A Complexity Based Model for Quantifying Forensic Evidential Probabilities

Richard E Overill, Jantje A M Silomon

Department of Computer Science, King's College London,
Strand, London WC2R 2LS

{richard.overill, jantje.a.silomon}@kcl.ac.uk

K P Chow

Department of Computer Science, University of Hong Kong,
Pokfulam Road, Hong Kong

chow@cs.hku.hk

Abstract - **An operational complexity model (OCM) is proposed to enable the complexity of both the cognitive and the computational components of a process to be determined. From the complexity of formation of a set of traces via a specified route a measure of the probability of that route can be determined. By determining the complexities of alternative routes leading to the formation of the same set of traces, the odds indicating the relative plausibility of the alternative routes can be found. An illustrative application to a BitTorrent piracy case is presented, and the results obtained suggest that the OCM is capable of providing a realistic estimate of the odds for two competing hypotheses. It is also demonstrated that the OCM can be straightforwardly refined to encompass a variety of circumstances.**

Keywords - operational complexity model, Bayesian posterior probability, odds, digital forensics, evidential probability, competing hypotheses

1. Introduction and Background

The basis of an operational complexity model (OCM) has recently been given by Overill & Silomon [1]. The model is based on the general observation that an inverse relationship holds between the difficulty and/or intricacy involved in performing a task in a specified manner, as measured by its intrinsic complexity, and the likelihood that the task in question was in fact performed in the specified manner. In essence, the idea underpinning the model is that the more complex a process is, the less likely it is to occur.

Thus, in the context of a digital forensic examination, the operational complexity of formation of a set of digital evidential traces by a specified route should in principle be susceptible to 'bottom-up' *ab initio* determination. The resultant complexity should then be inversely related to the probability of formation by that route.

There are many definitions of complexity. Lloyd [2] gives several complexity measures that in principle permit the complexity of formation of a set of digital evidential traces $\{E_i\}$ to be defined. These metrics include computational complexity [3], information based complexity [4], logical depth [5], thermodynamic depth (and dive) [6] and crypticity [7]. The data available in the problem space of digital forensic analysis appears to be most closely aligned with the computational complexity metric [3] since the remaining candidate metrics require additional information or knowledge that is not normally available during a digital forensic investigation. In addition, it is desirable to include in the model a component relating to the human (i.e. cognitive) complexity of the task. The GOMS (Goals, Operators, Methods, Selections) family of models offers a well-understood approach to the problem. In particular, the GOMS Keyboard-Level Model (KLM) [8] provides a tractable means of measuring the human involvement in the operational process.

The development of the OCM offers forensic examiners and expert witnesses the possibility of computing the probabilities that a given set of recovered digital evidential traces was formed via a number of alternative (mutually exclusive) routes, each corresponding to a different hypothesis (or narrative) about how the traces were formed. These hypotheses are understood to be in competition with one another and their relative plausibility will be an important factor in deciding whether a successful prosecution can be brought.

2. The Operational Complexity Model

The resulting model may be formalized as follows. The various ***feasible routes*** by which the recovered set of digital evidential traces could have been formed are first enumerated. For each feasible route k by which the set of digital evidential traces $\{E_i\}$ could have been formed the operational complexity of that route is given by:

$$C_k = KLM_k + CC_k$$

978-0-7695-3965-2/10 $26.00 © 2010 IEEE
DOI 10.1109/ARES.2010.42

where C_k comprises a cognitive complexity component specified by the GOMS-KLM model [8] and a suitably defined computational complexity (CC) [3] component.
The operational complexity C_k of each feasible route k and its probability of occurrence p_k are taken to be inversely related:

$$p_k \propto C_k^{-1}$$

The constant of proportionality is determined uniquely by the normalization condition on the sum of the probabilities over all feasible routes k:

$$\sum p_k = 1$$

The constant of proportionality α reflects the units in which the complexity of each of the feasible routes k is measured, and is given by:

$$\alpha = (\sum C_k^{-1})^{-1}$$

It should be noted here that while the OCM model makes use of a complexity metric it is not based on Shannon information theory [9], which would lead to an inverse exponential relation:

$$p_k \propto 2^{-Ck}$$

Bayesian statistics [10] (which deal with the probability of an event A given an event B) have been applied previously to generating plausible crime scenarios [11]. However, they have only recently been employed in a digital forensic analysis context [12], where a Bayesian network model was used to study the problem of missing evidence.

