To be published in the proceedings of the 12th International Conference on Privacy, Security and Trust (PST-2014)

Inferring Social Relationships from Technology-Level Device Connections

Jason R. C. Nurse†, Jess Pumphrey§, Thomas Gibson-Robinson†, Michael Goldsmith†, Sadie Creese†
† Cyber Security Centre, Department of Computer Science, University of Oxford, UK
†{*firstname.lastname*}*@cs.ox.ac.uk,* §*jmp@jmpumphrey.com*

***Abstract*—Technology is present in every area of our lives and, for many, life without it has become unthinkable. As a consequence of this dependence and the extent to which technology devices (computers, tablets and smartphones) are being used for work and social activities, a clear coupling between devices and their owners can now be observed. By coupling, we specifically refer to the fact that information present on a person's device, be it user-generated or created by the native OS, can produce great insight into their life. In this paper, we look to exploit this coupling to investigate whether connections between technology devices recorded in system log-files, can be used to make inferences about the social relationships between device owners. A key motivation here is to better understand and elucidate the privacy risks associated with the digital footprints that we as humans (often inadvertently) create. Our work draws upon Social Network Analysis and basic Computer Forensics to develop and achieve the inference goals. From our preliminary experimentation, we demonstrate that human social relationships can indeed be inferred even within our limited initial scope. To further investigate the level of privacy exposure from technology-level links, we outline a more comprehensive plan of experimentation that will be conducted in future work.**

***Keywords*-Privacy, technology coupling, social network analysis, forensics, device and systems logs, network visualisation**

I. Introduction

We live in a world built on technology. At work, we can see it in everything from the traditional computer workstations we use, through to the new smartcards adopted to streamline identification and authentication. In the social arena, technology has become essential as well, especially due to the ease of communication that it facilitates with family and friends across the globe. Recent surveys further evidence the proliferation of technology devices, as they estimate that in some cultures around 11 hours per day are spent using digital media of some sort [1]. There are two factors that have undoubtedly contributed to this growth. The first factor is the Internet and the worldwide connectivity it enables, and the second factor is the mobility of 'smart' devices such as phones and tablets, which allow online and offline use practically anywhere.

As these technology devices are increasingly being used to support and enrich people's lives, a very close coupling between devices and their owners has been emerging. Coupling here, refers specifically to the fact that information present on a person's device, be it user-generated or created by the native Operating System (OS), can produce great insight into their life. Examples of information generally collectable from technology metadata includes WiFi networks previously connected to, places visited, and languages device owners use [2–4]. In some ways, one might regard these devices as analogous to passive monitors or loggers that capture and store a rich set of metadata about their owners. This metadata can span: the owner's system preferences; applications and services that are accessed; when and how long devices are in use; current and historic device-location data; and information about other devices that were communicated with.

The aim and novelty of this paper therefore is to consider owner-device coupling in more detail, with special focus on investigating whether the metadata generated by devices can be used to make inferences about the social relationships between their owners. By social relationships, we refer to offline associations between individuals, and the features of those associations including strength, and any temporal and spacial constraints. This work expands on research into individual's digital footprints (e.g., [2,5,6]) to look at connections across devices and what they can reveal about social relationships. A key objective for us is to better understand the range of privacy risks accompanying the increasing use of technology.

To achieve our goals, we first sought insight from the Computer Forensics domain (and articles such as Ref. [7,8]) to determine exactly what metadata might be general available on devices and how easy it would be to gather. Our assessment led us to the consideration of system log-files (e.g., [9,10]) as these provided an ideal balance between content richness and availability of data at this stage of our research. Having identified logs as an informative source of data, we then analysed their parameters to determine exactly what log entries might be most useful to our task of inferring social relationships between device owners. Once this data was identified, we reflected on the use of Social Network Analysis (SNA) [11] to elucidate relationships between a set of devices that had this data available. SNA can be described as a set of approaches that allow the study of links between elements (e.g., people, devices, or things). The idea, therefore, was that if we could gather meaningful log data from a number of devices, then we might be able to use SNA techniques to quickly spot technology-level connections and links, which may allow further inferences to be made about the relationships between the owners of those devices.

The remainder of this paper is structured as follows: Section II reviews the literature related to our research with

§ Pumphrey was at the University of Oxford during this research.

special focus on the inferences pertaining to identity and social relationships. In Section III, we present a brief history of SNA, discuss what useful information might be discoverable from technology-device logs, and then outline our proposed approach to using such logs to make inferences. Section IV introduces the preliminary experiment that was conducted including the main hypotheses tested, while Section V presents our analysis and the current results. We then conclude the paper in Section VI, and outline the more comprehensive experiment that will be the basis of our immediate future work.

II. Related work

There is an enormous amount of data generated by individuals in modern-day society. This can be seen in everything from email traffic and instant messaging, to online social networks and dedicated lifeloggers (e.g., Saga [12]). In addition to this intentionally created data, unknown to many there is also a significant quantity of data and metadata being created on our behalf simply by the use of technologies. Recent articles have discussed this in depth while highlighting the many benefits (e.g., in personalisation and knowledge discovery) and privacy concerns (e.g., misuse of metadata and unfair targeting) surrounding this ever-increasing pool of data [13–15].

