Mind Your Probes: De-Anonymization of Large Crowds Through Smartphone WiFi Probe Requests

Adriano Di Luzio*, Alessandro Mei, and Julinda Stefa
Department of Computer Science, Sapienza University of Rome, Italy.
Email: *diluzio.1487872@studenti.uniroma1.it, {mei, stefa}@di.uniroma1.it.

***Abstract*—Whenever our smartphones have their WiFi radio interface on, they periodically try to connect to known wireless APs (networks the user has connected to in the past). This is done through WiFi Probe requests—special wireless frames that contain the MAC address of the sending device and, in most of the cases, the human-readable name-string (SSID) of the known AP. This semantic information, inherent to the network protocol, is sent in the clear and, if sniffed, can help discover important information and phenomena of people and human nature that have nothing to do with technology.**

In this paper we present the idea of exploiting WiFi probe requests to de-anonymize the origin of participants in large events. We make use of several, publicly available datasets containing more than 11M of probe requests collected in scenarios that are of citywide, national (two political meetings), and international religion-related relevance. We show how, by exploiting the semantic information brought by the relative WiFi probes, we are able to discover with high accuracy the provenance of the crowds in each event. In particular, the de-anonymization outcome of the two political meetings held few days before the election days in Italy match surprisingly well the official voting results reported for the two respective parties.

***Index Terms*—Privacy, WiFi probe requests, social sciences computing, social networks.**

I. Introduction

Posting on social platforms, contacting friends and family, online banking, emails, entertainment, almost anything we could need is, nowadays, just a touch of our thumb away. All thanks to our smartphones. They have undoubtedly changed the way we interact with technology. Not only do they allow us to navigate anytime and everywhere, but they are built to do so in the most efficient way. Take the wireless interface for example: If on, it automatically connects to WiFi networks, even if we are already covered by a 3G network (known to be more expensive and less energy efficient than WiFi connectivity). This is enabled by a type of special wireless frame called *WiFi probe request* [1] that our devices periodically send in the clear to discover the availability of known WiFi networks in range. As we will further discuss in Section II-A, the probe requests contain the MAC address that uniquely identifies the sending device. The MAC address can be the device's real universal address or a temporary address if MAC randomization is used—as it is done in the very latest versions of mobile operating systems. Probes can be of *broadcast* type—not specifically directed towards a particular WiFi network—or *directed*—specifying the SSID (string name identifier) of a particular WiFi network.

Directed probes grant a highly efficient, reliable, and automatic network discovery. Most of the current mobile OSs make use of the device PNL (Preferred Network List) and adopt directed probes to request the availability of WiFi networks to which the user has connected before. In this way our devices are able to automatically, in just a few seconds, switch to our "Home WiFi" as soon as we cross the doorstep. But there is much more than that: The SSIDs of the WiFi networks contained in these frames, inherent to the technological side of the connection protocol, is full of semantic information. This is what makes directed probes very valuable from a sociological point of view: They can help discover aspects of human nature that have nothing to do with technology. Indeed, by just *listening* to what smart-phones are *shouting* through their probes it is possible to draw a detailed picture of the people surrounding us. As we will also discuss in Section V, many insightful works have shown how wireless probes can be used to infer the relationships among people [2], predict who they will meet [3], [4], or even discover the welfare of large crowds [5].

In this paper we take a step forward towards the understanding of human nature through probe requests. Our goal is to uncover, with high accuracy, the geographical provenance of people in large gatherings. Our de-anonymization process, described in detail in Section III, is based on the probe request frames released by their mobile devices just by default; i.e., not requiring any intervention neither by the device owners, nor by any other. Upon the tiny pieces of information included in these frames we build an automatic methodology to de-anonymize the provenance of tens of thousands of people participating in gatherings of citywide, national (political meetings of two parties held around election days), and international (religion related) events, lasting just a few hours each. Finally, we test our de-anonymization methodology by comparing its outcome with ground truth data—the official election results for the two parties whose meetings, held around election days, we target in this paper. The comparison shows that the result of our de-anonymization match surprisingly well the official general election results of the two parties (Section IV).

To the best of our knowledge, this is the first work that shows how to achieve this amount of detailed knowledge on large crowds of people based solely on their probe requests. Knowledge whose impact and applications, as we further discuss in Section VII, can span from advertising of commercial

978-1-4673-9953-1/16/$31.00 ©2016 IEEE

activities to prediction and prevention of infection spreading at a neighborhood/city/nation level.

II. WiFi Probes: Background, Sniffing, and Datasets

A. Background

Wireless probes are a peculiar type of 802.11 Management frames [1] exploited by wireless cards of devices in a pre-connection phase. Devices use wireless *probe requests* to discover access points (APs) that are available in their vicinity. Upon the reception of a probe request, an AP replies with a *probe response* containing the rate of the data supported and other parameters of the wireless network station. In this work we will focus on probe requests.

Probe requests can be of two types: Broadcast (used to actively seek any AP in range) and directed (addressed to a specific AP). The header of these 802.11 management frames, depicted in Figure 1(a), contains the following information:

- Frame Control: indicates the subtype of the frame;
- Address 1: is the address of the destination (DA) of the frame;
- Address 2: equals the MAC address of the sender, also known as SA (source address);
- Address 3: in case of a directed probe request frame, the BSSID (MAC address) of the AP probed.
- Sequence Control: the (incremental) sequence number of the packet. Used to distinguish between re-transmissions.

The body of a probe request is instead shown in Figure 1(b). In particular, the value of the first field can either be the SSID of the AP probed (directed requests), or a null character (broadcast requests). Table I shows examples of broadcast and directed probe requests sent by the device with MAC address 10:9a:42:42:42:42.

