An Extensible Platform for the Forensic Analysis of Social Media Data

Huw Read[1,2(✉)], Konstantinos Xynos[1], Iain Sutherland[1,4], Frode Roarson[1,2], Panagiotis Andriotis[3], and George Oikonomou[3]

[1] University of South Wales, Pontypridd, CF37 1DL, UK
huw.read@southwales.ac.uk
[2] Noroff University College, 4608 Kristiansand S, Vest Agder, Norway
[3] University of Bristol, Bristol, BS8 1TH, UK
[4] Security Research Institute, Edith Cowan University, Perth, Australia

Abstract. Visualising data is an important part of the forensic analysis process. Many cell phone forensic tools have specialised visualisation components, but are as of yet able to tackle questions concerning the broad spectrum of social media communication sources. Visualisation tools tend to be stove-piped, it is difficult to take information seen in one visualisation tool and obtain a different perspective in another tool. If an interesting relationship is observed, needing to be explored in more depth, the process has to be reiterated by manually generating a subset of the data, converting it into the correct format, and invoking the new application. This paper describes a cloud-based data storage architecture and a set of interactive visualisation tools developed to allow for a more straightforward exploratory analysis. This approach developed in this tool suite is demonstrated using a case study consisting of social media data extracted from two mobile devices.

Keywords: Visualisation · Social media · Digital forensics · Mobile device

1 Introduction

Visualisation is a subject of growing importance to the field of computer forensics, in particular when dealing with large volumes of data from certain sources, such as social media sites. In an age where personal digital communication devices that can house information about our whereabouts, our conversations, or even current emotions are common [1], there are a multitude of ways that this information can be of value to a forensics investigation. Providing varied information on the different facets of an individual's activity [2]. The data sets can be extensive considering some of the new devices providing monitoring of personal fitness or health. It may be possible to query personal digital devices and social media for a wide range of data: What on-line names do they use? What information was sent to their contacts or received from their contacts? Even who had an elevated heart rate? The volume of data is further complicated by the process of trying to identify a conversation traversing multiple communication mechanisms, Facebook, Twitter, WhatsApp, iMessage, email, and SMS across multiple devices.

Existing research has made some progress in addressing this issue [3–6]. However the difficulty currently faced by forensic examiners is that the process involves extensive data sets, is time-consuming and requires a degree of manual processing. There are current tools that are able to visualise aspects of the data as part of a digital forensics investigation. Current visualisation tools provide only limited interactivity and an expert understanding of different visualisation tools, often with different input formats.

This paper presents an architecture that encourages direct interaction with social media data in the visualisations to compliment the exploratory and investigative mindset of an investigator. Social media application data is stored in an abstracted fashion into cloud-based storage and a defined API provides the interface for different visualisation tools to make requests for data. Visualisation tools retrieve data bound by their inherent data types and can send data directly to other visualisation tools depending on the pairing of compatible input/output data types. The key contributions of this work are summarised below:

- We present a visualisation tool framework encouraging data exploration between tools.
- We demonstrate how heterogenous social media data sources can be unified and exploited by investigators.
- We highlight a novel way of recording the visualisation trail to store the thought process of an investigator.

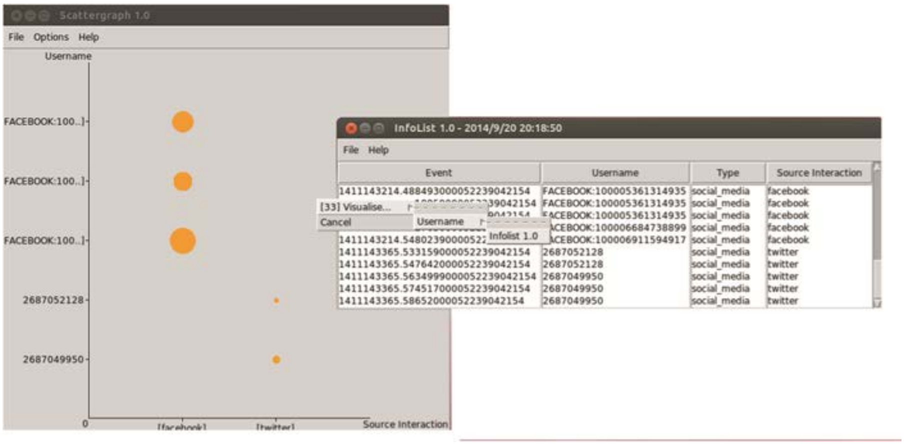
The rest of this paper is arranged as follows. Firstly, related work in the area of cell phone forensics visualisation is considered, secondly the architecture of the platform is described, thirdly a case study is presented to highlight the advantages of the platform, fourthly the outcome of the case study is analysed, finally the conclusions and future work are discussed.