The ***posterior probability*** of a feasible route k leading to the formation of a recovered set of digital evidential traces $\{E_i\}$ is given by $\Pr(H_k|\{E_i\})$, where H_k represents the hypothesis that feasible route k was taken. The ***posterior odds*** for two alternative routes k and k' leading to the formation of the same recovered set of digital evidential traces $\{Ei\}$ is then given by:

$$O(k{:}k') = \Pr(H_k|\{E_i\}) / \Pr(H_{k'}|\{E_i\})$$

Note that the odds are independent of the constant of proportionality since α appears linearly in both the numerator and the denominator.

In a digital forensics context, if H_k represents the prosecution's contention regarding the formation of $\{Ei\}$ and $H_{k'}$ is the defence's alternative contention, then the odds $O(k{:}k')$ provide a measure of the relative plausibility of the two competing hypotheses. More generally, if a total of n feasible routes are identified which are each capable of leading to the formation of the set of recovered traces, then the odds that feasible route k was taken are given by:

$$O(k) = \Pr(H_k|\{E_i\}) / \Pr(H^c_k|\{E_i\})$$

where H^c_k is the hypothesis that feasible route k was *not* taken, and involves summing the individual probabilities of the remaining *n-1* feasible routes.

3. Application to the BitTorrent Case

To illustrate the use of the operational complexity model outlined above we give here an application to the BitTorrent case described previously [12]. The prosecution case is taken to be exactly as described in [12], that is, the seized computer was used as the initial seeder to share the pirated file on a BitTorrent network; see also Figure 1 in the Appendix. The defence's alternative explanation for the presence of the recovered set of digital evidential traces $\{Ei\}$ has been constructed as follows. A Trojan horse program carrying the multimedia file as part of its payload installed itself on the defendant's unprotected computer and invoked the µTorrent client to upload the multimedia file to a peer-to-peer (P2P) file sharing website, before finally uninstalling itself.

We have made a number of simplifying assumptions for the purposes of this illustration. The computer was assumed to be running an MS Windows-like environment; the Trojan horse program is not equipped with its own life-support system; the computer is not protected by an operational firewall or anti-malware scanner. The basic unit of the GOMS-KLM characterization of cognitive information processing is taken to be the mouse button press or release; similarly, the basic unit of information processing used in characterizing the computational complexity is the byte. Disk accesses are assumed to take place autonomously and concurrently with CPU- and RAM-based processes. Given these assumptions and using documented typical or limiting values for all other quantities (see Tables 1-5 in the Appendix for full details) we obtain the following results:

$KLM_k = 510$

$KLM_{k'} = 0$

$CC_k = \mathrm{N} + 20\mathrm{N}/2^{19} + 1{,}844{,}346$

$CC_{k'} = (23/5)\mathrm{N} + 20\mathrm{N}/2^{19} + 9{,}938{,}941$

Taking a typical value for the size of the multimedia file as $N = 4\text{GB}$, we obtain:

$CC_k = 4{,}296{,}975{,}482$

$CC_{k'} = 19{,}766{,}952{,}343$

Hence, providing that route *k'* is the ***only*** feasible alternative to route *k,* we find $O(k{:}k') \approx 4.60$, indicating that the prosecution's case is 4.6 times more plausible than the defence's case, given the recovered evidence. Alternatively, in the absence of any other feasible explanations, the probability that the prosecution's case is correct is ≈ 82%.