By default, Mobile OSs keep memory of the networks which the user has connected to in the past. This information is stored within the so called Preferred Network List (PNL) of the device. This list contains, among other data, the SSIDs (or even BSSIDs in some cases) of the networks the user has connected to in the past, the type of cryptography involved in each connection, as well as the possible user passwords to access each of these networks. PNLs enable fast and energy-efficient connectivity switch to known networks. Indeed, mobile OSs use the entries of this list to periodically seek, through directed probe requests, if any of the PNL networks is in range. In the case of a positive probe response the device attempts to connect to the corresponding access point. Note that this happens even if the device is currently connected to a wifi access point, in the hope that a better connection (a stronger signal associated to a closer access point) is available.

The transmission frequency of directed probes typically varies from device to device. Among other factors, it depends on the type and the version of the OS installed and on the state of the device itself (either asleep, on standby, or associated to a wireless network). Devices that are asleep typically transmit a probe every minute. However, the transmission frequency can increase up to 10–15 times per minute for devices in standby and with their screen on[1].

WiFi network probing is the very first phase of a connection attempt. It becomes thus necessary to send probe requests in the clear. As a result, it is very easy to capture (*sniff*) and process these packets. Indeed, it is enough to place some equipment with a WiFi antenna set in monitor mode in the range of the target device [6]–[10].

B. Description of the datasets

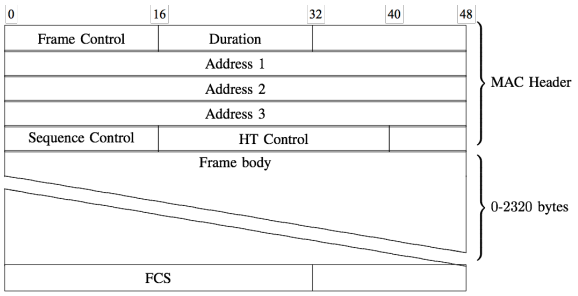
In this work we will be using the largest dataset of WiFi probes publicly accessible, made available online by the authors of [5]. The dataset contains around $11M$ probes of about $160K$ unique devices. It was the result of a collection campaign lasted 3 months in 2013 and performed in the Italian Capital, Rome. As described in [5], the campaign included scenarios of national, international, and citywide relevance. Here below we summarize shortly each scenario:

1) Nationwide scenarios: The national events targeted by Barbera et al. [5] were related to the 2013 Italian general elections (24–25 February, 2013). In particular, they focused on the meetings of two important political parties in Italy. The first is the meeting that closed the electoral campaign of the *M5S* party, held in Rome on February 22, 2013. The second event was the post-electoral meeting called by Silvio Berlusconi, ex Prime Minister of Italy and head of the *PDL* party, held in Rome on March 23, 2013. As Barbera et al. point out in their work [5], the local police declared that both events were attended by a mixed nation-wide audience. We will denote the relative datasets with *P1* (M5S) and *P2* (PDL).

2) International events: The second type of events targeted in [5] were of international audience. In particular, they include the farewell speech of Pope Benedict XVI after his resignation and the first public speech of his successor, the current Pope Francis. They were both held in Saint Peter's Square of Vatican City, on February 24th and March 17th, 2013, respectively. As described in [5], both events were two very important historical moments for the Catholic Church—the Vatican was literally occupied by pilgrims and tourists from all around the world. We will denote the datasets relative to these events with *V1* and *V2* respectively.

3) Citywide scenarios: These scenarios include data collected in the main train station of Rome (that we will denote with *Station*), a shopping center (that we will denote with *Mall*), and the biggest university campus in the city (that we will denote with *University*). The *Station* dataset contains probes collected for a total of 7 hours split in a time-range of four different days; the *Mall* scenario aimed at collecting data from local residents of Rome. It targeted one of the biggest malls of the Italian Capital in a particularly crowded day, the afternoon of the Holy Saturday (the day before Easter) in 2013. The data collection lasted 3 hours and a half. Lastly, for the *University* dataset the authors deployed an antenna at a fixed point located at the main entry of Sapienza University. This

[1] https://meraki.cisco.com/lib/pdf/meraki_whitepaper_cmx.pdf



(a) 802.11 Management frame format.

Order	Informations
1	SSID
2	Supported Rates
3	Request Informations
4	Extended Supported Rates
5	DSSS Parameter Set
6	Supported Operating Classes
7	HT Capabilities
8	20/40 BSS Coexistence
9	Extended Capabilities
10	SSID List
11	Channel Usage
12	Interworking
13	Mesh ID
Last	Vendor Specific

(b) Probe Request frame body.

Fig. 1. Management frame format and body of a probe request frame in the 802.11 protocol.

FC	Duration	DA	SA	BSSID	SEQ Ctl	SSID	FCS
...	...	ff:ff:ff:ff:ff:ff	10:9a:42:42:42:42	ff:ff:ff:ff:ff:ff	...	*null* (Broadcast)	...
...	...	ff:ff:ff:ff:ff:ff	10:9a:42:42:42:42	ff:ff:ff:ff:ff:ff	...	"Free-WiFi"	...
...	...	ff:ff:ff:ff:ff:ff	10:9a:42:42:42:42	ff:ff:ff:ff:ff:ff	...	"Home-WiFi"	...

TABLE I

EXAMPLE OF PROBE REQUESTS SENT BY DEVICE WITH MAC ADDRESS 10:9A:42:42:42:42: ONE BROADCAST PROBE (ON TOP) AND TWO PROBES DIRECTED TO THE NETWORKS WITH SSID "FREE-WIFI" AND "HOME-WIFI" RESPECTIVELY. SEQ CTL INDICATES THE INCREMENTAL SEQUENCE CONTROL NUMBER ASSIGNED TO SUCCESSIVE PROBES.

