Signals from the Crowd: Uncovering Social Relationships through Smartphone Probes

Marco V. Barbera, Alessandro Epasto, Alessandro Mei, Vasile C. Perta, and Julinda Stefa
Department of Computer Science, Sapienza University of Rome, Italy,
{barbera, epasto, mei, perta, stefa}@di.uniroma1.it.

ABSTRACT

The ever increasing ubiquitousness of WiFi access points, coupled with the diffusion of smartphones, suggest that Internet every time and everywhere will soon (if not already has) become a reality. Even in presence of 3G connectivity, our devices are built to switch automatically to WiFi networks so to improve user experience. Most of the times, this is achieved by recurrently broadcasting automatic connectivity requests (known as *Probe Requests*) to known access points (APs), like, e.g., "Home WiFi", "Campus WiFi", and so on. In a large gathering of people, the number of these probes can be very high. This scenario rises a natural question: "Can significant information on the social structure of a large crowd and on its socioeconomic status be inferred by looking at smartphone probes?".

In this work we give a positive answer to this question. We organized a 3-months long campaign, through which we collected around 11M probes sent by more than 160K different devices. During the campaign we targeted national and international events that attracted large crowds as well as other gatherings of people. Then, we present a simple and automatic methodology to build the underlying social graph of the smartphone users, starting from their probes. We do so for each of our target events, and find that they all feature social-network properties. In addition, we show that, by looking at the probes in an event, we can learn important sociological aspects of its participants—language, vendor adoption, and so on.

Categories and Subject Descriptors

C.2 [**Computer-communication networks**]: Network Architecture and Design—*Wireless communication*

Keywords

Smartphones; Wi-Fi probe requests; social networks.

1. MOTIVATION AND GOALS

WiFi access points (APs) are becoming increasingly ubiquitous in our homes, offices and public places. Initially, the APs were used to free our portable computers (laptops) from the ADSL/LAN cable. Nowadays, APs also represent a viable option for mobile devices to get fast and cheap connectivity. So, it is becoming more and more common for mobile devices to automatically switch to WiFi connectivity whenever possible. To facilitate this automatic process, current smartphone OSes store the list of the names (SSID) of the networks the user typically connects to. Periodically, many of our smartphones broadcast these SSIDs in the form of *Probe Request* to search for available networks [20, 11, 4]. This is done every few seconds, even when we are far from the WiFi access points we usually connect to. In a large crowd of people, a very high number of probe requests are sent every minute. In this paper we consider the following questions: "What information can we get on a large crowd from their probe requests?"; "Is it possible to infer important information on the crowd like its social structure or its socioeconomic status?"; and lastly: "If yes, can this analysis be done in a simple and automatic way?". To answer all these questions, we organized a campaign of probe collection: We targeted large gatherings of people at city-wide, national, and international events as well as a university campus. Our campaign lasted three months, and we managed to collect, using commodity hardware only, a total of 11,136,711 probes sent by 164,740 different devices.

Our main contribution and findings in this work are the following:

- We develop a simple and automated methodology that allows to extract, starting from our datasets, the existing social connections among the smartphone owners, and use it to uncover, for the first time, the underlying social network of the participants in each event;
- we analyze the properties of these social graphs and show that they all feature social-network attributes in many aspects such as diameter, clustering coefficient and degree distribution;
- we show that important information on the nature of large events can be learnt from the probes:
  - the distribution of the languages of SSIDs, as detected by our methodology, shows a clear relationship between the international nature of the event and the density of foreign participants;
  - the distribution of the smartphone vendors varies across the events and matches the expected socioecomomic characteristics of the participants.
- by using our datasets, we validate the well-known sociological theories of homophily and social influence in the context of smartphone vendor adoption;

IMC'13, October 23–25, 2013, Barcelona, Spain.
Copyright 2013 ACM 978-1-4503-1953-9/13/10 ...$15.00.
http://dx.doi.org/10.1145/2504730.2504742.

- we perform a temporal analysis of the data collected in our long-term university campus deployment that uncover strong correlation between the frequency of the co-occurrence of devices in the same time slot and the strength of the relationship inferred by our methodology;

An anonymized version of the dataset is available online [41]. The paper is organized as follows: Section 2 reviews the related work in this area. Section 3 introduces our data collection methodology and the events targeted by our analysis. Section 4 presents our main finding and contribution. Finally, Section 5 draws our conclusions and discusses possible future works.

2. RELATED WORKS

The use of probe requests as a way to discover nearby known WiFi networks has recently gained the attention of the community. In the last few years, the security community has focused on the potential perils of this technology. Zovi et al. [11], show how both Windows XP and OSX implementations of this feature exposes users to a man-in-the-middle attack, where a malicious access point (AP) pretends to be part of a trusted network (see also Klaus [20]). More recently, the same concern has been raised for mobile devices too [4]. In addition, probe requests have been shown to jeopardize users' privacy. Franklin et al. [15] show how the timings of passively collected probe requests can be exploited to fingerprint the wireless network interface driver. Loh et al. [12] expand on this idea by noting that not just the driver but also the network device and the OS can influence the timings between the probes. Finally, Pang et al. [28] show how, by combining information leaked by probe requests with other implicitly revealed identifiers, one can recognize users even when MAC addresses and names are replaced by pseudonyms.

Probe requests can also help WiFi monitoring and user tracking. Musa et al. [32] describe how to exploit probes to passively track smartphones and infer the trajectory of their users. Rose et al. [34] show how probes can be used to reveal past behavior of users. Cunche et al. [9] use probe requests to decide whether two devices potentially belong to socially linked users. They leverage the intersection between the SSIDs in the probe requests of the two devices and use a metric inspired by the Adamic-Adar [1] similarity to smooth out the influence of frequently used SSIDs. Later on, Cheng et al. [7] extended the previous analysis with spatial temporal information on probe requests on a small sample of users. These works focus on pairs of users, possibly recurrently observed over time, and aim at best characterizing their relationship.

In this work, we take a different approach: We focus on large scale events involving thousands of users. We capture the probe requests of the devices during the event, and, from that, we aim at building a snapshot of the society that the participants represent. A snapshot rich enough to reflect, as accurately as possible, the sociological features and peculiarities of the crowd in the event (e.g. language, wealth), and, simultaneously, the properties and features of the emerging social-network. Our datasets are collected at a large scale and target city-wide, campus-like, national, and international events.

Our methodology builds also on the work of [26, 25, 22, 40, 23], which analyze the social graphs induced by affiliation networks. To the best of our knowledge, we are the first to adapt and refine these techniques to large datasets of probes. This allows us to introduce an easily automated methodology through which it is possible to define and analyze large-scale social networks of mobile devices users. Finally, for the first time we perform language detection on the broadcast SSIDs, and exploit the vendor ID of the captured devices to validate the theory of homophily and social influence between smartphone users.

3. DATA COLLECTION METHODOLOGY

According to the 802.11 standard [37], a WiFi access point can announce its presence by broadcasting *Beacons*—frames containing network configuration parameters such as the service set identifier (SSID) and supported data rates. In particular, the SSID is a string that identifies the WiFi AP in a human readable format (e.g., "Home network", "Free WiFi"). Client devices—referred to as *Stations* by the standard—can use two methods to detect access points in range: *Passive* and *active* scanning. In the former, the client passively listens for beacons of nearby access points and uses them to decide which network to connect to. Conversely, with active scanning, it is the client that actively searches for available networks by sending *Probe Request* frames. The probe requests can be either *directed* to a specific network, by indicating its SSID, or *broadcast* to any network within range. An example of both types of probes is given in Table 1. Upon receiving a Probe Request, any AP belonging to the network the probe is directed to replies with a *Probe Response* enabling the client to initiate a connection.

Probe requests are an efficient way for energy-limited devices (like smartphones) to detect known and unknown WiFi networks within range. By sending active probes, a mobile device can keep the WiFi radio on for just a few milliseconds, the amount of time it takes for any response to be received. Directed probe requests are also useful for connecting to "hidden" networks whose APs do not broadcast its SSID. To further improve this mechanism, a *Preferred Network List* (from now on *PNL*) of networks a device has connected to in the past is maintained by the OS to transparently connect or switch between known networks whenever possible.

