# AN EMPIRICAL STUDY ON CURRENT MODELS FOR REASONING ABOUT DIGITAL EVIDENCE

Stefan Nagy[1], Imani Palmer[1], Sathya Chandran Sundaramurthy[2], Xinming Ou[2], Roy Campbell[1]

[1]Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana-Champaign, IL 61801, USA

[2]Department of Computing and Information Sciences
Kansas State University
234 Nichols Hall
Manhattan, KS 66506, USA

## ABSTRACT

The forensic process relies on the scientific method to scrutinize recovered evidence that either supports or negates an investigative hypothesis. Currently, analysis of digital evidence remains highly subjective to the forensic practitioner. Digital forensics is in need of a deterministic approach to obtain the most judicious conclusions from evidence. The objective of this paper is to examine current methods of digital evidence analysis. It describes the mechanisms for which these processes may be carried out, and discusses the key obstacles presented by each. Lastly, it concludes with suggestions for further improvement of the digital forensic process as a whole.

**Keywords**: digital evidence, forensic reasoning, evidence reliability, digital forensics

## 1. INTRODUCTION

As the use and complexity of digital devices continues to rise, the field of digital forensics remains in its infancy. The investigative process is currently faced with a variety of problems, ranging from the limited number of skilled practitioners, to the difficulty of interpreting different forms of evidence. Investigators are challenged with leveraging recovered evidence to find a deterministic cause and effect. Without reliable scientific analysis, judgments made by investigators can easily be biased, inaccurate and/or unprovable. Conclusions drawn from digital evidence can vary largely due to differences in their respective forensic systems, models, and terminology. This persistent incompatibility severely impacts the reliability of investigative findings as well as the credibility of the forensic analysts. Evidence reasoning is a fundamental part of investigative efficacy, however, the digital forensic process currently lacks the scientific rigor necessary to function in this capacity. This paper presents an overview of several recent methods that propose a deterministic approach to reasoning about digital evidence. Section 2 examines past discussion on the digital forensic process. Section 3 discusses the application of differential analysis. In section 4, we review several popular probabilistic reasoning models. Section 5 discusses the formalization of event reconstruction. In section 6, we consider a model that combines probabilistic reasoning with event reconstruction. Lastly, section 7 holds our conclusions and suggestions for additions to the field.

## 2. BACKGROUND

The standard for the admissibility of evidence stems from the *Daubert* trilogy, which establishes the requirements of relevancy and reliability [25]. NIST describes the general phases of the forensic process as: collection, examination, analysis and reporting [23]. Formalization is necessary to ensure consistent repeatability for all investigative scenarios. In recent years, literature has addressed the need for formalization of the digital forensic process, but primarily focused on evidence collection and preservation [2]. Ieong [24] highlights the need for an explicit, unambiguous representation of knowledge and observations. While a pedagogical investigative framework exists, there is yet to be a congruous system for digital evidence reasoning within the examination and analysis phases. Currently, digital forensic analysts use a variety of methods to develop conclusions about recovered evidence, yet the results are often marred with conflicting bias or are shrouded in a veil of uncertainty. There have been numerous proposed reasoning frameworks, typically relying on applied mathematics, statistics & probabilities as well as, logic. However, before we can employ any particular methodology, there is a need to examine, review and explore all options in order to carry out the investigative process with the utmost precision.

## 3. IFFERENTIAL ANALYSIS

Differential analysis is described as a method of data comparison used for reporting differences between two digital objects. Historically, it has been part of computer science for quite some time. Unix's diff command was implemented in the early 1970's, and is commonly used for fast comparison of binary and text files [3]. Continued advancements in hashing and metadata have since paved the way for more thorough differential analysis. It is flexible and adaptable to nearly all types of digital objects; Windows Registry hives, binary files, and disk images can all be compared for evidence of modification or tampering [4]. Non-forensic applications include security procedures of operating systems, such as Windows' use of file signatures to verify integrity of downloaded driver packages [5].