Latitude	**Longitude**	**SSID**	**BSSID**	**Last seen**	...
41.94156265	12.52643299	Stranger753	E0:91:F5:7C:D9:8E	2012-12-19 01:30:23	...
41.86511993	12.47039986	Alice-60652654	00:19:3E:40:40:F3	2011-11-05 20:25:06	...
41.88961792	12.50711727	TISCALI	00:90:d0:4b:78:28	2011-04-08 07:19:03	...

TABLE II

AN EXAMPLE OF THE RESULTS OF A WIGLE QUERY ON WIRELESS NETWORKS IN THE CITY OF ROME.

is the biggest university in Italy, and second big university in Europe. Differently from the other scenarios, in this case the collection was continuous and it lasted 6 weeks.

Altogether, these datasets contain as many as $11,136,711$ probes (both directed and broadcast) from $164,740$ uniquely identified devices [5]. As anticipated earlier in this paper, only directed probes are of relevance to our study. These add up to $5,345,083$ total probes from $59,684$ unique devices.

III. DE-ANONYMIZING EVENTS

Our goal in this work is to show how, through a tiny piece of datum that smartphones are routinely "shouting out" due to their inherent technological aspect, we can easily build up an amount of knowledge that allows us to uncover facts, details, and crucial information about large crowds. In particular, we aim at de-anonymizing the geographic region or the city the people participating in big events come from. In addition, we want our de-anonymization methodology to be automatic, to not involve collaboration from the user side, and to be highly accurate.

Our intuition is that people connect to WiFi networks of areas in which they spend long periods of time—homes, offices or schools, favorite coffee-shops, friends' houses, and so on. All these places are typically located in the same state/region/city people live in. As a result, so are the APs their devices connect to more frequently. Of course, there are some exceptions. The APs of the airports, for example, that we use when we travel outside the country or go on vacation. Even so, the amount of connections to these "outsider APs" cannot keep up with the continuous and recurrent connections to APs located in the area we live in, which dominate the PNL list of our smartphones. Therefore, towards our goal to de-anonymize the provenance of the people in a certain event, we make use of the geographic position (GPS coordinates) of the APs contained in the probe requests released by the devices of the participants. However, probe requests only contain the SSID of the AP, not their geographic position. So, the idea is to firstly link each SSID of the event's dataset with its corresponding geographic coordinates. Then, during a second step, to make use of this information in the de-anonymization process.

A. Linking SSIDs to geographic coordinates

In our de-anonymization process we aim at high-accuracy. Therefore, we make use of the Wigle.net (Wireless Geographic Logging Engine), one of the largest database mapping APs to GPS coordinates. Wigle.net is a crowd-sourced wireless position database. People can contribute to the database by installing the mobile app of Wigle.net. When launched, the app logs information like, SSID, BSSID, and GPS coordinates, for every AP in the range of the user device. The app also allows the user to upload the logged data on the servers of Wigle.net. In addition to the mobile app, users can also make use of larger devices, like desktops or laptops, to upload data to Wigle. In

City	Wigle APs per city	Dataset APs on Wigle
Genoa	11,913	238
Milan	119,653	2,428
Naples	25,994	624
Palermo	5,915	127
Rome	99,093	9,382
Total	**163,465**	**12,799**

TABLE III
DISTRIBUTION, PER CITY, OF THE WIGLE-COLLECTED APS AND DATASET APS THAT WERE UNIQUELY LOCATED IN WIGLE.

this case, they can use networking tools like *netstumbler*[2] or *kismet*[3] to collect the data. At the time of the writing of this paper, Wigle.net contains over $150K$ registered users and more than $136M$ wireless APs localized worldwide.

The Wigle dataset can be queried both through the web interface or through APIs. One can look up the position of an AP making use of the SSID (or BSSID) of the network. In addition, the system allows also to retrieve APs located within a geographical area of user's choice. The area is defined by 4 geographical points set by the user. More in details, given four pairs of GPS coordinates (latitude and longitude) in input, the system constructs the quadrilateral having these four geographic locations as its vertexes. It then queries the database for any wireless network whose GPS coordinates fall within the quadrilateral. Table II shows an extract of a query of this kind over a small area in Rome.

The events we intend to de-anonymize include meetings of national relevance, with a large number of participants from all over Italy. Unfortunately, Wigle limits both the number of queries that a registered user can perform (max $10,000$), both through the API as well as through the web interface, and the total number of results received per query (max $1,000$). The Wigle system limits not only the user ID, but also the IP from which the user connects to it. To overcome these difficulties we built a customized software that makes use of a large number of Wigle user accounts registered by us and continuously tunnels the connections to a large number of proxies. This way we were able to collect information on a massive number of networks spread throughout Italy—$163,465$ APs present on Wigle located in 5 major Italian cities: Milan, Genoa, Rome, Naples, and Palermo. Table III shows the distribution of the Wigle APs per each city, as well as the number of APs from the datasets [5] that were located on Wigle *in a unique* way. I.e., we filtered out APs whose SSIDs had more than one location in Wigle.

IV. DE-ANONYMIZATION: THE RESULTS

Combining the information of the PNLs revealed by devices' probes, together with the geographic position of the respective SSIDs, we are now able to investigate the provenance of the participants in the events of citywide, national, and international relevance described in Section II-B.