4. Refinement of the Model

It is possible to make straightforward refinements to the simple operational complexity model given in Section 2 above in a number of ways. For example, given the result obtained in Section 3 above for the BitTorrent case [12] with an unprotected computer, a forensic examiner or expert witness might wish to know how the computed odds would change if an up-to-date anti-malware scanner were in fact installed and operational on the computer. This can be determined by making use of published values for the success rates of Trojan horse interception by commercially available anti-malware scanners. Current average values quoted by commercial anti-malware providers are typically ≈ 98% [13], although independent comparative surveys do not appear to be publicly available. As a consequence, the probability of the Trojan horse narrative is reduced from ≈ 18% to ≈ 0.36% in this scenario.

A further refinement that has been considered is the most appropriate relative weighting between the cognitive and computational components of the operational complexity metric. It can be persuasively argued that the cognitive component should be scaled by the ratio of the processing rates of the human and the computer, typically $\approx 10^7$. In that case,

$CC_k = 9{,}396{,}975{,}482$

and $O(k{:}k') \approx 2.10$, decreasing the plausibility of the prosecution's case to ≈ 68%. Combining this result with the Trojan horse interception scenario yields a slightly increased probability of ≈ 0.64% for this narrative.

A further consideration is that of disk access. The model implicitly assumes that disk transfers are effectively 'hidden' by means of an autonomous concurrent process, such as that implemented by a DMA channel. An alternative approach would be to adopt a typical RAM-to-disk access times ratio for data transfers so that the DSK actions in the analysis (Tables 4 and 5 in the Appendix) can be treated explicitly. If a typical value of 10^5 is adopted for this ratio then the following results are obtained:

$CC_k = 9{,}398{,}875{,}482$

$CC_{k'} = 19{,}768{,}052{,}343$

yielding $O(k{:}k') \approx 2.10$, as previously.

The most appropriate levels of granularity for different parts of the complexity metric are still under consideration. One possible option is to use a more coarse-grained metric for repetitive byte-wise operations, such as copying a large multimedia file, such that these operations are represented as single macros. An advantage of this approach would be that copying a 2*N*-byte file would not be considered to possess twice the operational complexity of copying an *N*-byte file.

5. Summary and Conclusions

The recently developed operational complexity model enables the complexity of both the cognitive and the computational components of a process to be determined. From the complexity of formation of a set of traces via a specified route the probability of that route can be determined. By determining the complexities of alternative routes leading to the formation of the same set of traces, the odds indicating the relative plausibility of the alternative narratives of formation can be found. An illustrative application to the previously discussed BitTorrent case [12] has been presented, and the results obtained suggest that the proposed operational complexity model is capable of providing a realistic estimate of the odds for two competing hypotheses. It has also been demonstrated that the model is capable of straightforward refinement to encompass a variety of circumstances, such as relative rates of human *vs.* computer processing, RAM *vs.* disk access, and the presence of security measures. These features should provide valuable support for forensic examiners and expert witnesses seeking to assess the strength of a case given the recovered digital evidence.

Acknowledgement

The authors are most grateful to an anonymous referee for flagging two minor arithmetical errors in an earlier draft of this paper.

References

[1] R E Overill and J A M Silomon, A Complexity Based Model for Quantifying Forensic Evidence Probabilities, *Technical Report*, Department of Computer Science, King's College London (July 2009).

[2] S Lloyd, Measures of Complexity: a Non-exhaustive List, *IEEE Control Systems*, **21 (4)** August 2001, 7-8.

[3] C H Papadimitriou, *Computational Complexity*, Addison-Wesley (1994).

[4] J F Traub, G W Wasilkowski and H Wozniakowski, *Information-Based Complexity*, Academic Press (1988).

[5] C Bennett, Logical Depth and Physical Complexity, in *The Universal Turing Machine: a Half-Century Survey* (Ed: R Herken), Oxford University Press (1988) 227-257.

