturing, processing and delivering transport-related data," in *Proceedings of the IEEE consumer communications and networking conference*, 2008, pp. 687–691.

[7] A. Pfitzmann and M. Köhntopp, "Anonymity, unobservability, and pseudonymity—a proposal for terminology," in *Designing Privacy Enhancing Technologies*. Springer-Verlag, 2001, pp. 1–9. [Online]. Available: http://dx.doi.org/10.1007/3-540-44702-4_1

[8] C. Díaz, S. Seys, J. Claessens, and B. Preneel, "Towards measuring anonymity," in *Proceedings of the Workshop on Privacy Enhancing Technologies*, 2003, pp. 184–188. [Online]. Available: http://dx.doi.org/10.1007/3-540-36467-6_5

[9] A. Serjantov and G. Danezis, "Towards an information theoretic metric for anonymity," in *Proceedings of the Workshop on Privacy Enhancing Technologies*, 2002, pp. 259–263. [Online]. Available: http://dx.doi.org/10.1007/3-540-36467-6_4

[10] D. Cross, J. Hoeckle, M. Lavine, J. Rubin, and K. Snow, "Detecting non-discoverable bluetooth devices," *Critical Infrastructure Protection*, vol. 253, pp. 281–293, November 2007.

[11] T. Hurman, "Btscanner bluetooth scanner, version 2.0," August 2004.

[12] R. H. Warren and G. M. Zaverucha, "Bluetooth wireless security," in *Graduate Student Research Conference*, University of Waterloo, Ontario, Canada, April 2007, poster.

[13] W. E. Winkler, "Advanced methods for record linkage," Statistical Research Division, U.S. Bureau of the Census., Tech. Rep. rr945, 1994. [Online]. Available: http://citeseer.ist.psu.edu/winkler94advanced.html

[14] I. P. Fellegi and A. B. Sunter, "A theory for record linkage," *Journal of the American Statistical Association*, no. 64, pp. 1183–1210, 1969.

[15] H. B. Newcombe and J. M. Kennedy, "Record linkage: making maximum use of the discriminating power of identifying information," *Commun. ACM*, vol. 5, no. 11, pp. 563–566, 1962.