3.1 Probes collection: Technical details

Mobile devices periodically send probe requests with a frequency that is vendor specific but that we observed to be typically between 15 and 60 seconds, depending on the power state of the device. Since probe requests are used in the discovery phase that comes before the actual association to the access point, they are sent in the clear over all transmission channels in sequence. This makes intercepting (sniffing) probe requests an easy task, requiring just commodity hardware like the internal wireless card of a laptop set in monitor mode.

In our collection campaign we used the following hardware:

- 4 × MacbookPro equipped with a Broadcom BCM43xx card;
- 1 × ThinkPad X61 equipped with an Atheros network card;
- 1 × fixed external Ubiquity SuperRange Cardbus antenna[1].

Overall, we collected around 11M probes sent by around 160K unique devices. We collected data in more than eight different events and locations with large gatherings of people. After data collection we used the *tshark* network analyzer to filter out all corrupted probes with a bad checksum (field `FCS` in Table 1). Then, we built a database that associates each device, as identified by its MAC address (field `SA`), to the list of SSIDs (field `SSID`) derived from its probes.

3.1.1 Description of the datasets

As many metropolises, Rome hosts several important events with both national and international audience and is a recurrent venue of large political and religious gatherings. Some of these events became target of our study. In particular, in our data collection campaign we targeted the following scenarios: 1) Events of national

[1] `http://dl.ubnt.com/src_datasheet.pdf`

Frame Ctrl	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 1: Example of Probe Requests sent by device with MAC address 10:9a:42:42:42:42. One broadcast probe (on top), and two directed to the "Free WiFi" and the "Home WiFi" networks respectively.

audience; 2) events of international audience; 3) events with audience consisting of mostly local residents; 4) a train station; 5) a university campus; 6) other. The first four were one-shot events lasted from 40 minutes to 6 hours each. Data collection was physically carried out by a team of 5 researchers that joined the event equipped with their laptops. The fifth dataset was collected through a fixed hardware placed in a university campus over a continuous period of 6 weeks. The last dataset includes probes collected by the research group in other occasions (e.g. commuting from home to office). In the remaining of this section we describe in details each of the scenarios/events that we targeted. Details on the number of devices and number of probes of each datasets are given in Table 2.

3.1.2 National Events

In our collection campaign we targeted the political meetings of two very large parties in Italy: The Movimento Cinque Stelle (M5S), a recently-established progressive party that unexpectedly got 28% of the votes, and the conservative party Popolo della Libertà (PDL), one of the most important parties in the Italian political scene. The Movimento Cinque Stelle closed its electoral campaign in Rome, with a meeting held on February 22th 2013 in one of the largest square in Rome area ("Piazza San Giovanni"). This event is denoted in our dataset as *Politics 1*. The Popolo della Libertà called for a post-electoral meeting in Rome on March 23rd. The meeting was held in a famous square in the city center ("Piazza del Popolo"). This dataset is denoted as *Politics 2*. As the police reported, in both cases the event participants came from all over Italy.

3.1.3 International Events

On February 11th, 2013, Pope Benedict XVI announced his resignation, after 8 years of service as the Head of the Catholic Church. The day of his farewell Angelus[2], February 24th, Vatican City was literally invaded by pilgrims and tourists from all around the world, as police reported. The same happened on March 17th, when the newly elected Pope Francis delivered his first Angelus. The datasets relative to these two events are denoted as *Vatican 1* and *Vatican 2* respectively.

3.1.4 City-wide probes: The Mall

For this case study we aimed at collecting data from local residents of Rome. So, we targeted a location known for hosting many families and groups of friends all at once: One of the biggest malls in Rome (Porta di Roma). To make sure to get a relevant affluence of people, we chose to collect the data on a special afternoon, right before Easter (March 30th, 2013), when many Romans like to shop. The data collection lasted 3 hours and a half.

3.1.5 Train Station

In this case study we targeted Termini, the Rome central train station, and collected the probes for a total of 7 hours split in a time range of four days. The collection was carried out by positioning at vantage points in the station so as to maximize the area covered.

3.1.6 University

This dataset was collected by deploying an antenna in a fixed point located at one of the main entries of our university campus. Differently from the other datasets, in this case we collected probes continuously for a period of 6 weeks, collecting probes requests mostly transmitted by devices of students entering and leaving the campus.

3.1.7 Others

This dataset consists of probes collected individually by 3 members of our research group during several tours across the city performed in a time span of 4 weeks while, for example, commuting from home to work or shopping.

3.1.8 All

Our last dataset, denoted as *All*, is generated by merging all the previously described datasets together. We will refer to this dataset when giving global statistics on the aggregate data we collected.

4. DATA ANALYSIS

Our objective is to show how WiFi probe requests collected in big events or in a given area of a city reveal a number of insights on the sociological characteristics of the crowd in the event or on the target population. We do so by means of a methodology comprising social networks analysis techniques, sociological theories and natural language processing.

4.1 Vendors of the devices in the datasets

Today's market of mobile devices is very dynamic and lively. Periodically, the major vendors either launch newer versions of their flagship products or introduce a product line destined to a new segment of the market. We hereby investigate what type of influence this has on the data we collected. To do so, we grouped by vendor the over 160K devices we detected (see Table 2) and computed the percentage of devices of each vendor. The vendor of a device can be obtained by looking at the sequence of the first 3 bytes of the MAC address field of any of its WiFi probes and matching it against the IEEE Public OUI[3] list. This is the official database that lists the space of MAC addresses assigned to each vendor. The 6 most common vendors we found in our datasets are shown in Figure 1. The most common vendor in out datasets is Apple (57% of devices), followed by Samsung (17%), Nokia (6%), HTC (1%), Sony (1%) and RIM (1%). These results do not necessarily measure the market penetration of these brands—a number of factors may have influenced the distribution of the vendors like the range of the WiFi antennas, the probability that users leave the WiFi interface on, or the frequency with which devices scan for nearby networks. However, they are qualitatively similar to those obtained

[2] The Angelus is the Pope's speech and prayer delivered every Sunday at noon from his window overlooking Saint Peter's square.

[3] `http://standards.ieee.org/develop/regauth/oui/oui.txt`

Dataset	Devices	PNLs (%)		Total Probes	Directed Probes (%)		Broadcast Probes (%)		SSIDs
Politics 1 (P1)	16,695	4,677	(28.0%)	1,190,481	406,002	(34.1%)	784,479	(65.9%)	14,740
Politics 2 (P2)	12,619	4,144	(32.8%)	330,936	117,644	(35.5%)	213,292	(64.5%)	11,145
The Mall (M)	9,731	3,859	(39.6%)	820,806	394,184	(48.0%)	426,622	(52.0%)	10,451
Train Station (TS)	14,640	5,371	(36.6%)	393,143	218,234	(55.5%)	174,909	(44.5%)	17,295
University (U)	17,131	8,853	(51.6%)	5,349,894	2,803,104	(52.4%)	2,546,790	(47.6%)	14,751
Vatican 1 (V1)	23,430	7,631	(32.5%)	1,234,416	555,361	(45.0%)	679,055	(55.0%)	29,533
Vatican 2 (V2)	22,219	6,817	(30.6%)	507,945	208,028	(40.0%)	299,917	(60.0%)	23,345
Others	60,445	21,824	(36.1%)	1,309,090	642,526	(49.0%)	666,564	(51.0%)	42,105
All	164,740	59,684	(36.2%)	11,136,711	5,345,083	(48.0%)	5,791,628	(52.0%)	133,351

Table 2: Statistics on the probes captured in our target events. The column "PNLs" reports the number (and percentage) of devices that disclosed at least one entry of their Preferred Network List.