[6] S Lloyd and H Pagels, Complexity as Thermodynamic Depth, *Annals of Physics*, **188** (1988) 186-213; J P Crutchfield and C R Shalizi, Thermodynamic Depth of Causal States*, Physical Review B*, **59** (1999) 275-283.

[7] C H Bennett, Dissipation, Information, Computational Complexity and the Definition of Organization, in *Emerging Syntheses in Science* (Ed: D Pines) Santa Fe Institute Studies in the Sciences of Complexity, Vol. I, Addison-Wesley (1987) 215-234.

[8] D Kieras, Using the Keystroke-Level Model to Estimate Execution Times, University of Michigan (2001), available online at: http://www.cs.loyola.edu/~lawrie/CS774/S06/homework/klm.pdf

[9] C E Shannon and W Weaver, *The Mathematical Theory of Communication,* University of Illinois Press (1949).

[10] E T Jaynes, *Probability Theory: the Logic of Science*, Cambridge University Press (2003).

[11] J Keppens and J Zeleznikow, in Proc. 9th International Conference on Artificial Intelligence and Law (ICAIL'03), 24-28 June 2003, Edinburgh, UK, pp.51-59.

[12] M Kwan, K P Chow, F Law and P Lai. Reasoning About Evidence using Bayesian Network, *Advances in Digital Forensics IV*, International Federation for Information Processing (IFIP) January 2008, Tokyo, pp.141-155.

[13] For example, Take the Kaspersky Challenge: See what your current antivirus is missing, available online at: www.kaspersky.com/virusscanner/

Appendix

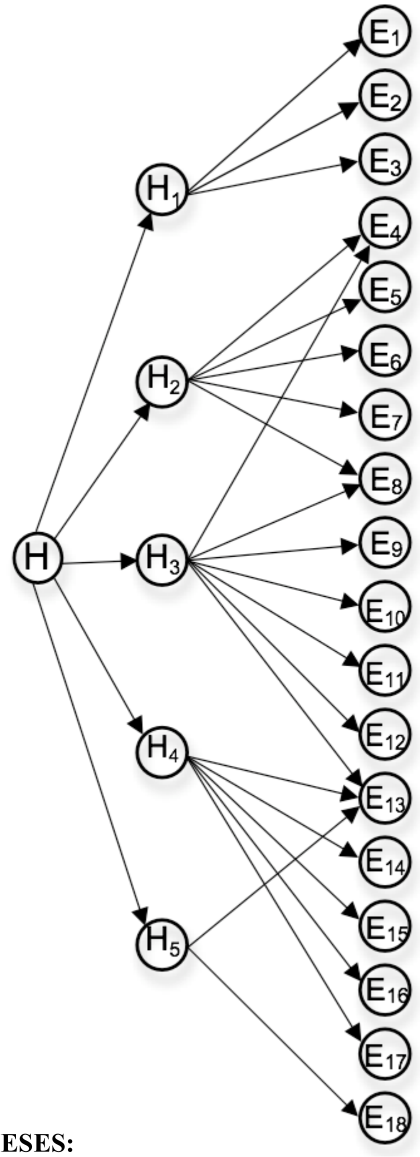

Figure 1 BitTorrent Network Diagram [10]

HYPOTHESES:

H The seized computer was used as the initial seeder to share the pirated file on a BitTorrent network

H_1 The pirated file was copied from the seized optical disk to the seized computer

H_2 A torrent file was created from the copied file

H_3 The torrent file was sent to newsgroups for publishing

H_4 The torrent file was activated, which caused the seized computer to connect to the tracker server

H_5 The connection between the seized computer and the tracker server was maintained

EVIDENCE:

E_1 Modification time of the destination file equals that of the source file

E_2 Creation time of the destination file is after its own modification time

E_3 Hash value of the destination file matches that of the source file

E_4 BitTorrent client software is installed on the seized computer

E_5 File link for the shared file is created

E_6 Shared file exists on the hard disk

E_7 Torrent file creation record is found

E_8 Torrent file exists on the hard disk

E_9 Peer connection information is found

E_{10} Tracker server login record is found

E_{11} Torrent file activation time is corroborated by its MAC time and link file

