Highlighting Relationships of a Smartphone's Social Ecosystem in Potentially Large Investigations

Panagiotis Andriotis, George Oikonomou, Theo Tryfonas, and Shancang Li, *Member, IEEE*

***Abstract*—Social media networks are becoming increasingly popular because they can satisfy diverse needs of individuals (both personal and professional). Modern mobile devices are empowered with increased capabilities, taking advantage of the technological progress that makes them smarter than their predecessors. Thus, a smartphone user is not only the phone owner, but also an entity that may have different facets and roles in various social media networks. We believe that these roles can be aggregated in a single social ecosystem, which can be derived by the smartphone. In this study we present our concept of the social ecosystem in contemporary devices and we attempt to distinguish the different communities that occur from the integration of social networking in our lives. In addition, we propose techniques to highlight major actors within the ecosystem. Moreover, we demonstrate our suggested visualization scheme, which illustrates the linking of entities that live in separate communities using data taken from the smartphone. Finally, we extend our concept to include various parallel ecosystems during potentially large investigations and we link influential entities in a vertical fashion. We particularly examine cases where data aggregation is performed by specific applications, producing volumes of textual data that can be analysed with text mining methods. Our analysis demonstrates the risks of the rising 'Bring Your Own Device' (BYOD) trend in enterprise environments.**

***Index Terms*—Network analysis, digital forensics, entity linking, sentiment analysis, BYOD, Enterprise Mobility Management.**

I. Introduction

Over the recent years mobile communications experienced a remarkable technological boom. Smart mobile devices are able to perform numerous tasks and they eventually replaced traditional, comparatively primitive cellular phones (feature phones). Contemporary smartphones are now equipped with stronger processors, larger storage capacities, precise sensors and more advanced operating systems (OS). Hence, the efficiency that these devices provide is increasing, making them important tools for their users. Smartphones and tablets can be used for e-commerce, they enhance educational experience, they can monitor one's fitness through their built-in sensors and also entertain people via streaming services. In addition, we are now able to exchange files and emails using wireless technologies and share information (sync) with other devices. Furthermore, mobile devices are used for connection with social media networks. These actions leave behind footprints of relevant activity in the form of locally stored evidence.

Social network analysis (SNA) is a scientific area which provides metrics and methods to describe internal relationships among network members and highlights distinct communities that exist in the network. To perform such an analysis based on evidence that are partially stored on phones and across communication networks, an investigator has to correlate metadata from the network (e.g. call data records) with data from the target devices' file systems.

Recent studies suggested that during a smartphone forensic investigation, derived data are able to describe telephone activity, SMS and email exchange history, calendar planning, application usage, social media activity and sensor-related actions [1]. Also, a critical task of the investigation is to specify interactions among different entities that exist in places like the *contact list* of the phone and the *'friends'* list of social media applications (apps). Existing literature demonstrates a variety of methods employed to successfully classify such interactions, utilizing SNA techniques ([2], [3]) combined with methods that extract account information from various sources [4]. However, there is no particular study that illustrates contact linking using data taken only from a target smartphone and especially using the traces from the syncing process provided by popular social media apps. Our proposed scheme points towards cases where the only source of evidence is the smartphone of the person under investigation. We specifically stress the fact that people nowadays bring mobile devices in their workplaces and use them inside their enterprise environments, thus data sharing in this context might introduce risks related to security and reputation.

In more details, in this paper we aggregate data preserved in the internal memory of the phone using database linking. We also construct visualization schemata, which illustrate interconnections among entities at different levels of the ecosystem. To achieve this, we use artefacts found in the phone to visualize how entities from the contact list are linked with social media accounts, utilizing the synchronization data left in the databases that store information in the phone. Our approach depends only on the information found in the phone, bypassing privacy policies that restrict other proposed methods. For instance, methods involving web crawlers and social media Application Programming Interfaces (API)[1] are restricted by third party privacy policies.

In this context, this paper contributes the following:

Panagiotis Andriotis, George Oikonomou, and Theo Tryfonas are with the Bristol Cryptography Group, University of Bristol, Department of Computer Science, Merchant Venturers Building, Bristol, BS8 1UB, U.K., e-mail: p.andriotis@bristol.ac.uk, g.oikonomou@bristol.ac.uk, theo.tryfonas@bristol.ac.uk

Shancang Li is with the Institute for Informatics & Digital Innovation, Edinburgh Napier University, Merchiston Campus, 10 Colinton Road, Edinburgh, EH10 5DT, U.K., e-mail: S.Li@napier.ac.uk

Manuscript received July 30, 2014; XXXXX XXXX XX, 20XX.

[1] e.g. the *Facebook* API

- We identify various sources of information that link smartphone users to the universal digital community and provide a structured description of the social ecosystems defined by smartphones.
- We provide formal definitions and metrics describing graphs formed by entities existing in social media communities inside the smartphone ecosystem. In addition, we propose algorithms that highlight major actors and reveal central entities using different facets in their social media activity. These techniques empower the development of an automated forensic tool able to read traces left by the application syncing process.
- We investigate the effect of the SMS integration that various apps[2] offer and we use the increased volume of textual data to perform sentiment analysis on short texts in order to reveal mood trends in the smartphone ecosystem.
- Finally, we extend our concepts to cover large investigation cases where entities from numerous parallel ecosystems are linked vertically within the same environment. We specifically stress the risks that exist by the use of personal smartphones in enterprise environments, where no BYOD policies are employed.

The rest of the paper is structured as follows. In Section II we discuss previous work in the fields of SNA and digital forensics and how it relates with enterprise risk management (ERM). A definition of the smartphone social ecosystem is given in Section III and our approach to visualize social activity is provided in Section IV. Section V presents the methodology we used to develop a tool, which presents the interactions of the central entity of the social ecosystem highlighting the important actors in a distinct way. In Section VI we are testing our assumptions in a case study, reconstructing the social ecosystem of one of this paper's authors. Section VII discusses the benefits of analysing textual data with text mining methods. In Section VIII we present the adoption of our concept in large investigations and we focus our analysis on risks that reside in environments, where ambiguous BYOD policies are utilized. We draw our conclusions in Section IX.

II. Related Work

The invasion of mobile devices in peoples' lives, changed their habits dramatically. Smartphones are integrated in our daily routines because we can use them in our social and professional lives. Consequently, employees tend to bring their personal devices in their workplaces. This fact might enhance their productivity because the technology they use is familiar to them. This is why Bring Your Own Device (BYOD)[3] policies are becoming very popular and necessary in advanced enterprise environments. However, the integration of several mobile devices in the workplace creates risks. Despite the security risks and the administrational chaos BYOD models might initiate, they cannot act as obstacles to business transformation [5]. However, mobile device management (MDM) measures for safer use of smartphones or tablets in the workplace ecosystem might be very restrictive and this fact could lead to BYOD failures [6]. The most common MDM schemes are based in app virtualization and containerization.

The acceptance of BYOD policies can make firms seem more attractive to their employees and customers but these policies are usually followed by ambiguous regulations [7]. For example, if an approved smartphone is allowed to access an app, is it also allowed to use location services or the contact list? From an ERM perspective [8] the actions of information exchange and data sharing among employees via several apps might introduce primarily non-financial issues, such as reputational risks. The data circulation might also initiate property-based risks related to interconnection and complexity [9]. Thus, we need to track and restrain the flow of data from the social networking activity of individuals inside interconnected environments in order to minimize risks.

In order to reveal social trends, contemporary research efforts study online social networks and report events using statistical methods to analyze data shared by users [10], [11]. SNA is also a critical asset that should be taken into account by forensic investigators, as Sparrow highlights in [12]. However, Garfinkel in his survey about the evolution of digital forensics underlines the difficulties forensic analysts face because of the plethora of diverse data types and the numerous devices that exist in the market [13]. A solution to this problem is the introduction of abstract data formats (using XML files) to describe and represent digital evidence. The concept of utilizing XML files to represent social media activity is also discussed in [2]. Furthermore, Read et al. in [14] proposed a different approach to visualize network traffic using a middleware scheme that connects diverse data with visualization tools via XML files.