Modern investigative tools such as EnCase [6], FTK [7] and SleuthKit [8] have incorporated modules for streamlining differential analysis of collected evidence, although each require significant training to become competent with the software features. Garfinkel et al. [3] formalize a model for differential analysis in the context of digital evidence; two collected objects – a *baseline* object and a *final* object – are compared for evidence of modification both before and after events of interest. Ideally, the process will highlight the most significant changes made from baseline $A$ to final $B$, assuming those transformations resulted from actions taken by the suspect in question. In this context, differential analysis is often used to detect malware, file and registry modifications [3].

While the strategy of differential analysis is fundamentally the same regardless of which system level is being examined, each level possesses a certain degree of noise. In discussing differential analysis, will define "noise" as information resulting from comparison between baseline and final that is wholly irrelevant to the investigation. As the context of investigation is expanded, so does the difficulty to identify noise [9].

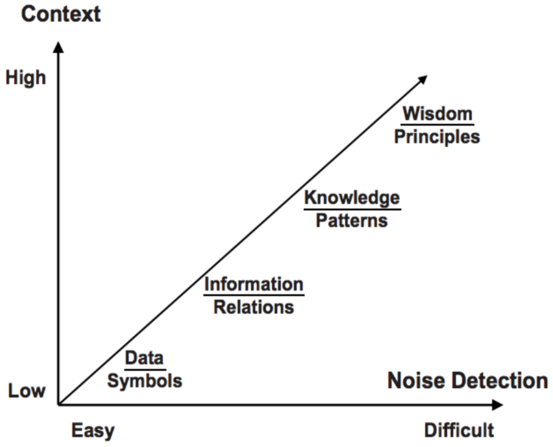

Figure 1. Knowledge management understanding hierarchy [9].

A potential form of noise presents itself as benign modifications made to digital objects resulting from normal operation of a system. For example, an investigator may wish to examine the presence of a suspicious binary on a particular system apart of an enterprise network. The investigator selects a disk image of an identical, unmodified system from the same enterprise network to serve as the baseline for comparison. Differential analysis may reveal that the image of the system in question is incredibly anomalous compared to the baseline. This could potentially lead to the injudicious assumption that "the most anomalous system is the most malicious" [4], when in reality, it might have only been the result of benign modifications arising from differences in installed software. While files at the kernel level are generally protected from tampering, files in user directories are much more vulnerable to modification.

Although noise is often assumed to be unintentional, it is very possible that it could be inserted on purpose. When dealing with instances of steganography, differential analysis compares objects that are known to be hiding information with those that do not. Fiore [10] describes a framework by which "selective redundancy removal" can be used to prepare HTML files for carrying out linguistic steganography. Since the information is being hidden through the otherwise normal process of HTML file optimization, differential analysis will only appear to reveal benign occurrences, such as differences in HTML tag styling.

Future research is needed to expand metrics for identifying and accounting for different forms of noise in digital evidence. Mead [1] explains the National Software Reference Library's effort to create a library of hashes of commercial software packages. Through combining hashing with differential analysis, investigators can drill-down the scope of inquiry by cross-referencing evidence with a database of known hash values. Eliminating evidence matching existing hashes can reduce the amount of noise arising from benign objects that is commonly problematic when dealing with larger systems, and better isolates the few remaining questionable objects. Further improvement of such databases, robust hashing algorithms, and perhaps a formal technique would be of benefit to investigators.

## 4. PROBABILISTIC MODELS

Conventional forensic analysis has long included models of statistical inference to assess the degree of certainty for which hypotheses and corresponding evidence can be causally linked [11]. This casual linkage is expressed by the following: if a *cause* $C$ is responsible for *effect* $E$, and $E$ has been observed, then $C$ must have occurred [12]. For example, researchers know that the probability of two identical DNA fingerprints belonging to two different individuals is close to one in one billion [13]. If holding an item leaves fingerprints on it, and fingerprints found on the weapon at a murder scene match the suspect's own, then investigators can conclude there is over 99% certainty that the suspect held that weapon. Because criminal investigations are ultimately abductive, probabilistic techniques have become widely accepted in the forensic reasoning process [14] [12].

