Overcast: Forensic Discovery in Cloud Environments

Stephen D. Wolthusen

Information Security Group
Department of Mathematics
Royal Holloway, University of London, UK

and

Norwegian Information Security Laboratory
Department of Computer Science
Gjøvik University College, Norway

and

Security Technology Department
Fraunhofer-IGD
Darmstadt, Germany

September 16, 2009

Introduction

Digital forensics (also: computer forensics) can be defined as

> *Approaches and techniques for gathering and analyzing traces of human and computer-generated activity in such a way that it is suitable in a court of law.*

The objective is to perform a structured investigation into past and ongoing data processing and transmission whilst maintaining a documented chain of evidence.

- Existing, immature research has thus far concentrated on storage and, more limited, network flow data
- Focus here is on cloud and distributed systems

Research Challenges

Research on digital forensics, particularly for distributed systems is of necessity an endeavor which draws upon multiple disciplines and sub-disciplines:

- Computer science
  - Networks, databases, and operating systems
  - Computer graphics, visual analytics, and multimedia systems
  - Signal processing and pattern recognition
  - Inference and deduction systems
- Mathematics
  - Distributed algorithms, graph theory, statistics
- Ancillary disciplines:
  - Information retrieval, psychology, visual sciences, law

Discovery of Computational Structure

In distributed systems even identifying the computational and storage structures relevant for an investigation is non-trivial.

- Systems (and services) cannot be seized – this would require cessation of unrelated services, cross jurisdictional boundaries, and may need to occur clandestinely
- Data volumes defy easy seizure or duplication

A key research question is therefore how to derive the scope of the computation, documents, or services to be captured.
This requires the consideration of several sub-aspects.

Discovery: Temporal Extent

As complete capture of all data related to an event under investigation is not possible, a snapshot must typically suffice. Several questions must be addressed for this to be satisfactory:

- Start and end intervals must be established, these may differ significantly for different event components
- Data may be ephemeral, or staged through storage and processing hierarchies
- As in any distributed system, a truly global clock does not exist, limiting available information (typically in non-cooperative environments) to partial orderings

Relevant data is often transient and volatile — establishing of a consistent state closure is critical as subsequent events can further alter structure and semantics

Discovery: Spatial Extent

Location transparency is one of the key aspects of cloud and distributed systems

- Digital forensics may at times require breaking this abstraction

Understanding location(s) of data may determine ability to seize data or have it subjected to injunctions

- Even more importantly, collection and discovery of some data sets may be legal in one jurisdiction, but be inadmissible or even illegal in another jurisdiction
- In other cases, authorities and entities involved may have limited interest in cooperative behavior

Circumstantial information from location abstraction and delivery performance optimization mechanisms may permit establishment of direct or circumstantial evidence

Discovery: Dependency Analysis

Establishing a complete understanding of an event's dependencies (or constituent elements, which may also be generated dynamically

- Distribution across different systems and components
- Semantic dependencies must capture not only data for reconstructing a view or document or reconstruct a process, but also the semantics at the point in time of the event
- This requires capturing constituent components, ontological information, data dictionaries, and also includes structures and components that are implicit or potentially off-line at the time of analysis

A further challenge to all discovery aspects lies in the need to conduct some investigations in a clandestine manner — adversaries detecting investigations may take steps to obfuscate and erase evidence

Attribution of Data

Ascertaining the provenance of data, and wherever possible providing attribution is a key requirement in any forensic investigation

- Even when data is cryptographically protected, longer-term attribution can become tenuous if key material becomes (perhaps deliberately) invalid or unavailable
- Attribution often relies on ephemeral entity authentication, not available in evidential data — this requires the creation of circumstantial evidence both directly and through indirect evidence of service usage
- Key issue is the derivation and deduction of communication and interactions which are not immediately observable

Quality of attribution may differ among data elements, making trust and reputation systems as well as causality and evidence models crucial

Semantic Integrity

Distributed systems and ephemeral structures present a potentially unlimited and unknown number of permutations and variations of data structures, formats, and implementations