The aim of SNA [15] is to demonstrate structural properties in social networks utilizing theoretical and statistical methods. We therefore represent social networks with graphs connecting nodes with edges [16]. In social network forensics the nodes (entities) are usually persons or accounts defined by various attributes such as names, email addresses or telephone numbers. The edges are strings that connect the entities. Newman in [17] reviews developments in the field of SNA illustrating concepts like clustering, degree distributions, network correlations and graph models. These models were derived by empirical studies on networked systems and describe techniques that were developed to understand, and presumably predict, the behavior of such systems. Adam Edwards et al. in [18] discuss the evolution of traditional social research methods and they stress the arising necessity to re-orientate them to include social media analysis in order to explain social life in the new digital era.

Forensic analyses on smartphones usually produce large quantities of data. Subsequently, various visualization schemata have been proposed in order to present the data in an efficient way. The common attribute of these frameworks is usually a timeline interface that depicts activity collected by hard drives or disk images [19]. For example, in [20] the authors present a tool that scans file systems such as NTFS and FAT32 and visualizes files using their timestamps. They also suggest that this type of data representation leads to faster investigations. Another example is the FATKit framework [21],

[2] e.g. *Google Hangouts*

[3] or Bring Your Own Technology (BYOT)

which extracts and analyses data from volatile system storage, offering visual information of *low level* data structures.

The proposed automated solution in [22] deals with *high level* data and it is focused on unstructured texts, such as webpages and chat logs. The authors apply data mining techniques to extract useful information and discover criminal networks (communities). These clusters are further visualized following two different visual representations. Further work has been done with textual data recently in the area of Natural Language Processing (NLP) using machine learning algorithms [23].

Sentiment analysis is a popular subject among NLP researchers. They aim to extract mood trends for important people, products and services. Their sources vary from short messages (SMS) to forum data [24]. Twitter is a microblogging service where users can post publicly short messages and it has been used extensively as a corpus for sentiment analysis [25]. Twitter posts and SMS are similarly structured [26], thus Twitter feeds can be used to perform sentiment analysis on SMS [27], [28]. The results from related studies show that the accuracy of unsupervised lexicon based methods is approximately 0.69.

Law enforcement agencies and forensic investigators frequently face the challenge of data extraction from smartphones and mobile devices [2]. Prior work in the field of digital forensics focused on the multitude of acquisition techniques that can be applied to devices, which run distinct operating systems. Physical and logical acquisition methods have been proposed for the popular Android and the iOS operating systems. Mylonas et al. in [29] identified potential evidence that may be collected by a seized phone during a forensic examination and proposed their *Proactive Smartphone Investigation Scheme*, which takes into account any legal implications an ad hoc acquisition of smartphone evidence might cause. For Android-based smartphones, there are some well-known data acquisition methods that require root access to the phone and utilize familiar Linux tools like 'dd' to acquire physical digital images of the devices' partitions [1].

Lessard and Kessler in [30] studied the artefacts in the 'contacts' database. They linked entities from the contact list with individuals in *Facebook*, using logical digital images from smartphones. However, their study described techniques for manual extraction of information. Our concept proposes an automated way to visualize these data. Our visual metaphor also highlights the connections that link existing entities. Mutawa et al. in [31] conducted forensic analysis on three popular social networking applications using artefacts left on three phones running different operating systems. They performed logical acquisition on the phones and analysed the digital images, concatenating on a reference table the information obtained from the specific social networks. Among the retrieved data, the table featured traces from user names, contact details and chat messages.

Jun-Ki Min et al. in [3] adopted a different computational model and used SMS and call logs to classify contacts according to their interpersonal relations. With their machine-learning method they were able to distinguish contacts in categories like 'family', 'work' or 'social'. This classification scheme is able to provide specific information about the relationships among entities. The aforementioned study was motivated by previous work on SNA that used supervised and unsupervised machine learning approaches ([32], [33]) to define ties among entities.

Mulazzani et al. in [4] introduced an automated approach to identify important data for network forensics. Futhermore, they integrated this information and produced useful visualizations describing entity connectivity and social network activity. Their methods do not need collaboration from the social network operator. In their study they used the *Facebook* API to demonstrate data acquisition procedures. Huber et al. [34] also used crawling methods to collect data from *Facebook* accounts and these data were visualized as graphs. However, Andriotis et al. in [35] stressed ethical issues that might occur from the use of the aforementioned approach. They also highlighted potential hurdles that might arise from strict privacy policies, which restrict the use of crawlers.

SNA is also used by other representative tools related to telephony. These are usually able to highlight clustered neighborhoods within abstract networks. Catanese et al. in [2] introduced 'LogAnalysis', a tool which employs SNA to explore cohesive groups, main figures and influential entities, using log files from mobile phone telecom networks. The authors stress that their proposed methods focus on the visual representation of the relationships among entities. They state that their tool is able to analyze large-scale mobile telecom networks utilizing centrality measures, such as 'degree', 'closeness' and 'betweenness'. However, the proposed model uses log files provided by network operators, which poses the additional requirement for the investigator to first obtain them. Our scheme depends only on the smartphone and the data that exist internally. This feature makes the scheme independent, flexible and unique.

Another interesting study combines SNA and forensics and presents a detailed narration of the methods used for malicious user detection in networks [36]. The study also employs a confidence system among entities in social media networks, based on trust evaluation. For smartphones, a multi-layer model is deployed to unify multiple instances of the same entity that exist in various social networks. One of the approaches to achieve this goal is through database matching; the address book of the phone is also utilized. The database linking is demonstrated using an iPhone 3Gs, but there is no discussion as to whether the same technique can be applied to more recent versions of the operating system or if it is also applicable to other platforms, such as the Android OS.

In the current study we aim to demonstrate the underlying risks of data sharing when IT administrators do not set adequate security-related BYOD policies in enterprise environments. We use the paradigm of a post mortem forensic analysis on a smartphone that might be held as evidence. We assume that the analysts do not have access to the user's credentials or any written clearance to crawl social media accounts of the person under investigation. They also have no access to metadata from network providers, thus their investigations are focused on smartphones only. We focus our analysis on Android devices and we target databases that exist in all versions of the popular mobile OS.

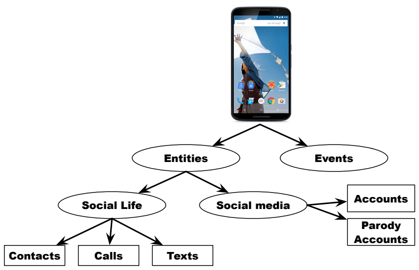


Fig. 1. The smartphone ecosystem

III. Taxonomy of the smartphone ecosystem

Modern smartphones contain numerous personal data of their users. From a forensic point of view, smartphones are considered to be valuable sources of information and they are now used as digital evidence in courts. They can be seen as ecosystems describing two different areas of interest. The ecosystem describes *events* occurring between *entities* [37]. The main entity that exists in this ecosystem (could be characterized as the *epicentre* of the smartphone universe) is the phone's user. Smartphones are as powerful as personal computers and they can connect to the Internet and therefore, to cloud services. Thus, contacts from the social life of the device owners, in addition with 'friends', 'followers' and 'connections' that exist in their social media accounts are gathered together under unified digital environments (ecosystems).

