Computer Communications 117 (2018) 1–12

Contents lists available at ScienceDirect

Computer Communications

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

Building up knowledge through passive WiFi probes

Alessandro E.C. Redondi, Matteo Cesana[*]

Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano Milan, Italy

ABSTRACT

Inexpensive WiFi-capable hardware can be nowadays easily used to capture traffic from end users and extract *knowledge*. Such knowledge can be leveraged to support advanced services like user profiling, device classification. We review here the main building blocks to develop a system based on *passive* WiFi monitors, that is, cheap and viable sniffers which collect data from end devices even without an explicit association to any Wi-Fi network. We provide an overview of the services which can be enabled by such approach with three practical scenarios: user localization, user profiling and device classification. We evaluate the performance of each one of the three scenarios and highlight the challenges and threats for the aforementioned systems.

1. Introduction

Recent studies on the growth rate of wireless traffic have predicted that Wi-Fi traffic will account for more than half of total IP traffic by 2019, with the total public Wi-Fi hotspots growing sevenfold from 2015 to 2020, from 64.2 million in 2015 to 432.5 million by 2020 [1]. This means that a good deal of IP traffic generated by end users to post on social networks, interact with friends, get access to entertainment and other services will go through a wireless first-mile connection.

In this context, capturing and properly processing data from such networks does provide a goldmine to build up value-added services. As a matter of fact, WiFi Internet Service Providers and system integrators are already broadening their commercial offer beyond the simple provision of internet connectivity to include advanced services based on WiFi data analytics. In such a way, commercial WiFi deployments may be transformed into powerful tools for conducting market research and gauging insights from customers, as WiFi traffic can reveal information on first time vs. frequent visitors at shops, customer loyalty, dwell times, walking paths, real-time heat-maps, customer gender and age.

The aforementioned services are generally offered by leveraging either *active* or *passive* WiFi measurements. Active WiFi measurements capture and analyze the traffic of end users which are associated to the specific WiFi hot spots. Such measurements are generally very rich in terms of available information (uplink/downlink traffic exchanged, total connection time, etc.). Furthermore, they generally include the identity of the accessing user, since the vast majority of WiFi hot spots around the world require some type of authentication (e.g., through captive portals). Conversely, passive measurements occur when the data is collected from end user devices which are not associated to any WiFi hot spots. Clearly, such measurements are generally “less informative” than active ones since they are based only on WiFi management frames which are exchanged by WiFi devices regardless their association status (e.g., association request/response, probe request/response, etc.). Still, insightful information can be extracted from passive measurements with the clear advantage of being less intrusive (and less expensive) than active approaches.

In this work, we showcase the potentials of leveraging passive wifi measurements to extract value-added knowledge. Namely, we focus here on the analysis of WiFi *probe request* management frames, which are broadcast by end devices to probe for available WiFi hot spots. Starting from the availability of millions of probe request frames, we provide three different contributions: (i) we propose a thorough analysis of localization-based services built on top of probe request frames; (ii) we propose a method to identify groups of people having similar behaviors in the way they visit a particular area and (iii) we show how to leverage the information contained into probe request frames to automatically detect if the sending device is a smartphone or a laptop, an information that can be used to optimize the network configuration and/or implement services such as management of wide WiFi network or smart content caching approaches.

Referring to the first contribution, only very recently some attention has been given to the problem of exploiting probe request frames to localize users in a passive way. The majority of the works in this area focus on creating location-based heat maps or to track mobile users in a coarse way, rather than focusing on fine-grained localization. Differently, we try to push the localization accuracy of systems based on probe request frames to its limit and we evaluate two localization techniques based on probe request: in the first one, we resort to

[*] Corresponding author.
E-mail addresses: alessandroenrico.redondi@polimi.it (A.E.C. Redondi), matteo.cesana@polimi.it, cesana@elet.polimi.it (M. Cesana).

https://doi.org/10.1016/j.comcom.2017.12.012
Received 7 March 2017; Received in revised form 23 October 2017; Accepted 18 December 2017
Available online 19 December 2017
0140-3664/

parametric model-based triangulation approaches, whilst in the second one we use fingerprinting.

For the second contribution, we propose a set of features derived from the analysis of probe requests capture time that are later used to cluster the users in different groups. We show that with our approach groups of users with very different behaviours can be highlighted and separated. Moreover, we show that probe request messages can be used to infer the geographical features of users (provenience and attitude to travel).

As for smartphone/laptop classification, we show that it can be be performed by collecting (and parsing) only probe requests Wi-Fi management frames, in contrast with those systems that resort to invasive deep packet inspection techniques to read out application layer information in the exchanged packets. Our proposed classification framework first characterizes each device with a set of *features* extracted from the probe request frames; the reference set of feature captures information on the temporal process of probe request transmission (how frequently probe requests are transmitted) and the power levels used in the probe request transmission. Then, a supervised learning approach is used to train different classifiers able to predict the type of transmitting device just by looking at its corresponding features.