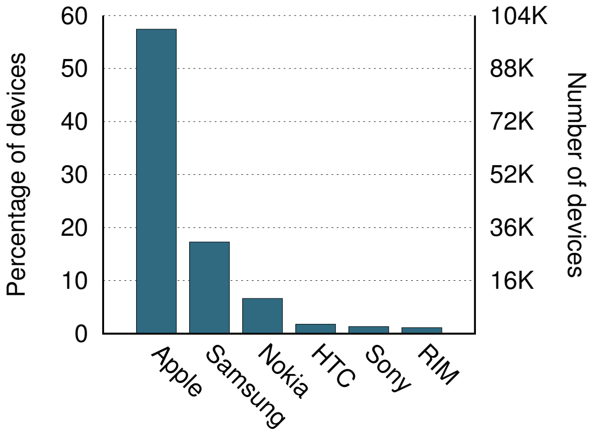


Figure 1: Vendors distribution (All datasets).

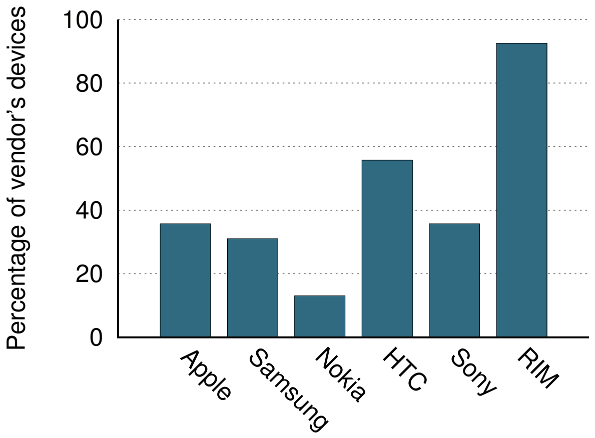


Figure 2: Vendor percentage of devices that exposed part or all the SSIDs in their Preferred Network List (dataset All).

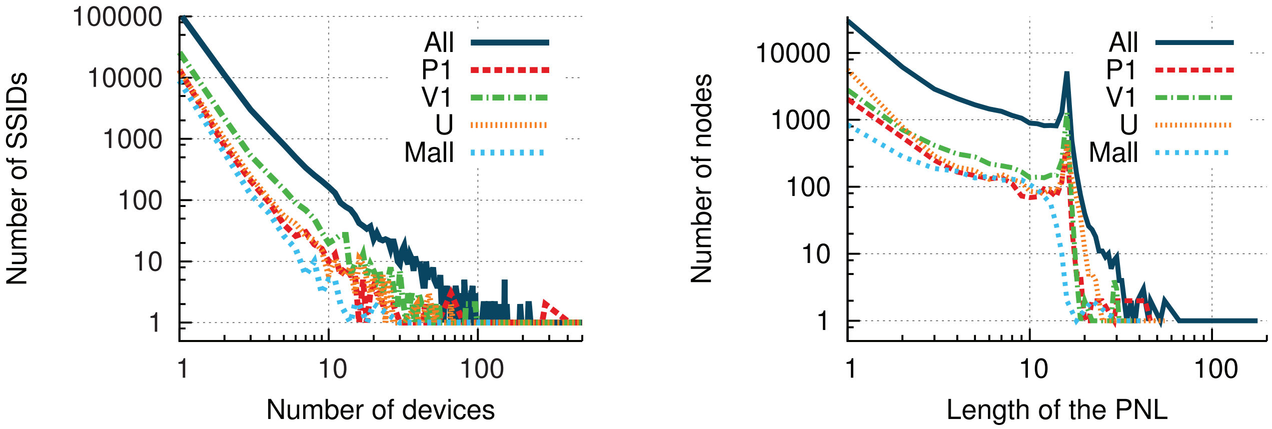
by Musa et. al. in [32], and we can observe that all vendors in our list are amongst the market leaders of mobile devices, such as smartphones and tablets. This suggests that our datasets indeed capture a representative sample of the most commonly used mobile devices.

A question that we raise is how the choice of a particular vendor affects the probability that all or part of the PNL is exposed by means of directed probe requests. This is relevant for our analysis, and may also be an important information from a security and privacy point of view. An adversary, by using just commodity hardware, might collect the SSIDs in the PNL of a user's device and perform an Evil Twin man-in-the-middle attack [4] against the user. Also, the names of the networks may reveal sensible information on the user, such as the place where she lives or works, the communities she belongs to, and the places she likes or frequently visits. The percentage of devices of each of the most common vendors in the datasets is reported in Figure 2: 92% of RIM devices disclosed part of their PNL, followed by HTC (55%), Sony (35%), Apple (35%), Samsung (31%) and Nokia devices (13%).

4.2 SSIDs Analysis

Today it is becoming more and more common to find WiFi access points not only in homes and workplaces, but also in public environments such as restaurants, hotels, and pubs. As a result, the probability that users connect to the WiFi networks of the places they visit is increasing too. Intuitively, this should produce two side effects: First, SSIDs of WiFi networks of public places that are popular among the people that participated to an event should be found in many of their devices. Second, a significant fraction of the PNLs of these devices should store more than one SSID—e.g. home WiFi, office WiFi, and so on. We investigate whether the datasets we collected confirm these two intuitions. To do so, we reconstructed the PNLs of the devices in our datasets by collecting all the directed probe requests we logged. The total number of devices with a non-empty PNL is around 59K. The results of our analysis for Politics 1 (P1), Vatican 1 (V1), University (U) and the Mall, as well as for all the datasets together (All), are shown in Figure 3. The other datasets show similar trends.

The distributions of the popularity of the SSIDs are reported in Figure 3(a) for a selection of the datasets, one for each type of event we targeted. As the figure shows, in our datasets the distributions are heavy-tailed, with very few highly popular access points coexisting with hundred of thousands of SSIDs stored in a handful of PNLs. It is also interesting to notice that this result is consistent across different datasets independently on the data collection methodology. In fact, the datasets collected in local events (P1, Mall), those collected in international events (V1), and the long term dataset (U) show the same type of distribution. As expected, many popular SSIDs are relative to the WiFi network of public places, such as airports, tourist attractions, university campuses and so on. Among them, there are also SSIDs of a number of city-wide WiFi networks, such as "Provincia WiFi", serving around



(a) SSIDs

(b) PNLs

Figure 3: Distribution of the popularity of the SSIDs (a) and of the sizes of the PNLs (b) relative to the All, Politics 1 (P1), Vatican 1 (V1), University (U), and The Mall (Mall) datasets.

300K users in Rome, and "Guglielmo", a WiFi network with more than 9K hotspots across Europe. Although we found SSIDs that are popular only because they have very generic names (e.g., "Hotspot" or "Dlink"), our results suggest that popular, public networks can indeed be detected by our data collection methodology.

The distributions of the lengths of the PNLs are shown in Figure 3(b). All the distributions present a peculiar shape: Up to about 16, the curves approximately follow a heavy-tailed distribution, then we observe a peak and a very steep drop. This behavior is easily explained by the fact that many vendors limit the number of different networks to which send directed probe requests to 16. For instance, we found that in the Android OS this limit is hard-coded as a constant with value 16 in the wireless driver source code[4]. These results, which are consistent across all the datasets, confirm the intuition that a significant fraction of users have more than just 1 SSID in the PNL of their device. More in detail, for the All dataset, 50% of the PNLs of all the devices store one SSID only, around 30% store between two and 10 SSIDs, and the remaining 20% store more than 10 SSIDs.

4.3 Uncovering the underlying social network

A contribution of our work is to show how WiFi probes can be used as a new lens to look at a crowd and uncover important information about it. One relevant information is the social structure of the set of people in the crowd.

We can regard the PNL of a device as a list of significant places visited by the user—significant enough that the user spent some time to connect to the access point. Therefore, the fact that two users share one or more SSIDs in the PNL of their devices should intuitively provide some information on the existence of a social relationship between the two. This intuition is supported by recent findings on the spatio-temporal properties of human behavior that have shown how social relationships can be correctly inferred between people sharing similar location trails [14, 10]. We investigate whether, by considering social links inferred from WiFI probes, we can uncover a *social network* that underlies the crowd that participated to the events we targeted. We do so by describing an automatic method for inferring social links between users starting from the PNLs of their mobile devices. We then apply this method to show how full-fledged social networks emerge from our datasets.