To give an example, consider the case of a photograph taken with smartphone. To the layman, this captures an image of a scene and when it was shot, but to the trained eye, a photograph file also stores camera make and model, device type (e.g., iPhone), camera settings, location where the photo was taken (if geo-tagging was enabled), and even the saved name and address of the camera owner (this is unlikely to be the case for smartphones, but is more probable with high-spec cameras) [16]. This metadata is automatically added by devices, often without the owner's knowledge. Assuming the individual were then to post the photograph online, they would be sharing potentially sensitive identity data and would be oblivious to it; we do note, however, that some sites (e.g., Facebook) strip metatadata to reduce file size. This highlights some of the many privacy risks associated with the use of these devices and the unchecked publishing of information; further details are available in several articles (e.g., [17, 18]).

The topics of (systems) fingerprinting and user profiling are also of interest to our research. In fingerprinting, the general aim is to use certain data or properties of a system or device to allow it to be uniquely identified again in the future. The application of this technique to the Web, in particular, has been heavily deliberated as some institutions attempt to use it to track individuals (e.g., for targeted marketing campaigns), and privacy advocates aim to the contrary and to preserve some sense of anonymity online [19]. Given its success on the Web (typically via browser fingerprinting), there have been several other areas where fingerprinting has been applied, including identifying smartphone devices [20] and operating systems (OSs) [5]. The relevance of these contributions is that they highlight the fact that through inadvertently created data, much about devices and user identities can be inferred, thereby having a direct impact on a user's privacy.

User profiling is another approach that seeks to draw on the vast pool of data created by individuals to create purposeful profiles. In the literature [21–24], profiling can be witnessed on Web users, network users, and even criminals, all based on their generated metadata (e.g., browsing behaviour, typing patterns, and network and computer usage).

Possibly one of the most concerning privacy-related inferences using technology-device metadata is that proposed in Cunche *et al.* [2] and expanded on later in Cheng *et al.* [25] and Barbera *et al.* [3]. These research articles exploit the fact that wireless-enabled devices, such as smartphones, often use active WiFi probe requests in their search for familiar networks. That is, once a device's WiFi is turned on, it may occasionally broadcast a list of all of the WiFi points that it has previously been connected to, to check if any of them are available. If they are available, a connection is then made.

As found in Cunche *et al.*, as a result of these active probes, many devices are broadcasting data (e.g., MAC addresses, WiFi names) that could be used to fingerprint them, thus placing their owner's privacy at risk [2]. To enable the inference, Cunche *et al.* propose similarity between fingerprints as a metric, and thus, that devices with similar fingerprints are likely to be linked, as are their owners. Their more recent research [25] has even sought to use the data gathered from WiFi probes and physical location data associated with the access points (from sites such as `wigle.net`) to make more accurate inferences. Other work has adopted this general technique and further applied it to infer social relationships between crowds of people, as well as sociological characteristics such as language and smartphone vendor adoption [3]. In summary, these articles serve to illustrate the value of technology-level data and how it can be used to infer a variety of private information pertaining to an individual.

III. Applying SNA to Device Logs

A. Background and use of SNA

Social Network Analysis (SNA) is the study of social links between elements (e.g., people, devices, or things), a field which has been approached from many different angles and for various purposes over the last century. Freeman describes studies carried out in the 1930s by Moreno, Jennings, Warner and others which investigated the social networks which form in settings such as schools, prisons and workplaces, citing these as the origins of the field [26]. A famous example of SNA is the 1969 Small World Experiment that has established what is now commonly known as the "six degrees of separation" phenomenon [27].

Today, SNA continues to be used in a wide variety of applications, both online and offline. These span healthcare [28], identification of enterprise experts [29], social good [30] and even, law enforcement [31]. Typically, SNA is used to create comprehensive network graphs that can then be mathematically or visually assessed to identify influential figures in the network, significant links between nodes (individuals), and noteworthy clusters or groups in the network.

B. Discovering information from device logs

As the use of technology grows, so does the digital footprint that we, as individuals, create. This introduces us, and those we interact with, to a range of new privacy risks. To better understand the breadth of these risks, this paper proposes and investigates an approach where a model of a social network between individuals, is derived from reports of connections to other devices found in the system log-files of their devices. Native logs, a key tool in Computer Forensics, present a novel and largely untapped source through which we posit that data for SNA purposes can be attained. Of course, we do appreciate that these logs are much more difficult to gather than an individual's social-media presence or server-side HTTP logs for instance, but nonetheless, they do constitute an increasingly prevalent part of our digital footprints.

There are two factors which make these logs of particular interest. Firstly, they are present on practically all technology devices, especially those that tend to be in use by and coupled closely with individuals, such as computers, mobile phones and tablets [9, 10, 32, 33]. Secondly, they record a vast range of actions and events that occur on a system as a part of their role as a diagnostic tool for both hardware and software activities. As it pertains to our research, the aspect of most relevance is the record of connections with external devices, such as other computers, USB devices, Bluetooth devices or broader networks (e.g., WiFi or Local Area Network (LAN)).

Take as an example these lines from a Ubuntu-style *syslog*:

```
Mar 12 09:40:06 FITH NetworkManager[744]: <info> WiFi
  now disabled by radio killswitch
Mar 12 09:40:06 FITH kernel: [495.672113] usb 7-2: >USB
  disconnect, device number 2
```

The first point of note is that each line begins with a timestamp, followed by the computer name (`FITH`) and process which generated the line, an identification number, and a message. In this example, two processes – `NetworkManager` and the `kernel` – are reacting to the user deactivating the radio transmitter and receiver. NetworkManager is noting the loss of WiFi, and the kernel is noting the loss of the radio device itself. This exchange is evidence of the computer communicating with another device, albeit a built-in one.