### 4.1 CLASSICAL PROBABILITY

Several recent criminal investigations have seen classical probability used to reason about contradicting scenarios regarding the presence of incriminating digital evidence. Examining two cases originating in Hong Kong, Overill et al. [15] reasoned the likelihood that the respective defendants intentionally downloaded various forms of child pornography versus accidentally downloading it among other benign content. In each case, the amount of child pornography seized was very small compared to the total amount of miscellaneous benign content, and in both instances were found to have been downloaded over a long period of time. In each case, it was determined that the probability of unintentionally downloading a small amount of child pornography is significantly below 10% [15].

While this method can indeed provide a quantitative assessment of the likelihood of guilt, it is limited to investigations where only few characteristics of the evidential traces are known. In both examples above, the defendants pleaded guilty, and thus metadata was disregarded [15]. It was assumed that the incriminating files had been downloaded over long periods of time, but had metadata been collected, the original hypothesis may have changed entirely. An example would be the offending content timestamped to a one-hour browsing period, thus invalidating the original hypothesis of accidental download. The growing importance of preserving metadata creates the need for probabilistic models that can integrate it into reasoning.

### 4.2 BAYESIAN NETWORKS

In the last decade, Bayesian inference has gained popularity in the scientific community. Unlike Frequentist inference that reasons with frequencies of past events, Bayesian inference reasons with "subjective beliefs estimations", and allows room for new evidence to revise these beliefs [12]. Kwan et al. [14] introduced the idea of reasoning about digital evidence in the form of Bayesian networks: directed acyclic graphs whose leaf nodes represent observed evidence and interior nodes represent unobserved causes. The root node represents the central hypothesis to which all unobserved causes serve as sub-hypotheses. The model uses Bayes' theorem to determine the conditional probability of evidence $E$ resulting from hypothesis $H$:

$$P(E|H) = P(E)P(H|E);$$

$P(E)$ is the prior probability of evidence $E$; $P(H)$ is the prior probability of $H$ when no evidence exists; $P(H|E)$ is the posterior probability such that $H$ has occurred when $E$ is detected.

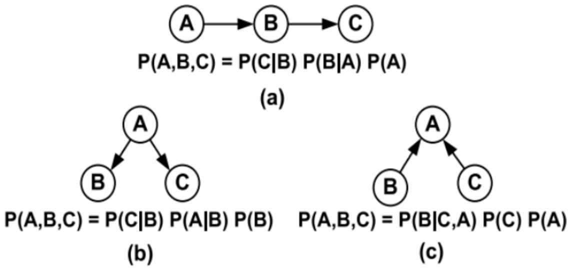
A
B
C
P(A,B,C) = P(C|B) P(B|A) P(A)
(a)
P(A,B,C) = P(C|B) P(A|B) P(B)
(b)
P(A,B,C) = P(B|C,A) P(C) P(A)
(c)

Figure 2. Bayesian network connections: (a) Serial; (b) Diverging; (c) Converging [14].

The construction of a Bayesian model begins with the defining of a root hypothesis. An example would be "The seized computer was used to send this malicious file." The possible states of the hypothesis – *Yes*, *No*, and *Uncertain* – are assigned equal probabilities. As more evidence is discovered, sub-hypotheses and their corresponding probabilities are added beneath the root hypothesis. The process is repeated until refinement produces a most likely hypothesis.

However, Bayesian networks are dependent on the assignment of prior probabilities to posterior evidence [14]. In scenarios where uncertainty is present, *fuzzy logic* methodology is incorporated to quantify likelihood as a value between 1 (absolute truth) and 0 (false) [16]. The case study presented in [14] based its prior probabilities on results from questionnaires sent to several law enforcement agencies. Since human-computer interactions are non-deterministic, there is no systematic way to reason posterior evidential probabilities with complete certainty; conditional probabilities inferred from demonstrably normal behavior of one network might differ with those from another. Discrepancies in prior evidential probabilities can significantly impact the overall outcome of the Bayesian network, and thus, there is difficulty in soundly applying this method to digital forensic investigations.

### 4.3 DEMPSTER-SHAFER THEORY

One of the limiting factors of using Bayesian analysis in security is that it requires the assignment of prior and conditional probabilities for the nodes in the reasoning model. Often times, the numbers are very hard to obtain. For example, how does one compute the prior probability for a particular registry key being modified? As another example, how does one compute the conditional probability of a particular registry key being modified given that the malware did not gain privileged access? Bayesian analysis works very well when the reasoning structure is well known and the probabilities are easy to obtain. In the real world, it is very hard to obtain those numbers and there is a high degree of uncertainty in the obtained evidence.