A smartphone device can reveal information about the social life of the owner through the calls that were made and received or via the short message service (SMS) texts that were sent and received by the user. In addition, social media interactions can be derived by traces left in the internal structure of a phone. These traces include messages exchanged through Instant Messengers or short texts posted via social media accounts. Hence, during a forensic examination on a smartphone it is possible to come up with data, which describe interactions between entities. Also, the data describe events that happened in social life and in the social media life. However, it is not always clear how these entities are connected together. In order to depict the relationships between entities, we need visualizations to demonstrate the different nature of contacts and interactions.

Figure 1 shows a conceptual view of the ecosystem discussed in this section. We can see that a smartphone contains information about entities and events. Entities are instances from the contact list or from social media accounts. Additionally, the figure illustrates that we usually communicate via calls and texts (SMS) with people from our social life. This paper does not deal with events and it is mainly focused on the entities that constitute the ecosystem. However, there is a short discussion at the last Section upon our fututre plans to distinguish events and connect them with the entities.

IV. Social activity visualization

In this study we aim to design social maps of smartphone owners (epicentres) using the traces left on their devices. We thus need to classify the different forms of social activity and the importance of social media interactions. Looking back at the ecosystem in Fig. 1 (Section III) we can see that it is possible to derive data describing aspects of the phone owner's actual social life. Entities interacting with the ecosystem's epicentre using SMS or making telephone calls represent usually existing people. These entities are stored in the contact list and the nomenclature that is used for them depends on the smartphone owner. There will be contacts assigned a real name and there will be contacts with a nickname. With the exception of interactions with automated systems, a telephone call or a text can be assumed as communication between living humans. In other words, these connections should appear to be stronger in the smartphone ecosystem and consequently the visualization of these interactions will stress the importance of these relationships.

Complexity increases when we are dealing with social media activities. Social media networks can be diverse and the interactions between entities do not always depict relations between friends and close relatives. For example, *Twitter* is a powerful network where entities post short texts publicly. It became popular when celebrities started to use it as a very easy and efficient way to communicate with their fans. Through the years the concept of the social network hasn't changed dramatically. We still have people interacting with the public through directed relationships (followers), but we also have entities (using nicknames) impersonating people that do not exist. These entities could be called 'parody accounts' and two characteristic examples would be "The Dark Lord" @Lord_Voldemort7 or "William Shakespeare" @ShakespeareSays. Interactions between the epicentre entity and a parody account should not be valued as very important during the social media network mapping.

On the other hand, social networks like *LinkedIn* promote professional links and they are considered reliable, because participants tend to use their real names and identities. Members of such networks use their profiles as their official professional image and therefore they tend to provide more accurate information in order to achieve a better representation of their skills.

There also exist applications that could be classified as highly relevant to the real social life of the epicentre entity. Such applications provide the ability to the end user to make voice or video calls without having to pay any fee to their provider. Popular examples of such applications are: *Skype*, *Viber* and *WhatsApp*. The aforementioned apps describe relationships between different entities that exist in the unique ecosystem of each smartphone user. Additionally, smartphone users might choose to enrich their phones' contact list with more details to describe each entity. These details might be the entity's address, email, second telephone number, photo and so on. These attributes further characterize the entity's identity and play a major role to the syncing process between different apps. The importance of linking separate actors in the ecosystem can be highlighted by the fact that modern smartphones provide the ability to the owner to link manually or automatically real contacts with social media accounts.

The final aspect of the social integration within the ecosystem is the *contact syncing* capability provided by several social media apps. Most of the aforementioned applications provide the ability to the central entities to automatically sync their contact lists. The result of such an action will be the linking of different entities that exist in the various areas of the ecosystem. The most interesting part of the process is that the application itself performs the linking automatically. Each application (e.g. *Twitter*, *Facebook* or *LinkedIn*) performs this syncing based on their individual implementation, but at the end of this procedure, it is up to the user to determine whether the linking was successful. Therefore, the central entity is responsible to accept or reject the linking proposed by the app. This action makes the final linking outcome reliable and provides strong indications that accounts from different aspects of the ecosystem are strongly interconnected. Hence, contact syncing can be used as a tool to depict the way social life and social media life are linked together within the smartphone hyper-community.

V. Implementing a Visualization Scheme

For this study we used a Samsung Galaxy Y (GT-S5360) smartphone running the Android OS (version 2.3.5). Our aim is to represent different communities producing efficient visualizations able to highlight the most important interactions between the entities of the smartphone ecosystem. Automated visualizations provide valuable information about the communication between the entities. This paper describes procedures and methods focused on the Android OS (open source), thus we do not have to deal with any issues that might arise from the use of other proprietary platforms (e.g. iOS). Additionally, the latest reports show that Android dominates the global market[4], thus our research is related with the choices of the majority of users.

We assumed that the smartphone was seized by authorities and can be held as evidence. We used open source tools to proceed to the forensic analysis and focused on the artefacts of social networking activities. The data acquisition method we used is described in [38]. We performed physical acquisition of the phone's data partition utilizing well-known Linux tools like 'dd'. In order to do that, the analyst needs a computer running a Linux distribution to act as the investigation machine. The communication between the computer and the smartphone takes place over USB using the 'adb' tool, which is provided in the Android SDK by the official Android Developers web page. As soon as we acquire the data partition image we can 'pull' it in the investigation machine and 'mount' it to the system. After the image has been successfully mounted, the analyst can start the investigation of the data. The basic source of information is the *data* folder, which contains SQLite databases and other files of interest.

The concept of our automated visualization scheme depends on the ecosystem analysis presented previously. Therefore, the telephone calls, the exchanged short text messages and the social media circles are considered to be vital aspects of the user's social ecosystem. The automated visualization scheme will present the most critical interactions between entities that happened by SMS exchange and telephone calls and it will also point out relationships in social media. (In this paper we demonstrate our findings about *LinkedIn* and *Facebook*.) To achieve this goal we used weighted directed and undirected graphs. We implemented our algorithms to produce *Gexf* files [39]. Moreover, we plotted these files using the *Gephi* visualization tool [40], which also provides network analysis algorithms. Finally, in order to produce clusters (communities) and make complex graphs look well-ordered, we used the Yifan Hu algorithm [41].

A list of the databases we queried is provided below (from the */data/data/* folder). Note that this list can grow according to the extent of the investigation.

- *com.facebook.katana/databases/contacts_db2*
  This is the database where the 'friends' information is stored. Names, nicknames, email accounts are some of the data that can be found there.
- *com.linkedin.android/databases/linkedin.db*
  In this database we can find names and additional information about our *LinkedIn* contacts.
- *com.android.providers.contacts/databases/contacts2.db*
  This is the contact list of the phone. The data we will find here depend on the user's input. Users can provide a variety of information for each entity, such as name, telephone numbers, addresses, email accounts and they can also link entities with photos.
- *com.sec.android.provider.logsprovider/databases/logs.db*
  This database keeps track of activities related to the (specific) phone, such as a received call or a sent SMS.
- *com.android.providers.telephony/databases/mmssms.db*
  In this database we can find attributes which describe the SMS activity (sent or received messages, the actual content of the message, the telephone number that sent or received a text message).

A. SMS component

For the SMS visualization component we need to query the database *com.android.providers.telephony/databases/mmssms.db*. We will use the 'threads' table of the database because it provides information about messages that were sent and received under a *single* thread. This characteristic helps us to easily plot the most important written interactions between entities. The number of messages in each thread is used to determine each thread's importance (w_{t_r}).