As mentioned above, connections to external devices are also logged. Consider the example below taken from a different system, but one that has the same Ubuntu/Debian-style *syslog* file as well:

```
Mar 13 12:33:42 OXON kernel: [121.39] usb 3-12: New USB
  device found, idVendor=0781, idProduct=5567
Mar 13 12:33:42 OXON kernel: [121.61] usb 3-12: Product:
  Cruzer Blade
Mar 13 12:33:42 OXON kernel: [121.68] usb 3-12:
  Manufacturer: SanDisk Corp.
Mar 13 12:33:42 OXON kernel: [121.75] usb 3-12:
  SerialNumber: XXXX1111
```

From this log, it is apparent what device was connected to the computer system (i.e., a USB drive) and details about that device itself, including vendor ID and name, product ID and name, and crucially the device's serial number (here anonymised). In terms of our approach, this serial number is one of the data points that provides particular insight, because it can be used, in addition to some of the other device details, to link computers – that is, computers using this same device would also have the device's make and ID details in their system log entries. A similar analysis method can also be applied when searching for device associations formed over Bluetooth, WiFi or a LAN; with WiFi, even though network names (SSIDs) might not be unique, MAC addresses of their access points (typically BSSIDs) are likely to be.

Since these log files are written to by the device OS, and by other programs, they may well differ across devices. From a small experiment that we conducted across Linux (Ubuntu 13.10), Apple Mac OS (10.9) and Windows (7 and 8) platforms however, we were able to confirm that systems do capture the aforementioned details on connected devices (albeit in varying formats) and therefore, we strongly believe that links can be made. We would note, however, that some of this assessment, especially for Windows, had to be done manually because of how the log information is stored in the registry and system event viewer.

C. The approach

Following on from the discussions above, we summarise our approach to applying SNA to device logs in three steps:

1) Gather device system data and metadata — In our case, in the shape of log-files from compliant participants.

2) Isolate and extract the data and metadata of most use for SNA — Although in some cases one might be able to use the full dataset, it is more common that the data will need to be preprocessed to find and extract the most relevant device log entries. These would typically include records of connections to external devices, such as USB drives, Bluetooth devices, WiFi access points, or Gateways.

3) Apply SNA techniques and metrics to the extracted dataset to identify relationships between devices and thus, potentially device owners — This task applies the range of SNA methods with the aim of discovering device relationships. Once connections have been noted at this technology-device layer, we then exploit the coupling phenomenon introduced in Section I, to allow associations to be made to, and between device owners.

In the next section, we present the preliminary experiment that was conducted to test this approach, but also to determine the ease (or difficulty) with which the steps above could be achieved. At this point, it was important to identify, document and understand any issues, as the next step in our work involves a more comprehensive study to thoroughly assess the scope of this inference research, and exposure to privacy risks.

IV. Preliminary experiment

A. Setup

The first task was to gather system data from a set of individuals' devices. In terms of scope, at this stage we focused on *computer* log-files and on devices with *Unix-based OSs*. Moreover, given that we were interested in inferring relationships between individuals, we targeted a participant

cohort likely to have some real-world social relationships. Consequently, the participants recruited were from two general groups within the university's Computer Science department, namely an undergraduate and a researcher cohort. There were 11 undergraduates and 6 researchers surveyed, resulting in a total of 23 computer devices; these encompassed their own personal devices and devices used for work. As is common with all studies dealing with personal information, participants were fully informed of the goals and process of the experiment, and required to sign ethics consent forms before involvement.

Collecting log data from participant's devices was relatively straightforward, once we had the necessary privileges to access the appropriate folders (e.g., */var/logs*). To automate this and the subsequent (and more complicated) data isolation and extraction task, we developed several Python scripts. These scripts were also central in protecting the privacy of the participants, by replacing each of the device identifiers (serial numbers, MAC addresses, etc.) with a salted SHA-1 digest. For the data-isolation task, we designed a set of regular expressions based on known patterns within log files that would allow for efficient searching of the desired system-connection events (such as those presented in Section III-B). These expressions were able to capture records for USB, WiFi, Bluetooth and LAN connections. Once connections and connected device details were identified, these were then extracted and placed in separate files, one per scanned device.

The final task was the application of SNA techniques to the device files to facilitate further analysis. This involved the generation of graphs in Graph Exchange XML Format and then importing these into the network-analysis and visualisation tool, Gephi (`gephi.org`), to allow a more user-friendly SNA assessment. Graphs were created such that devices were represented by vertices (nodes) and links or connections between them were shown as graph edges. We defined each edge to have a small set of properties common to all connections, namely, type of connection, start and end times, and source of the connection information. These would be used for later analyses.

B. Hypotheses

The hypotheses of the experiment were as follows:

H1) It is possible to identify and produce social network graphs using the log data gathered — This aims to validate our thinking that technology devices can be linked by USB, Bluetooth, WiFi access point, and LAN connections mediated by Dynamic Host Configuration Protocol (DHCP).

