Digital Investigation 16 (2016) S55–S64

Contents lists available at ScienceDirect

Digital Investigation

journal homepage: www.elsevier.com/locate/diin

DFRWS 2016 Europe — Proceedings of the Third Annual DFRWS Europe

A method and a case study for the selection of the best available tool for mobile device forensics using decision analysis

Shahzad Saleem [a], [*], Oliver Popov [b], Ibrahim Baggili [c]

[a] School of Electrical Engineering and Computer Science, National University of Science and Technology, H-12, Islamabad, Pakistan
[b] Department of Computer and Systems Sciences, Stockholm University, Postbox 7003, SE-164 07, Kista, Sweden
[c] Tagliatela College of Engineering, University of New Haven, 300 Boston Post Road, West Haven, CT, 06516, USA

Keywords:
Digital forensics
Mobile device forensics
Mobile device forensics tools
Evaluation
Multi-criteria decision analysis
Digital evidence
Digital investigation
Expected utility
Total ranking
Hypothesis testing

ABSTRACT

The omnipresence of mobile devices (or small scale digital devices – SSDD) and more importantly the utility of their associated applications for our daily activities, which range from financial transactions to learning, and from entertainment to distributed social presence, create an abundance of digital evidence for each individual. Some of the evidence may be a result of illegal activities that need to be identified, understood and eventually prevented in the future. There are numerous tools for acquiring and analyzing digital evidence extracted from mobile devices. The diversity of SSDDs, types of evidence generated and the number of tools used to uncover them posit a rather complex and challenging problem of selecting the best available tool for the extraction and the subsequent analysis of the evidence gathered from a specific digital device. Failing to select the best tool may easily lead to incomplete and or improper extraction, which eventually may violate the integrity of the digital evidence and diminish its probative value. Moreover, the compromised evidence may result in erroneous analysis, incorrect interpretation, and wrong conclusions which may eventually compromise the right of a fair trial. Hence, a digital forensics investigator has to deal with the complex decision problem from the very start of the investigative process called preparatory phase. The problem could be addressed and possibly solved by using multi criteria decision analysis. The performance of the tool for extracting a specific type of digital evidence, and the relevance of that type of digital evidence to the investigative problem are the two central factors for selecting the best available tool, which we advocate in our work. In this paper we explain the method used and showcase a case study by evaluating two tools using two mobile devices to demonstrate the utility of our proposed approach. The results indicated that XRY (Alt_1) dominates UFED (Alt_2) for most of the cases after balancing the requirements for both performance and relevance.

Introduction

The overarching goal of this work is to help investigators select the best available tool for mobile device forensics. The selection is based on both the performance of the forensics tools and relevance of the digital evidence in solving

[*] Corresponding author. Tel.: +92 321 5051723.
E-mail addresses: shahzad.saleem@seecs.edu.pk (S. Saleem), popov@dsv.su.se (O. Popov), IBaggili@newhaven.edu (I. Baggili).

http://dx.doi.org/10.1016/j.diin.2016.01.008

or furthering a specific case. The outcome will facilitate proper extraction, valid analysis, correct interpretation, right conclusions and the increased possibility for a fair trial. The selection is based on a formal method called Multi Criteria Decision (MCD) analysis. Performance and relevance are the two factors for MCD analysis in our proposed work.

ICT facts and figures" released by ITU (International Telecommunication Union, 2012; International Telecommunication Union (ITU), 2013) indicate deep penetration and wide acceptance of mobile devices in our society. These devices are versatile in nature and are used for various extensive daily activities. Consequently, a user will leave traces of digital activities (digital footprints) whenever he/she interacts with a mobile device. These digital footprints transform the mobile device to a personal digital behavioral archive. These behavioral archives are typically important to an investigation because they not only reveal digital evidence but behavioral patterns of its user as well. Moreover, around 80% of court cases have digital evidence linked to them (Rogers, 2004; Baggili et al., 2007). In the past years dozens of murder cases have been settled with the help of digital evidence found on the murderer's and or victim's mobile devices (Baggili et al., 2007).

The forensics community has appreciated the importance of mobile devices by acknowledging a separate branch of digital forensic science called “Mobile Device Forensics” (Casey, 2011). Private sector has also responded by developing numerous dedicated tools to perform mobile device forensics.

The problem however, is that the number of forensics tools is quite large and their performance varies for different types of digital evidence. For example one tool will perform better for recovering SMS while the other will perform better for recovering standalone files. Therefore, during the preparatory phase it becomes difficult for an investigator to select the best available tool. Therefore, as a general guideline, experienced digital forensic scientists and examiners typically cross-validate their results by using a variety of tools, which in turn leads to longer investigative time.

Preservation and protection are the two umbrella principles stipulated by the extended abstract process model with 2PasU (Saleem et al., 2014a). Selection of the best tool is one of the requirements of the model during preparatory phase. Failing to select the right tool may easily lead to incomplete and or improper extraction, which eventually may violate the integrity of the digital evidence and diminish its probative value and hence admissibility. Moreover, the compromised evidence may lead to erroneous analysis, incorrect interpretation and wrong conclusions, with an eventual consequence of a compromise in the litigating party's right of a fair trial.

In the past, vendor evaluation results were the only results available for use when selecting appropriate tools for a particular investigative scenario. The National Institute for Standards and Technology (NIST) realized the need for evaluating the forensics tools as an independent third party. Therefore, they published Smartphone Tool Specification (National Institute of Standards and Technology (NIST), 2010a) and Smartphone Tool Test Assertions and Test Plan (National Institute of Standards and Technology (NIST), 2010b). Later on, NIST used these specifications and test plans to evaluate forensics tools. Evaluation reports were published on the NIST website (National Institute of Standards and Technology (NIST), 2013) with free public access.