Dempster-Shafer theory (DST) is a reasoning technique that provides a way to encode uncertainty more naturally [17]. Contrasting with Bayesian analysis, DST does not require one to provide a prior probability for the hypothesis of interest. DST also does not require the use of conditional probabilities thus addressing the other major limitation of Bayesian analysis techniques. The presence of certain evidence during forensic analysis does not necessarily indicate a malicious activity. For example, a change in registry key could be either due to a malware or by a benign application. There is always a degree of uncertainty in the obtained evidence at any given stage of the forensic analysis process. DST enables one to account for this uncertainty by assigning a number to a special state of the evidence "don't know". For example, a sequence of registry key modifications might indicate that a malware of specific family might have been downloaded. Based on empirical evidence, let us assume one believes that with 10% confidence. A probabilistic interpretation would then mean that one would believe that there is a 90% chance that the malware was not downloaded—which is not intuitive. When using DST one would assign 10% to the hypothesis that the malware was downloaded and 90% to the hypothesis that *I am not sure.*

One can explain the difference between DST and probability theory using a coin toss example. When tossing a coin with unknown bias probability theory will assign a probability value of 0.5 to both the outcomes *Head* and *Tail.* This representation does not capture the inherent uncertainty in the outcome. DST, on the other hand, will assign 0 to the outcomes *{Head}* and *{Tail}* while assigning a value of 1 to the set *{Head, Tail}*. This exactly captures the reasoning process of a human in that when you toss a coin (with unknown bias) the only thing you are sure about the outcome is that it could be either *Head* or *Tail.* In general, when calculating the likelihood of a hypothesis DST allows admittance of ignorance on the confidence of evidence. DST provides rules for combining multiple evidences to calculate the overall belief in the hypothesis. The challenge of using DST is analogous to Bayesian analysis, though much better, in that no prior values have to be assigned to evidences.

## 5. EVENT RECONSTRUCTION MODELS

The ability to reconstruct events is of great importance to the digital forensic process. Al-Kuwari and Wolthusen [18] proposed a general framework to reconstruct missing parts of a target trace. This can be used for various areas of an investigation. This algorithm graphs a multi-modal scenario, determining all of possible routes connecting the gaps of a specific trace. Additional information may be included in the graph and marked appropriately. The broadcast algorithm used to determine all possible routes may require exponential time, suggesting that the search area should be bounded [18].

This approach relies on a specific target and would best be used to determine if an attack on a system occurred. However, this approach poses problems for the algorithm if a specific target is not identified. Event reconstruction is not unique to digital forensics, and the ability to apply existing techniques could yield effective results.

### 5.1 FINITE STATE MACHINES

Modern computer systems are often modeled as a series of finite states, graphically presented as a *Finite State Machine (FSM).* It is expressed as the quintuple $M=(Q, \Sigma, \delta, s0, F)$, where:

- $Q$ is the finite, non-empty set of machine states
- Σ is the finite, non-empty alphabet of event symbols
- $\delta$: $Q \times \Sigma \rightarrow Q$ is the transition function mapping events between machine states in Q for each event symbol in Σ
- $s0 \in Q$ is the starting state of the machine
- $F \subseteq Q$ is the set of final machine states
- Nodes represent possible system states
- Arrows represent transitions between states [19]

Gladyshev and Patel [20] introduced a formalization of this model into digital forensics. By back-tracing event states, investigators are presented with a reconstruction of events and can thus select the timeline most relevant to the available evidence.

For finite state machine models to perform accurately comprehensive event reconstruction, investigators must be able to account for all possible system states. Complex events, such as those resulting from advanced persistent threats, are incredibly difficult to analyze. In addition, changing factors such as software updates may affect the resulting machine states. Carrier [19] proposes the development of a central repository for hosting information about machine events. Likening it to existing forensic databases on gun cartridges, an exhaustive, continuously updated library of system events would be of invaluable aide to investigators performing event reconstruction. However, an investigator may wish to explore other characteristics of events, such as the odds of a particular investigative hypothesis, or the real time distributions of reconstructed events. To compute answers to such questions, the formalization of event reconstruction must be extended with additional attributes that describe statistical and real-time properties of the system and incident [20].