H2) The prevalence of certain connection types exceeds others and this provides insight into the individuals' relationship network — Here we aim to test two sub-hypotheses: (a) Shared WiFi access is the most common kind of link. This looks at devices connecting to the same WiFi access point. (b) Connections via USB and Bluetooth devices are considerably rarer and hint towards a stronger social bond (potentially even trust) between device owners.

H3) The properties of the device graphs resulting from our analysis reflect those of an accurate model of the social network between device owners — In detail: (a) Given that participants study or work in the same department, we might expect that a majority of their devices are connected; (b) In the graph, the most connected devices (individuals), and strongest links between devices (individuals) in the sampled network become apparent; (c) It is possible to differentiate the two main groups of participants (i.e., undergraduates and researchers) by assessing the clustering of nodes in the device graphs.

V. Analysis and results

A. SNA concepts and metrics supporting our analyses

Here we introduce some relevant SNA concepts and metrics.

Degree centrality of a node is a measure of how connected a node is to other nodes in the network [26]. It is typically used as an indicator for node influence or popularity.

Giant components are informally defined as occurring when a single connected component contains a significant fraction of all network nodes [34]. This is used to provide general insight into the network's connectivity.

Affiliation graphs connect actors to foci such as activities, or groups, rather than to each other [35]. Such graphs are bipartite, meaning their nodes can be partitioned into two sets (here, people and foci), with no two nodes in either set adjacent. If the network we find is bipartite, there may be some value in viewing it as an affiliation graph.

B. The network of devices

Using the approach detailed in Section IV, we collected and processed system log-files on the 17 study participants to produce the graph displayed in Figure 1, called Network A. In total, there were 23 scanned systems, and 249 others (e.g., USB, Bluetooth, WiFi access points, Gateways) discovered in their log files. These are colour-coded according to the key in Figure 1, with scanned systems labelled according to their computer type and shown in red and yellow nodes, and discovered systems presented in the other coloured nodes. Below are some initial findings and comments.

The first general point as it pertains to the stated hypotheses is that it is indeed very possible to identify and produce social networks using the data found in system logs (thus validating *H1*). As mentioned before, this can be particularly useful in defining some initial relationships between devices that might then be further used to link owners of such devices. It was also apparent that some systems were richer in data than others. This could be because they are used more frequently, or connected to other devices more often. Alternatively, devices' connections could have been missed due to different logging styles. Known omissions at this point are: (i) Arch Linux (`archlinux.org`) systems do not log USB serial numbers (thus, USB devices are not uniquely identifiable); and (ii) Apple Macs do not always seem to log connections to DHCP servers, Gateways, or Nameservers.

We also found that Bluetooth connections and ad-hoc networks seem very rare in practice. This is contrary to our initial thoughts, but perhaps should be unsurprising – we noted that both were rather difficult to account for in testing, and

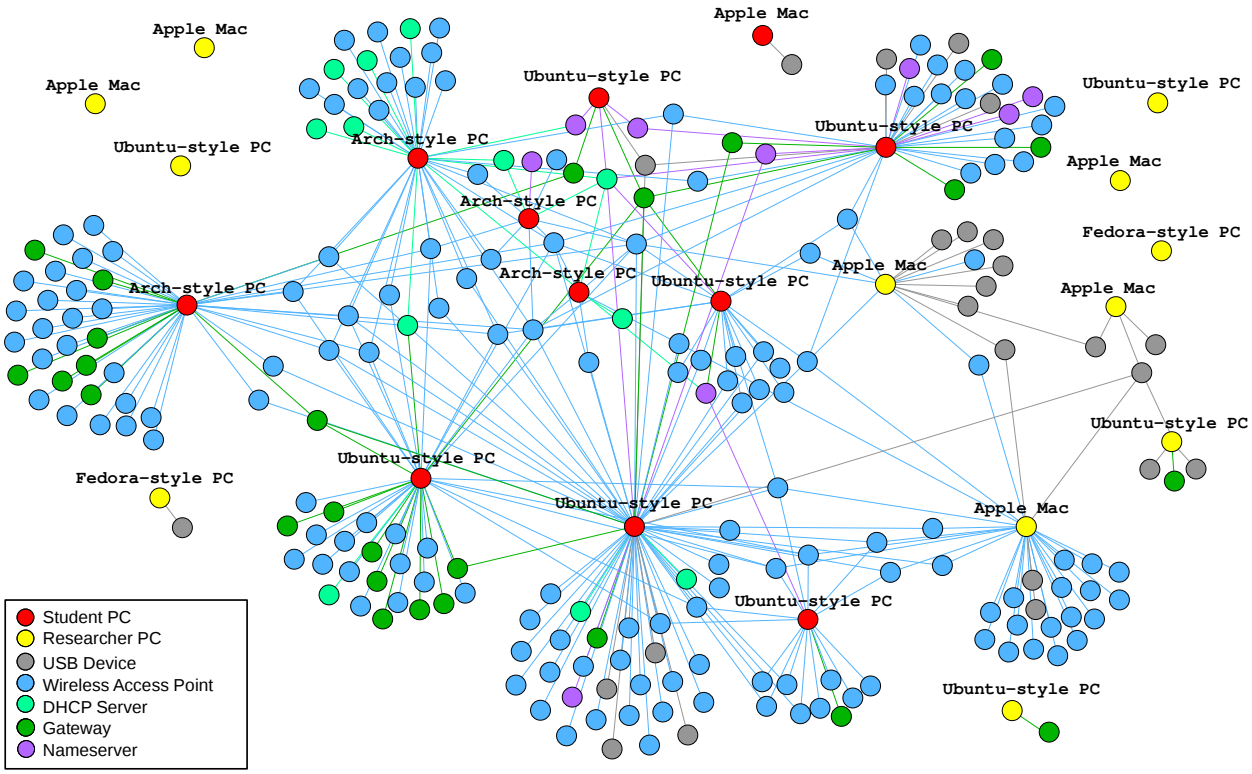

Fig. 1. Network A showing a graph of the connections between the scanned and discovered technology devices