NIST has evaluated the forensics tools by using different mobile devices. So, the evaluation results cannot be generalized and used to compare different forensics tools. To solve this problem the same mobile devices were used to evaluate different forensics tools and the results were published in Kubi et al. (2011). But the comparison in Kubi et al. (2011) was not formal and automatic. The evaluation process was moved further to formally compare the forensics tools by using quantitative analysis (Saleem et al., 2013).

In Saleem et al. (2013) the tools were formally compared only with respect to their performance. Every type of digital evidence is equally important and relevant in a given scenario was the underlying assumption. However research has illustrated that different types of digital evidence extracted out of a mobile device are not equally relevant to understand and solve the case at hand (Saleem et al., 2014b). The work presented in this paper extends the prior research and proposes a formal method for to select the most appropriate tool for a particular investigative scenario. It is based on multi-criteria decision analysis with performance and relevance as the two critical factors. We further present as a case study two forensics tools that were evaluated with the help of two mobile devices to demonstrate the utility of our proposed formal method.

Performance measurements of nineteen potential sources of digital evidence were already published (Kubi et al., 2011). These measurements provided the base to connect the alternative forensics tools while building the MCD model.

Relevance is the second factor for the MCD model. It is case dependent, e.g. an SMS can be more important than call logs in one case type and vice versa in another. Our work uses seven different types of investigations having an associated digital side, as identified by Maxwell (Anobah, 2013). With that said, the method we present is extensible to inherit other types of crimes and is not limited to the seven types used when writing this paper.

This research actually builds on the idea presented in a short paper (Saleem and Popov, 2014), measures the factor of relevance by conducting a survey and concludes by producing the formal results. Relevance, in the survey, was measured on a linear scale from zero to ten points. It provided a formal association of the relevance factor to each type of digital evidence. This association was the final necessary prerequisite to build the MCD model and to perform the subsequent analysis.

Expected value graphs at different levels of contraction, cardinal and total rankings were computed using the MCD model with the help of an in house developed tool called DecideIT (Preference AB). Visual representation of the results in the form of graphs and charts helped in an obvious and formal selection of the best tool for a particular type of investigation. Theoretical and mathematical background of decision analysis, MCD model, total and cardinal ranking and expected utility graphs are discussed in the next section.

Theoretical background

Cogito ergo sum (I think, therefore I am) is René's first principle of philosophy (Dupre and Mansfield, 2007), with which he concluded the existence of free will in a non-deterministic world. We can choose from different alternatives with different consequences by this free will.

The freedom of choice necessitates a sense of responsibility that asks for discrimination between right and wrong. Sometimes the discrimination process for choosing the right thing is complex with major consequences – such as selecting the right forensics tool. Such a case necessitates for a structured formal approach before selecting the best investigative tool and "Decision Analysis" is one of these approaches (Preference, 2011).

Decision analysis

The term "decision analysis" was coined by Professor Ronald A. Howard (Howard et al., 1966). It is based on subjective probability and the appropriate measure of preference under uncertainty or subjective utility (Preference, 2011; Miles et al., 2007). Foundations of subjective probability were laid down by Ramsey (Ramsey, 1931) and Finetti (De Finetti, 1931; De Finetti, 1937) while Neuman, Morgenstern (Von Neumann and Morgenstern, 1947) and Savage (Savage, 1972) provided the basis of appropriate measure of preference and subjective utility (Miles et al., 2007).

Probability theory

Earlier, the forensics experts were selecting the tools on the basis of heuristics (based on their experience) about the performance of the forensics tools. We posit that most examiners select the forensics tools without following a formal quantification method of performance and relevance. Thus they were acting under uncertainty.

In our work, performance is measured in terms of probability of successful extraction of a particular type of digital evidence by a specific forensics tool using the equations below:

$$P(S) = p_S,\ \text{Where } \mathrm{p_S} \in [0, 1] \tag{1}$$

$$p_{S=}\frac{x}{n} \tag{2}$$

Where:

x = total number of objects extracted successfully
n = total number of objects (of that type) populated

Subsequently, these proportions were used to perform hypothesis testing (Saleem et al., 2013) to relate the performance of the forensics tools together. Relational comparison helped to connect the forensics tools to each criterion in the MCD model.

Utility theory

Saint Petersburg's Paradox (Weiss, 1987; Translated from Die Werke von Jakob Bernoulli, 2013) is a problem involving a theoretical lottery game with an infinite expected value/payoff as indicated by Equation (3) (Preference, 2011). In reality the game will yield only a small amount of payoff. Denial Bernoulli (1738) solved the paradox by introducing a utility function based on a logarithmic function of the gambler's total wealth. This function implicitly has the notion of diminishing marginal utilities as shown in Equations (4) and (5) (Wikipedia).

$$E(MV) = \sum_{n=1}^{\infty} \frac{1}{2^n} * 2^n \tag{3}$$

$$U(w) = \ln(w),\ \text{where w} = \text{total wealth} \tag{4}$$

$$E(U) = \sum_{k=1}^{\infty} \frac{\ln\left(w + 2^{k-1} - c\right) - \ln(w)}{2^k} \tag{5}$$

Utility can be regarded as a measure of some degree of satisfaction, while the utility function maps the outcome w.r.t. that degree (Preference, 2011). Experts in the field were surveyed to capture their degree of satisfaction for the relevance of all the types of digital evidence in furthering or solving a specific case using Equations (3)–(5). In our prior work, we received 5772 responses from which the degree of satisfaction was captured in the form of weights (Saleem et al., 2014b). Weights were subsequently normalized in such a way that the sum of the weights of all the children of the same parent node in our MCD tree is equal to 1. Consequently the model was used for multi-criteria decision analysis. It is discussed in the following section.

