Pervasive and Mobile Computing 11 (2014) 56–69

Contents lists available at ScienceDirect

Pervasive and Mobile Computing

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

Fast track article

Linking wireless devices using information contained in Wi-Fi probe requests

Mathieu Cunche [a,*], Mohamed-Ali Kaafar [a,b], Roksana Boreli [b]

[a] INRIA, Rhône-Alpes, Grenoble, France
[b] National ICT, Australia

ARTICLE INFO

Article history:
Available online 19 April 2013

Keywords:
Link prediction
Privacy
Social matching systems
Wi-Fi
IEEE 802.11

ABSTRACT

Active service discovery in Wi-Fi involves wireless stations broadcasting their Wi-Fi fingerprint, i.e. the SSIDs of their preferred wireless networks. The content of those Wi-Fi fingerprints can reveal different types of information about the owner. We focus on the relation between the fingerprints and the links between the owners. Our hypothesis is that social links between devices' owners can be identified by exploiting the information contained in the fingerprint. More specifically we propose to consider the similarity between fingerprints as a metric, with the underlying idea: similar fingerprints are likely to be linked. We first study the performances of several similarity metrics on a controlled dataset and then apply the designed classifier to a dataset collected in the wild. Finally we discuss potential countermeasures and propose a new one based on geolocation. This study is based on a dataset collected in Sydney, Australia, composed of fingerprints belonging to more than 8000 devices.

1. Introduction

The huge popularity of mobile and portable wireless devices, including smart phones, tablets and laptops, has further increased the widespread of one of the most used wireless technologies, IEEE 802.11 or Wi-Fi. Included in over a billion of mobile and portable devices in use worldwide, Wi-Fi is provided by Access Points (APs) from a vast majority of homes in the developed and developing world, businesses and by around 750,000 worldwide hotspots. The extensive availability of Wi-Fi connectivity, together with a growing popularity of community networks, has resulted in mobile and portable devices establishing connections to an increasingly large number of APs in various locations.

The open nature of Wi-Fi connectivity has raised a number of privacy concerns in both the media and research literature [1–3], including the well publicized Google Street View collection of traffic from home Wi-Fi APs.

Mobile and portable devices are likely to move across the coverage areas of different APs, e.g. while a person is traveling between home and work. In this likely scenario, low-latency service discovery is a highly desirable feature, as it will maximize the amount of connection time for such devices. We note that the Wi-Fi passive discovery mode, in which the AP periodically sends beacons announcing the AP's Service Set IDentifier (SSID), and where devices listening to these probes select the desired specific SSID, can lead to high discovery delays. In the active discovery mode, the device periodically and actively probes the neighborhood for known APs. This mode, which is supported by operating systems and Wi-Fi chipset drivers on the majority of devices and often activated by default, allows a much lower discovery delay. However, the active discovery mode also raises privacy concerns. Indeed, while in active discovery mode, the mobile or portable device is periodically broadcasting, in the clear, the SSIDs of the APs to which this device has previously been connected to. The

* Corresponding author. Tel.: +33 0651237682.
E-mail address: mathieu.cunche@inria.fr (M. Cunche).

1574-1192/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.pmcj.2013.04.001

device also broadcasts its MAC address, a globally unique identifier. As a consequence, the device is not only advertising its presence to any eavesdropping equipment, but, as an associated issue, this also makes the owner vulnerable to location tracking and possibly profiling attacks.

In this paper, we introduce and study a new privacy threat caused by the Wi-Fi active discovery mode, that is the possibility to infer a relation between devices (and hence the owners of these devices) based on their device's "publicly" announced preferred networks. We refer to the list of preferred networks stored on a device, either partially or globally collected by monitoring wireless probes, as a *device Wi-Fi fingerprint*.

Various sources of Wi-Fi related potential for privacy loss have been addressed in the research literature. The timing pattern of the probe messages can be used to identify the wireless interface's driver [4], or to create a unique device identifier [5]. Our work considers the list of preferred networks as the fingerprint, rather than the timing of probes. We use this fingerprint to link selected devices with similar fingerprints, with the aim of grouping devices and, by association, the owners of these devices.