informally, most participants did not report using them often in everyday life. Other connection types including WiFi and USB do appear as hypothesised – Figure 1 clearly demonstrates the vast prevalence of WiFi as a connector in comparison to USB or Bluetooth – thus providing some validation for *H2.a*.

To assess the extent to which connection types are shared, we defined a *share rate* metric. This metric captures the fraction of nodes of that type that have a *degree* (i.e., association with other devices) of at least 1. From our analysis of share rates across device types, the most significant finding was that USB devices have one of the lowest (at 9% as compared to the 37.5% of WiFi access points), suggesting that although they are being used to share data, they are mainly used as private devices. This finding provides some support for the hypothesis (*H2.b*) that sharing a USB device may be a useful indicator for friendship between users; we do, however, acknowledge that some systems do not log unique identifiers for USB devices, so that may have affected this initial result.

Another finding was that there was a recognisable 'giant component' in Network A, in the sense described, and that all the other components (i.e., sub-graphs not connected to the giant component) contained only one scanned system each. To an extent, this supports the hypothesis (*H3.a*) that a majority of devices would be connected given their shared affiliation with the Computer Science department – in most cases where devices have connections, they are joined to the giant component. From the graph, yet another notable finding is that there seems to be a clear distinction between devices that 'socialise' (i.e., interact a lot) and those which do not, even within the giant component. This and its relation to *degree centrality* are discussed further in the next section.

C. Determining relationships between the scanned devices

Network A gives a clear picture of which devices are acquainted with each other, but as it is bipartite, it does not directly link together any devices belonging to specific people. Can it be viewed as an affiliation network in which the scanned devices act as actors, and the discovered devices as foci? Feld defines a focus as any social, psychological, or physical entity around which joint activities between individuals are organised [36]. This raises the question: how can technological devices and connections act as foci, or evidence for the presence of foci, in a social network of humans? Borgatti and Halgin [35] describes several methods for analysing affiliation networks. Only some involve drawing a co-affiliation network (in our case, a network of the scanned devices), but all require foci that provide suitable evidence of social interaction.

We are going to need a weaker definition than Feld's. While it could be argued that, for example, a meeting to exchange data might be organised around a storage device that changes hands, Feld's definition does not encompass many of the links that we believe might be found in our current network. Therefore, we define a technological focus as *a technologically detectable event which indicates that some social interaction has occurred between individuals*. This is a broad definition

encompassing, for example, phone calls, friendships on social-networking sites, in-person meetings recorded on CCTV, etc.

Which of the interactions in Network A are, or could be argued to be, technological foci in this sense? Consider the connection types that may occur between two devices: (i) *Sharing a WiFi access point* suggests that the two devices have both visited the same physical location. We argue that this does not constitute a focus since there is no evidence that the users of the two devices ever met; (ii) *Sharing a WiFi access point at the same time* however, does suggest that the two users were at least in close proximity, so we treat such events as foci; (iii) *Sharing a USB or Bluetooth device* is strong evidence for deliberate collaboration and sharing of data, and thus constitutes a focus; (iv) *Sharing a DHCP server, gateway or nameserver* is not considered to constitute a focus, as these may be shared by large sections of a network.

It is straightforward to identify the scanned devices in Network A which have shared a USB or Bluetooth device (by node colour), but simultaneous access to a WiFi access point is much harder to visibly detect. For 19,041 out of the 22,467 connections (85%) to such points, the end time of the connection could not be determined. Even when the connection was reported as successful, a significant proportion last under a minute. We are not convinced that this accurately represents the amount of time that a user would typically be expected to be connected, so for now, we take the median time of 36 minutes, and assume this time for unknown connections. While this is not ideal, and indeed may seem to be an overestimation, we believe this to be a conservative estimate, as (i) the University timetable is divided into 60-minute slots, and (ii) it is possible that one physically static session of use involves multiple connections due to reboots or connection errors (the authors experienced many such erratic connections). Future work will need to address this problem.

Our creation of the co-affiliation device graph is as follows:
1) Let each scanned device be a graph node.
2) A graph edge is created between two scanned devices if a direct link via USB, Bluetooth or WiFi exists between them.
3) To determine the weight (i.e., significance) of WiFi-link edges, we compute the total number of minutes that both devices were simultaneously connected to the WiFi point. For USB or Bluetooth-link edges, as we posit that these connections are an indicator of a real relationship between device owners, these are assigned a higher edge weighting; a value of 60 was chosen, equivalent to an hour's co-location. If two devices have connected by both USB or Bluetooth and WiFi, their values as calculated above are summed to give a final edge weight.

The resulting graph is presented in Figure 2 with larger edge weights highlighted by thicker edges between nodes.