- Semantics of the data set can rely on directly obtained schemata and file formats, but more often must make inferences
- For databases and markup languages, this can sometimes be obtained, but may often require separate capturing
- In other cases inferences will be indirect or circumstantial, and may not provide a close match to actual semantics – this is particularly problematic if a data model evolves after capturing

Clear identification of any inferences and deductions made both in obtaining (often also second-order primary) evidence must be made

- This also applies to presentation since reasoning and analysis must be performed transparently

Stability of Evidence

A further challenge close to those faced in long-term archiving of digital data is preserving the stability of evidence

- Retaining the extensional semantics of a data set over longer periods of time as well as documenting uncertainties and inconsistencies arising
- For distributed data sets and partial data sets collected from distributed systems, it will not always be possible to obtain individual specifications and data dictionaries (“the implementation is the specification”)
- Establishing and documenting extensional information upon capture is thus critical, as is describing semantic information, often deduced from restricted, proprietary data

Stability of Ephemeral and Endpoint Data

A further issue arises with time-sensitive ephemeral data such as streaming media or control systems data flow. This necessitates capturing of context information

- Generally, information on (potentially inaccessible, or unknown) endpoints may be necessary to obtain aggregate state or context
- Especially in case of transactional data sets, state information is often implicit, not all data relevant for a transaction may be retained on each participating node
  - This is particularly relevant where data is only processed into its final presentation format on the endpoint (as is typically the case for Web 2.0 applications, etc.)

Presentation and Visualization of Evidence

Main challenges for presentation and visualization of evidence collected in cloud and distributed systems:

- Complexity of data sets
  - Ill suited for linear presentation of simple visualization techniques, even for static (non-ephemeral data): Need for exploratory, interactive analytical mechanisms
  - Limited confidence in source data, its interpretation and semantics: Requirement to understand limitations of data sets, their origin, and interpretation

Any presentation mechanism must be neutral and allow unambiguous reproduction and validation by competent third parties

- This is particularly true for exploratory and hypothesis-based analysis

Presentation of Well-Known Formats

Even for relatively static documents and document structures consisting of separate files and records, several challenges emerge for presentation and visualization

- Documents, or (intricately) linked sets of documents need to be presented and analyzed
- Components may be based on different formats, versions, and presentation mechanisms
- Incomplete data sets must be characterized as such while still permitting further analysis
- Data volume presents considerable challenges for (semi-) interactive presentation

Presentation of Non-Standard Data and Processes

Presenting, documenting, and investigating non-standard data sets (e.g. databases, work flows, processes, transactions) in an efficient and re-usable manner is less straightforward:

- Gaps in data sets, uncertainties about semantics and interpretation, limitations of the collection mechanisms must be both documented and also compensated for
- Critical particularly for exploratory investigative processes including hypothesis formation and validation

Data types may often exhibit complexity itself, and may result from multiple sources (e.g. multiple concurrent sensor or media streams), which must be synchronized or aligned where exact reconstruction is not possible

- Links to pattern matching, statistical analysis, inference mechanisms, and data fusion

Cross-Jurisdictional Aspects

Forensic data from distributed systems is likely to transcend national borders in at least some of their constituting elements during the processing or also storage of data sets.

- Ongoing European and international harmonization efforts notwithstanding, it is still crucial to identify applicable legal frameworks and constraints
- Different standards for collection, retention, and presentation exist both with regard to jurisdiction but also regarding the type of proceedings — neglect may lead to jeopardizing admissibility of an overall data set
- Any research on digital forensics, particularly for distributed environments must therefore aim to cooperate closely with relevant domain experts

Conclusion

Digital forensics is still very much a practitioner-driven craft rather than scientifically founded, and this presumably even more the case for distributed and cloud forensics, where even feasibility is questioned. This talk has aimed to highlight a selection of some of the most pressing fundamental and, to a lesser extent, applied open research questions and the potential for a intra- and interdisciplinary research approach to perform both

- novel and innovative blue-skies research
- research with clear application potential both in forensics and in operating on large, ill-structured and ill-characterized document and data sets

Beyond research potential, this provides significant opportunities for establishing national and international procedures and standards, providing professional and consulting services, and generating IPRs