1) Definition and Relation Density Calculation: If a is the number of threads that were found in the 'threads' table and T is a set such that $T = \{t_r : r, t_r \in \mathbb{N}^+ \; and \; 1 \leq r \leq a\}$ and $\|t_r\|$ is the number of messages in the thread, then the relation density w_{t_r} that characterizes the relationship between the epicentre and the individual entity interacting with it, is calculated by equation 1.

$$w_{t_r} = \frac{\|t_r\| \cdot 100}{\sum\limits_{i=1}^{a} \|t_i\|} \tag{1}$$

2) Algorithm to Produce the SMS Visualization: A problem we should handle in the SMS case is that the 'threads' table does not provide nominal information about the entities exchanging messages (e.g. names or telephone numbers). For this

[4] http://www.idc.com/getdoc.jsp?containerId=prUS25450615

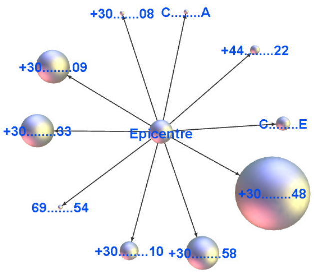


Fig. 2. SMS interactions visualization

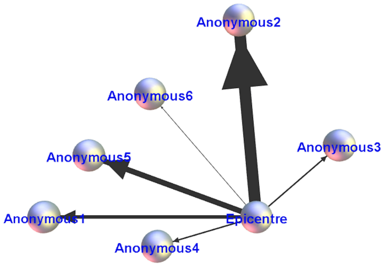


Fig. 3. Discussions visualization

reason we must also query the 'sms' table, in order to obtain the name or telephone number of the entity that exchanged messages with the epicentre of the ecosystem (algorithm 1). Figure 2 visualizes the outcome of algorithm 1. Here we can see the epicentre in the middle of the SMS interactions. The spheres represent threads of SMS that were found in the internal memory of the phone. The size of the sphere shows the size of the thread. Thus, large spheres represent large threads. The (anonymized) numbers on the spheres are the telephone numbers that interact with the epicentre. In some cases we might know the names of the people that interact with the epicentre as Fig. 2 shows. At this point we should note the possibility to encounter spam messages in the 'sms' database. Therefore, data cleansing might be needed before we perform algorithm 1 using one of the existing spam detection algorithms [42].

Algorithm 1 Produce SMS visualization

```
Prepare and Open SQLite DB
SELECT * FROM THREADS
**while** tuples exist **do**
    **if** new thread was found **then**
        Store info internally from 'data' and 'message_count' attr.
    **end if**
**end while**
SELECT * FROM SMS
**while** tuples exist **do**
    Identify entities
**end while**
Open .gexf file for 'w'
SELECT * FROM THREADS
**while** tuples exist **do**
    Create nodes & edges from THREADS
    Calculate relation density from equation 1
    Write info to .gexf file
**end while**
Close DB and files & Free space
```

B. 'Discussions' component

Our reference phone stores data about all the calls that were made to and from the device in the database *com.sec.android.provider.logsprovider/databases/logs.db*. The design concept of the component that extracts the activities considering the telephone calls is quite similar to the previous one.

1) Definition and Relation Density Calculation: If b is the number of discussions with different (distinct) contacts that were found in the 'logs' table in the logs provider database and D is a set such that $D = \{d_s : s, d_s \in \mathbb{N}^+ \mathit{and}\ 1 \leq s \leq b\}$ and $\|d_s\|$ is the number of calls with a single entity, then the relation density w_{d_s} that characterizes the relationship between the epicentre and the individual entity interacting with it, is calculated by equation 2.

$$w_{d_s} = \frac{\|d_s\| \cdot 100}{\sum\limits_{i=1}^{b} \|d_i\|} \tag{2}$$

2) Algorithm to Produce the 'Discussions' Visualization: Here we need to parse the 'logs' table of the database. The algorithm takes into account only the telephone calls, counting at the same time their frequency. At the end of this process we print a Gexf file with weighted edges (algorithm 2). This type of visualization distinguishes written communications from oral. Figure 3 depicts the telephone calls between the epicentre and the contacts that exist in the contact list[5]. Thicker arrows represent more frequent telephone calls (higher density w_{d_s}).

Algorithm 2 Produce 'Discussions' Visualization

```
Prepare and Open SQLite DB
SELECT * FROM LOGS
**while** tuples exist **do**
    **if** 'message' attribute = NULL **then**
        Store info about frequency from 'name' & 'duration' attr.
    **end if**
**end while**
Open .gexf file for 'w'
Create nodes & edges from internal info
Calculate relation density from equation 2
Write info to .gexf file
Close DB and files & Free space
```

The graph can further depict interactions stored in applications like *Skype* by simply extending the procedures we presented above. Applications of this category allow written and oral communication. *Skype* data are stored primarily in the database *com.skype.raider/files/USERNAME/main.db* and 'Calls', 'Chats' and 'Messages' are the main tables of interest.

C. Social component

Finally, the ecosystem includes relationships between social media entities. In this work we used *Facebook* and *LinkedIn* to demonstrate the concept of our proposed visualization scheme. As discussed earlier, we consider *LinkedIn* entities to be more important because of the social network's profile. On the other hand, networks such as *Facebook* might contain parody

[5] The contact names are anonymized here.

or secondary accounts. Thus, forensic investigations become really interesting if the smartphone owners have synced their contacts at the past, because it is highly possible to see connections among entities that exist in their social life and their social media life. The value of the syncing process is notable because the end user (epicentre) chooses to accept or reject connections proposed by apps. This action makes the connections reliable and adds validity to the links between abstract entities and living humans.

1) Definitions: Let m, n, l be the contacts that were found in *Facebook*, contact list and *LinkedIn* databases respectively. We thus define the following sets:
$C = \{c_i : i, c_i \in \mathbb{N}^+ \ and\ 1 \leq i \leq n\}$ is the contact list, $L = \{l_j : j, l_j \in \mathbb{N}^+ \ and\ 1 \leq j \leq l\}$ represents *LinkedIn* connections and $F = \{f_k : k, f_k \in \mathbb{N}^+ \ and\ 1 \leq k \leq m\}$ represents *Facebook* friends. We also set c_0, l_0, f_0 to be the elements that represent the epicentre entity in the respective networks. The social graph G we will produce is an undirected graph such that: $G = (V, E)$, where $V = \{\text{set of nodes}\}$ and $E = \{\text{set of edges}\}$ with $|V| = m + n + l + c_0 + l_0 + f_0$ and $|E| = m + n + l + p + q$. Note that p is the number of synced *Facebook* accounts and q is the number of synced *LinkedIn* accounts. $|V|$ stands for the order of the graph (the number of nodes) and $|E|$ stands for the size of the graph (the number of edges).

Algorithm 3 Produce interconnected visualization

```
Prepare and Open all SQLite DBs
SELECT * FROM CONTACT_SUMMARIES
while tuples exist do
    Copy names internally from ‘display_name’ attribute
end while
SELECT * FROM CONNECTIONS
while tuples exist do
    Copy names internally from ‘display_name’ attribute
end while
SELECT * FROM CONTACTS
while tuples exist do
    if link_type1 - link_type5 attributes $\neq$ NULL then
        Store info internally from ‘name_raw_contact_id’ attribute
    end if
end while
SELECT * FROM RAW_CONTACTS
while tuples exist do
    Identify entity (id) from ‘display_name’ or ‘sort_key’ attributes
end while
Open .gexf file for ‘w’
Write Facebook, LinkedIn, Contacts list into gexf & add nodes.
SELECT * FROM CONTACTS
while tuples exist do
    Connect entities from ‘name_raw_contact_id’ and internal info
end while
Close all DBs and files & Free space
```

2) Creating the Interconnected Graph: The tool queries three databases for this component. Links between the central entity and *Facebook* are described in the ‘contact_summaries’ table of the *com.facebook.katana/databases/contacts_db2* database. To identify relations between the epicentre and *LinkedIn* accounts we must query the ‘connections’ table of the *com.linkedin.android/databases/linkedin.db* database. Our intention in this context is to visualize evidence that already exist in the internal memory. We do not want to construct and visualize relationships based on ambiguous and complex algorithms, because the digital forensics science is valid only if it highlights the existing data.