In terms of relationships within device groups, 8 of the 11 undergraduate computers were connected to another node and 3 were solo. In the research cohort, it was somewhat to the contrary as 4 computers were connected and 8 were solo. There is evidence to support the perspective that the nature of participants' roles had an influence on their ties

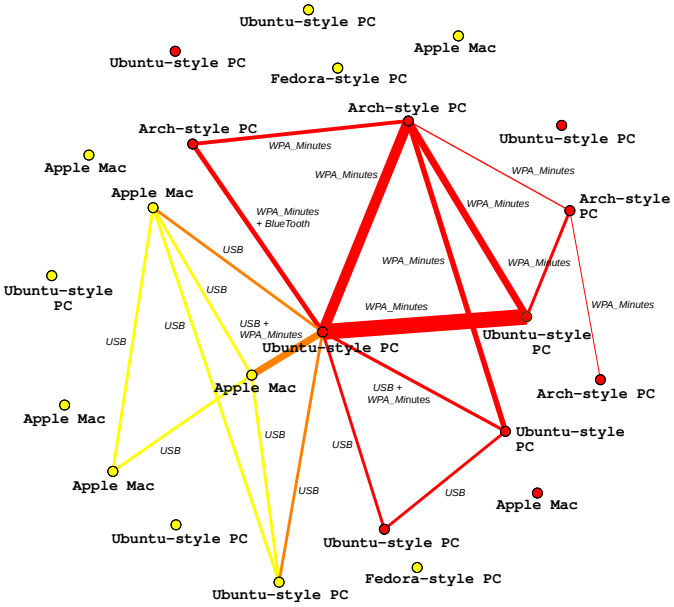

Fig. 2. Network B showing undergraduates (as red nodes), researchers (as yellow nodes) and the types of connections present between them (edge labels)

and connections. That is, undergraduates informally reported being more social generally and in their university lives via group projects, LAN (gaming) parties and so on. Researchers, however, seemed more isolated and even if they did work together, tended to opt for mechanisms such as email or instant messengers to communicate and share data and files.

Other important graph features worth noting include the strength of relationships between individual devices and general connectivity of devices in the network; these speak to *H3.b*. In Figure 2 for instance, it is possible to quickly identity that the strongest link between devices is that between the *Ubuntu-style PC* in the centre of the graph and other *Ubuntu-style PC* to its immediate right; we note here that these PCs do belong to different individuals. Such an association could be indicative of human friendship, potentially even a close one. Considering the question of how connected a device is to other devices, we use the *degree centrality* metric introduced in Section V-A and scale the node sizes accordingly; see Figure 3.

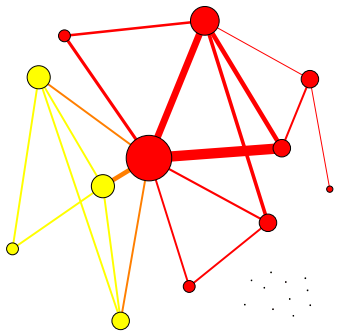

Fig. 3. Network B with the size of undergraduate and researcher nodes scaled according to how well connected they are in the device network

From the new graph, we can quickly spot the most connected device in the network, and incidentally, also the device (and potentially individual) acting as the bridge between the undergraduate and researcher cohorts. This paragraph therefore validates *H3.b*.

The final question pertaining to our hypotheses is: could the network in Figure 2 be used to infer the two main social groups? To test, we ran the highly-referenced clustering algorithm of Blondel *et al.* [37] on the graph, with resolution set at `1`. As apparent in the new network shown in Figure 4, it is indeed possible to automatically identify the two predominant groups (thus validating *H3.c*), with half of researchers devices occupying the blue cluster and most of undergraduates' in the green cluster; the network modularity of `0.218` is not ideal (the closer to `1`, the better) but can be accounted for by the number of disconnected graph nodes. Our clustering analysis largely ignores nodes that are not connected, i.e., 27% of undergraduate and 50% of researcher devices. This high percentage highlights a potential issue for our wider study, in that, on occasion, device logs may simply not be maintained or certain devices may never be connected.

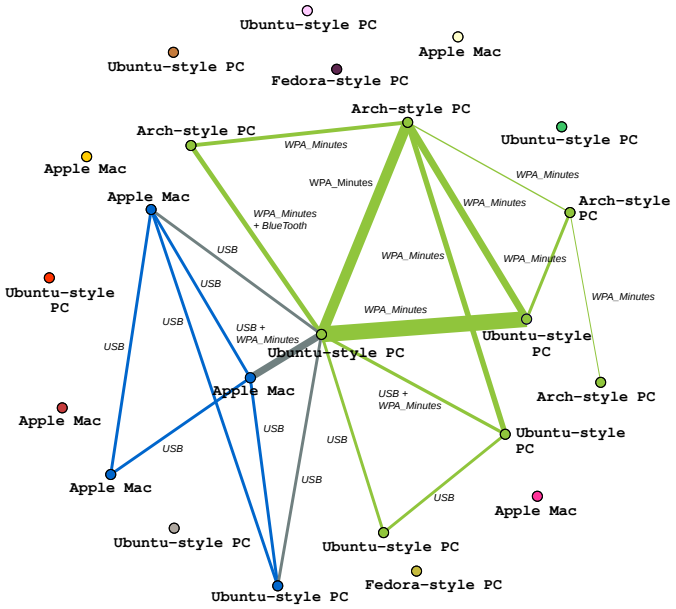

Fig. 4. Network B with the clusters of devices identified, each in a different colour. Here we can spot two device (individual) clusters in blue and green.