2 Related Work

As stated by Garfinkel [7] forensic visualisations tend to serve two purposes, presentation and discovery. The former specialising in describing what has happened to a courtroom, the latter helping an investigator reach conclusions about some criminal activity. Discovery tools should be [7] data driven, have fixed, predictable, static output and interactive for producing the visualisation. In addition to [7] we should also be seeking to record the data-mining activities the investigator uses in order to retrace through the thought/decision making process (Fig. 1). The visualisations should indeed be repeatable so prosecution and defence can view the same output however a means of visualising how an investigator derived their conclusions would also be beneficial when describing the process to a jury for example.

Social media should be an area of considerable interest to the forensics investigator as it provides an insight into user actions and activities although these need to be interpreted with care [8].

Mutawa [9] has suggested when considering smartphones and social media that these devices are "…a goldmine for forensic investigators" and demonstrated that social



Event	Username	Type	Source Interaction
1411143214.488493000052239042154	FACEBOOK:100005361314935	social_media	facebook
39042154	FACEBOOK:100005361314935	social_media	facebook
9042154	ACEBOOK:100005361314935	social_media	facebook
	ACEBOOK:100006684738899	social_media	facebook
1411143214.5480239000052	ACEBOOK:100006911594917	social_media	facebook
1411143365.533159000052239042154	2687052128	social_media	twitter
1411143365.547642000052239042154	2687052128	social_media	twitter
1411143365.5634999000052239042154	2687049950	social_media	twitter
1411143365.574517000052239042154	2687049950	social_media	twitter
1411143365.58652000052239042154	2687049950	social_media	twitter

Fig. 1. A scattergraph plot sends a list of usernames from gathered social media data to a listing tool

media activity is retained on a number of cell phones, while recognising that this type of communication takes place across multiple devices [9]. A number of efforts have been made to examine the connections between mobile devices to better understand user activities [3]. Other tools have been applied to the visualization of social media including NodeXL [10] an extension of Microsoft Excel to enable a ‘network graph’ visualisation of data. There have also been efforts made into visualizing other aspects of social networks [4] and visualising the results. There are therefore existing tools, techniques and solutions of both commercial and open source forensics tools that can be applied to an investigation involving social media data. There are also a number of big data tools that can be used for social media analysis, including those tools used for e-discovery [11–15]. The tool proposed in this paper bridges the gap between these tools providing an extensible social media analysis tool capable of visualisations that are appropriate for forensic analysis.

2.1 Cell Phone Forensics Visualisations

There are existing tools available for forensic visualisation; these can provide some visual representation of the data. How can we visualise cell phone social media communications? Commercial software is already looking at social-based data. However, as can be seen below, commercial applications are not yet able to visualise much social networking information, instead relying on more manual query languages to obtain answers in investigations. Social links tend to be made from more traditional digital communications, emails between individuals, SMS/MMS messages, telephone calls, etc. Open source software tends to follow a more generic approach, towards visualising any structured data rather than anything focusing on a specific niche like cell phone forensics.

The following considers the state of cell phone forensics and how recovered social media data is being incorporated into visualisations.