The SSID of the probed network can also reveal sensitive information about the user, as shown in [3,6,7]. In particular, SSIDs can reveal geographical location of users [8] when combined with specific databases containing APs' location coordinates, like [9]. Our work is complimentary to the research showing how user location can be derived from AP locations, as e.g. users who are detected by our methodology as linked socially or professionally, may also reside in a similar location. This enables additional location detection of users who have APs which may not be in publicly available data bases, or which do not have globally unique SSIDs. In regard of the various problems caused by the current probing process, a privacy preserving access-point discovery has been proposed by Lindqvist et al. [7].

In general, discovering links between any two devices is a challenging task. In [10] the geographical proximity sensing capabilities of bluetooth technology have been used to suggest potential friendship links between users. This approach relies on an always-on application running on the user's phone, which constantly monitors surrounding bluetooth-enabled devices. In [11] friendship links are predicted using a large set of information collected from mobile phones. In both research works, discovering a relationship is seen from the perspective of one of the partners', and assumes the collection of a sustainable amount of complete information both in terms of temporal data and number of attributes. In our case, the amount of information is quite limited and often partial (Wi-Fi fingerprints consist of small pieces of information, and the information is sparse). On the other hand, the datasets we consider are composed of a much larger set of samples, and thus contain much more possibilities, which in turn makes the task of linking the devices more challenging.

We propose to exploit the device's Wi-Fi fingerprint to group devices and the owners of those devices. The techniques presented hereafter are widely applicable to all devices that publicly expose the SSIDs to which they have been previously associated. Our purpose is to show that the evaluation of fingerprint similarity enables simple, yet efficient tracking and profiling of mobile devices users. In light of these findings, it is desirable that the research community investigates the capabilities and limits of these techniques. This will, in turn, allow for the design of appropriate countermeasures to protect users' privacy.

The contributions of this paper are manifold. First, we introduce the problem of linking devices relying on monitored wireless probes. This problem is, to the best of our knowledge, novel and has not been explored in the literature. Second, we consider and adapt several record linkage techniques as fingerprint similarity metrics, to build a classifier that, given two fingerprints, can determine with high accuracy whether the two devices belong to individuals with an established relationship. This tool, when used with suitable record linkage techniques, validates an intuitive observation: the two separate devices belonging to two socially linked individuals most likely share common SSIDs. However, only the existence of a large and/or rare overlap in SSIDs between the two fingerprints results in establishing the links between individuals with an established social relationship. These results are tested and validated on real world data. By passively monitoring broadcast Wi-Fi probes in the city of Sydney, Australia for a period of 100 days, we have collected fingerprints of more than 8K unique devices and more than 24K different SSIDs. Additionally, we have collected a control dataset comprising of device fingerprints representing 30 social relationships, which we use to validate the classifier. We then apply selected metrics to the public dataset and analyze the characteristics of the detected links.

Finally, we introduce and examine the potential of several possible countermeasures that can be implemented to circumvent the privacy threat identified in this paper. We propose a geolocation-assisted service discovery that can be easily implemented at the level of users' devices. We analyze our proposed solution and discuss some implementation details by considering the Android operating system as a use case.

We envision several possible uses of these techniques, not all of them malicious. In particular, in cases of criminal investigations, where the Wi-Fi device fingerprint of a known suspect is collected. Forensics analysts can for instance, collect the fingerprints of users in the suspect's neighborhood, and our classifier could allow them to distinguish if a set of individuals are likely to have a social relationship with the suspect. As a second use case, wireless service providers could gather valuable information across a specific geographical area to send extremely targeted advertisements to users they deem to be socially linked to their (known) customers. For example, when a group of people have been identified as linked through a friendship relation, the knowledge of a single individual's age can allow the service provider to infer the whole group's age category. This would enable very targeted and efficient ad campaigns.

This paper extends a prior report [12] in many aspects. First, we provide a deeper analysis of the data collection process. We introduce practical challenges for Wi-Fi fingerprints collection, and describe the main characteristics of our collected data by analyzing for example the rate at which Wi-Fi fingerprints can be collected in practice and the time required to collect