E_{12} Internet history record about the publishing website is found

E_{13} Internet connection is available

E_{14} Cookie of the publishing website is found

E_{15} URL of the publishing website is stored in the web browser

E_{16} Web browser software is available

E_{17} Internet cache record about the publishing of the torrent file is found

E_{18} Internet history record about the tracker server connection is found

KLM Operator	Normalised Value
K (key press & release)	2
P (point mouse)	11
B (button press/ release)	1
H (hand to/from keyboard)	4
M (mental preparation)	12

Table 1 KLM Operators and Normalised Values

Action		M	P	B	K	H	Total
1	Drag and Drop	2	2	2	0	0	48
2	Double click	1	1	4	0	0	27
3	Single click	1	1	2	0	0	25
4	Create torrent	5	6	10	0	0	136
5	Upload torrent	5	5	10	0	0	125
6	Type URL	2	1	4	16	2	79
7	Log in (username/pw)	4	2	4	16	4	122

Table 2 Frequent KLM Actions and Values

Var-iable	Description	BT specific value
N	no. data bytes in file to be shared	4GB
N_{THC}	no. data bytes of Trojan Horse Code	128KB
N_{THD}	no. data bytes in Trojan Horse Dropper program	$(N+N_{THC})/IFL$
N_{TC}	no. data bytes in Torrent Client	7MB
N_{TCI}	no. data bytes in Torrent Client Installation file	276KB
N_{DI}	no. data bytes in Desktop.ini file	47B
TD	TimeDate read or write	8B
TFN	Torrent File Name	256B
TFL	Torrent File Length	4B
TPL	Torrent Piece Length	4B
TPS	Torrent PieceSize	512KB
TAU	Tracker Announce URL	35B
IFL	Inflation factor (unzip)	1.25
DSK	Disk access (assumed autonomous)	0
PCI	Peer Connection Information process	52B + 3TD + DSK
TSL	Tracker Server Login process	60B + 3TD + DSK
PG	Page creation process (webpage)	600KB + TD + DSK
CO	Cookie creation process	256B + TD + DSK
CA	Cache creation process	16B + TD + DSK

Table 3 Definitions and Values of Variables used in the Analyses

Route *k* (Criminal)			
Evidence (see [10])	KLM action	DSK	Bytes
E_1 (incl. $E_{2, 3, 5, 6}$)	1	3	N + 7TD
E_4	0	0	0
E7 (incl. E_8)	2+4	5	20N/TPS + TFN + TFL + TPL + TAU + 11TD
E9 (incl. $E_{10\text{-}12, 14, 15, 17, 18}$)	2+3+5+7	11	3PG + CO + 3CA + N_{DI} + PCI + TSL + 17TD
E_{13} , E_{16}	0	0	0

Table 4 Complexity of the Criminal Route

Route *k'* (Trojan)			
Evidence (see [10])	KLM action	DSK	Bytes
DSI (Dropper S/W Install)	0	1	N_{THD} + 3TD
DSU (Dropper S/W Uninstall)	0	1	N_{THD} + 3TD
TSI (Trojan S/W Install, payload copy incl. $E_{1, 2, 3, 5, 6}$)	0	4	N_{THD}*IFL + N + 10TD
E_4	0	1	16 + N_{TCI} + N_{TC} + 3TD
E_7 (incl. E_8)	0	2	20N/TPS + TFN + TFL + TPL + TAU + 4TD
E9 (incl. $E_{10\text{-}12, 14, 15, 17, 18}$)	0	11	3PG + CO + 3CA + N_{DI} + PCI + TSL + 17TD
E_{13} , E_{16}	0	0	0
TSU (Trojan S/W Uninstall)	0	1	N_{THD}*IFL + 3TD

Table 5 Complexity of the Trojan Route