Multi-criteria decision analysis

A decision maker is obliged to evaluate and balance multiple, usually conflicting, criteria and maximize the overall gain/output while making his/her decision. Our decision problem is also not different in terms of the requirements for evaluation and balance. Such a decision problem can be best studied by multi-criteria decision analysis (MCDA) because:

1. The cornerstones of the problem are uncertainties and utilities associated with different criteria (types of digital evidence) and alternatives (Forensics Tools).
2. An in house developed software (Preference AB) (DecideIT) to assist in evaluation and complex mathematical operations is freely available.

In this solution, individual utility functions are defined for each criterion to measure the utility of each alternative by Equation (6) (Sutinen, 2010).

$$u_i(x) = \frac{x - x_i^-}{x_i^+ - x_i^-} \tag{6}$$

Where: $\mathrm{U_i(x)}$ is the utility of "x" in criterion "i". "x" is a measured value and is the result of hypothesis testing with its corresponding z-score (Saleem et al., 2013). The individual utility was used to relationally connect the performance of the forensics tools for the particular criterion with the following expression (Equation (7)).

$$Alt1 > Alt2 + z_i \bigg/ \sum_{i=1}^{n} |z_i| \tag{7}$$

Where: z_i is the z-score computed for an alternative during hypothesis testing for a particular criterion (Saleem et al., 2013). For example if the quantitative analysis of criterion SMS indicated that Alt_1 is better than Alt_2 with 95% confidence and its z-score is 14.345 then the two alternatives were connected with the following expression.

$$XRY > UFED + 14.345 \bigg/ \sum_{i=1}^{n} |z_i| \tag{8}$$

Relational connection in the form of $Alt_1 > Alt_2$ is vague in describing the level of preference and leaves the room for subjective interpretation. To solve this problem, normalized z-scores were used in conjunction with the simple relational equation as described in Equations (7) and (8). It further quantified the level of preference and gave us better granularity in its description.

In MCD analysis, cumulative utility is measured by combining all the individual utility functions. Therefore, an additive global utility function (Preference, 2011) was derived by aggregating all the individual utility functions using Equation (9).

$$U(x) = \sum_{i=1}^{n} w_i u_i(x) \tag{9}$$

Where: w_i is the weight representing the level of relevance of criterion "i". The weights were also normalized and thus hold the expression $\sum w_i = 1$. Moreover $u_i{:}X_i \rightarrow [0,1]$ is the individual utility of criterion "i" with its state space "X_i" ranging from zero to one. Individual utility of a criterion "i" is one for the best possible state and zero for the worst (Preference, 2011). Eventually these utility functions were used to model the problem with multi-criteria model (MCM).

Multi-criteria model

MCM was constructed in the form of a tree with all the criteria expressed as nodes. Each node carries one type of digital evidence (criterion) needed to be evaluated against performance and relevance. Its graphical representation can be seen in Saleem and Popov (2014).

Each node has an individual utility function measuring the performance of different alternatives using hypothesis testing. The edges between the nodes carry their respective weights. Global additive utility function adds all the individual utility functions using these weights.

DecideIT only shows multiple criteria and their associated weights while hiding the details about the performance of alternatives and their connections thus making the tree neat and simple for better understanding (Saleem and Popov, 2014). Fig. 1 depicts the mental model of one node in the MCM with its full detail.

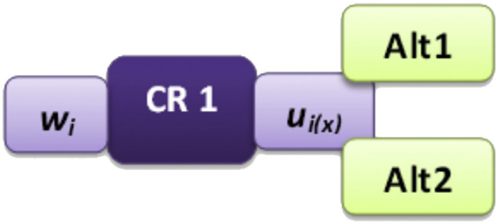

Fig. 1. Mental model of a criterion with its associations.

Where: *CR 1* is the first criterion, w_i is the weight assigned to *CR 1*, depicting its relevance. Alt_1 and Alt_2 represent alternative forensics tools *1* and *2* respectively. u_i is the individual utility function relationally tying the performance of both the forensics tools to *CR 1*. The following section discusses the methodology for capturing and tagging both the factors of performance and relevance with each criterion.

Performance and relevance

Nineteen types of digital evidence were identified in Kubi et al. (2011). Therefore we used nineteen criteria to be evaluated and balanced against performance and relevance.

Performance

Performance can be measured from historical data or from the results of carefully designed experiments. Historical data included performance evaluation results by both the vendors and a trusted third party. The problems however were that: (i) vendor evaluation lacked trust and (ii) trusted third party's evaluation used different mobile devices to evaluate the forensics tools. The tools were not evaluated on equal grounds and thus the results cannot be generalized for comparing their performance.

Experiments were designed to formally evaluate the performance of the tools by using the same mobile devices (Kubi et al., 2011; Saleem et al., 2013). The solution presented here extends these performance measures, normalizes them using Equation (7) and computes the individual utility function to relationally connect the alternatives with each criterion (SMS: $Alt_1 > Alt_2 + z_i$).

The results are summarized in Tables 1 and 2, which is actually an extension to the results discussed in Saleem et al. (2013).

Relevance

Every criterion has a distinct level of importance in furthering or solving a particular type of case. Some important types of investigations having association with mobile device forensics were identified by Maxwell (Anobah, 2013), which include:

1. Drug Trafficking (DT)
2. Rape (RP)
3. Murder (MD)
4. Credit Card Fraud (CC)

5. Harassment (HMT)
6. Espionage/Eavesdropping (EE)
7. Child Pornography (CP)

Experts were surveyed to measure this level of relevance. A linear ***scale*** of eleven points was used as a way to order every criterion with respect to its relevance to a particular case. The scale starts with zero and ends at ten. The criterion is not relevant if it scores zero points and carries maximum relevance if it scores ten points. The difference between two points on the scale is constant. Moreover, if CR_1 scores "x" points and CR_2 scores "$2x$" points then it implies that CR_2 is twice as relevant as CR_1 (Saleem et al., 2014b).