4.3.1 From affiliation networks to social networks

The SSIDs in the PNLs of a set of devices can be represented in the form of an *affiliation network* [23, 5]. An affiliation network, denoted as $G = (V_1, V_2, E)$, is a bipartite graph that connects a set V_1 of *actors* and a set V_2 of *groups* they belong to. In our case, V_1 is the set of devices that disclosed at least one entry of their PNL, V_2 is the set of the network SSIDs we collected, and an edge $(v_1, v_2) \in E$ represents v_1 having v_2 in its PNL. Statistics of the affiliation networks relative to each of our datasets are reported in Table 3.

Starting from an affiliation network of devices and SSIDs we can build a social network $\bar{G} = (V_1, \bar{E})$ between devices as follows: First, we define a *similarity measure* $f : V_1 \times V_1 \to \mathbb{R}$ that, given two devices u and v, yields the strength of the social relationship between their respective users. Then, we impose a minimum threshold τ and place an edge $\{u, v\} \in \bar{E}$ only if $f(u, v) > \tau$. A first possible choice of a similarity measure would be one based the size of the intersection between the PNLs of the two devices. In order for an edge to be placed, we would require at least $\tau = k$ common SSIDs. Although widely used in the literature [22, 23], this similarity measure would not work in our case as it gives the same importance to all SSIDs regardless of their popularity. In fact, in Section 4.2 we have seen how very popular SSIDs correspond to city-wide free networks, or to networks with very generic names (e.g., "Dlink") that, intuitively, should not produce strong social links. On the other hand, less popular SSIDs of home or small private networks clearly denote a potential strong relationship between users that connect to them. We therefore need a similarity measure that takes into account *both* the intersection of the PNLs *and* the popularity of the SSIDs. The similarity measure we found out to be the one best matching our needs is the Adamic-Adar [1] which penalizes SSIDs of popular networks in favour of those of networks shared by few people only (see Figure 4). More formally, the Adamic-Adar measure is defined as follows:

$$f_{ADA}(u, v) = \sum_{w \in N(u) \cap N(v)} \frac{1}{\log_2(|M(w)|)} \qquad (1)$$

where $N(u)$ is the PNL of the device u, and $M(w)$ is the set of devices with the SSID w in their PNLs. In other words, the Adamic-

[4] http://androidxref.com/4.1.1/xref/external/wpa_supplicant_8/src/drivers/driver.h

Dataset	$\lvert V_1\rvert$	$\lvert V_2\rvert$	$\lvert E\rvert$	d_1	d_2
Politics 1 (P1)	4,677	14,740	24,494	5.24	1.66
Politics 2 (P2)	4,144	11,145	17,722	4.28	1.59
The Mall (M)	3,859	10,451	19,374	5.02	1.85
Train Station (TS)	5,371	17,295	27,515	5.12	1.59
University (U)	8,853	14,751	32,608	3.68	2.21
Vatican 1 (V1)	7,631	29,533	48,498	6.36	1.64
Vatican 2 (V2)	6,817	23,345	37,149	5.45	1.59
Others	21,824	42,105	80,502	3.69	1.91
All	59,684	133,351	277,214	4.64	2.08

Table 3: Statistics on the affiliation networks of the probes detected in the various experimental settings. The values $|V_1|$ and $|V_2|$ refer, respectively, to the number of devices and SSIDs extracted from the directed probe requests. The value $|E|$ is the number of links between the devices and SSIDs. The average number of SSIDs each smartphone is linked to, and the average number of smartphones announcing an SSID are given in the columns d_1 and d_2 respectively.

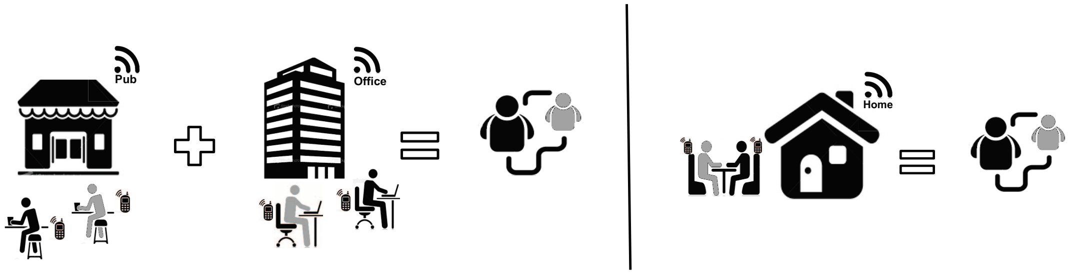


Figure 4: Adamic-Adar similarity: Several networks (e.g., pubs, workplaces) shared by a moderate or large amount of people are necessary for a relationship between two users to be correctly inferred. On the other hand, a single, private home network used by just a handful of persons may be enough.

Adar measure discounts the importance of an SSID by a factor that is logarithmic in the number of the devices that connected to it, which also well adapts to the heavy-tail distribution of the SSIDs popularities reported in Figure 3(a).

Among the other possible graph similarity measures found in the literature [21, 17, 1, 19, 27], those based on the Jaccard Coefficient [27] would not work because they do not reduce the importance of popular SSIDs. Measures based on random walks and nodes distance [27, 19] are instead not directly applicable in our case as affiliation networks are bipartite. Finally, measures based on textual similarity of SSIDs, determined by applying standard techniques in information retrieval [29], would not be suited in our context, as two nodes with exactly the same single (and very common) SSID would receive the maximum possible similarity score.

4.3.2 Topological properties of the social networks

We studied a number of structural properties of the social networks we extracted from our datasets and compared them with those of commonly studied online social networks [24]. We found that, consistently in all our datasets, Adamic-Adar with threshold values close to $\tau = 0.3$ generates social networks with structural properties that are similar to those of other well-known social networks. We therefore discuss the results obtained with this threshold. Note that in our discussion we ignore nodes without edges as they are irrelevant to our study.

Table 4 reports, for each of our social networks: The number $|V|$ and $|E|$ of nodes and edges in the network, the average node degree $\bar{d}$, the number of connected components *NC*, the size of the biggest connected component *BCC*, the diameter *D*, and the effective diameter D_{eff} of the *BCC* (the 90th percentile of the length of the shortest paths between nodes of the biggest connected component), the triadic closure t_c, and the clustering coefficient c_c. As we can observe, although our networks feature a large number of connected components (column *NC*), the biggest one (column *BCC*) always includes between 75.9% and 94.2% of all the nodes. The length of the longest shortest path (column *D*) and the 90th percentile of the shortest paths lengths (column D_{eff}) of the BCCs of our networks are close to those of popular online social networks. For comparison, those computed on a publicly available Facebook dataset are equal to 8 and 4.7 respectively [30]. We found structural similarities between our networks and popular online social networks also when we measured their density by means of the clustering coefficient (column c_c) and the triadic closure [39, 16] (column t_c). These are close to those computed in a publicly available Twitter dataset, which are 0.56 and 0.06 respectively [30]. Consistently with other social networks, in our networks too the distribution of the nodes degrees follows a power law [3]. These distributions are reported in Figure 5 for a group of datasets spanning the various types of events we targeted: Politics 1, Vatican 1, the Mall, and the long-term University dataset. Finally, for the same datasets, we report in Figure 6 the distributions of the connected components sizes. These show that most of the connected components, excluding the biggest component, contain between 1 and 10 nodes. The same property can be found in the other datasets too.