One final point regarding these two clusters is that while researchers exhibit a high preference for USB sharing, undergraduates are much more likely to connect to the same WiFi points. This might be indicative of higher interpersonal trust between researchers, or simply the result of researcher connections being mostly via DHCP, while undergraduates use email, the cloud, and so on, to interact. At the very least however, these clusters do highlight the fact that digital footprints can result in some level of privacy exposure to individuals, in the context of their real-world human social relationships.

D. Limitations

The main limitation of this preliminary study is intrinsic to the use of log-files and the now-apparent fact that computer OSs vary considerably in what is logged, and some even allow users to control what gets logged. While this is positive from a privacy perspective, it has limited the extent to which we could properly test what can be inferred from this particular digital footprint. Without useful data, our approach would suffer significantly and result in a network of orphaned nodes; this may have happened with some of the nodes in Network A, possibly resulting in a graph that is not fully representative of the offline social network.

VI. Conclusion and Future Work

In this paper, we have sought to exploit the notion of owner-device coupling to investigate whether low-level metadata maintained by devices can be used to make inferences about the social relationships between their owners. To this end, we have proposed and experimented with an approach based on SNA and basic Computer Forensics through which this could be achieved. From our preliminary experimentation we found that this inference task is possible, and that there is notable insight (regarding device connectivity, link strength, and clusters of individuals) to be gleaned from such analyses. This provides cause for concern from a privacy perspective as it highlights yet another way that the digital footprints, which we inadvertently create through the use of technology devices, can expose us to privacy risks; here, the exposure of our real-world social relationships.

As it relates to future work, although our experiment was successful and yielded positive results, our aim was always to use it as a pilot study to better inform the planning of a larger, more thorough experiment. This future experiment would seek to rigorously assess the privacy risks to individuals based on what can be found in, and inferred from technology-level data and metadata; initially, our focus is log-files but this will broadened as the research progresses. Reflecting on our preliminary study, there are several areas to be further explored and a number of challenges to be overcome as we go forward. Here, we present three of these, starting with the scope of data collection.

In this work, we focused on Unix-based OSs due to their prevalence in our immediate environment. In the wider world, however, the Windows OS is much more popular and therefore, a more comprehensive experiment should seek to encompass these systems as well. We envisage a technique similar to the Linux *syslog* analysis but this time targeting the Windows registry (especially folders such as *USBSTOR*, *Devices*, *NetworkList*) and *Event Viewer* to extract information on connections to external devices. Moreover, as we concentrate more on coupling, expanding our scope to mobile devices (e.g., Android, iOS, Windows) is imperative. An initial issue here which we are yet to tackle is that some system logs appear to be frequently purged in order to save on precious internal storage space.

Another area for further consideration is the reality that different versions of the same OS can have different logging formats. This issue was noticed between some Apple Mac OS versions in particular. To ensure that our wider study captures all the relevant data therefore, the logging format of each OS version would need to be thoroughly studied and the regular expressions updated and extended as appropriate. This would hopefully address the issues regarding missing device connections and disconnections encountered in this initial experiment. Also, a more complete dataset would undoubtedly benefit our SNA analysis and lead to more representative, and informative findings.

The final area concerns the participant cohort. The preliminary study used a small, convenience sample of individuals and aimed to characterise the main relationships and groups. As we seek to thoroughly assess the exposure to privacy risks from technology-level data in our next study, it would be advantageous to extend this scope in two ways. Firstly, enlarging the sample size of participants to determine whether it is possible to identify previously unknown relationships – a challenge to tackle here would be finding sufficient individuals willing to grant us the necessary privileges and share their system log-files. Second, formally assessing real-life associations, friendships and trust relationships between participants (via self-report questionnaires, for instance) and creating a social network from these, that could then be compared against the technology-device network to assess how closely that corresponds to the reported ground truth. Experimenting with these should provide insight into the privacy risks faced and complement other work assessing the risks to privacy as a result of digital footprints.

REFERENCES

[1] Mashable. (2014) U.S. Adults Spend 11 Hours Per Day With Digital Media. [Online]. Available: http://mashable.com/2014/03/05/american-digital-media-hours

[2] 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 *IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks*. IEEE, 2012, pp. 1–9.

[3] M. V. Barbera, A. Epasto, A. Mei, V. C. Perta, and J. Stefa, "Signals from the crowd: uncovering social relationships through smartphone probes," in *International Conference on Internet Measurement*. ACM, 2013.

[4] Sophos: NakedSecurity. (2012) What is your phone saying behind your back? [Online]. Available: http://nakedsecurity.sophos.com/2012/10/02/what-is-your-phone-saying-behind-your-back/

[5] T. Matsunaka, A. Yamada, and A. Kubota, "Passive OS Fingerprinting by DNS Traffic Analysis," in *27th International Conference on Advanced Information Networking and Applications*. IEEE, 2013, pp. 243–250.

[6] D. Irani, S. Webb, K. Li, and C. Pu, "Large online social footprints–an emerging threat," in *International Conference on Computational Science and Engineering*, vol. 3. IEEE, 2009, pp. 271–276.