[2]http://www.netstumbler.com/

[3]http://www.kismetwireless.net/

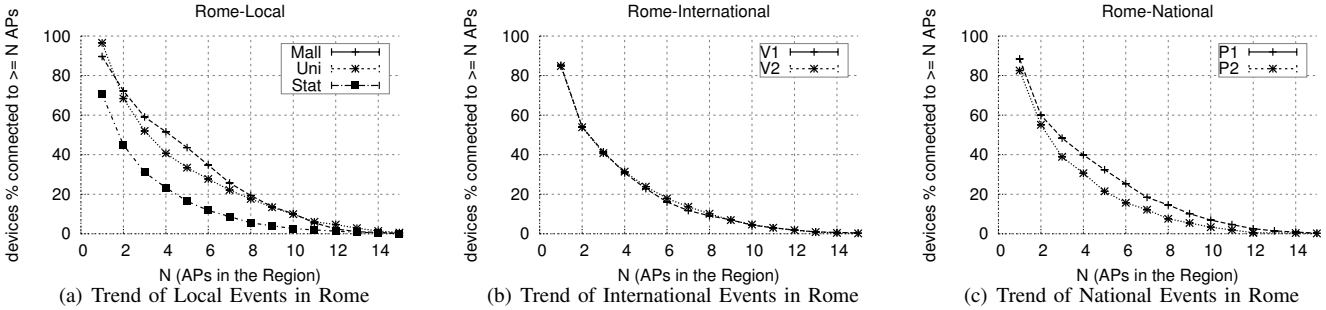
A. PNLs vs Cities

A first, interesting aspect to explore is the number of APs of a given city present in the PNL of the event participants. In particular, we study for each event the distribution of the users requesting at least N APs of a city in dependence of N. Here we focus on three representative cities—Rome, Milan and Naples. The results, for Local (citywide), National, and International events are shown in Figures 2, 3, and 4. In the figures the x axis denotes the number N of APs of a city while the y axis shows the percentage of devices that requested, through their probes, at least N APs in that specific city.

Obviously, all distributions present a decreasing shape and have a relatively high starting point when $N = 1$. This is expected: It is relatively easy for a device to have at least one AP from a certain city, especially when this city is Rome where the collection of probes has been performed (see Figure 2). As soon as the number N of APs from Rome within the same device's PNL increases, the devices that meet this requirement naturally drops down. In addition, the number of different APs that can be stored in our devices' PNLs is limited, and the entries are updated dynamically. The limit depends both on the vendor and on the operating system; e.g., on Samsung Galaxy S+ with Android 2.3 the PNL can contain up to 16 entries. Once the PNL limit is reached and as we change our connection habits, APs which we do not connect to anymore are replaced by those that are most recently used.

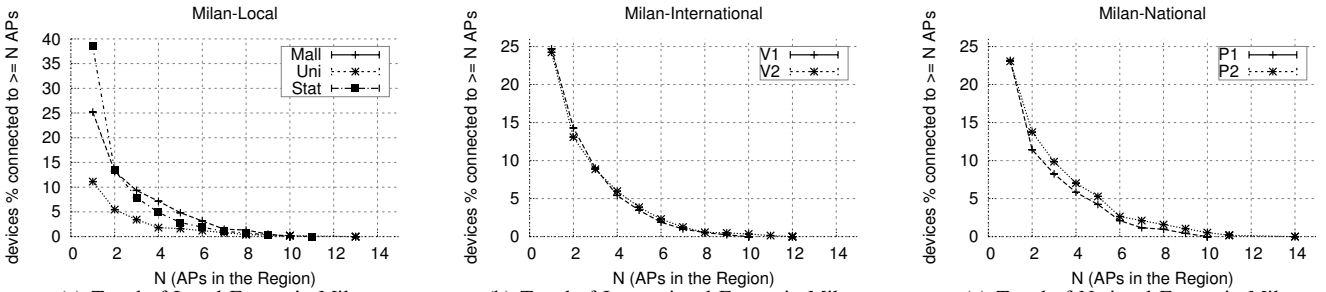
In addition, the distributions for Rome present higher values than those of the other cities for the whole distribution, independently from the event type. Indeed, although some of the events were of international and national relevance, they took place in the Italian Capital. Therefore, it is reasonable to believe that most of the participants were from Rome. As such, their PNLs contained more APs located in Rome than in other cities.

Now, let us concentrate on the events. As far as local events are concerned, we note that the distribution of the Mall dataset is prominent in Rome, while it is taken over from the Station event in other cities. This confirms the intuition that most of the visitors of the shopping center during the evening of the Holy Saturday are, most probably, families and groups of friends living in the metropolitan area of Rome. The Station dataset gains prominence in Naples and Milan, while it is won over by the University dataset in Rome. This also is to be expected: While Sapienza University is one of the biggest institutions in Italy with students from many other cities of Italy and Europe, most of them move to Rome for the duration of the studies (except for a small fraction of commuters from the geographical area around Rome). Therefore, the PNLs of their devices are dominated by APs located in Rome. Aside from this, being Rome located in central Italy and being the Capital, it is frequented by people from all around the country for business, tourism, or personal reasons. This effect is especially amplified in areas like Airport, Train Stations, and so on. Accordingly, the Station dataset presents low values in Rome, while it takes over both the University and Mall



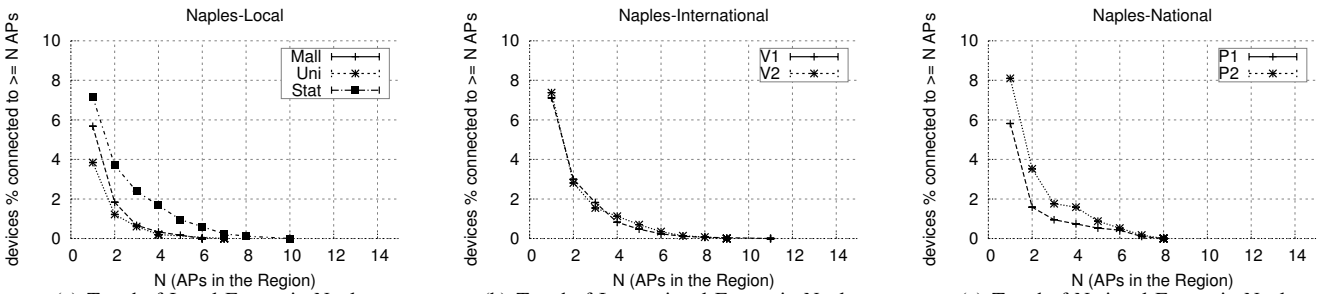
(a) Trend of Local Events in Rome (b) Trend of International Events in Rome (c) Trend of National Events in Rome

Fig. 2. For each dataset, the percentage of devices that connected to at least N APs in *Rome*.