Overall, these results show that the Adamic-Adar metric allows to bring to light meaningful social structures from all our datasets. We show experimentally why the same cannot be said about a similarity measure based just on the size *k* of the intersection between

Dataset	$\|V\|$	$\|E\|$	$\bar{d}$	NC	BCC (%)	D	D_{eff}	t_c	c_c
Politics 1 (P1)	2,119	28,250	13.33	89	1,896 (89.4%)	9	3.88	0.144	0.491
Politics 2 (P2)	1,566	9,452	6.04	77	1,375 (87.8%)	10	4.57	0.154	0.505
The Mall (M)	1,742	33,835	19.42	74	1,533 (88.0%)	8	3.66	0.189	0.537
Train Station (TS)	2,397	16,045	6.69	90	2,164 (90.2%)	10	4.56	0.127	0.484
University (U)	2,448	96,498	39.42	59	2,307 (94.2%)	8	3.49	0.200	0.549
Vatican 1 (V1)	4,337	44,502	10.26	159	3,936 (90.7%)	9	4.47	0.145	0.453
Vatican 2 (V2)	3,423	32,239	9.42	172	3,003 (87.7%)	10	4.56	0.149	0.469
Others	8,770	134,188	15.30	798	6,662 (75.9%)	10	4.44	0.132	0.507
All	26,410	572,519	21.68	1,244	23,241 (88.0%)	11	4.70	0.132	0.460

Table 4: Structural properties of the social networks induced by using the Adamic-Adar measure with threshold $\tau = 0.3$.

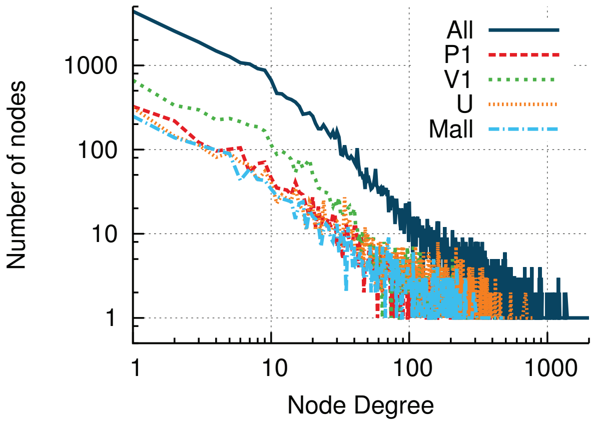


Figure 5: Degree distribution of the social networks induced by Adamic-Adar with threshold $\tau = 0.3$ from the All, Politics, Vatican 1, University and The Mall datasets.

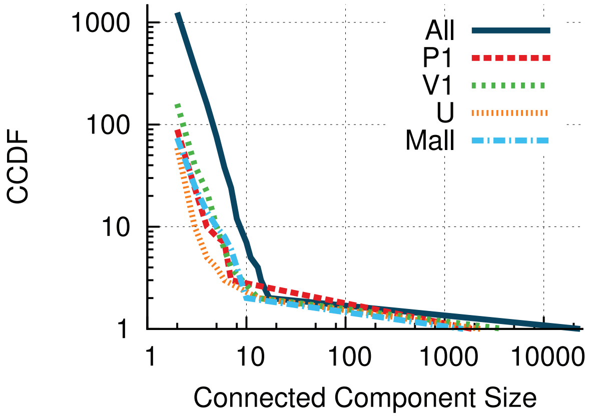


Figure 6: Distribution of the sizes of the connected components in the social networks induced by Adamic-Adar with threshold $\tau = 0.3$ from the All, Politics, Vatican 1, University and The Mall datasets.

two PNLs. To do so, we use a KNC-Plot, a tool for analysing the macroscopic properties of the graphs generated from our affiliation networks defined by Kumar et al. [22]. A KNC-Plot provides a visual indication on how the number of connected components and the size of the biggest connected component (BCC) change according to the minimum number k of common SSIDs required for an edge to be placed. Figure 7 shows a sample of the KNC-Plots of the Politics 1, Vatican 1, the Mall and University datasets. According to the figure, when the threshold k is equal to 1 almost all the nodes are connected. But increasing just slightly k produces a steep degradation in the connectivity structure of the graphs. For instance, with $k = 2$ the BCCs shrink to around 50% of the nodes, whereas with $k = 3$ they shrink to about the 25%. In other words, the similarity measure based on the size of the intersection of the PNLs generates an all-or-nothing effect that makes it hard to gain any insight on the social structure that underlies our datasets. A similar result was observed in a user-interest graph derived from Flickr [22].

4.4 Homophily and social influence in vendor adoption

Social networks based on WiFi probes may be a useful tool for studying the effects that physical proximity has on people in our society. Indeed, the way we derived our social networks leverages the intuition that users that connect to the same WiFi access point are likely to be socially connected to each other—the higher the Adamic-Adar measure that generates a link, the higher the chance that two users meet regularly or even live in the same place. According to a widely studied sociological theory known as *homophily* [31], physical proximity should cause interconnected users to be related in terms of, for instance, interests, social extraction, age, or gender. Starting from this observation, we evaluate whether our large scale data collection methodology may be used to experimentally confirm theories like that of homophily and social influence. We do so by measuring the homogeneity in device vendors adoption in groups of socially interconnected people. In fact, the choice a user makes of a particular vendor over another results from a number of factors, such as the user's wealth or age or social influence. We therefore expect people closely related to each other to tend to use devices of the same vendor.

As a measure of the homogeneity in device vendor adoption between socially connected users, we use the *assortativity* [33]. In our case, the assortativity quantifies the extent to which users of devices of a given vendor are likely to be connected to each other rather than with users of devices of different vendors. More formally, given a social network $G = (V, E)$ and a partition C of its nodes according to their respective vendors, the assortativity of G is defined as follows:

$$a(G) = \frac{\sum_{i \in C} e_{ii} - \sum_{i \in C} c_i^2}{1 - \sum_{i \in C} c_i^2}$$

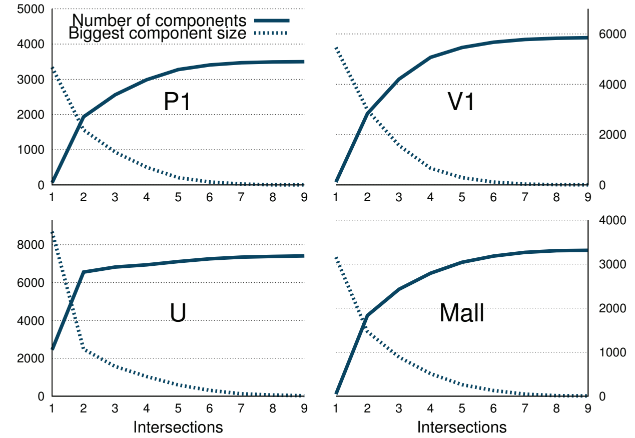


Figure 7: KNC-Plot for the Politics 1 (P1), Vatican 1 (V1), University (U), and The Mall datasets.

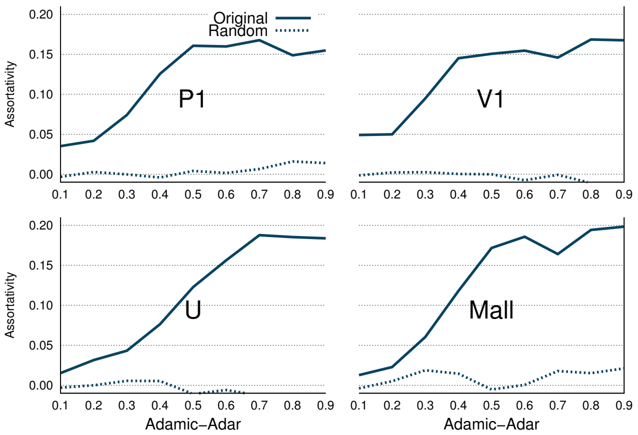


Figure 8: Vendor assortativity of the social network ("Original") vs. vendor assortativity of the randomized social network ("Random") for different Adamic Adar threshold τ values.

where c_i is the fraction of nodes that belong to vendor i and e_{ij} is the fraction of edges connecting nodes of vendor i to nodes of vendor j. The assortativity gets values in the $[-1,1]$ range. If most of the social relationships are between users who chose the same device vendor, then assortativity is positive, as $\sum_{i \in C} e_{ii}$ is close to 1. In the opposite case, that is, if most of the social relationships are between users who chose different device vendors, then assortativity is negative, as the values e_{ii} are close to 0. Finally, if social relationships are independent of the device vendor, then assortativity has values close to 0, as $e_{ii} \approx c_i c_i$.