2.1.1 AccessData MPE+

MPE+ is the mobile phone forensics application from AccessData, producers of FTK and FTK Imager. MPE+ itself specialises in data extraction from mobile phones [16], and has its own visualisation component [17]. It uses SMS, MMS, and call histories to look for relationships, and has a social analysis that looks for relationships via email addresses. Mention is made of an SQL Builder feature, which allows investigators to run custom queries on app data, but nothing specific is mentioned about correlations from social media sources.

2.1.2 Micro Systemation XRY and XAMN

Micro Systemation focuses on the data extraction with XRY and the visualisation with XAMN [18]. XAMN provides a number of different views including timeline, list, connections and geographic. Although XRY provides the means to extract application data, XAMN does not currently visualise social media communications.

2.1.3 CelleBrite UFED Link Analysis

CelleBrite provides presentation and discovery visualisations through the Link Analysis software [19]. Link Analysis provides the means to view commonalities across multiple cell phones in terms of phone calls, SMS, MMS, email, chat and a few others. Social media communications do not presently feature in Link Analysis.

2.1.4 Oxygen Forensics

Oxygen Forensics suite contains several different visualisation tools [20], timeline, web connections & locations, links & stats, but of particular interest is the Social Graph. The Social Graph looks at relationships between device owners and their contacts, i.e. how much time was spent using different communication mechanisms. Metrics include call length, number of messages sent and times of communication [20].

2.1.5 Open Source Software (OSS) Cell Phone Tools

OSS tools tend to concentrate on the extraction of data from the standard sources (call history, SMS, MMS, email [21]) and then on an application-by-application basis (e.g. WhatsApp [22], Skype [23]). Some of the tools generate lists as a visual aid for presentation. There are a number of OSS tools that lend themselves expertly to big-data social network visualisation [24] but are not designed for the type of focus required of the types of digital forensic investigations described in this paper.

To the best of the authors' knowledge no one has addressed the discovery side of visualisation for cell phone forensics in the open source community with a focus on the analysis and correlation of social media activity.

3 Architecture and Implementation

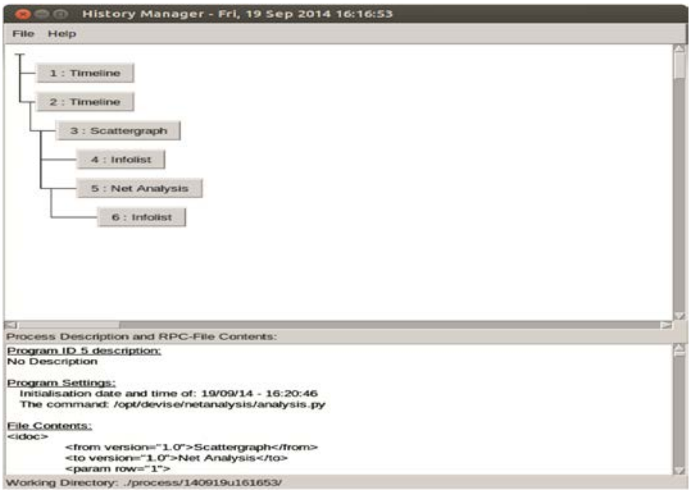
The architecture of the platform discussed in this paper is introduced below, highlighting the different core components of the storage, tool-to-tool data exchange and the audit process of the examiner.

3.1 Cloud Storage

SQlite is a database format found on most cell phones, as it is portable (i.e., store as files) and provides the advantages that come with relational database architectures. Mobile applications on Google's Android and Apple's iOS primarily make use of SQlite.

One of the main issues with these relational databases is the way the information is stored. This is usually in a standard format, as per the SQlite specification, although the actual table and column layouts are bespoke on the needs of the application. This poses a challenge to investigators and any tools they may rely upon. Applications are updated very frequently and it would not take much for a database design to change overnight.