Table 1
Relationally connecting performance of mobile device forensics tools using Xperia.[a,b]

ID	Criteria	Relational connection (Equation (7))
1	Phonebook/Contacts	Alt1 = Alt2
2	Calendar entries	Alt1 > Alt2 + 0.12472
3	Memo/Notes	Alt2 > Alt1 + 0.01156
4	Tasks/To-Do-Lists	Alt1 > Alt2 + 0.11483
5	SMS	Alt1 > Alt2 + 0.02105
6	EMS	Alt1 > Alt2 + 0.03867
7	MMS	Alt1 > Alt2 + 0.04998
8	Audio calls	Alt1 > Alt2 + 0.09732
9	Video calls	Alt1 = Alt2
10	Emails	Alt1 > Alt2 + 0.23780
11	URLs visited	Alt1 = Alt2
12	Bookmarks/Favourites	Alt1 > Alt2 + 0.12480
13	Audio	Alt1 = Alt2
14	Video	Alt2 > Alt1 + 0.09048
15	Graphics/Pictures	Alt2 > Alt1 + 0.08895
16	Word	Alt1 = Alt2
17	Excel	Alt1 = Alt2
18	PowerPoint	Alt1 = Alt2
19	PDF	Alt1 = Alt2

[a] Sony Xperia X1.
[b] Some columns in Tables 1 and 2 are hidden to make space for the required column(s). Hidden columns can be viewed in Saleem et al. (2013).

Table 2
Relationally connecting performance of mobile device forensics tools using Nokia.[a]

ID	Criteria	Relational Connection (Equation (7))
1	Phonebook/Contacts	Alt1 > Alt2 + 0.00648
2	Calendar entries	Alt1 > Alt2 + 0.07240
3	Memo/Notes	Alt1 > Alt2 + 0.05686
4	Tasks/To-Do-Lists	Alt1 > Alt2 + 0.06519
5	SMS	Alt1 = Alt2
6	EMS	Alt1 = Alt2
7	MMS	Alt1 > Alt2 + 0.08072
8	Audio calls	Alt1 > Alt2 + 0.10663
9	Video calls	Alt1 > Alt2 + 0.04106
10	Emails	Alt1 > Alt2 + 0.13805
11	URLs visited	Alt1 = Alt2
12	Bookmarks/Favourites	Alt1 = Alt2
13	Audio	Alt1 > Alt2 + 0.14452
14	Video	Alt1 = Alt2
15	Graphics/Pictures	Alt1 = Alt2
16	Word	Alt1 > Alt2 + 0.08313
17	Excel	Alt1 > Alt2 + 0.04004
18	PowerPoint	Alt1 > Alt2 + 0.07919
19	PDF	Alt1 > Alt2 + 0.08573

[a] Nokia Xpress Music 5800.

DecideIT can take weights in normalized form. We had the measurements of relevance from fifty-five respondents who were surveyed. The outliers were removed and then a weighted average of these levels of relevance for each criterion in the context of every type of investigation was computed (Saleem et al., 2014b). These weights were then normalized before being fed into DecideIT. Detailed discussion about the survey, methodology, outlier removal and the results can be found in Saleem et al. (2014b).

Table 3 represents the weights in light of our survey. Table 4 represents the normalized weights of all the digital evidence and Table 5 represents the normalized weights of all the classes of digital evidence for every type of digital investigation.

To save space, criteria IDs from Table 1 are used in Tables 3–5. Normalization was performed on two levels:

1. *Intra class*: Weight of individual digital evidence was divided by the total weight of all the types of digital evidence belonging to that particular class. Mathematically it is represented by Equation (10).

$$w_d = \frac{w_i}{\sum_1^k w_i} \tag{10}$$

Where:

w_d is the normalized weight of a digital evidence belonging to a specific class of digital evidence.
w_i is the weight of a digital evidence
k is the total number of the types of digital evidence in a particular class of digital evidence.

Table 3
Weights/relevance of digital evidence.

Criteria (ID)	DT	RP	MD	CC	HMT	EE	CP
1	9.56	9.08	9.64	8.55	9.51	9.32	8.82
2	6.30	6.13	8.48	6.88	7.11	7.51	6.08
3	6.31	4.93	7.92	7.23	6.85	7.79	5.98
4	5.83	4.44	7.03	6.85	6.41	7.49	5.31
5	9.68	9.33	9.68	8.84	9.84	9.16	9.05
6	8.83	9.03	9.17	8.21	9.59	8.58	9.03
7	7.62	7.51	8.20	7.26	8.59	7.80	8.16
8	9.09	8.77	9.23	8.03	9.50	9.37	7.95
9	6.36	6.84	6.82	5.92	7.97	7.47	7.38
10	8.65	7.46	8.87	8.82	9.38	9.13	9.08
11	6.20	5.47	7.36	8.44	6.84	8.39	9.28
12	5.30	4.38	6.18	8.03	6.11	7.55	9.18
13	5.42	5.87	6.00	5.69	7.41	8.67	6.08
14	7.04	7.65	7.13	5.92	8.00	8.50	9.61
15	8.77	9.11	8.56	7.36	9.00	8.79	9.83
16	4.35	3.58	5.38	7.29	5.11	7.92	5.95
17	4.98	2.90	4.93	7.64	3.00	7.63	5.03
18	3.11	2.27	4.35	5.05	3.45	7.11	5.64
19	3.57	2.66	5.00	6.00	3.21	7.58	4.97

Table 4
Normalized relevance of digital evidence for every type of digital investigation.