For each dataset, we study how the assortativity of the corresponding social network varies according to Adamic-Adar threshold τ, so that we can verify whether mutual influence increases with the strength of the social links. We also perform a significance test to rule out the possibility that the assortativity of the graph depends only on the distribution of vendors popularity (shown in Figure 1). The test consists in generating, for each dataset, a randomized version of the corresponding affiliation network with the same in-degree and out-degree distribution as the original one. This is done by iteratively switching the endpoints of pairs of randomly selected edges until the affiliation network converges to a random bipartite graph. Then, we compare the assortativity of the social networks obtained from the original and randomized affiliation network with a given Adamic-Adar threshold.

The results of our measurements are reported in Figure 8 for a sample of our datasets. The other datasets present similar characteristics. As the figure shows, the vendor assortativity of our networks is not only always positive, but also significant, and the assortativity of the randomized network approximates zero. Second, as the Adamic-Adar threshold increases, the assortativity increases too, meaning that stronger social links are associated to stronger mutual influence.

4.5 Social Analysis

Collecting probe requests of a large number of mobile devices is an effective way to take a social snapshot of a crowd that participated to an event or that live in a certain area. So far, we have proved it by showing how it is possible to infer the social relationships between people in the crowd by leveraging the SSIDs in the PNLs of the devices. The social networks we extracted share the main properties of those emerging from other contexts, which confirms that our analysis methodology is sound. We now take a further step. We show that certain characteristics of the population that participated to the different events might be inferred from the probe requests sent by their devices. In particular, we focus on users' language and on vendors popularity.

4.5.1 Inferring user languages

We observed that a large fraction of the SSIDs stored in directed probe requests consist of natural language strings. In fact, as already discussed in Section 4.2, the SSIDs of many public WiFi networks reveal the name of touristic attractions (e.g., "Tour Eiffel"), hotels ("manhattan hotel"), bars ("Caffe Barberini"), and so on. On the other hand, we found that many broadband subscribers customized the SSID of their WiFi network in a number of different ways. For instance, some of them use their name. Some others use their SSID as a way to communicate something to their neighbors ("Please don't steal our WiFi"), or even to cheer for their favorite football club ("Forza Roma"). Intuitively, this should make it possible to get a hint on the language of the social context where a user lives by just looking at the SSIDs in the PNL of her device. The language would correspond either to the nationality of the user, if she lives in her country, or to the language of the country where she spends most of the time. Following this intuition, we defined an automatic and scalable user language identification procedure based on the name of the networks in the PNLs. When applied to one of our datasets, this technique helps deduce the national or international nature of the event, and the composition in terms of nationality of the crowd in the event.

Automatic language detection.

Inferring the language of an SSID is not always an easy task as SSIDs are very short (about 13 characters on average) and typically difficult to analyze due to the lack of white spaces or the use of special characters. This makes even the state-of-the-art methods for language identification of short texts [2, 6] unsuitable for our task. We therefore opted for a simple, ad-hoc methodology that turned out to be very effective. Given an SSID, we tokenize it after removing special characters and stop-words (including common words such as "WiFi" and "aDSL"). Each of the words is then

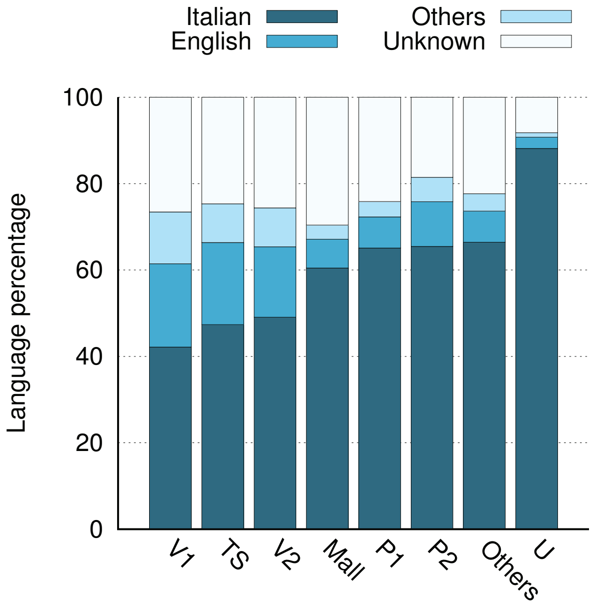


Figure 9: Distribution of languages in the different scenarios. Datasets are sorted by percentage of Italian devices.

searched in a large corpora[5] of texts in 5 languages (Italian, English, French and German) and assigned the language where it appears with highest frequency. The language of an SSID is then given by the language of the majority of its words. If no language is detected, or in case of ties, the language of the SSID is classified as unknown. Similarly, the language of a user is set to be that of the majority of the SSIDs in the PNL of her device, classifying it as unknown in case of ties. We also improved the accuracy of our automatic classification by complementing it in two ways. First, we manually annotated the language of the first 2K most popular SSIDs ($\sim 2\%$ of the total) in our All dataset. Due to the skewed distribution of SSIDs popularity (see Figure 3), these SSIDs appear in a very large fraction of PNLs ($\sim 75\%$), thus greatly improving the detection accuracy with a manual task that requires just a few minutes to be completed. Second, we associated to SSIDs containing the name of popular broadband providers the language of the country where the provider operates. For instance, "FASTWEB" is assigned to Italian, "Orange" to French, "Verizon" to English and so on. This is particularly useful to classify the language of those broadband WiFi networks that were left with their default SSID.

We checked the accuracy of our classification method over a randomized sample of 1000 devices manually annotated by a panel of three independent judges. On average, the judges found the percentage of correctly classified devices to be 92%. To measure the level of agreement between the judges, we used a standard NLP approach known as Free-Marginal k agreement [38]. The resulting agreement value of $k = 82\%$ validates the reliability of our manual review process.

Results.

Figure 9 reports the distribution of the languages detected in our datasets. For clarity, in the figure we explicitly show the Italian and English languages only, aggregating all the remaining ones as "Others". Our results show a strong correlation between the percentage of Italian devices and the international nature of the events. More in detail, Vatican 1 (V1) and Vatican 2 (V2)—described as international events in Section 3.1.1—are amongst the datasets with the lowest percentage of Italian devices. The Train Station (TS) dataset too shows characteristics typical of an international event. This is a direct consequence of the fact that the central train station in Rome is located in one of the most international areas of the city. The train station also happens to be the main hub connecting the city to its airports, which makes it an almost forced stop of any tourist that visits Rome. The Mall, Politics 1 (P1), Politics 2 (P2), and Other datasets show, instead, a high (i.e., $> 50\%$) percentage of Italian devices, confirming the fact that these events mostly consisted of an Italian crowd. Nevertheless, the percentage of Italian devices that were detected in these datasets is still significantly lower than that of the University (U) dataset. This is because the surroundings of the university entrance where we positioned our fixed antenna (see Section 3.1.1) were almost exclusively frequented by students. As in our university there are very few courses that are taught in English, these students are either Italian, or foreigners speaking Italian and living in an Italian social context.

Assortativity of SSIDs' languages.

Our SSID language detection method can also be used to verify the intuition that people tend to connect to networks of the same nationality. To do it, we first build the graph G_{SSID} of WiFi networks that share common users. More specifically, we add an edge between two networks if there is at least one device that connected to both of them. Then, we compute the language assortativity of G_{SSID} with the same technique described in Section 4.4 but, this time, by partitioning the nodes of the graph according to their language. The result confirms our intuition as we found a positive and significant language assortativity of 0.20. By comparison, the language assortativity of the randomized graph of SSIDs is -0.01, showing that the result does not depend on the distribution of language popularity.