[7] J. Wiles and A. Reyes, *The Best Damn Cybercrime and Digital Forensics Book Period*. Elsevier Science, 2011.

[8] W. H. Allen, "Computer forensics," *IEEE Security & Privacy*, vol. 3, no. 4, pp. 59–62, 2005.

[9] Ubuntu Wiki. (n.d.) LinuxLogFiles - Community Help Wiki. [Online]. Available: https://help.ubuntu.com/community/LinuxLogFiles

[10] Microsoft. (n.d.) What information appears in event logs? (Event Viewer). [Online]. Available: http://windows.microsoft.com/en-gb/windows/what-information-event-logs-event-viewer

[11] D. Knoke and S. Yang, *Social network analysis*. Sage, 2008, vol. 154.

[12] Saga. (n.d.) SAGA. [Online]. Available: http://www.getsaga.com

[13] Guardian US. (2011) A Guardian guide to your metadata. [Online]. Available: http://www.theguardian.com/technology/interactive/2013/jun/12/what-is-metadata-nsa-surveillance

[14] P. Alvarez, "Using extended file information (EXIF) file headers in digital evidence analysis," *Journal of Digital Evidence*, vol. 2, no. 3, 2004.

[15] I. Arroyo and B. P. Woolf, "Inferring learning and attitudes from a bayesian network of log file data," in *International Conference on Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology*, 2005, pp. 33–40.

[16] How To Geek. (2011) How to use EXIF data to learn from master photographers. [Online]. Available: http://www.howtogeek.com/68085/how-to-use-exif-data-to-learn-from-master-photographers/

[17] D. Rosenblum, "What anyone can know: The privacy risks of social networking sites," *IEEE Security and Privacy*, vol. 5, no. 3, pp. 40–49, 2007.

[18] S. Creese, M. Goldsmith, J. R. C. Nurse, and E. Phillips, "A data-reachability model for elucidating privacy and security risks related to the use of online social networks," in *11th IEEE International Conference on Trust, Security and Privacy in Computing and Communications*. IEEE, 2012, pp. 1124–1131.

[19] P. Eckersley, "How unique is your web browser?" in *Privacy Enhancing Technologies*. Springer, 2010, pp. 1–18.

[20] T. Stöber, M. Frank, J. Schmitt, and I. Martinovic, "Who do you sync you are?: smartphone fingerprinting via application behaviour," in *6th Conference on Security and Privacy in Wireless and Mobile Networks*. ACM, 2013, pp. 7–12.

[21] Y. C. Yang, "Web user behavioral profiling for user identification," *Decision Support Systems*, vol. 49, no. 3, pp. 261–271, 2010.

[22] P. Chairunnanda, N. Pham, and U. Hengartner, "Privacy: Gone with the typing! Identifying web users by their typing patterns," in *3rd IEEE International PASSAT and SocialCom Conferences*, 2011, pp. 974–980.

[23] C. M. McDowell, "Creating profiles from user network behavior," Ph.D. dissertation, Monterey, California: Naval Postgraduate School, 2013.

[24] M. Rogers, "The role of criminal profiling in the computer forensics process," *Computers & Security*, vol. 22, no. 4, pp. 292–298, 2003.

[25] 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.

[26] L. C. Freeman, "The development of social network analysis – with an emphasis on recent events," in *The SAGE Handbook of Social Network Analysis*, J. Scott and P. J. Carrington, Eds. SAGE, 2011, pp. 26–54.

[27] J. Travers and S. Milgram, "An experimental study of the small world problem," *Sociometry*, vol. 32, no. 4, pp. 425–443, 1969.

[28] L. Mercken, C. Steglich, P. Sinclair, J. Holliday, and L. Moore, "A longitudinal social network analysis of peer influence, peer selection, and smoking behavior among adolescents in british schools." *Health Psychology*, vol. 31, no. 4, p. 450, 2012.

[29] K. Ehrlich, C.-Y. Lin, and V. Griffiths-Fisher, "Searching for experts in the enterprise: combining text and social network analysis," in *International Conference on Supporting Group Work*. ACM, 2007, pp. 117–126.

[30] D. Ediger, K. Jiang, J. Riedy, D. A. Bader, C. Corley, R. Farber, and W. N. Reynolds, "Massive social network analysis: Mining twitter for social good," in *39th International Conference on Parallel Processing*. IEEE, 2010, pp. 583–593.

[31] D. A. Bright, C. E. Hughes, and J. Chalmers, "Illuminating dark networks: a social network analysis of an australian drug trafficking syndicate," *Crime, Law and Social Change*, vol. 57, no. 2, pp. 151–176, 2012.

[32] eLinux. (2012) Android logging system. [Online]. Available: http://elinux.org/Android_Logging_System

[33] J. Pond. (2013) OSX log files. [Online]. Available: http://pondini.org/OSX/Logs.html

[34] D. Easley and J. Kleinberg, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

[35] S. P. Borgatti and D. S. Halgin, "Analyzing affiliation networks," in *The Sage Handbook of Social Network Analysis*, J. Scott and P. J. Carrington, Eds. SAGE, 2011, pp. 417–433.

[36] S. L. Feld, "The focused organization of social ties," *American journal of sociology*, pp. 1015–1035, 1981.

[37] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.