In order to support the tools that have been developed, the database design proposed in [5, Fig. 2] provides a unifying database design for storing social media information. Individual parsers for Twitter and Facebook were created to convert the SQlite information into XML and then these were transformed then into the relevant social media XML format which is inserted into the cloud-based data store. Apache Cassandra was used to store the information and an XML middleware was then used to store and extract information when required by the visualisations.



```
Process Description and RPC-File Contents:
Program ID 5 description:
No Description

Program Settings:
   Initialisation date and time of: 19/09/14 - 16:20:46
   The command: /opt/devise/netanalysis/analysis.py

File Contents:
<idoc>
            <from version="1.0">Scattergraph</from>
            <to version="1.0">Net Analysis</to>
            <param row="1">
Working Directory: ./process/140919u161653/
```

Fig. 2. History Manager records an investigators progress.

```
<?xml version="1.0" encoding="UTF-8" ?>
   <idoc>
       <from version="1.0">Scattergraph</from>
       <to version="1.0">Net Analysis</to>
       <param row="1">
           <object name="From Account Relationship">
               <data type="from_account_relationships">1347865813.632551000052239042154</data>
           </object>
           <object name="To Account Relationship">
               <data type="to_account_relationships">1347865813.632551000052239042154</data>
           </object>
       </param>
</idoc>
```

Fig. 3. Recording the visualisation progression can reveal much about an investigation when expert testimony is required.

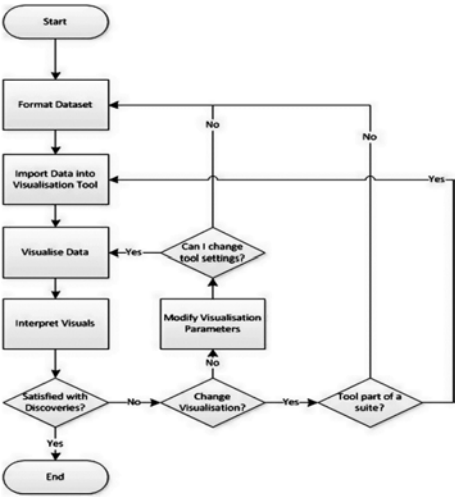

Start
Format Dataset
No
No
Import Data into Visualisation Tool
Yes
Visualise Data
Yes
Can I change tool settings?
Interpret Visuals
Modify Visualisation Parameters
No
Satisfied with Discoveries?
No
Change Visualisation?
Yes
Tool part of a suite?
Yes
End

Fig. 4. Rapid visual investigation: Timeline (Top Left) to Scattergraph (Top Right) to InfoList (Bottom).

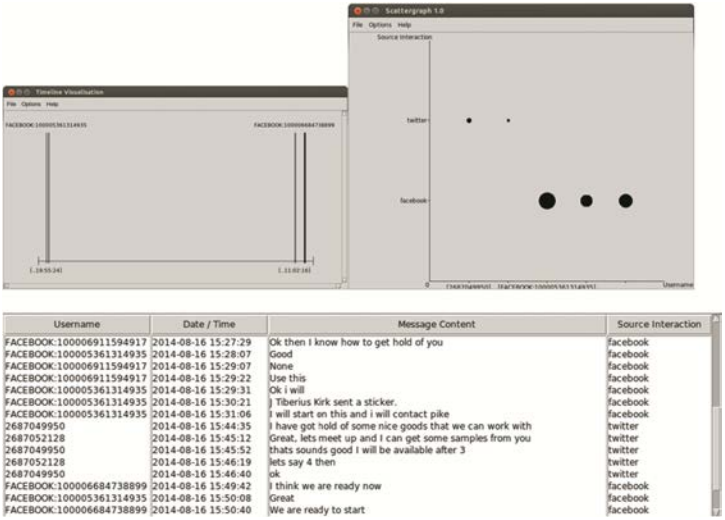
3.2 Interacting with Different Visualisation Tools

To facilitate data exchange between visualisations, the tools must incorporate the following components:

A configuration file (CDOC) that describes the quantity and type of data it can accept as input. These XML CDOCs are used whenever an investigator selects a subset of data within one visualisation tool. A list is presented to the investigator of other visualisation tools that can accept the subset as input.

An information file (IDOC) that is used to transport the data between tools.