4.5.2 Demographics of brand penetration

We now focus on device vendors. Our objective is to understand what the distribution of vendors in a dataset reveals about the categories and the socioeconomic status of people that participated to the corresponding event. In Figure 10 we show the distributions of the device vendors of each of our datasets. Notice how the distribution of vendors varies in a less marked way from event to event with respect to that of users languages. This is expected, as language characterizes people in a much stronger way than the choice of a device vendor. However, note that the vendor distribution is computed over the total number of devices in each dataset, which can be up to 4 times higher than the number of devices we could infer the language of (compare column "Devices" with column "PNLs" on Table 2). Therefore, small variations in the distributions can be considered significant in this case. That said, there are two trends that emerge from our results. First, the Vatican 1, Vatican 2 and Train Station datasets all feature a very similar vendor distribution, which is characterized by a high ($\sim 62\%$) percentage of Apple devices and a low ($\sim 15\%$) percentage of Samsung devices with respect to most of the other datasets. Interestingly, these datasets correspond also to the events where we observed a significant presence of tourists. Based on the observation that Apple devices are typically more expensive that the others, these results can suggest that foreigners visiting Rome may represent a sample of people that are, on average, wealthier than those that participated to the local

[5] Available at `http://wacky.sslmit.unibo.it/doku.php?id=corpora`

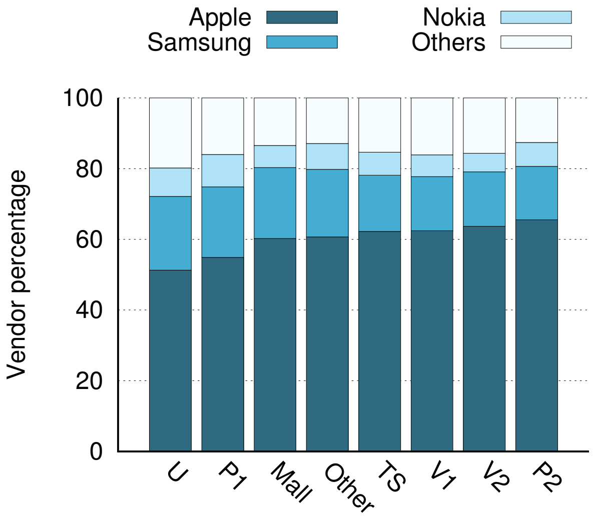


Figure 10: Distribution of vendors in the different scenarios. Datasets are sorted by percentage of Apple devices.

events. This is reasonable, since a trip to a foreign country, especially in a historical city like Rome, can be rather expensive. Local events, on the opposite, have the potential to attract an audience which is wider both in terms of social extraction and economic status. Second, we noticed that the differences in vendor popularity between local events is correlated with the difference we observed in the average age and social status of the people that participated. Again, this may be caused by Apple devices being more expensive. More in detail, we can notice a steep drop in the percentage of Apple devices when comparing the Politics 2 ($\sim$ 65%) to the Politics 1 ($\sim$ 54%) dataset, and the Mall ($\sim$ 60%) to the University ($\sim$ 51%) dataset. In fact, the people that participated to the Politics 2 event organized by the conservative party were, on average, older and wealthier than those that participated to the Politics 1 event. Also, the Mall is frequented by a wider range of people with respect to the University, which is mainly frequented by students who are typically on a budget. Finally, notice how the Others ($\sim$ 60% of Apple devices) dataset shows a distribution very similar to that of the Mall dataset. Both datasets, in fact, represent a rather uniform sample of population. Overall, these results suggest that difference in vendor distribution between two events might reveal a significant difference in the people that participated to them. The most compelling evidence of this fact is that the difference in age and socioeconomic status we observed between supporters of Politics 1 and Politics 2 parties is well represented by their difference in terms of popularity of Apple devices.

4.6 Temporal analysis on the University dataset

As opposite to the other datasets, which have been collected during one-shot events, the University dataset is the result of a 6 weeks long observation period performed from a fixed vantage point at the campus entrance. In this section we show how this long-term dataset allows to characterize the social dynamics of an observed area and to get further insights on its target population. Also, we study the correlation between co-occurrence of people and the strength of their social relationship as inferred by using the Adamic-Adar metric. Our studies are related to those regarding the characteristics of human social behavior [18, 35].

4.6.1 Recurrent patterns

Figure 11(a) reports the number of new (solid column) and known (dashed column) devices that were detected in each day of observation. Figure 11(b) reports, instead, the number of detected devices in a sample day. Our observations are consistent with the intuition that students' life is very predictable. Indeed, according to Figure 11(a), the number of detected devices abruptly decreases in correspondence of both week-ends and days when courses are suspended due to seasonal vacations (e.g., Week 1). Plus, consistently across all working days, the number of detected devices has a peak around lunch time (Figure 11(b)) and decreases in the evening. Also the repetitive schedule typical of students' life is captured in this dataset, as it takes only a few days for the number of newly detected devices to drop down to 30% of the total (Figure 11(b)). This suggests that the probability that a student does not visit the campus for more than a couple of working days in a row is small. In fact, most of them visit the campus with a very high frequency, which is confirmed by Figure 12 where we plot the distribution of the number of times the devices were detected in different time slots of 1 hour. According to the distribution, 40% of the devices were detected more than 100 times ($\sim$ 2 times per day on average).

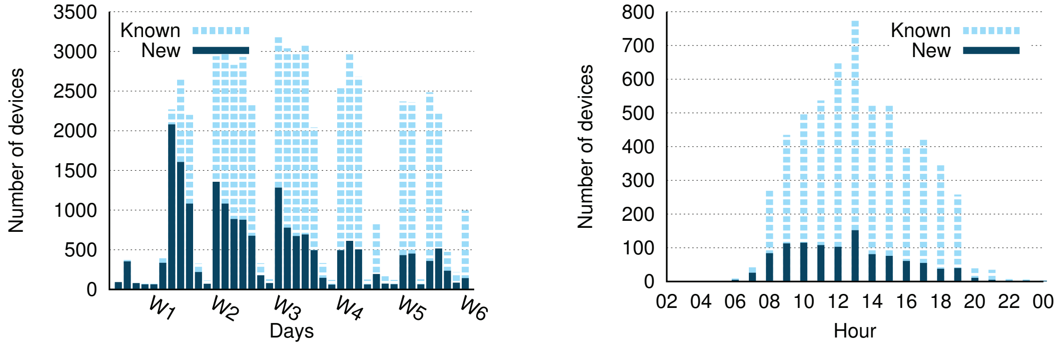
Overall, these results show how a long-term collection of probes requests adds a further dimension to the characterisation of a target population. In our case, we found a number of properties that match the typical students' movements patterns. This complements the other insights we inferred on students' language and socioeconomic status.

4.6.2 Smartphone co-occurrences

As observed in other contexts [8, 13, 36], the simultaneous presence of pairs of people in a given place denotes the existence of a possible social relationship between them. At the same time, the Adamic-Adar similarity we defined in Section 4.3 is based on a similar intuition, as it assigns a higher strength to social links connecting people that are more likely to frequent the same places. Intuitively, there should be a correlation between these two ways of inferring social links. The long-term University dataset allows to confirm this by verifying whether increasing values of the Adamic-Adar similarity between two devices correspond to a higher probability of their simultaneous occurrence. To do so, we first divide the time in slots of 120 seconds, which are sufficiently large to allow most of the devices to transmit their probes. Then, we group pairs of devices together in buckets according to their Adamic-Adar similarity. Finally, we average the number of co-occurrences observed for the pairs in each bucket. The results, shown in Figure 13, confirm our hypothesis, as we can observe a positive correlation between the Adamic-Adar similarity and the average number of co-occurrences (Pearson coefficient 0.858. 1-tailed p-value < 0.005).

5. CONCLUSIONS AND FUTURE WORK

Probe requests, both broadcast and directed ones, are a useful tool that enables energy-limited mobile devices to efficiently and transparently discover available access points and switch between them. The goal of this work was to show that, besides from their basic utility, smartphone probes actually bring insightful and interesting information about a crowd at an event or about a target population. To fully investigate this idea, we organized a three months long campaign of probe collection. During our campaign we targeted scenarios differing in both terms of scale and characteristics of the population involved: From campus and city-wide, to national



(a) Daily number of devices discovered. W1, W2, W3, W4, W5, and W6 mark the beginning of each week. Irregularities in W1, W4, and W5 correspond to classes breaks due to national holidays.