ID	DT	RP	MD	CC	HMT	EE	CP
1	0.34	0.37	0.29	0.29	0.32	0.29	0.34
2	0.23	0.25	0.26	0.23	0.24	0.23	0.23
3	0.23	0.20	0.24	0.25	0.23	0.24	0.23
4	0.21	0.18	0.21	0.23	0.21	0.23	0.20
5	0.37	0.36	0.36	0.36	0.35	0.36	0.35
6	0.34	0.35	0.34	0.34	0.34	0.34	0.34
7	0.29	0.29	0.30	0.30	0.31	0.31	0.31
8	0.59	0.56	0.57	0.58	0.54	0.56	0.52
9	0.41	0.44	0.43	0.42	0.46	0.44	0.48
10	1.00	1.00	1.00	1.00	1.00	1.00	1.00
11	0.54	0.56	0.54	0.51	0.53	0.53	0.50
12	0.46	0.44	0.46	0.49	0.47	0.47	0.50
13	0.26	0.26	0.28	0.30	0.30	0.33	0.24
14	0.33	0.34	0.33	0.31	0.33	0.33	0.38
15	0.41	0.40	0.39	0.39	0.37	0.34	0.39
16	0.27	0.31	0.27	0.28	0.35	0.26	0.28
17	0.31	0.25	0.25	0.29	0.20	0.25	0.23
18	0.19	0.20	0.22	0.19	0.23	0.24	0.26
19	0.22	0.23	0.25	0.23	0.22	0.25	0.23

2. *Inter class*: To calculate normalized weight of a class, Equation (11) was used.

$$w_c = \frac{\sum_1^k w_i}{\sum_1^n w_j} \tag{11}$$

Where:

w_c is the normalized weight of a class of digital evidence.
k is the total number of the types of digital evidence in a particular class of digital evidence.
n is the total number of digital evidence in all the classes i.e. $n = 19$

Evaluation

Evaluation starts after structuring the problem with the MCD model and estimating the required variables of performance and relevance. DecideIT converts point estimates of these variables into a range by a user-defined value using Equation (12). The variables are contracted over the range to understand the impact of each variable on the decision problem.

$$\left[p - \frac{p}{20}, p + \frac{p}{20}\right] \tag{12}$$

Table 5
Normalized relevance of classes of digital evidence for every type of investigation.

Class	DT	RP	MD	CC	HMT	EE	CP
PIM entries	0.22	0.21	0.24	0.21	0.22	0.21	0.18
Messages	0.21	0.22	0.19	0.18	0.20	0.16	0.18
Call logs	0.12	0.13	0.11	0.10	0.13	0.11	0.11
Emails	0.07	0.06	0.06	0.06	0.07	0.06	0.06
Internet history	0.09	0.08	0.10	0.12	0.09	0.10	0.13
Standalone files	0.17	0.19	0.16	0.14	0.18	0.17	0.18
Application files	0.13	0.10	0.14	0.19	0.11	0.19	0.15

In our case the evaluation process is composed of the visual representation of expected utility, cardinal and total rankings in the form of graphs. Graphs help in selecting the most appropriate forensics tool.

Expected utility graph

DecideIT, by default, takes 5% indifference interval and 20% evaluation steps to calculate cutting hull values of an expected utility graph. Two tools are being evaluated, so a pair wise analysis for the alternatives on their expected utility is performed (Fig. 2). Mathematically speaking, during evaluation, the values are not fixed but are scattered along the entire range obtained by Equation (12). So the expected utility yields a quadratic objective function of the following form (Preference, 2011).

$$EU(A_i) = p_{i1}v_{i1} + p_{i2}v_{i2}\ldots\ldots p_{in}v_{in} \tag{13}$$

DecideIT maximizes the expected utilities by varying the values of the variables on their range. Relative strengths of different alternatives are computed by using Equation (13) to evaluate, compare and rank different alternatives (Preference, 2011).

$$mid(\delta_{12}) = \frac{(max(\delta_{12}) + min(\delta_{12}))}{2} \tag{14}$$

Where:

$\delta_{12} = EU(A1) - EU(A2)$
$max(\delta_{12})$ means the difference of expected utilities of *Alt₁* and *Alt₂* when *Alt₁* is made as good as possible in relation to *Alt₂*. Where $max(\delta_{12}) \in [-1,1]$.
$min(\delta_{12})$ means the difference of expected utilities of *Alt₁* and *Alt₂* when *Alt₁* is made as bad as possible in relation to *Alt₂*.

Relative strength is shown by the middle line in the expected utilities graph (Fig. 2). To interpret the results we use the concept of ***dominance***. *Alt₁* ***strongly dominates*** *Alt₂* if $min(\delta_{12})>0$, it ***markedly dominates*** if $mid(\delta_{12})>0$ and finally it ***weakly dominates*** if $max(\delta_{12})>0$.

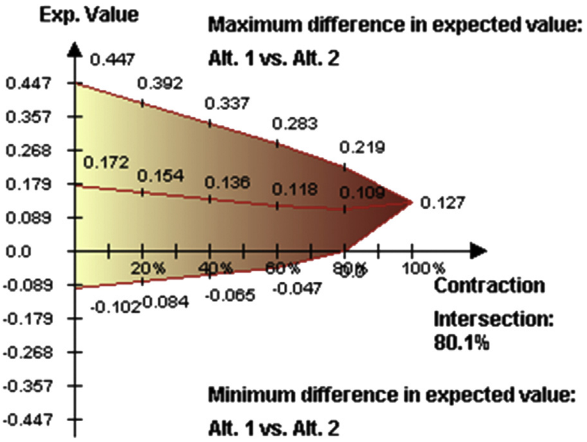

Fig. 2. Expected utility graph for Xperia X1 in credit card investigation.

Fig. 2 is an expected utility graph[1] for Xperia X1 while used for an investigation of credit card fraud by both the forensics tools. Both max(δ_{12}) and mid(δ_{12}) are greater than zero at zero percent contraction. So, at this level of contraction, we can say that Alt_1 is markedly better than Alt_2 since mid(δ_{12})>0. At 80.33% contraction level the hull is cut and all the max(δ_{12}), mid(δ_{12}) and min(δ_{12}) are greater than zero. So at this point we can say that Alt_1 strongly dominates Alt_2 since min(δ_{12})>0.