It is an XML file that describes where the relevant data is in the underlying data source that the receiving tool needs to query. The structure stores the source visualisation tool, the destination visualisation tool, the type of information to be visualised, and the identifier of the row entry that stores the information. In the following sample, we can see two tools exchanging information relating to the relationships between social media accounts (Fig. 5).



Username	Date / Time	Message Content	Source Interaction
FACEBOOK:100006911594917	2014-08-16 15:27:29	Ok then I know how to get hold of you	facebook
FACEBOOK:100005361314935	2014-08-16 15:28:07	Good	facebook
FACEBOOK:100006911594917	2014-08-16 15:29:07	None	facebook
FACEBOOK:100006911594917	2014-08-16 15:29:22	Use this	facebook
FACEBOOK:100005361314935	2014-08-16 15:29:31	Ok i will	facebook
FACEBOOK:100005361314935	2014-08-16 15:30:21	) Tiberius Kirk sent a sticker.	facebook
FACEBOOK:100005361314935	2014-08-16 15:31:06	I will start on this and i will contact pike	facebook
2687049950	2014-08-16 15:44:35	I have got hold of some nice goods that we can work with	twitter
2687052128	2014-08-16 15:45:12	Great, lets meet up and I can get some samples from you	twitter
2687049950	2014-08-16 15:45:52	thats sounds good I will be available after 3	twitter
2687052128	2014-08-16 15:46:19	lets say 4 then	twitter
2687049950	2014-08-16 15:46:40	ok	twitter
FACEBOOK:100006684738899	2014-08-16 15:49:42	I think we are ready now	facebook
FACEBOOK:100005361314935	2014-08-16 15:50:08	Great	facebook
FACEBOOK:100006684738899	2014-08-16 15:50:40	We are ready to start	facebook

Fig. 5. Sample IDOC XML.

With these incorporated, the actual exchange of data becomes a straightforward task for the forensic investigator. As in Fig. 1, an analyst selects the grouping or pattern they wish to investigate further, selects another visualisation tool that can receive such input from the library, and the new visualisation tool is invoked with the subset. The investigator can even create a new instance of the same visualisation tool (albeit with the reduced dataset) if they wish.

As the process of identifying compatible visualisation tools based on their data-types is performed in real-time, new visualisation tools can be integrated without closing down the entire system and disrupting the investigative process.

3.3 Recording the Thought Process

When a forensic practitioner is using visualisation as a mechanism to discover facts relating to an investigation, the process is typically unidirectional. Large datasets are visualised, groupings of interest are investigated, subsets are extracted for further visualisation until a discovery is made (Fig. 3). With the architecture, we log the exchange that takes place between visualisation tools to let an investigator reinvoke existing visuals at a later date. As mentioned above, each visualisation tool is started with an IDOC XML file that describes the location of the data the tool needs to request data from the cloud data store.

As the investigator moves from one visualisation to another, we store the IDOC files and use a GUI tool called the History Manager (Fig. 2) to visualise the investigator's trail. Any views the investigator had seen previously are stored with the IDOC. The History Manager provides the facility to the investigator to select an existing IDOC and launch the visualisation tool, with the same inputs, and produce the same visual output again. The History Manager simply stores the files and other descriptive information in a hierarchical directory structure, which can easily be stored and recalled for later use or archiving.

4 Case Study

There are a range of potential sources for social media data. Access to corporate servers may be a lengthy legal process leading investigators to consider other possible sources [5, 25]. Therefore the case study detailed in this paper uses data collected directly from the mobile phones.

A simple scenario was developed to test the system between three individuals typifying communication usage across several platforms such as instant messages, voice over IP and different devices. As part of the study the individuals also shared documents using Google drive and email. The simulated criminal activity is the selling of illegally obtained information and software. Much of the discussion between the phone owners take place over Facebook and Twitter; in particular making use of Twitter's direct messaging function and Facebook's Messenger application.

