Chapter 7

HYPOTHESIS-BASED INVESTIGATION OF DIGITAL TIMESTAMPS

Svein Willassen

Abstract Timestamps stored on digital media play an important role in digital investigations. However, the evidentiary value of timestamps is questionable because timestamps can be manipulated or they could refer to a clock that is erroneous or improperly adjusted. This paper presents a formalism for defining clock hypotheses based on historical adjustments to clocks, and for testing the consistency of the hypotheses with respect to stored timestamps. Two consistency tests are proposed for justifying clock hypotheses without having to rely on timestamps from external sources.

Keywords: Digital investigations, timestamps, causality, clock hypothesis testing

1. Introduction

A timestamp is a recorded representation of a specific moment in time. In digital computing, a timestamp is a recorded representation of a specific moment in time in a digital format. This representation is either stored on digital media or is transmitted on a network designed to convey digital data.

Timestamps play an important role in digital investigations. They are traditionally used to place the timestamped event at a specific moment in time, thereby facilitating event reconstruction. The identification that a certain event on a computer took place at a specific time makes it possible to correlate the event with other events occurring outside the computer system. These external events may have occurred in another digital system or in the physical world. A Windows system hard drive in a typical digital investigation can contain tens or hundreds of thousands of timestamps.

Stored timestamps may not accurately reflect the times that the events occurred. A timestamp is always relative to the setting of the clock that generates it. Unfortunately, clocks are not completely reliable. They may drift, generating timestamps that are increasingly different from those generated by other clocks. Clocks may also fail or may produce incorrect timestamps. Furthermore, clocks on most systems can be adjusted by users intentionally or accidentally. Consequently, timestamps generated by the same clock cannot be reliably compared unless it can be shown that the clock was not adjusted during the time period between the creation of the timestamps. Timestamps generated by different clocks are reliably compared by computing the difference between the clocks and verifying that the clocks were not adjusted.

Timestamps are vital to reconstructing events in digital forensic investigations. But they cannot be relied upon as evidence without considering all the factors that may lead to errors. This paper describes a formalism for defining and testing the consistency of clock hypotheses. Carrier's hypothesis-based investigation model [2] is used to test the evidentiary value of timestamps. In this model, the history of the medium under investigation is the complete set of configurations, states and events that have occurred during the lifetime of the medium. The data directly observable by the investigator is the final state of the medium, and it includes observations of all timestamps stored on the medium and the clock. The ability to test clock hypotheses increases the evidentiary value of timestamps even when clocks are erroneous, improperly adjusted or are known to have failed.

2. Related Work

The problem of timestamp interpretation has been studied by several researchers. Schatz and colleagues [6] have analyzed clock synchronization in enterprise computer networks. They suggest that clock drift can be mitigated by correlating timestamps stored in the web cache with records obtained from web servers. Other researchers [1, 7] also advocate the use of correlation methods for timestamps stored on target computers that were created by other clocks (e.g., timestamps in dynamically generated web pages). These methods provide correlations for the periods during which the cached data exists on the target computers. They are able to confirm or refute hypotheses about a clock in the period for which correlation data exists, but they may be unable to provide reasonable evidence to refute certain hypotheses (e.g., that the timestamps have been changed or that the clock has been adjusted during the period for which no correlation data exists). Correlation with

server records is only possible when such data exists and the forensic investigator has legal access to this data.

Gladyshev [4] studied the use of causality properties to establish a time period during which an event may have occurred. In his approach, time boundaries can be established when an event that occurred at an unknown or uncertain time is causally preceded and succeeded by events whose times of occurrence are known. When investigating a target computer, the events whose occurrence times are known must come from external sources. Our approach also uses the notion of causality, but it does not require time references from external sources.

3. Hypothesis-Based Timestamp Investigation

This section discusses the main concepts underlying hypothesis-based timestamp investigation.

3.1 Causality

Causality – the relationship between cause and effect – can be formally expressed as a mathematical relation between events. Lamport [5] was the first to use the *happened-before* relation ($\rightarrow$) for ordering events pertaining to executing processes and message passing. Lamport's definition was generalized by Fidge [3] to encompass process creation and termination as well as synchronous and asynchronous message passing.

Gladyshev [4] proposed an extended definition of *happened-before* for digital investigations. According to Gladyshev, $e_1 \rightarrow e_2$ if e_2 uses the result of e_1 or e_1 precedes e_2 in the usual course of business of some organization or during the normal operation of a machine. This definition is useful because digital investigations require the reconstruction of events both within and external to computer systems.

Gladyshev's definition of the *happened-before* relation uses the terms, "usual course of business" and "normal operation," which are open to interpretation. In contrast, our definition of the relation directly captures the notion of causality.

Definition. Let e_1 and e_2 denote events and let $\rightarrow$ represent the *happened-before* relation. If $e_1 \rightarrow e_2$, then the occurrence of e_1 is necessary for e_2 to occur because e_2 depends on the effects of e_1.

Examples of causality captured by the *happened-before* relation are:

- "e_1 produces an item that is a necessary input for e_2."
  This is equivalent to Gladyshev's definition "e_2 uses the result of e_1." The definitions of Lamport and Fidge are also covered by this example.