Total and cardinal ranking

Total ranking gives us an opportunity to concisely observe the direct relationship between different alternatives at a specified indifference interval. Indifference interval is the difference in terms of the percentage of expected utilities where one alternative is considered better than the other.

Alternative tools are evaluated, ranked and represented in a total ranking graph (Left in Fig. 3). All the alternatives are positioned vertically; starting from the highest ranked alternative placed at the top and the lowest ranked alternative placed at the bottom most position.

With cardinal ranking we can see the detailed relationship between different alternatives at a specific contraction level and contraction mode. Contraction level is the level of contraction of the range of the point estimates (Equation (12)). Contraction mode specifies the type of values that must be contracted. In normal contraction mode both the relevance weights and performance measures are contracted. Expected utilities for all the alternatives are computed at the specified contraction level and contraction mode.

The results are shown in the form of a column graph (Right in Fig. 3). This type of analysis also shows the overlapping areas (if any) of expected utilities i.e. the areas where different alternatives have similar utility.

Results and discussion

Evaluation was performed on all the fourteen MCD models, seven case types for both the mobile devices (7 × 2 = 14) using both the factors of performance and relevance. The evaluation was based on expected utility graphs, cardinal and total rankings. All the results represented different levels of detail to help in selecting the most appropriate tool. As a case study, we had two alternatives; so expected utility graphs were presented in this paper. It can compare two alternatives while comprehensively showing the relative strengths of the alternatives on the entire range of values. Total and cardinal rankings are more helpful when concise information is required for more than two alternatives. Figures below graphically represent the outcome.

In terms of performance Alt_1 performs better than Alt_2 in most of the cases as is evident from Tables 1 and 2. So, we can expect that Alt_1 will come out as a preferred choice for most of the types of digital investigations. The evaluation results of DecideIT also confirmed the same.

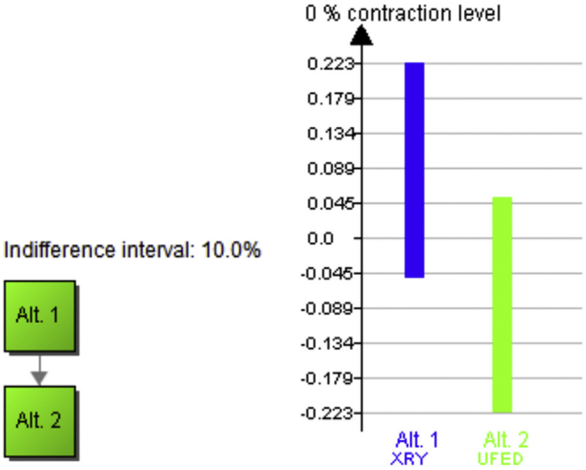

Fig. 3. Total (left) and cardinal (right) ranking for Xperia X1 in credit card investigation.

Fig. 4 indicates that Alt_1 strongly dominates Alt_2 for all the types of investigations and when Nokia 5800 is the device. min(δ_{12}) > 0 for all the consistent assignment of value variables, in this case.

Similarly Alt_1 moderately dominates Alt_2 for all the types of investigations when Xperia X1 is the device. In this case, mid(δ_{12}) > 0 for all the consistent assignment of the value variables. However, we also have an intersection point at around 80% level of contraction for all cases. At this point, a sub-set of decision frame, min(δ_{12}) > 0 for all the consistent assignment of value variables. It indicates that Alt_1 strongly dominates over Alt_2 at the intersection point for all the types of digital investigations.

Cardinal and total ranking is a better method to select the best tool from more than two alternatives. Both these methods will provide a concise visual representation to help in the selection of the best tool. The best tool will be placed at the top position in case of total ranking. The best tool in cardinal ranking will have a higher expected utility bar graph. Fig. 3 shows this phenomenon for our case study with two forensics tools.

The outcome of the survey to tag the factor of relevance for each type of digital evidence was a healthy seized data set. Relevance based best practices guide for mobile device forensics was proposed by generalization of the data set and results were published in Saleem et al. (2014b).

Conclusion and future work

From the results of our case study, we can conclude that Alt_1[2] strongly dominates over Alt_2 when Nokia 5800 has to be investigated for any type of digital investigation involving a mobile phone. While Alt_1 moderately dominated over Alt_2 when Xperia X1 has to be investigated for any type of mobile digital investigation. Alt_1 strongly

[1] DecideIT, in its graphs, uses the terms Expected Value to represent both expected utility and expected value.

[2] Alt1 = XRY 5.0 and Alt2 = UFED Physical Pro 1.1.3.8.