The investigator, as part of his research, wants to look into the social media communications to identify any evidence of nefarious activity. However, it appears the thread of communication occurs over several social media types. In particular the following are found during an investigation of applications found on the cell phones:

- com.twitter.android/db/xxxxxxxxxx.db
- com.facebook.orca/db/threads_db2

These SQLite files contain private messages between the phone owner and others. The investigator begins by ingesting the files of interest into the data store. Once these have been parsed, he opens the Timeline tool to obtain an appreciation of the timescales in the data.

The investigator then tries to look for correlations between the different usernames and social media formats found. To do this he interacts with the Timeline tool (Fig. 4, Top Left) by clicking the background, which initiates a search for other tools in the architecture that can take these visualised data-types as input. The Scattergraph tool is selected, and the dataset is transferred as an IDOC XML file. This is stored in the HistoryManager and can be accessed at a later date if required.

The Scattergraph tool (Fig. 4, Top Right) initially shows the two types it was provided (Date/Time and Usernames). The axes are altered to show the Source Interaction against the Usernames. This highlights how users sent messages via different social media formats. From this visualisation, the investigator now wants to look at the actual messages sent, across all social media types, ordered by time.

The Infolist application (Fig. 4, Bottom) provides a view that lets the investigator see the messages sent in time order. It includes all social media formats from this case study and is not focused purely on one type.

5 Case Study Results and Evaluation

The case study highlights the advantages of the architecture. An investigator is able to easily transition from one tool to another, and rapidly data mine to gain an understanding of what has happened in their investigation. It should be clear from the case study that the emphasis in this paper is on the facilitation of interactivity rather than any perceived novelty of the visualisations themselves.

The investigator has the ability to choose which visualisation tools they want to use when visually mining the data. In this fashion they are able to make discoveries more rapidly as the data mining process is effectively tailored to suit.

By looking for commonalities across social media formats we are able to process the data collectively. Using the right combination of visualisation tools conversation threads across different communication mediums can be aligned and interpreted in a more straightforward fashion than existing mechanisms.

The analysis process in the case study was kept short for brevity. However, true of both a quick triage and a full investigation, the History Manager stores all interactions between visualisation tools. The IDOCs are effectively a log or audit trail of an investigators visualisation thought process.

6 Conclusions

Visualisation of social media is of major importance to computer forensic investigators in an age where personal digital communication occurs over various applications on multiple platforms. Existing tools and techniques for analysis include a range of commercial and open source forensics applications and Big Data visualisation tools. This paper has highlighted the lack of forensics tools that are appropriate for interactively visualising social media from cell phones. This paper describes the testing of an open source platform capable of large scale forensics visualisation. Data is stored in an abstracted fashion into cloud-based storage and a defined API provides the interface for multiple interactive visualisation tools to access the data. Visualisation tools retrieve data and can send data to other tools. This platform supports a forensic investigators analysis by enabling multiple simultaneous interactive visualizations to be generated: While capturing the investigators actions to record and log the process of the investigation. Future work will concentrate on the possibility of integrating other third-party visualisations into the platform, and the exploration of other potential evidence sources in a unified fashion to provide a more holistic view of an investigation.

Acknowledgements. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement Number: HOME/2010/ISEC/AG/INT-002.

References

1. Samsung Galaxy S5. http://www.samsung.com/global/microsite/galaxys5/features.html
2. Cellebrite: Cellebrite' s outlook for the mobile forensics industry 2014. White Paper (2014). http://www.cellebrite.com/collateral/OUTLOOK_FOR_THE_MOBILE_FORENSICS_INDUSTRY_2014_WP.pdf