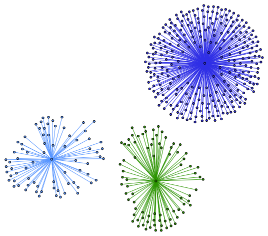


Fig. 4. The unassociated smartphone ecosystem

The main reference point of our methodology is the database *com.android.providers.contacts/databases/contacts2.db* which contains footprints from the linking process between the contact list and various social media entities. We are interested in the ‘contacts’ and the ‘raw_contacts’ tables. The ‘contacts’ table provides information about links between one person from the contact list with another social media entity, demonstrating its *name_raw_contact_id*, but not its name. We can learn the contact’s name by investigating the ‘raw_contacts’ table. If we have a match between a person from the contact list and a social media entity, the visualization connects the node from the contact list with the social media entity, underlining the achieved correlation (algorithm 3). The Gexf files that will be produced by algorithm 3 can be plotted in various visualization tools. However, Gephi does not provide yet the capability to differentiate the shape of the nodes to make them more distinctive. For this reason we use different node colours to achieve this distinction in Fig. 5 and 6.

VI. Case study

We conducted an experiment to demonstrate the visualization concept we proposed in this paper. The highlighted entities are presented as anonymous accounts. For the first part of the experiment we did not use account syncing. We performed a data factory reset on the device, restarted the phone and logged in to a Google account. We were using the same SIM card and our contact list was automatically backed up. We planned and executed a sequence of activities for more than a week including telephone calls and text message exchanges. The epicentre entity that used the phone was logged into his *Facebook* and *LinkedIn* accounts having the contact syncing features disabled. When the list of activities was completed, we acquired a physical image of the data partition using the method we mentioned previously. An image of the social ecosystem defined by this experiment is shown in Fig. 4. Here the contact list is drawn at the middle in green, *LinkedIn* connections are presented at the left hand side (in light blue) and *Facebook* friends are shown at the right hand side (in blue). We can see that the entities seem to have no relationship with each other, because no account syncing has been performed. Thus, people from the contact list seem to have no relationship with entities from social media accounts. In an enterprise environment for example, this figure could possibly show that the phone’s owner has no connections with colleagues.

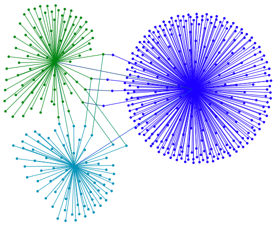


Fig. 5. The interconnected smartphone ecosystem

At the second phase of our experiments we followed the same procedure. We reset the phone and logged into the same accounts. This time when we were prompted by the apps to sync our contact list, we accepted to sync them all. The result of this action returned some suggestions by the apps, asking if we knew various people. The interaction with both apps enriched the social ecosystem. We have to note that our epicentre entity did not provide any information to the contact list apart from each entity's name and telephone number. There was no manual linking between accounts. Also the use of the phone could be characterized as limited and social media interactions were conservative and hypotonic. However, while the syncing took place, we were able to obtain an ecosystem that was more informative than the previous one (Fig. 5).

In more detail, the contact list size remained the same after the sync (82 contacts). However, the 'data' table, which is also used for storing information about the entities, contained 3 more rows of data after the automatic syncing process. The 'contacts' table showed that after syncing, 6 *LinkedIn* and 4 *Facebook* accounts were linked with the contact list, respectively. The fact that more *LinkedIn* than *Facebook* accounts were connected with the contact list might have appeared because *LinkedIn* users tend to provide their real names and identities during registration. Finally, 4 *Facebook* entities were connected with 4 *LinkedIn* accounts. These might look separate at first glance (Fig. 4), but in reality they are not. The aforementioned experiment leads us to the assumption that if the smartphone owner provides more information to the contact list (such as secondary email accounts), the synchronization among different social media entities will be more accurate, achieving better correlations.

In Fig. 5 we can see again the 3 different communities after contact syncing. At the top (on the left hand side) the contact list of the epicentre is shown in green colour. The cluster at the right (blue colour) represents the *Facebook* community and the last one aggregates *LinkedIn* entities. The figure shows links among the three communities, derived from the traces left on the phone. The 4 links at the top (between the contact list and the *Facebook* communities) are those that connect contacts, *Facebook* friends and *LinkedIn* connections. The remaining 2 links (bottom left between the contact list and the *LinkedIn* communities) do not share *Facebook* accounts.

We intentionally do not show any names because we wanted to highlight the connections between different communities in the ecosystem. However, visualization tools like Gephi provide the capability to their users to zoom in and out and investigate the connections closely. In this example we avoided to provide a very 'noisy' graph. Hence, we only used the most esteemed professional network and the most popular social networking service. Our results showed that the linking suggestions we see as users by the apps are probably the outcome of advanced algorithms running on the networks' servers.

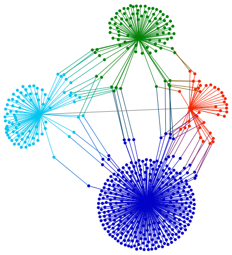


Fig. 6. An extended interconnected smartphone ecosystem

We repeated the last experiment after a few weeks and we also included another popular social network (*Instagram*). The new (extended) interconnected ecosystem can be seen in Fig. 6. It depicts the *Instagram* network on the right with red colour. It also demonstrates a fundamental advantage of our scheme; The use of algorithm [41] separates communities in distinct clusters and provides the capability to easily recognize linked entities. The figure also depicts the evolution of the social ecosystem (compared to Fig. 5). If the user syncs often the apps, more connections will probably occur.

VII. Expanding the SMS component functionality

In Section V-A we highlighted important message exchanges between entities that interact with the epicentre. However, the scheme does not provide any further information, which is based on the SMS textual content. This section is focused on the advantages we could draw from the functionality provided by apps such as *'Google Hangouts'*.

A. Instant messenger and messaging app integration

The introduction of the second version of the *Hangouts* app (which is the main instant messenger shipped with Android devices) on April 2014, offered a choice to the users to integrate their SMS messages with the specific application. Users would benefit from this integration because they could use the app as their main texting service. This feature allows the user to keep all messages and chats within a single environment. Thus, the app can be used as an instant messenger (IM) as well as an SMS client simultaneously. As discussed previously, we can retrieve the actual texts (SMS) and also other information about the parties that exchange the messages through the *mmssms.db*. However, the fact that apps like *Hangouts* offer text integration features, urged us to further examine the impact of these actions to the internal memory of a smartphone. The hypothesis is that if users choose to enable such functions on their devices, then the corresponding app databases would probably contain interesting information. Also a plethora of texts (chats and SMS) would be available to enhance our knowledge about the specific ecosystem.

Our reference phone for the current experiment was a Samsung Fame (GT-S6810P) running the Android OS, version

4.1.2. The device was used for a short period of time having the *Hangouts* SMS feature enabled. However, we should underline that *Hangouts* was not used as the main texting (SMS) app. The data partition acquisition was performed using the methodology described in Section V. The analysis of the data folder showed that *Hangouts* store critical information in the *com.google.android.talk/databases/babel1.db* SQLite database. In addition, the hypothesis that there will be various related data (chat logs and SMS, recipients and senders) inside the corresponding database was confirmed.

To be more specific, the database contained information related to *chat* and *SMS* activity. SMS actions were easily recognizable by the fact that the *'sms_raw_sender'* or the *'sms_raw_recipients'* attributes were not NULL, when the tuple described a received or sent SMS message. Additionally, all recorded actions in the database were time stamped. The actual text messages were not encrypted and open source management tools like 'SQLiteman' could retrieve them without any particular problem. Moreover, the analysis of the specific database revealed that geolocation data were also available in some tuples. In other words, if the app users decide to embed geographic information (latitude and longitude) in their chat messages, these data will be recorded in the database. These actions enrich the information we can get in a post mortem analysis because, without the app integration functionality, such data could not be embedded in an SMS.