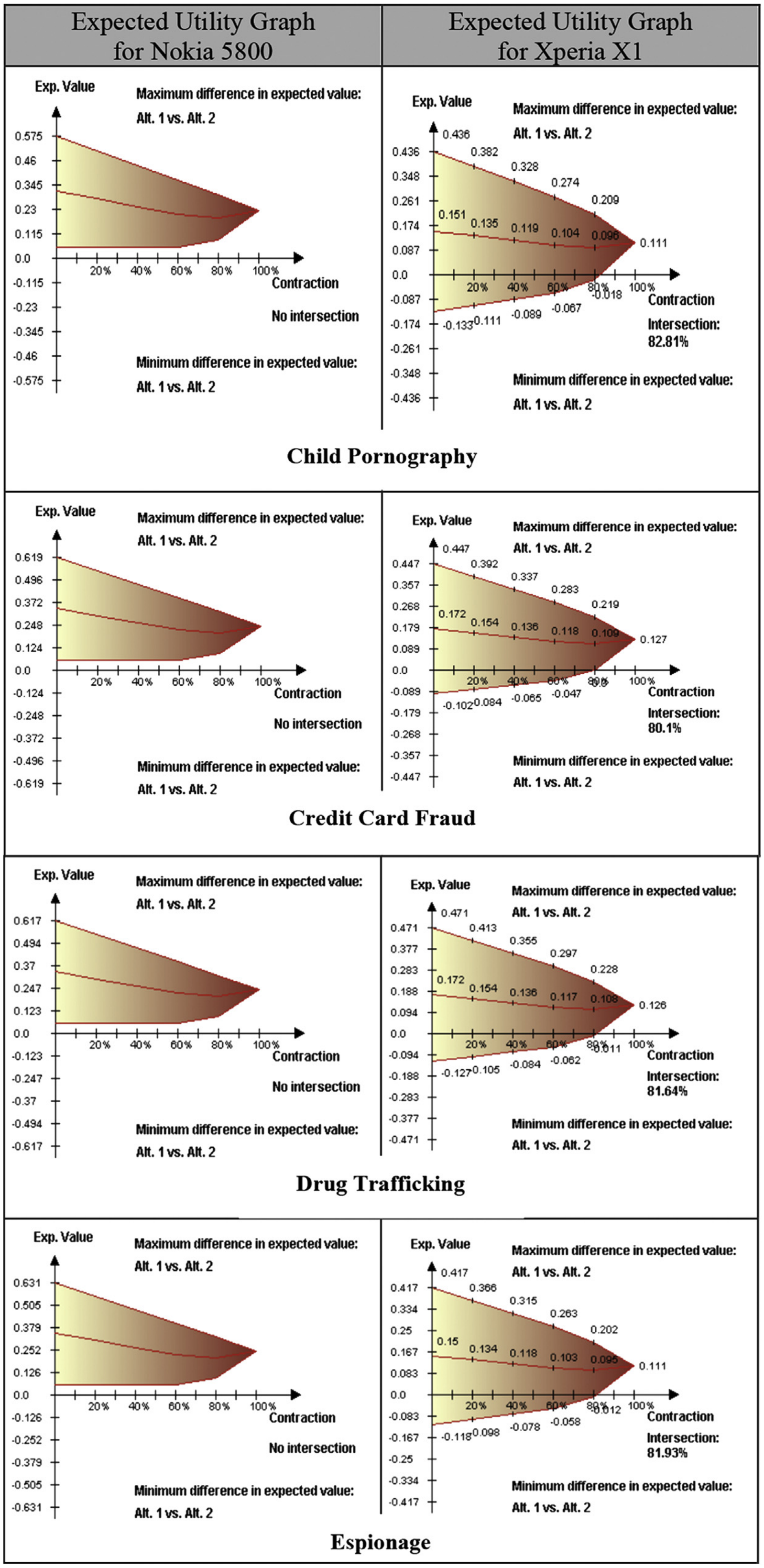

Fig. 4. Expected utility graphs for all the fourteen sets of evaluation.

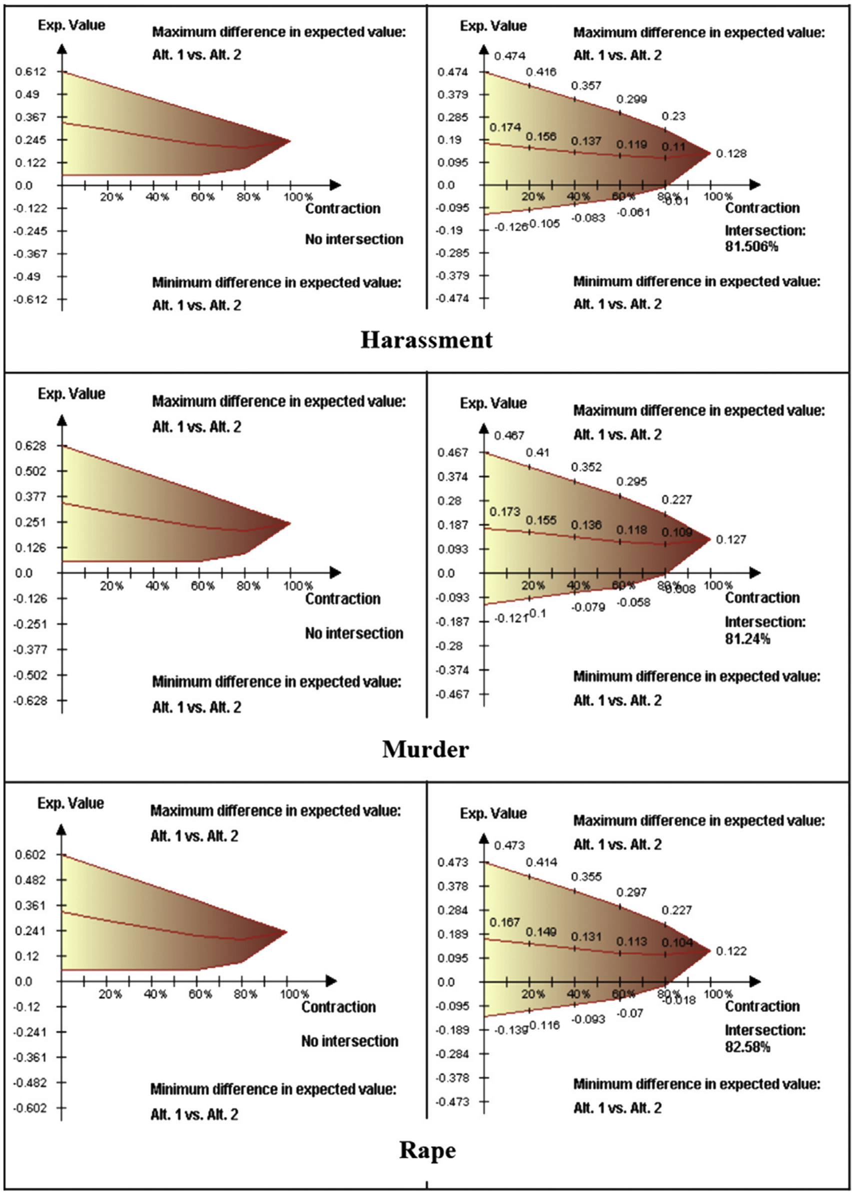

Fig. 4. (*Continued*)

dominates Alt_2 at the intersection point i.e. around 80% contraction level for all the types of investigations.