## 6. COMBINING PROBABILITY WITH EVENT RECONSTRUCTION

*Attack graphs* are typically used for intrusion analysis, where each path represents a unique method of intrusion by a malicious actor. It is possible to use attack graph techniques in the event reconstruction process. Attack graphs are directed graphs where nodes represent *pre* and *post* conditions of machine events, and directed edges are conditions met between these nodes; the root node represents the singular event of interest to which all other nodes serve as precursors [21].

While attack graphs are helpful in identifying mechanisms of intrusion, their lacking of any probabilistic inference hinders their usefulness in quantitative evidential reasoning. Investigators presented with attack graphs must select the most probable attack scenarios, but there are currently no clear metrics for assessing likelihood. To address this, Xie et al. [22] combined attack graphs with

Bayesian networks. By transferring attack graphs into acyclic Bayesian networks, this method utilizes conditional probability tables for nodes with parents, and prior probabilities for nodes without parents.

Like in regular Bayesian networks, this approach relies on the investigator supplying accurate conditional and prior probabilities for each event. Estimating prior probabilities has traditionally relied on feedback from the community in the form of surveys. This becomes incredibly difficult as scale increases; a large attack graph would require that the investigator survey and obtain probability information for every unique event, making analysis costly.

## 7. FUTURE DIRECTION AND CONCLUSIONS

Evidence reasoning models are an important part of the forensic process. Unlike traditional forensic sciences, digital forensics deals almost exclusively with objects of nondeterministic nature; there is great difficulty in analyzing and scrutinizing digital evidence. Fundamental flaws hinder current evidence analysis models in their ability to assess accurately the likelihood of crime occurrence. Furthermore, conclusions based on probabilities complicate explanations in the courtroom, as demonstrated in the legal arguments surrounding *Shonubi I-V* [26]. These flaws must be identified and understood to avoid the possibility of injudicious assumptions resulting from the forensic process.

Differential analysis of digital evidence becomes difficult when the scope of investigation is widened; unintentional noise in the form of benign modifications may lead to dubious conclusions about system integrity. Furthermore, recent obfuscation techniques have successfully averted detection by traditional methods. Event reconstruction models are limited in their ability to provide investigators with clear attack scenarios, because they rely on the exhaustive identification of possible machine states; there is yet to be a resource providing such information. Probabilistic reasoning models rely on prior probabilities known to the investigator, which have so far mainly been determined from surveying others in the field. Besides the obvious expenditure of time and effort in conducting such surveys, it is reckless to underestimate the potential for entropy and reason that small samples of observed probabilities hold true for all investigations. It can be concluded that each of these techniques is only applicable to a small niche of forensic scenarios.

The increasing rate of software development places a burden on forensic examiners to keep up with the latest software packages, both commercial and free. Each of the models discussed in this paper lacks a comprehensive database of information to conduct analysis with the highest accuracy. We highlight the need for a community-driven, updated catalogue of file hashes, machine states, and probability metrics for use in forensic analysis. The changing nature of technology and software necessitates that researchers and law enforcement collaborate to ensure the digital forensic process is as reliable as possible.

## 8. ACKNOWLEDGEMENT

This research is partially supported by the National Science Foundation under Grants No. 1314925 and 1241773. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

This material is based on research sponsored by the Air Force Research Laboratory and the Air Force Office of Scientific Research, under agreement number FA8750-11-2-0084. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

## REFERENCES

[1] Mead, S. Unique File Identification in the National Software Reference Library. *Digit. Investig. 3, 3 (September 2006), 138-150.*

[2] Stallard, T., Levitt, K. 2003. Automated Analysis for Digital Forensic Science: Semantic Integrity Checking. in *Proceedings of the 19th Annual Computer Security Applications Conference (ACSAC '03).* IEEE Computer Society, Washington, DC, USA, 160-.

[3] Garfinkel, S., Nelson, A., Young, J. A General Strategy for Differential Forensic Analysis. in

*Digital Forensics Research Workshop 2012,* August 2012, pages S50--S59.