Subsequently, if the user enables the app's SMS integration feature, the information we can gather from the relevant database during a forensic analysis can possibly show activities and relations among entities that were previously residing in different areas in the ecosystem. In addition, the invasion of wearable technology in the market (aka smart watches - *Android Wear*) and the introduction of sophisticated voice recognition systems like *Apple Siri* or *Microsoft Cortana* might urge customers to integrate their SMS with apps like *Hangouts*, in order to be able to create short messages using phonetic instructions. Thus, ecosystem integrations of this kind seem to be a likely scenario in the near future. An indication that we are probably heading towards this direction is the elimination of the *Email* app from the brand new, fifth version of the Android OS and its integration with the *Gmail* app. As a matter of fact, the default SMS app on some devices[6] running the Android OS (version 5.1) is now the *Hangouts* app.

B. Short message sentiment analysis

So far, the study we conducted was focused on highlighting major actors performing various activities in the smartphone ecosystem. Moreover, the aforementioned amalgamation of textual information offers the chance to perform automated analysis tasks on the obtained short texts from a single database. The information synthesis, indeed, provides a large pool of short texts, which could be analysed using text mining methods. The textual content can eventually describe emotions and sentimental trends among the ecosystem's entities.

In our final experiment we evaluated the ability of three algorithms to classify short messages (such as SMS and chat texts) considering their emotional polarity. The evaluation was done using open source tools (*Weka* [43]) and publicly available datasets. Messages expressing joy, love, happiness, gratitude or interest were marked as 'positive' and texts that expressed anger, sadness, disgust and fear were marked as 'negative'. Hence, the short messages were classified (using supervised learning) in two categories. The classifiers (Multinomial Naïve Bayes, Maximum Entropy and Support Vector Machine) were trained on approximately 4,600 manually labelled monograms and random Twitter feeds (positive and negative) from the Sentiment140[7] dataset using 10-fold cross-validation. We tested our classification models on 2,670 manually labelled random SMS from an English SMS Corpus [44]. The results can be seen in TABLE I.

TABLE I
SMS SENTIMENT ANALYSIS WITH SUPERVISED LEARNING

Training on manually labelled Twitter feeds from Sentiment140 dataset				
Correctly Classified	Classes	TP Rate	FP Rate	F-Measure
MNB: 79.587%	Positive	0.640	0.089	0.728
	Negative	0.911	0.360	0.837
	Weighted Avg.	0.796	0.244	0.790
SVM: 78.653%	Positive	0.605	0.079	0.707
	Negative	0.921	0.395	0.832
	Weighted Avg.	0.787	0.260	0.779
MaxE: 74.028%	Positive	0.586	0.144	0.658
	Negative	0.855	0.413	0.791
	Weighted Avg.	0.740	0.299	0.734
Testing on manually labelled SMS from SMS dataset				
MNB: 74.000%	Positive	0.778	0.335	0.799
	Negative	0.665	0.222	0.633
	Weighted Avg.	0.740	0.297	0.743
SVM: 72.096%	Positive	0.743	0.322	0.779
	Negative	0.678	0.257	0.621
	Weighted Avg.	0.721	0.300	0.726
MaxE: 57.191%	Positive	0.575	0.421	0.643
	Negative	0.566	0.423	0.472
	Weighted Avg.	0.572	0.422	0.585

Weka uses the following metrics to estimate accuracy (F-Measure). *TP Rate* is the fraction of texts classified as class x, among all texts that indeed have class x. This is also known as *Recall*. *FP Rate* is the fraction of texts classified as class x (but belong to another class), among all texts which are not of class x. *Precision* is the fraction of texts that indeed have class x, among all those classified as class x. Finally *F-Measure* is equal to $2 * Precision * Recall/(Precision + Recall)$[8].

TABLE I shows that the Multinomial Naïve Bayes (MNB) algorithm achieved better results on our training data than the other two classifiers (it correctly classified more text instances). The Support Vector Machine (SVM) we utilised, which is in fact the sequential minimal optimization (SMO) algorithm, presented similar F-score with MNB on the training data. This can be seen at the upper part of TABLE I at the Weighted Average tuples.

On the SMS testing data (bottom tuples of TABLE I), MNB once again outperformed the other two methods. It correctly classified 74% of the SMS in two categories ('positive', 'negative') and the Weighted Average F-Measure was

[6] e.g. Google Nexus 7 (2012) tablets

[7] http://help.sentiment140.com/home

[8] http://weka.wikispaces.com/Primer

0.743. The classifier also outperformed current state of the art systems. The accuracy they achieve is estimated to be approximately 0.721 [45]. In addition, MNB seems to be a better choice for sentiment analysis on SMS, given the fact that it produces lower False Positive Rates (FP). Thus, mood analysis can be achieved quite efficiently in large sets of short texts revealing behavioural trends and emotional peaks. It can also underline regions of interest where the sentiment changes and the smartphone ecosystem radiates positive or negative mood biases [28]. Positive mood is important in an enterprise environment because it is related with productivity [46].

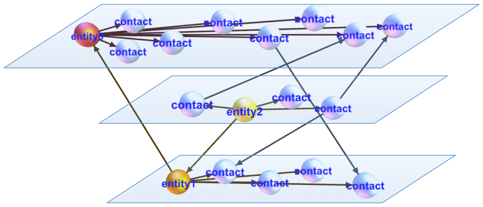


Fig. 7. Vertical entity assosiation in distinct ecosystems

VIII. Vertical Ecosystem Integration and Mobile Enterprise Management

Previously we presented an analysis derived by the syncing capability of social media apps. We demonstrated that this action links different communities within the same ecosystem. Moreover, we focused our interest on textual data aggregation, which produces large volumes of information. These data can be analysed with text mining methods. The aforementioned principles can be further extended to cover cases, where large investigations should take place. Figure 7 depicts the *conceptual* representation of such ecosystem integrations.

We assume that forensic analysts have seized a plethora of smartphones related to a particular case. Hence, they are able to build a cosmos that includes all the single ecosystems from each phone, deployed in parallel layers as Fig. 7 illustrates. For clarity and presentation reasons, the figure demonstrates three simplified ecosystems and their parallel formation. The epicentres of the three ecosystems are marked as *entity1*, *entity2* and *entity3*. Moreover, the other nodes represent entities existing in smartphone contact lists. Additionally, because of the fact that seized smartphones belong to people that interact within a specific closed environment, the epicentres of each ecosystem can be vertically linked with entities that belong to different ecosystems. An enterprise environment could be a cosmos of this kind. In Fig. 7 for example there is a direct link between entity2 and entity1. These epicentres could be employees of the some department in a commercial environment. Relationships of this kind extend initial data derived by one device and offer a deeper insight of the internal social structures in a certain environment (cosmos). The functionality of our visualization scheme enters a deeper level of connectivity and reveals important actors that might influence the universal system described by the parallel planes.

Influential entities might be also discovered using text mining on the SMS and chat logs that reside in databases like those that have been already discussed in this study (*mmssms.db* and *babel1.db*). Furthermore, sentiment analysis can be more efficient in larger environments because we are able to mix trends from different ecosystems and perceive a better understanding about mood inclinations during a day or specific timeframes, which could be of interest. This is possible because we have more data to analyse, thus there are potentially more resources to take into account.

So far, we focused our study on the Android OS. This choice was made because there is a plethora of devices in the market running the highly popular OS. Also the Android project is open-source, hence it is more preferable for research (it is non-proprietary). Compared to Apple's iOS, the Android framework has a basic structural difference; its security policies depend on permissions. If users accept these permissions when they install an app, they have no means to restrict the content that this app can reach. iOS provides the ability to the users to choose if an app has the right to access specific data, e.g. the contact lists of their phones. This feature makes our concept redundant for devices running iOS, if they restrict the contact list access to social media networks. These differences stress that mobile enterprise management policies should be applied to enhance security in workplaces.