From Tables 1 and 2, it is evident that Alt_1 performs better than Alt_2 for most of the types of mobile digital evidence. Final results of our evaluation using DecideIT also showed that Alt_1 is dominant over Alt_2. Therefore, from these results, we can conclude that with the given measurements of performance and relevance, performance is more important in our current decision frame.

The purpose of this paper is to present a technique which is generic and firmly rooted in the decision theories, probability and utility. The study will certainly be extended to include more tools, devices and types of evidence. So, a project to evaluate performance of more tools against a broader set of mobile devices will be launched which will help in the selection of the most appropriate forensics tool for a particular scenario.

The wider set of evaluation results, balanced in terms of both performance and relevance can be used to create a reference manual. This reference manual can help in selecting the most appropriate forensics tool for a particular type of mobile device in a specific type of digital investigation during the preparatory phase of mobile device forensics process. It can help in proper extraction, better conclusions and appropriately holding the litigating party's right of a fair trial. Moreover, if the manual comes from an independent third party then the community can potentially place more trust in it.

Appendix

Excel sheets with raw survey data capturing the factor of relevance are publically available at http://www.unhcfreg.com/#!datasetsandtools/c18k6 The data contains details about the experts, all the responses, the details of ANOVA and normality testing and evaluation process in the form of fourteen (14) MCD models which can be used to generate expected utility, total and cardinal ranking graphs using DecideIT.

References

Anobah M. Testing framework for Mobile forensic investigation tools. Stockholm University; 2013.

Baggili I, Mislan R, Rogers M. Mobile phone forensics tool testing: a database driven approach. Int J Digit Evid 2007;6.

Casey E. Digital evidence. In: Digit. evid. comput. crime forensic sci. comput. internet. 3rd ed. 2011. p. 37–8.

De Finetti B. Sul significato soggettivo della probabilità. Fundam Math 1931;17:298–329.

De Finetti B. La prévision: ses lois logiques, ses sources subjectives. Institut Henri Poincaré; 1937.

Dupre B, Mansfield K. 50 philosophy ideas (You really need to know). Quercus; 2007.

Howard RA. Decision analysis: applied decision theory. In: B HD, J M, editors. 4th Int. Conf. Oper. Res. New York: Wiley-Interscience; 1966. p. 55–77.

International Telecommunication Union. Key statistical highlights: ITU data release June 2012. 2012.

International Telecommunication Union (ITU). ICT facts and figures. 2013.

Kubi A, Saleem S, Popov O. Evaluation of some tools for extracting e-evidence from mobile devices. In: Appl. Inf. Commun. Technol. Baku: IEEE; 2011. p. 603–8. http://dx.doi.org/10.1109/ICAICT.2011.6110999.

Miles Jr Ralph F. The emergence of decision analysis. In: Edwards W, Miles Jr Ralph F, Von Winterfeldt D, editors. Adv. decis. anal. from found. to appl. 1st ed. Cambridge University Press; 2007. p. 13–31.

National Institute of Standards and Technology (NIST). Smart phone tool specification, Version 1.1. 2010.

National Institute of Standards and Technology (NIST). Smart phone tool test assertions and test plan, Version 1.1, test. 2010.

National Institute of Standards and Technology (NIST). Computer forensics tool testing program: Mobile devices. 2013.

Preference AB. DecideIT User's manual. 2011.

Preference AB, Preference Calculated Risks - Rational Decisions, (n.d.).

Ramsey F. Truth and probability (1926). Found. Math. Other Log. ….. 1931.

Rogers MK. DCSA: a practical approach to digital crime scene analysis. 2004.

Saleem S, Popov O. Formal approach for the selection of a right tool for Mobile device forensics. In: Digit. forensics cyber crime lect. notes inst. comput. sci. soc. informatics telecommun. eng, vol. 132; 2014.

Saleem S, Popov O, Kubi A. Evaluating and comparing tools for Mobile device forensics using quantitative analysis. In: Digit. forensics cyber crime lect. notes inst. comput. sci. soc. informatics telecommun. eng, vol. 114; 2013. p. 264–82. http://dx.doi.org/10.1007/978-3-642-39891-9_17.

Saleem S, Popov O, Baggili I. Extended abstract digital forensics model with preservation and protection as umbrella principles. Procedia Comput Sci 2014a;35:812–21.

Saleem S, Baggili I, Popov O. Quantifying relevance of mobile digital evidence as they relate to case types: a survey and a guide for best practices. J Digit Forensics, Secur Law 2014b;9:19–50.

Savage LJ. The foundation of statistics. New York: Dover Publications; 1972.

Sutinen M. How to support decision analysis with software-case Förbifart Stockholm. Aalto University; 2010.

Translated from Die Werke von Jakob Bernoulli, Correspondence of Nicolas Bernoulli Concerning the St. Petersburg Game, (2013) 1–9.

Von Neumann J, Morgenstern O. Theory of games and economic behavior. 2nd ed. Princeton University Press; 1947.

Weiss MD. Measurable utility on the real line. In: Concept. found. risk theory (Technical Bull., U.S. Dept. of Agriculture, Economic Research Service; 1987. p. 31–49.

Wikipedia, St. Petersburg Paradox, in: Wikipedians (Ed.), Parad. Situations Which Defy Intuit., n.d.: pp. 80–88.