The manuscript is organized as follows: Section 2 provides a survey on the reference literature exploiting passive measurements within WiFi networks; Section 3 provides a quick background on WiFi active scanning procedures and further describes the reference system set up used to collect and analyzed the passive WiFi traces; in Section 4 we show how to perform localization, user profiling and smartphone/laptop classification based on passive WiFi measurements. Section 5 concludes the work.

2. Background and related work

The IEEE 802.11 standard defines three types of layer-2 frames which are exchanged among WiFi devices: *control frames, management frames*, and *data frames* [2]. Passive measurement systems generally leverage *management frames* which are exchanged by Wi-Fi enable devices. Note that such devices do not need to be associated to any WiFi access point in order to exchange management frames.

More specifically, we are interested in the management frames transmitted by end devices during the *Active Scanning* phase, that is, the phase in which they search for WiFi networks (access points) in the surroundings to connect to. In such phase, each end device broadcasts a *probe request* management frame to stimulate in-range APs to manifest themselves (replying with a probe reply management frame). Such probe requests are usually broadcasted in sequence on all the available WiFi channels (1–14). The set of information which is contained in (or can be easily extracted by) probe request frames include the Medium Access Control (MAC) address of the sending device, the Received Signal Strength Indicator (RSSI) out of the transmission of the frame and the Preferred Network List (PNL), that is a list of Service Set IDentifiers (SSID) of the WiFi networks which are already known by the sending device. Fig. 1 reports the standard format of probe request frames

Recent studies have shown that properly collecting, processing and possibly coupling such basic information with other external data allows for building up build up some knowledge on the reference population/scenario the probe requests are collected from. The proposed system to accomplish this process are generally composed of three main elements: (i) a collection front-end based on WiFi-enabled hardware to collect probe requests frames; such hardware can be composed of either commercial Access Points or by “home-made” solutions based on low-power embedded devices; (ii) a data processing engine which operates on the collected data to extract the target context knowledge, and optionally (iii) an external service providing side-information on the reference scenario which can be coupled in the data processing phase with the information extracted from the field.

Table 1
Knowledge extracted from passive WiFi sensor systems.

Context	References	
	Indoor	Outdoor
Localization/Tracking/Density	[3–8]	[9–12]
De-Anonymization	[13–17]	[13–17]
Users/Device Profiling	[16–25]	[16–17] [20–25]

The available systems based on passive WiFi probes can be classified according to the specific target information/knowledge which is extracted. Generally speaking, three broad classes can be identified in this respect: (i) systems targeting localization and tracking of end users; (ii) systems willing to associate a specific identity (real or cyber) to any given captured device; and (iii) systems targeting end user profiling with respect to technical and social parameters. Table 1 reports a classification of the reference literature with respect to the aforementioned guidelines, further distinguishing among contributions targeting indoor and outdoor environments.

2.1. Localization and tracking

Systems of this type generally exploit the information on the RSSI and the proximity to WiFi sensors to infer the geographical position of end users. The work in [3] targets pedestrian flow estimation across the security check in an airport. Several non-supervised learning approaches are proposed and qualitatively compared against a proxy measure for the flows and density, that is, the number of boarding pass scans performed at the security check at given time intervals.

Along the same lines, Fukuzaki et al. propose in [4] a pedestrian flow estimation service in shopping malls; the proposed system first estimates the position of a given device by leveraging the Received Signal Strength Indicator (RSSI) out of a probe request message captured by multiple probes, and then builds up a statistical characterization of the users flows. The proposed system is also used to count the number of people in the reference environment, namely, the authors use a simple linear model which returns the estimated number of people out of the total number of perceived MAC addresses; the model is trained against a secondary people counting system based on motion detection sensors at the entrances of the shopping mall.

Very recently, there have been some works targeting the problem of passive indoor localization using setups similar to the one proposed in this paper. In [5], a system composed of eight WiFi sniffers is deployed in an area of about 5000 m^2. A triangulation-based algorithm is used to localize coarsely the users in eight areas of the experimental area, however no details on the performance of the localization algorithm are reported. In [6], twelve WiFi sniffers are deployed in an area of about 340 m^2. Fingerprint localization through k-Nearest Neighbour classifier is performed, with a reported median error of about 4.5 m. The work in [7] presents *Probr*, an open source software solution to capture and process probe requests in order to support several on-line analysis tasks,

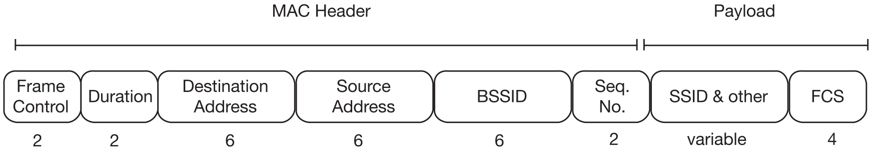

Fig. 1. Probe request frame format. Numbers represent the field size in bytes.