The analysis we presented in this paper illustrates arising risks from data sharing in large interconnected social networks. For example, if there are no BYOD policies in enterprise environments, contact syncing through social apps might expose critical social links among colleagues. Contact syncing also involves email address and personal information sharing. As we mentioned previously, smartphone users might choose to enrich their contact lists with more information for each entity. This action shares several personal data to a large network of (alien) servers. Thus, there might be personal data leakage from the enterprise environment to other sources. BYOD policies must take into account these problems and secure the digital social infrastructure of corporations. Another example that indicates the necessity of regulations that restrict data sharing among social vendors is the SMS integration. We mentioned in Section VII-A that the SMS integration with chatting apps makes location sharing possible. All these embedded data can be distributed in alien ecosystems if restriction policies are not applied.

However, there are also risks that originate from the *adoption* of BYOD policies in enterprise environments. The complexity for IT departments under these schemes is arising and the need for qualified personnel to manage problems introduced by such policies is apparent. BYOD is a contract that bonds employers and employees and there might be people that feel displeased by regulating the use of their devices. On the other hand, information management in unified ecosystems is likely to be a very important priority for large enterprise environments in the near future. For this reason we have seen recently the introduction of mobile enterprise management tools from *Google*[9], *Apple*[10] and their competitors.

[9] https://www.android.com/work

[10] https://www.apple.com/iphone/business

IX. Discussion and Conclusions

In this paper we demonstrated an automated method to highlight important entities within the smartphone ecosystem. We assumed that the epicentre's contact list should be the main reference point of the investigation, because it represents entities from its social life. We also discussed the role that social media networks play within our scheme, considering their importance. Moreover, we proposed a concept that aggregates information from online sources and visualizes the connections of digital entities and humans. Our approach is based on the fact that the user of the mobile device chooses to accept or reject the app recommendations during the syncing process. Thus, despite the fact that the account correlations are automatic, the final approval is always made by the epicentre entity. This action offers a higher level of validity compared to other methods that simply correlate information shared by users on different social media.

Furthermore, we used database linking to produce visualization schemata, which could enhance forensic examinations providing indications of major actors within the social ecosystems defined by smartphones. For example, assume that a person under investigation is connected with some colleagues via a professional network (e.g. *LinkedIn*) and at the same time they are also linked on other social media networks. These links might indicate that the specific group of people has a higher degree of connectivity in the enterprise environment. A significant advantage of our methodology is that we do not need to know the credentials of the persons under investigation, in order to produce the social mapping of their phones. Also, our approach does not involve the use of APIs, thus we do not crawl accounts and do not violate privacy policies that might restrict, delay or detain the forensic investigation. In addition, assuming that the epicentre entity used the contact syncing feature, we are able to create a visualization of an interconnected ecosystem by querying multiple databases. Thus, our concept can be extended to cover all the popular apps that use contact syncing.

Given that our approach results in fast, efficient and automated representations of singular ecosystems, we further proposed the construction of a vertical system that will be able to accumulate information derived by several smartphones. These systems will aggregate data from multiple social ecosystems related to a specific case. In this study we also used text mining on short texts (SMS) to distribute them in two sentiment classes. Furthermore, our experiments exposed artefacts derived from the SMS app integration with the *Hangouts* app. The textual information we can get from such integrations increases the volume of data that can be analysed by experts. Additionally, the vertical linking of ecosystems multiplies the data in large investigations. The large quantity of data provides more opportunities for a deeper insight into the patterns they follow.

Moreover, this paper demonstrated privacy issues that occur from data sharing in unified environments. BYOD models are basically abstract contracts between employees and employers and they must address risks related to both parties via acceptable MDM regulations. Proactive forensic schemes or other monitoring frameworks might provide affordable solutions that do not violate the parties' rights and restrict vulnerabilities in enterprise environments.

Smartphone databases also include information related to events that occurred in the ecosystem (Fig. 1). The aggregation of 'entities' and 'events' in a common framework will be the subject of our future work and it will complete the representation of the ecosystem, describing which entity participated in distinct events. Graph databases (NoSQL database schemata) can be utilized to link nodes with edges that will represent events that connect entities.

References

[1] T. Vidas, C. Zhang, and N. Christin, "Toward a general collection methodology for android devices," *Digital Investigation*, vol. 8, no. 1, pp. S14–S24, 2011, 11th Annual DFRWS Conference, New Orleans, LA, Aug 01-03, 2011.

[2] S. Catanese, E. Ferrara, and G. Fiumara, "Forensic analysis of phone call networks," *Social Network Analysis and Mining*, vol. 3, no. 1, pp. 15–33, 2013.

[3] J.-K. Min, J. Wiese, J. I. Hong, and J. Zimmerman, "Mining smartphone data to classify life-facets of social relationships," in *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 2013, pp. 285–294.

[4] M. Mulazzani, M. Huber, and E. Weippl, "Social network forensics: Tapping the data pool of social networks," in *Eighth Annual IFIP WG*, vol. 11, 2012.

[5] G. Thomson, "Byod: enabling the chaos," *Network Security*, vol. 2012, no. 2, pp. 5–8, 2012.

[6] P. Steiner, "Going beyond mobile device management," *Computer Fraud & Security*, vol. 2014, no. 4, pp. 19–20, 2014.

[7] B. M. Gaff, "Byod? omg!" *Computer*, vol. 48, no. 2, pp. 10–11, 2015.

[8] D. D. Wu and D. L. Olson, "Computational simulation and risk analysis: An introduction of state of the art research Preface," *MATHEMATICAL AND COMPUTER MODELLING*, vol. 58, no. 9-10, pp. 1581–1587, NOV 2013.

[9] D. D. Wu, S.-H. Chen, and D. L. Olson, "Business intelligence in risk management: Some recent progresses," *INFORMATION SCIENCES*, vol. 256, pp. 1–7, JAN 20 2014.

[10] E. Kalampokis, E. Tambouris, and K. Tarabanis, "Understanding the predictive power of social media," *Internet Research*, vol. 23, no. 5, pp. 544–559, 2013.

[11] P. De Meo, E. Ferrara, D. Rosaci, and G. M. Sarné, "Trust and compactness in social network groups," *Cybernetics, IEEE Transactions on*, vol. 45, no. 2, pp. 205–216, 2015.

[12] M. K. Sparrow, "The application of network analysis to criminal intelligence: An assessment of the prospects," *Social networks*, vol. 13, no. 3, pp. 251–274, 1991.

[13] S. L. Garfinkel, "Digital forensics research: The next 10 years," *Digital Investigation*, vol. 7, pp. S64–S73, 2010.

[14] H. Read, A. Blyth, and I. Sutherland, "A unified approach to network traffic and network security visualisation," in *Communications, 2009. ICC'09. IEEE International Conference on*. IEEE, 2009, pp. 1–6.

[15] S. Wasserman, *Social network analysis: Methods and applications*. Cambridge university press, 1994, vol. 8.

[16] A. Bezerianos, F. Chevalier, P. Dragicevic, N. Elmqvist, and J.-D. Fekete, "Graphdice: A system for exploring multivariate social networks," in *Computer Graphics Forum*, vol. 29, no. 3. Wiley Online Library, 2010, pp. 863–872.

[17] M. E. Newman, "The structure and function of complex networks," *SIAM review*, vol. 45, no. 2, pp. 167–256, 2003.

[18] A. Edwards, W. Housley, M. Williams, L. Sloan, and M. Williams, "Digital social research, social media and the sociological imagination: surrogacy, augmentation and re-orientation," *International Journal of Social Research Methodology*, vol. 16, no. 3, pp. 245–260, 2013.

[19] C. Hargreaves and J. Patterson, "An automated timeline reconstruction approach for digital forensic investigations," *Digital Investigation*, vol. 9, pp. S69–S79, 2012.