(a) Trend of Local Events in Milan (b) Trend of International Events in Milan (c) Trend of National Events in Milan

Fig. 3. For each dataset, the percentage of devices that connected to at least N APs in *Milan*.



(a) Trend of Local Events in Naples (b) Trend of International Events in Naples (c) Trend of National Events in Naples

Fig. 4. For each dataset, the percentage of devices that connected to at least N APs in *Naples*.

datasets—typically with a higher percentage of people from Rome—in Milan and Naples.

That said, it is interesting to notice how the two International events held in Vatican City present almost identical distributions in all cities. After all, they were very similar events gathering people of the same faith in the same location, Saint Peter's Square. In addition, both events were of a historical importance for the Catholic church, happening just a few weeks one from the other, and with participants not only from all over Italy but from all over the world. Quite surprisingly, however, the two political events, also happening with a few weeks one from the other and both around the 2013 Italian political elections (February 24–25), present different results in all cities. As we will discuss in the next section, a further investigation revealed that this difference matches exactly the election results of the respective parties in each of the cities considered in this work.

B. Determining the geographical provenance of crowds from WiFi probes

In order to de-anonimyse the events we targeted in this work, each device (user) needs to be assigned, based on the APs in her own PNL, one home-town among the cities considered. As we already argued in Section III, a user (device) must live in the city X whose APs dominate her PNL (i.e., most of the AP (SSID) entries in the PNL are located in city X). However, the number of the APs of a given city only is not enough to determine the provenance of a user. Indeed, take for example a businessman from a small town from outside Rome (e.g. Latina) who commutes from Latina to Rome very frequently. It is possible that his PNL contains a similar number of APs from both cities, even though his hometown is actually Latina. However, the list of the APs from Rome is very likely to contain networks to which lots

of other people connect to (office AP, hotel AP, the AP of the restaurant he has lunch in during office hours, and so on). On the contrary, the list of the APs from Latina is more likely to include smaller, less popular APs (home AP, friend's home AP, and so on), and that intuitively are more meaningful into determining the geographical provenience of the user. For this reason, in order to assign a given device to a location, we have to take into account a given APs popularity as well, defined as the total number of devices present in the dataset that have the given AP in their PNL list. To do so we group the APs of a given user's PNL list according to their geographical location (city). Then, we build a corresponding list of cities $L_u = \{cd_1, \ldots, cd_n\}$ that represents possible candidates for the user's provenance. Then, we assign each location candidate in L_u a provenance rank value that reflects, not only the number of the APs from that location present in the user's PNL, but also how meaningful these APs are, in a simple way as follows:

$$provRank(cd_i) = \frac{\#APs \in cd_i}{\Sigma_{AP_k \in cd_i} popularity(AP_k)}, \quad (1)$$

where $popularity(AP_k)$ denotes the number of devices in the dataset whose PNL contains AP_k. Finally, we assign to a given user the candidate city $cd_i \in L_u$ with the highest $provRank(cd_i)$ value.

Intuitively, Equation 1 gives more prominence to those cities that are represented by many APs in a given user's PNL list that are also meaningful for the users' geographical provenance. However, although unlikely, it might happen that Equation 1 assigns the same provenance rank value to two different candidate cities in a user's L_u. This can happen, for example, when a given user from a city X (of which he has few unpopular APs in her PNL list) seldom travels for work to a city Y where he frequents many popular locations (restaurants, hotels, offices, and so on). In these cases we break ties selecting the city with the fewer but more meaningful APs in the user's PNL list (city X in the above example).

The results of the assignment discussed above are depicted in the Figures 5 (for the city of Rome) and 6 (for the other cities considered). First note that, as already discussed in the previous section, Rome, being the area in which the collection actually took place, is the city with the highest participation in each of the events. In addition, the gap between the two highly local events (Mall and the University) and the Station dataset is more prominent. Again, this is expected: Intuitively, the main Train Station of Rome is frequented by citizens from all over Italy. In particular, from all the cities considered, we note that the train station is more frequented from Milan and Naples citizens—consider that Milan and Naples have the highest number of inhabitants, besides Rome, among the cities investigated. Indeed, they present the highest values also for other types of events (see Figure 6(b) and 6(a)).

The two Vatican events present very similar results in all cities. Recall that these two events were very similar, and not that distant in time from one another. One would expect that, for the same reason, also the two nationwide political

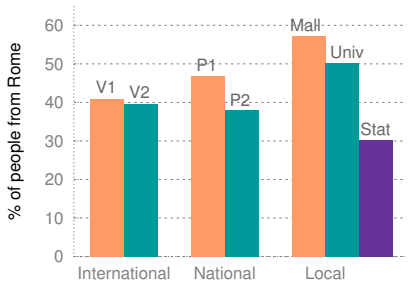


Fig. 5. Participation in the events of citizens from Rome. In the figures V1, V2, P1, P2, Mall, Univ, and Stat denote respectively the datasets of the M5S meeting, the PDL meeting, the last speech of Pope Benedict II, the first speech of Pope Francis, the pre-easter collection in the shopping center, the data collection at the university campus, and the data collection at the train station.