[4] Gielen, M.W. 2014. *Prioritizing Computer Forensics Using Triage Techniques.* University of Twente.

[5] 2015. *Microsoft Windows*. Microsoft.

[6] 2015. *EnCase*. Guidance Software.

[7] 2015. Forensic Toolkit (FTK). Access Data.

[8] 2012. *The Sleuth Kit.* Carrier, D.

[9] Nunamaker, N.J.J., Romano J., Briggs, R. A Framework for Collaboration and Knowledge Management. in *Proceedings of the 34th Annual Hawaii International Conference on System Sciences,* January 2001.

[10] Fiore, U. Selective Redundancy Removal: A Framework for Data Hiding. in *Proceedings of Etude de la notion de pile application à l'analyse syntaxique..* 2010, 30-40.

[11] Overill, R.E., Silomon, J.A.M. Digital Meta-Forensics: Quantifying the Investigation. in *Proceedings of the Fourth International Conference on Cybercrime Forensics Education and Training,* 2010

[12] Huygen, P.E.M. Use of Bayesian Belief Networks in Legal Reasoning. in *17th BILETA Annual Conference,* Amsterdam 2002

[13] Overill, R.E. Quantifying Likelihood in Digital Forensics Investigations. *Journal of Harbin Institute of Technology,* Vol.21, No.6, 2014

[14] Kwan, M., Kam-Pui Chow, Law, F., Lai, P. Reasoning About Evidence Using Bayesian Networks. in *Advances in Digital Forensics IV, Fourth Annual IFIP WG 11.9 Conference on Digital Forensics,* Kyoto University, Kyoto, Japan, January 28-30, 2008

[15] Overill, R.E., Silomon, J.A.M., Kam-Pui Chow, Tse, H. Quantification of Digital Forensic Hypotheses Using Probability Theory. in *Systematic Approaches to Digital Forensic Engineering (SADFE),* 2013 Eighth International Workshop on , vol., no., pp.1,5, 21-22 Nov. 2013

[16] Stoffel, K., Cotofrei, P., Han, D. Fuzzy Methods for Forensic Data Analysis. in *Soft Computing and Pattern Recognition (SoCPaR),* 2010 International Conference of , vol., no., pp.23,28, 7-10 Dec. 2010

[17] Shafer, G. Probability Judgment in Artificial Intelligence and Expert Systems. *Statistical Science,* Vol.2, No.1 (Feb., 1987), pp. 3-16

[18] Al-Kuwari, S., Wolthusen, S.D. Fuzzy Trace Validation: Toward an Offline Forensic Tracking Framework. in *Systematic Approaches to Digital Forensic Engineering (SADFE).* 2011 IEEE Sixth International Workshop on, pages 1 – 4, IEEE, 2011.

[19] Carrier, D. 2006. *A Hypothesis-based Approach to Digital Forensic Investigations.* Purdue University.

[20] Gladyshev, P., Patel, A. Finite State Machine Approach to Digital Event Reconstruction. *Digit. Investig.,* 1(2):130–149, June 2004.

[21] Liu, C., Singhal, A., Wijesekera, D. Using Attack Graphs in Forensic Examinations. in *Availability, Reliability and Security (ARES),* 2012 Seventh International Conference on , vol., no., pp.596,603, 20-24 Aug. 2012

[22] Xie, P., Li, J.H., Ou, X., Liu, P., Levy, R. Using Bayesian Networks for Cyber Security Analysis. in *Dependable Systems and Networks (DSN),* 2010 IEEE/IFIP International Conference on , vol., no., pp.211,220, June 28 2010-July 1 2010

[23] NIST. Guide to Integrating Forensic Techniques into Incident Response. http://csrc.nist.gov/publications/nistpubs/800-86/SP800-86.pdf

[24] Ricci S. C. Ieong. 2006. FORZA - Digital forensics investigation framework that incorporate legal issues. *Digit. Investig.* 3 (September 2006), 29-36.

[25] Vickers, A. Leah. "Daubert, Critique and Interpretation: What Empirical Studies Tell Us About the Application of Daubert." *USFL Rev.* 40 (2005): 109.

[26] Izenman, J.A. *Introduction to Two Views on the Shonubi Case*. Temple University.