[20] J. Olsson and M. Boldt, "Computer forensic timeline visualization tool," *Digital Investigation*, vol. 6, pp. S78–S87, 2009.

[21] N. L. Petroni Jr, A. Walters, T. Fraser, and W. A. Arbaugh, "Fatkit: A framework for the extraction and analysis of digital forensic data from volatile system memory," *Digital Investigation*, vol. 3, no. 4, pp. 197–210, 2006.

[22] R. Al-Zaidy, B. Fung, A. M. Youssef, and F. Fortin, "Mining criminal networks from unstructured text documents," *Digital Investigation*, vol. 8, no. 3, pp. 147–160, 2012.

[23] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

[24] D. D. Wu, L. Zheng, and D. L. Olson, "A Decision Support Approach for Online Stock Forum Sentiment Analysis," *IEEE TRANSACTIONS ON SYSTEMS MAN CYBERNETICS-SYSTEMS*, vol. 44, no. 8, pp. 1077–1087, AUG 2014.

[25] E. Martinez-Camara, M. Teresa Martin-Valdivia, L. Alfonso Urena-Lopez, and A. Montejo-Raez, "Sentiment analysis in Twitter," *Natural Language Engineering*, vol. 20, no. 1, pp. 1–28, JAN 2014.

[26] G. Laboreiro, L. Sarmento, J. Teixeira, and E. Oliveira, "Tokenizing micro-blogging messages using a text classification approach," in *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*. ACM, 2010, pp. 81–88.

[27] S. M. Mohammad, S. Kiritchenko, and X. Zhu, "Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets," in *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, Atlanta, Georgia, USA, June 2013.

[28] P. Andriotis, A. Takasu, and T. Tryfonas, "Smartphone message sentiment analysis," in *Advances in Digital Forensics X*. Springer, 2014, pp. 253–265.

[29] A. Mylonas, V. Meletiadis, B. Tsoumas, L. Mitrou, and D. Gritzalis, "Smartphone forensics: A proactive investigation scheme for evidence acquisition," in *Information Security and Privacy Research*. Springer, 2012, pp. 249–260.

[30] J. Lessard and G. C. Kessler, "Android forensics: Simplifying cell phone examinations," *Small Scale Digital Device Forensic Journal (SSDDFJ)*, vol. 4, no. 1, Sep. 2010.

[31] N. Al Mutawa, I. Baggili, and A. Marrington, "Forensic analysis of social networking applications on mobile devices," *Digital Investigation*, vol. 9, pp. S24–S33, 2012.

[32] M. Chau and J. Xu, "Mining communities and their relationships in blogs: A study of online hate groups," *International Journal of Human-Computer Studies*, vol. 65, no. 1, pp. 57–70, 2007.

[33] M. E. Newman, "Detecting community structure in networks," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 38, no. 2, pp. 321–330, 2004.

[34] M. Huber, M. Mulazzani, M. Leithner, S. Schrittwieser, G. Wondracek, and E. Weippl, "Social snapshots: digital forensics for online social networks," in *Proceedings of the 27th annual computer security applications conference*. ACM, 2011, pp. 113–122.

[35] P. Andriotis, Z. Tzermias, A. Mparmpaki, S. Ioannidis, and G. Oikonomou, "Multilevel visualization using enhanced social network analysis with smartphone data," *International Journal of Digital Crime and Forensics*, vol. 5, no. 4, pp. 34–54, December 2013.

[36] C. Perez, B. Birregah, and M. Lemercier, "A smartphone-based online social network trust evaluation system," *Social Network Analysis and Mining*, vol. 3, no. 4, pp. 1293–1310, 2013.

[37] A. Yazidi, O.-C. Granmo, and B. J. Oommen, "Learning-automaton-based online discovery and tracking of spatiotemporal event patterns," *Cybernetics, IEEE Transactions on*, vol. 43, no. 3, pp. 1118–1130, 2013.

[38] P. Andriotis, G. Oikonomou, and T. Tryfonas, "Forensic analysis of wireless networking evidence of android smartphones." in *WIFS*, 2012, pp. 109–114.

[39] S. Heymann, M. Bastian, M. Jacomy, C. Maussang, A. Rohmer, J. Bilcke, and A. Jacomy, "Gexf file format," http://gexf.net/format/index.html, accessed Februay 16, 2014.

[40] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: an open source software for exploring and manipulating networks." in *ICWSM*, 2009, pp. 361–362.

[41] Y. Hu, "Efficient, high-quality force-directed graph drawing," *Mathematica Journal*, vol. 10, no. 1, pp. 37–71, 2005.

[42] C. Wang, Y. Zhang, X. Chen, Z. Liu, L. Shi, G. Chen, F. Qiu, C. Ying, and W. Lu, "A behavior-based sms antispam system," *IBM Journal of Research and Development*, vol. 54, no. 6, pp. 3–1, 2010.

[43] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[44] T. A. Almeida, J. Maria Gomez, and A. Yamakami, "Contributions to the Study of SMS Spam Filtering: New Collection and Results," in *Proc. 2011 Symposium on Document Engineering (DOCENGd 2011)*, 1515 Broadway, New York, NY 10036-9998 USA, 2011, pp. 259–262.

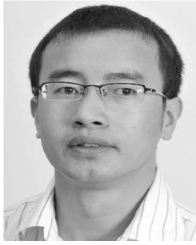
[45] S. Rosenthal, P. Nakov, A. Ritter, and V. Stoyanov, "Semeval-2014 task 9: Sentiment analysis in twitter," *Proc. SemEval*, 2014.

[46] U. K. Bindl, S. K. Parker, P. Totterdell, and G. Hagger-Johnson, "Fuel of the self-starter: How mood relates to proactive goal regulation." *Journal of Applied Psychology*, vol. 97, no. 1, p. 134, 2012.

Panagiotis Andriotis received the BSc degree in Mathematics from the National Kapodistrian University of Athens, Greece, in 2006, and the MSc degree with Distinction in Computer Science from the University of Bristol, U.K., in 2012. He is currently a PhD student at the University of Bristol. His research interests include Digital Forensics with a special focus on the Android OS, data mining, data hiding, steganalysis, social network analysis and he is also interested in the human aspects of Information Security, Privacy and Trust. Previously, he worked in the banking sector and also as a Mathematics teacher in Greece.

George Oikonomou received the MSc and PhD degrees in computer science from the Athens University of Economics and Business, Athens, Greece, in 2002 and 2009 respectively. Previously, he worked as a Research Associate at the Computer Science department, Loughborough University, UK. He is currently a Research Associate with the Faculty of Engineering at the University of Bristol and a member of the Cryptography research group. His current research focuses on digital forensics for emerging technologies, such as smartphones, wireless sensor networks and the Internet of things. He is also interested in security and IPv6 networking for low-power, severely constrained devices. Dr Oikonomou is an active developer of the Contiki open source embedded operating system for the internet of things.

Theo Tryfonas is an expert in Cybersecurity and Systems engineering with research work focused on assurance and resilience of critical infrastructures including transportation, utilities, healthcare and government. He worked in particular on systems for maritime safety and port security, public transport security, protection of UAVs, telecom revenue and system assurance, information security risk analysis as well as assisted in the investigation of computer crimes. His current interests extend to modelling cyber-capability with system dynamics and applications of game theory to the analysis of cyber attacks. He is a Chartered IT Professional member of the BCS and a Certified Information Systems Auditor.



Shangcan Li received the BSc and MSc degrees in mechanical engineering and PhD degree in computer science from Xi'an Jiaotong University, China, in 2001, 2004 and 2008, respectively. He is currently a lecturer at the Edinburgh Napier University and a member of the Institute for Informatics & Digital Innovation. His current research interests include mobile security, wireless sensor networks, Internet of Things, and applications of wireless technologies.