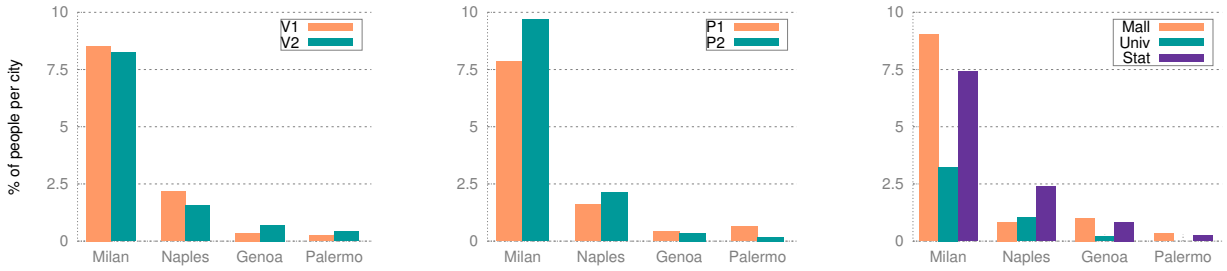
events P1 and P2 would present not-so-different trends from one-another in all cities. These two events were very close to the Italian election days in 2013. Therefore, we realized that there might have been a correlation between the origin of the participants in these two political meetings and the actual election result for the two respective parties (M5S and PDL). To confirm our intuition we checked the official election results published on the web-site of the Italian Ministry of Internal Affairs for these two parties in each of the cities under study[4]. The comparison among the official votes and our de-anonymization results, included in Table IV is impressive: The de-anonymization outcome succeeds in predicting the party which got most votes for all cities considered. Most importantly, the absolute ratio of the results predicted de-anonymizing the events is extremely close to the ratio of the official election results. This phenomena is particularly evident for the 3 largest cities considered—Rome, Milan, and Naples. Note that the final dataset we were able to work on included only about $20,000$ different users; i.e., about 0.3% of the $+6.4M$ total inhabitants of the 5 cities considered[5]. This makes us believe that the results of our methodology can further improve with datasets that approach the real population and within other contexts related to people inclinations and trends.

V. Related Work

WiFi probe requests have been targeted by many research groups worldwide. The creative and profound ideas and solutions of the works based on these tiny data packets cover issues from a multitude of areas—security, privacy, social networks, sociological studies, and so on. In this section we discuss the works that, in our opinion, have most influenced the area.

As we already discussed in Section II-A, probe requests help devices to efficiently investigate whether they are in range of networks to which they have connected before. In case of a positive response, the devices attempt to automatically connect to them. This process, however, opens the way to a myriad of

[4]http://elezionistorico.interno.it/index.php
[5]http://www.citypopulation.de/Italy-Cities.html



(a) De-anonymization of International Events. (b) De-anonymization of National Events. (c) De-anonymization of Local Events.

Fig. 6. Participation in the events of citizens from Milan (Mil), Genoa (Gen), Naples (Nap), Palermo (Pal). In the figures V1, V2, P1, P2, Mall, Univ, and Stat denote respectively the datasets of the M5S meeting, the PDL meeting, the last speech of Pope Benedict II, the first speech of Pope Francis, the pre-easter collection in the shopping center, the data collection at the university campus, and the data collection at the train station.

City	De-anonymization Winner	Votes Ratio (%) (P1 / P2)	Election Winner	Votes Ratio (%) (P1 / P2)
Rome	M5S (P1)	1.23 (46.70 / 38.03)	M5S (P1)	1.22 (26.74 / 21.91)
Milan	PDL (P2)	0.80 (7.83 / 9.68)	PDL (P2)	0.77 (15.90 / 20.70)
Naples	PDL (P2)	0.75 (1.60 / 2.12)	PLD (P2)	0.78 (23.02 / 29.30)
Genoa	M5S (P1)	1.20 (0.42 / 0.35)	M5S (P1)	1.60 (31.00 / 19.00)
Palermo	M5S (P1)	3.50 (0.63 / 0.18)	M5S (P1)	1.23 (31.01 / 25.20)

TABLE IV
COMPARISON AMONG THE DE-ANONYMIZATION OUTCOME AND THE OFFICIAL ELECTION RESULTS FOR THE TWO POLITICAL MEETINGS.

attack scenarios. For example, the work [6] shows how an attacker could monitor the probes released by the victim and set up honey–pot ad-hoc APs with the same SSID as those released by the probes to attract the user device to connect to it. Once the device is hooked and connects to the malicious AP, the attacker can cause severe damage to the unaware user: Redirecting any DNS request of the victim to a malicious server, logging credentials and other sensitive user data, or even dumping the whole session to have a full copy of the victim's (unencrypted) browsing history. This attack known as the *Evil Twin Attack* [6] is efficient only for SSIDs of networks which do not involve cryptography in the communication with the user device. However, because of the proliferation of public access hot-spots, it still can be largely exploited. The production and the sending of probe requests does not stop when the device is associated with an AP, making the protocol even more vulnerable. Indeed, as shown in [7], [8], an attacker can still eavesdrop user probe requests, send particular 802.11 packets (called disassociation frames) to explicitly trigger the termination of the association with the network currently in use, and then launch the Evil Twin Attack.

Aside from inducing the security threats mentioned above, it is evident that probe requests menace the privacy of users. At the time of the writing of this paper, for example, London's trash bins [11] are able to sniff smartphone probes and gather information on citizens walking on the city streets. Recall that probes reveal the users' PNLs. From it, an adversary can deduce sensitive information like the name of a user's workplace, the bar/restaurant she usually goes to, and so on. Over and above that, probe requests can be exploited in more sophisticated attacks like tracking users' path [12], [13], discover their indoor location [14], estimating crowd volumes [15], or even predict their movements [16] or discover social relationships among people [2]–[4]. Lastly, the probes allow to fingerprint and uniquely identify user devices. Indeed, as noted by the authors of [9], [10], the implementation of the probing algorithms varies from manufacturer to manufacturer. So, by analyzing statistically the time interval among two consecutive probes, it is possible to discover the device the manufacturer, the OS, and the drivers used by the wireless network interface [9], [10]. This type of attack is completely passive. So, it is undetectable and hardly preventable. What is worse, it enables the adversary to take advantage of a particular vulnerability of the devices' wireless interface, drivers, the OSs, and so on, to severely damage the user.