3. Catanese S.A., Fiumara, G.: A visual tool for forensic analysis of mobile phone traffic. In: Proceedings of the 2nd ACM workshop on Multimedia in Forensics, Security And Intelligence, pp. 71–76 (2010). ISBN:978-1-4503-0157-2, doi:10.1145/1877972.1877992, http://dl.acm.org/citation.cfm?id=1877992
4. Perer, A., Shneiderman, B.: Balancing systematic and flexible, exploration of social networks. IEEE Trans. Visual. Comput. Graphics **12**(5), 693–700 (2006). http://hcil2.cs.umd.edu/trs/2006-25/2006-25.pdf
5. Andriotis, P., Tzermias, Z., Mparmpaki, A., Ioannidis, S., Oikonomou, G.: Multilevel visualization using enhanced social network analysis with smartphone data. Int. J. Digit. Crime Forensics **5**, 34–54 (2013)
6. Andriotis, P., Tryfonas, T., Oikonomou, G., Li, S., Tzermias, Z., Xynos, K., Read, H., Prevelakis, V.: On the development of automated forensic analysis methods for mobile devices. In: Holz, T., Ioannidis, S. (eds.) Trust 2014. LNCS, vol. 8564, pp. 212–213. Springer, Heidelberg (2014)
7. Garfinkel, S.L.: Forensics Visualizations with Open Source Tools (2013). http://simson.net/ref/2013/2013-11-05_VizSec.pdf
8. Browning, J.G.: Digging for the digital dirt: discovery and use of evidence from social media sites. SMU Sci Tech L Rev. **14**, 465 (2010). http://heinonline.org/HOL/LandingPage?handle=hein.journals/comlrtj14&div=26&id=&page=
9. Al Mutawa, N., Baggili, I., Marrington, A.: Forensic analysis of social networking applications on mobile devices. In: Proceedings of the 2012 Digital Forensic Research Workshop. http://www.dfrws.org/2012/proceedings/DFRWS2012-3.pdf
10. Smith, M. A., Shneiderman, B., Milic-Frayling, N., Mendes Rodrigues, E., Barash, V., Dunne, C., Capone, T., Perer, A., Gleave, E.: Analyzing (social media) networks with NodeXL. In: Proceedings of the Fourth International Conference on Communities and Technologies, pp. 255–264, New York, NY, USA (2009). ISBN:978-1-60558-713-4, doi:10.1145/1556460.1556497, http://hcil2.cs.umd.edu/trs/2009-11/2009-11.pdf
11. Afentis Facebook Forensics Tool. http://www.facebookforensics.com/features.html
12. Forte, D., Power, R.: Electronic discovery: digital forensics and beyond. Comput. Fraud Secur. **2006**(4), 8–10 (2006)
13. Ringtail E-discovery Tool. http://www.ftitechnology.com/Products-Services/Software-and-Services/Ringtail/Ringtail.aspx
14. Attenex E-Discovery Software. http://www.ftitechnology.com/Products-Services/Software-and-Services/Attenex.aspx
15. Xera I-conect. http://www.iconect.com/
16. AccessData Mobile Phone Examiner. http://www.accessdata.com/solutions/digital-forensics/mobile-phone-examiner
17. AccessData, Mobile Device data Visualization with MPE+ (2012). https://www.youtube.com/watch?v=bjcLDjju-kU
18. Micro Systemation, XAMN (2014). https://www.msab.com/xry/xamn
19. Cellebrite, UFED Link Analysis (2014). http://www.cellebrite.com/mobile-forensics/products/applications/ufed-link-analysis
20. Oxygen Forensics, Social Graph Tool (2014). http://www.oxygen-forensic.com/en/features/analyst/social-graph
21. ViaForensics, Aflogical OSE (2014). https://viaforensics.com/resources/tools/android-forensics-tool/#aflogical-ose
22. Ztedd, Whatsapp Xtract: Backup messages extractor (2012). http://forum.xda-developers.com/showthread.php?t=1583021
23. Garronski, N.: Skype Xtrator v.0.1.8.8 (2014). http://www.skypextractor.com/

24. Forensics WIKI: Graph and (Social) Network Visualization (2013). http://www.forensicswiki.org/wiki/Tools:Visualization#Graph_and_.28Social.29_Network_Visualization

25. Mulazzani, M., Huber, M., Weippl, E.R.: Social Network forensics: tapping the data pool of social networks. In: Eighth Annual IFIP WG 11.9 International Conference on Digital Forensics (2012). https://www.sba-research.org/wp-content/uploads/publications/socialForensics_preprint.pdf