(b) Hourly number of devices discovered in a sample day.

Figure 11: Number of devices discovered in the University dataset for the whole collection period (a) and for a sample day (b). 'New' refers to the number of devices seen for the first time, 'Known' to the other ones.

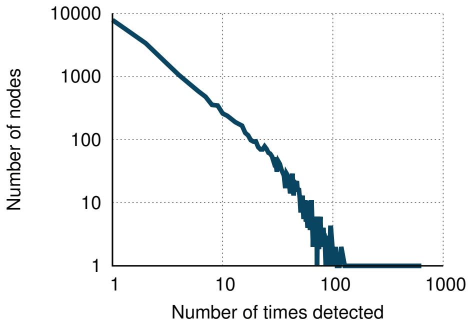


Figure 12: Distribution of the number of times the devices have been detected in the University dataset.

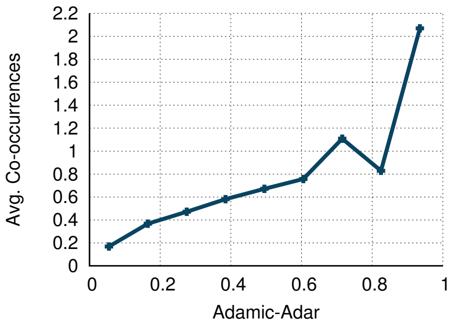


Figure 13: Expected number of co-occurrence per couple of nodes vs. their Adamic-Adar similarity in the University dataset.

and international scenarios. As a result we collected more than 11M probes sent by more than 160K different smartphone devices overall.

In this paper, we presented an in-depth study of these novel and large-scale datasets. The most important findings of our study are the following: First, we developed an automated methodology through which to derive the underlying relationship graphs between the users in each scenario, and showed how all these graphs feature properties that are typical of social networks—user and SSID degrees distributed as a power law, short diameter, high clustering coefficient and so on. Then, we studied how, in our social networks, groups of interconnected people tend to choose the same device vendor with a probability that increases with the strength of their social relationship. These results support the theory of homophily and social influence between people living in geographical proximity. We also performed, for the first time, language detection on the broadcast SSIDs, and exploited the vendor ID to show how the probes can directly reflect the sociological aspects of the people involved in each scenario like nationality, age, and socioeconomic status.

We believe this is just a first step towards a new, non invasive methodology for uncovering non-online social networks. As a future work, we plan to include information regarding the location of the networks referred to by the probes. This could be done, for instance, by using crowd sourced databases of 802.11 networks observed around the world, like Wigle[6]. This extra information will help us to achieve an even stronger characterization of the sample of the population contained in our datasets.

Acknowledgments

This work is partly supported by the Italian Government under the PRIN program (Progetti di rilevante interesse nazionale). We thank Sokol Kosta for his contribution to the data collection campaign.

6. REFERENCES

[1] L. A. Adamic and E. Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003.

[2] T. Baldwin and M. Lui. Language identification: The long and the short of the matter. In *Human Language Technologies*. Association for Computational Linguistics, 2010.

[3] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[4] J. Bhardwaj. Naked security blog: What is your phone saying behind your back?

[6] `https://wigle.net/`

[5] R. L. Breiger. The duality of persons and groups. *Social forces*, 53(2):181–190, 1974.

[6] S. Carter, W. Weerkamp, and M. Tsagkias. Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation*, pages 1–21, 2013.

[7] N. Cheng, P. Mohapatra, M. Cunche, M. A. Kaafar, R. Boreli, and S. Krishnamurthy. Inferring user relationship from hidden information in wlans. In *Military Communications Conference*. IEEE, 2012.

[8] D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, 107(52):22436–22441, 2010.

[9] 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 *Proceedings of the International Symposium on World of Wireless, Mobile and Multimedia Networks (WoWMoM)*. IEEE, 2012.

[10] J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh. Bridging the gap between physical location and online social networks. In *Proceedings of the 12th International Conference on Ubiquitous Computing*. ACM, 2010.

[11] D. A. Dai Zovi and S. A. Macaulay. Attacking automatic wireless network selection. In *Proceedings from the Sixth Annual SMC*. IEEE, 2005.

[12] L. C. C. Desmond, C. C. Yuan, T. C. Pheng, and R. S. Lee. Identifying unique devices through wireless fingerprinting. In *Proceedings of the first conference on Wireless Network Security*. ACM, 2008.

[13] N. Eagle and A. Pentland. Reality mining: sensing complex social systems. *Personal and ubiquitous computing*, 10(4):255–268, 2006.

[14] N. Eagle and A. S. Pentland. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–1066, 2009.

[15] 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 *Proceedings of the 15th USENIX Security Symposium*, 2006.

[16] M. S. Granovetter. The strength of weak ties. *American journal of sociology*, pages 1360–1380, 1973.

[17] M. Heymans and A. K. Singh. Deriving phylogenetic trees from the similarity analysis of metabolic pathways. *Bioinformatics*, 19(suppl 1):i138–i146, 2003.

[18] T. Karagiannis, J.-Y. Le Boudec, and M. Vojnovic. Power law and exponential decay of intercontact times between mobile devices. *Transactions on Mobile Computing*, 9(10):1377–1390, 2010.

[19] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[20] C. Klaus. Wireless 802.11b security faq. `https://lwn.net/2001/1011/a/wlan-security.php3`.

[21] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[22] R. Kumar, A. Tomkins, and E. Vee. Connectivity structure of bipartite graphs via the knc-plot. In *Proceedings of the international conference on Web search and web data mining*. ACM, 2008.

[23] S. Lattanzi and D. Sivakumar. Affiliation networks. In *Proceedings of the 41st annual symposium on Theory of computing*. ACM, 2009.

[24] J. Leskovec. Stanford large network dataset collection. `http://snap.stanford.edu/data/`.

[25] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, 1(1):5, 2007.

[26] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2, 2007.

[27] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.

[28] J. Pang, B. Greenstein, R. Gummadi, S. Seshan, and D. Wetherall. 802.11 user fingerprinting. In *Proceedings of the 13th annual international conference on Mobile computing and networking*. ACM, 2007.

[29] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge, 2008.

[30] J. McAuley and J. Leskovec. Learning to discover social circles in ego networks. In *Advances in Neural Information Processing Systems 25*, pages 548–556, 2012.

[31] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.

[32] A. Musa and J. Eriksson. Tracking unmodified smartphones using wi-fi monitors. In *Proceedings of the 10th Conference on Embedded Network Sensor Systems*. ACM, 2012.

[33] M. E. Newman. Mixing patterns in networks. *Physical Review E*, 67(2):026126, 2003.

[34] I. Rose and M. Welsh. Mapping the urban wireless landscape with argos. In *Proceedings of the 8th Conference on Embedded Networked Sensor Systems*. ACM, 2010.

[35] M. Salathé, M. Kazandjieva, J. W. Lee, P. Levis, M. W. Feldman, and J. H. Jones. A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences*, 107(51):22020–22025, 2010.

[36] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo. Socio-spatial properties of online location-based social networks. *ICWSM*, 11:329–336, 2011.

[37] I. C. Society. Ieee standardd 802.11, 2012.

[38] M. J. Warrens. Inequalities between multi-rater kappas. *Advances in data analysis and classification*, 4(4):271–286, 2010.

[39] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *nature*, 393(6684):440–442, 1998.

[40] E. Zheleva, H. Sharara, and L. Getoor. Co-evolution of social and affiliation networks. In *Proceedings of the 15th international conference on Knowledge discovery and data mining*. ACM, 2009.

[41] M. V. Barbera, A. Epasto, A. Mei, S. Kosta, V. C. Perta, and J. Stefa. CRAWDAD data set sapienza/probe-requests, 2013.