Very recently, work has been done to propose alternative ways to send probe request in such a way that privacy is protected [17]. The authors of this work propose that SSIDs are associated to a GPS position, and probes are sent to an access point only when, by looking at the current position of the device, we know that the access point may be in range. While this technique is an interesting approach and helps reduce the privacy leakage, it may be hard to use indoor, where GPS does not work reliably, and it is not clear whether keeping the GPS interface always on is actually too energy demanding for most users.

Among all the insightful works dealing with probe requests we believe that [5] is the most related to our study. Similarly to us, the authors aim at studying sociological aspects of large crowds. After performing a collection campaign of probe requests lasted about 3 months and targeting events of International, National, and Citywide relevance, they focus on

discovering social-related phenomena like: the distribution of languages of the people participating in the events, the vendors of the devices they use, and, based on the cost of the different brands, give insights on the wealth of the population.

In this work, however, we showed how to de-anonymize the provenance of large crowds of people. Our de-anonymization process relies on the small piece of information included in the WiFi probe requests that mobile devices release, due to the technological characteristics of the 802.11 protocol, just by default; i.e., not requiring any intervention neither by the user, nor by an outsider. We build up upon the information contained in WiFi probe requests and deliver an automatic de-anonymization methodology to undercover the home-town (provenance) distribution of tens of thousands of people participating in big events lasting just a few hours each. Most importantly, the outcome of our de-anonymization methodology match accurately the trend of Italy's 2013 official political election results. To the best of our knowledge, this is the first work to attain this sort of outcome from wireless probe requests.

VI. Discussion

WiFi probe requests incontestably open up many issues with respect to security and privacy. Actually, the only way to prevent these issues is to turn off the device's WiFi interface. However, very few users prefer to do so. Indeed, typically people choose usability over security and/or privacy (think of the billions of people that actively use Facebook to post private pictures, sensitive information, and so on, on a daily basis). In addition, not all the users are tech-friendly and aware of these risks. What is worse, mobile OSs do not always make it easy for users to block their devices from sending probes. For example, Google's Android versions successive to 4.3 (*Jelly Bean*) have a WiFi-location related option, called *scanning always available*, enabled by default[6]. This option, whose purpose is to make the *Google Location Services* able to work without relying on the GPS or the cellular site position, leaves the probing mechanism operate, even with the wireless interface being explicitly switched off by the user.

Directed probe requests are particularly used in mobile OSs, which intentionally prefer them over broadcast probe requests. The rationale of this choice is energy-efficiency. In fact, after sending a directed probe request the device receives at most one probe in response—the one of the AP solicited and only if the AP is in range. On the contrary, broadcast probes cause all APs in the device's range to send a response. Obviously, broadcast probes are more privacy savy, but they generate many conflicts on the MAC layer due to the number of response packets involved. This is especially true in places with many people and many access points, which is very common nowadays. Not only that, broadcast probes force the device to receive and process all of these packets, often in vain if the user is far from his own WiFi access points, with the unavoidable energy costs associated. Indeed, typically, most of these responses are sent from APs not belonging to the user PNL and can thus not be used to automatically switch the connection. Therefore, WiFi directed probe requests are undoubtedly the most energy efficient tool that enable user devices to discover and associate with reliable known networks as soon as they are available. As a result, they are still highly used by most devices equipped with a wireless network interface. Actually, we believe that it will continue to be so, despite the security and privacy issues they bring and despite the fact that users are more and more privacy concerned, since users and so OS developers still value energy efficiency as one of the top priorities with mobiles.

The very recent randomized MAC address technique that is being introduced in the latest versions of mobile operating systems consists in using temporary and locally managed MAC addresses when sending probes. These addresses are used, however, only under some circumstances that depend on the version of the operating system. For example, with iOS 8.1.3 they are only used when the device is in sleep mode. Even though this technique helps somewhat avoid tracking the MAC address over time, recent works have shown that it can easily be defeated [18]. As a matter of fact, the authors in [18] analyze the problem and point out that, in order o develop products using WiFi according to the ISO/IEC 8802 standards, an organization must register to IEEE MAC Address Block Large [19]. On top of this, randomized MAC addresses correspond to non assigned ones. As a result, randomized MAC addresses can be promptly differentiated from the real ones. In addition, the probes released from the same device contain an incremental sequence number SEQ as well as brand-related information. Putting all this information together, it becomes easy to link packets to a same device [18].

That said, in this work we are interested in instant snapshop of crowds. Therefore, randomized MAC addresses have no influence in the kind of work described in this paper—our ability to deanonimize crowds depend on whether directed or broadcast probes are used, not on the linkabiliity of MAC addresses over time. Of course, it will be possible in the future to use simple crypto techniques to hide both the MAC address and the SSID of known networks still maintaing the energy efficiency of directed probes and small overhead. However, these techniques will probably need to change the protocol on the device as well as on the access points, and this is something that takes time to happen and most importantly to be fully deployed, especially on the infrastructure side. In the meantime, we believe that it is very interesting to see what are the consequences of a network protocol designed without privacy concerns in mind.

VII. Conclusions and Future Work

In our thinking, technology and ICT systems are just mere tools that make our life easier. What we are totally unaware of is that, by only observing the small pieces of information, accidentally leaked from some of these very pervasive systems, we can learn a lot on the people surrounding us. This is the case, for example, of WiFi directed probe requests. At

[6]Section "Location" of http://www.android.com/about/jelly–bean/.

first sight, the information they contain (MAC address of the sending device and SSID of the probed network) is purely related to their technological goal they were built for. But, if properly exploited, the potential of this information and the application scenarios it enables are immense.

The purpose of this paper is to give evidence of the huge possibilities that arise from collecting probe requests to discover human-related phenomena. We do so by showing how one can build up, starting from probes, a large amount of detailed knowledge on big crowds of people. In this work we exploit this knowledge to automatically de-anonymize (discover the hometown of) tens of thousands of people participating in gatherings of citywide, national (political meetings of two parties held around election days), and international (religion related) events, lasting just a few hours each. Most importantly, we show that the de-anonymization through probes can be accurate at the point to precisely match ground truth information: The outcome of our de-anonymization methodology on the two political meetings happened in the Italian Capital around Italy's 2013 election days succeeds in predicting the party which got most votes for all 5 Major Italian cities considered in this paper. To the best of our knowledge, this is the first time that wireless probes are being exploited to discover, with high accuracy, information of this kind on large crowds.

De-anonymization of large events is just one possible application of wireless probes. We believe that these tiny 802.11 packets can open up the way to a myriad of new opportunities. (Here we list a couple of them, that we will start pursuing in the recent future.) Advertising of commercial activities is one of the examples. Indeed, owners of businesses (e.g. grocery stores), can set-up an automatic system to understand, from the SSIDs of the probes collected recurrently in their store, the provenance (neighborhood) of the clients, their inclination, hobbies, ad so on. Accordingly, they can use this information to build up client profiles and start a more targeted advertising campaign in the areas they live. In addition, business owners can also decide to push more on advertising their activity in areas whose inhabitants do not frequent the store.

Another possible application of WiFi probe requests is the prediction and prevention of infection spreading at neighborhood/city/nation level. Indeed, one might imagine a system setup across e.g. a large city, including public places like Stations, Hospitals, etc., that recurrently logs WiFi probes and track the APs used by people. The system could then use this information to compute the possible rate of people infected with e.g. an influenza virus that is spreading in that period of time. Most importantly, it can collaborate with Hospital IoT systems to discover the areas of the city that are playing a major role in the spreading of the infection, and signal this information to appropriate authorities. Another possible feature of the system can be that of notifying people about the chances they have to get the virus by frequenting a certain location in the city, e.g., a bar they're about to enter.

De-anonymization of crowds, advertising, and prediction of infection spreading in large metropolitan areas are just a few examples of the application scenarios of WiFi probes. We believe that there are many more scenarios, with impressive impacts on our everyday life, yet to be discovered.

REFERENCES

[1] "Wireless lan medium access control (mac) and physical layer (phy) specifications," ANSI/IEEE Std 802.11, 1999 Edition.

[2] N. Cheng, P. Mohapatra, M. Cunche, M. A. Kaafar, R. Boreli, and S. Krishnamurthy, "Inferring user relationship from hidden information in wlans," in *Proc. of MILCOM*, 2012.

[3] M. Cunche, M. A. Kaafar, and R. Boreli, "I know who you will meet this evening! linking wireless devices using wi-fi probe requests," in *Proc. of IEEE WoWMom*, 2012.

[4] M. Cunche and M. A. Kaafar and R. Boreli, "Linking wireless devices using information contained in Wi-Fi probe requests," *Pervasive and Mobile Computing*, vol. 11, pp. 56 – 69, 2014.

[5] M. V. Barbera, A. Epasto, A. Mei, V. C. Perta, and J. Stefa, "Signals from the crowd: uncovering social relationships through smartphone probes," in *Proc. of ACM IMC*, 2013.

[6] C. C. Klaus, "WLAN Security FAQ," https://lwn.net/2001/1011/a/wlan-security.php3, 2001.

[7] M. Moser, "Hotspotter: Automatic wireless client penetration," http://www.wirelessdefence.org/Contents/hotspotter.htm, 2010.

[8] D. A. D. Zovi and S. A. Macaulay, "Attacking automatic wireless network selection," in *Proc. of IEEE SMC IAW*, 2005.

[9] J. Franklin, D. McCoy, P. Tabriz, V. Neagoe, J. V. Randwyk, and D. Sicker, "Passive data link layer 802.11 wireless device driver fingerprinting," in *Proc. 15th USENIX Security Symposium*, 2006.

[10] L. C. C. Desmond, C. C. Yuan, T. C. Pheng, and R. S. Lee, "Identifying unique devices through wireless fingerprinting," in *Proc. ACM WiSec*, 2008.

[11] L. Vaas, "Businesses are building shopper profiles based on sniffing phones' WiFi — Naked Security Blog," https://nakedsecurity.sophos.com/2014/01/16/businesses-are-building–shopper-profiles–based–on–sniffing–phones-wifi/, 2014.

[12] L. Vaas, "Nordstrom tracking customer movement via smartphones' WiFi sniffing, Naked Security Blog," https://nakedsecurity.sophos.com/2013/05/09/nordstrom-tracking-customer-smartphones–wifi–sniffing/, 2013.

[13] A. Musa and J. Eriksson, "Tracking unmodified smartphones using wi-fi monitors," in *Proc. of ACM SenSys*, 2012.

[14] P. Bahl and V. N. Padmanabhan, "Radar: An in-building rf-based user location and tracking system," in *Proc. of IEEE INFOCOM*, 2000.

[15] L. Schauer, M. Werner, and P. Marcus, "Estimating crowd densities and pedestrian flows using wi-fi and bluetooth," in *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, ser. MOBIQUITOUS '14, 2014.

[16] I. Rose and M. Welsh, "Mapping the urban wireless landscape with argos," in *Proc. of ACM SENSYS*, 2010.

[17] Y. S. Kim, Y. Tian, L. Nguyen, and P. Tague, "Lapwin: Location-aided probing for protecting user privacy in wi-fi networks," in *IEEE Conference on Communications and Network Security (CNS)*, Oct 2014.

[18] J. Freudiger, "How talkative is your mobile device?: An experimental study of wi-fi probe requests," in *Proceedings of the 8th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, ser. WiSec '15, 2015.

[19] IEEE, "MA–L PUBLIC LISTING. ISO/IEC 8802 standards," 2015.