- "e_1 and e_2 are events in a computer program where e_2 uses data produced by e_1."
  Since events that occur when computer programs execute use items produced by other events in the same program (e.g., variables, data stored in memory, registers and stack pointers), many events that occur during program execution can be expressed using the *happened-before* relation. This is a special case of "e_1 produces an item that is a necessary input for e_2." The definition of *happened-before* also captures events related to processes modeled by Lamport and Fidge with the exception of events that do not use results from each other. This exception makes the definition suitable for modern computer systems in which the execution order of program statements can be modified by compilers and processors when the instructions do not depend on the results of each other.

3.2 Time

Time is considered to be a fundamental quantity because it is not defined in terms of other quantities. However, it is measurable via comparisons with periodic events such as those occurring in clocks. Examples of periodic events are the swings of a pendulum (pendulum clock), movement of the earth (sundial) and microwave emission (atomic clock). We assume that every event has a moment in time associated with it and these moments in time can be ordered using the $<$ and $=$ relations.

Definition. Let e be an instantaneous event in the domain of events E, and let T be the domain of time. The function $t(e) : E \mapsto T$ provides the moment in time at which event e occurred.

We assume that causality is preserved in time, i.e., no event can causally depend on an event occurring at the same time or at a later time than itself. This notion is expressed explicitly using the *happened-before* relation ($\rightarrow$):

$$t(e_i) \leq t(e_j) \Rightarrow e_j \not\rightarrow e_i. \tag{1}$$

This assumption captures the intuitive relationship that exists between causality and time. If such causal relationships were allowed, then events in the future would affect events in the past, which has not been shown to occur in the real world.

For two events that satisfy the *happened-before* relation ($\rightarrow$), Equation 1 implies that:

$$e_i \rightarrow e_j \Rightarrow t(e_i) < t(e_j). \tag{2}$$

This equation imposes an ordering in time on events related via the $\rightarrow$ relation. However, it does not imply any ordering in time for events not related by $\rightarrow$. Also, $t(e_i) < t(e_j)$ does not imply that $e_i \rightarrow e_j$. Events may occur at different moments in time without being related by $\rightarrow$. On the other hand, if two moments in time, $t(e_1)$ and $t(e_2)$, are ordered such that $t(e_1) < t(e_2)$, events occurring at those moments in time cannot be causally connected in reverse such that $e_2 \rightarrow e_1$.

3.3 Clocks

A clock is a device designed to provide its owner with an approximation of time that is sufficiently coherent to allow the owner to measure and compare time periods. Also, a clock is sufficiently consistent with other clocks to allow its owner to perform actions concurrent with other clock owners without continuous coordination. The definition of a clock should reflect the possibility of clock drift and adjustment discussed in Section 1.

Definition. Let V be the domain of time values produced by a clock. A clock function is defined as $c(t) : T \mapsto V$.

The definition of a clock function does not impose any restrictions on clock values as a function of time. For example, even if $t_1 < t_2$, it may well be the case that $c(t_1) > c(t_2)$. Also, even if $t_1 < t_2 < t_3$, the relationship $c(t_1) = c(t_2) = c(t_3)$ may hold. The latter situation could occur if the events at t_1, t_2, t_3 are so close together in time that the clock is unable to differentiate between them.

3.4 Timestamped Events

A timestamped event is an event for which there exists a timestamp value in domain V of time values. The timestamp value can be represented as a function of the event. A timestamp is created when an event makes a copy of the value provided by a clock. The timestamps in a set of timestamped events are not necessarily related to the same clock.

Definition. Let E be a set of timestamped events and let V be the domain of time values. The function $\tau_c(e) : E \mapsto V$ is defined such that $\tau_c(e_i) = c(t(e_i))$, where $\tau_c(e_i)$ is the timestamp associated with the event e_i relative to clock c.

A timestamp in the above definition is the value of the producing clock at the time of the event. The timestamp thus reflects the clock's representation of time at that particular moment. The definition of

timestamps as a function of events and clocks makes it possible to reason about timestamps and clocks.

3.5 Ideal and Non-Ideal Clocks

An ideal clock is one that can only go forward. A non-ideal clock is a clock that is not ideal.

Definition. Let I be the set of ideal clocks. An ideal clock $c(t) \in I$ satisfies the properties:

$$\begin{aligned} \forall i \forall j (t(e_i) < t(e_j) &\Rightarrow c(t(e_i)) \leq c(t(e_j))) \\ \forall i \forall j (t(e_i) = t(e_j) &\Rightarrow c(t(e_i)) = c(t(e_j))). \end{aligned}$$

An ideal clock has a monotonically increasing clock function. However, note that the values, $c(t(e_i))$ and $c(t(e_j))$, produced for two different moments in time, $t(e_i)$ and $t(e_j)$ (where $t(e_i) < t(e_j)$), may be equal. Many clocks express moments in time as discrete values. A discrete clock with limited resolution may represent two moments that are close in time using the same clock value.

Theorem 1. *Timestamps produced by all ideal clocks $c \in I$ satisfy the property:*

$$e_i \rightarrow e_j \Rightarrow \tau_c(e_i) \leq \tau_c(e_j).$$

Proof: An ideal clock satisfies the property:

$$\forall i \forall j (t(e_i) < t(e_j) \Rightarrow c(t(e_i)) \leq c(t(e_j))).$$

That is, for events e_i and e_j occurring at times $t(e_i)$ and $t(e_j)$:

$$t(e_i) < t(e_j) \Leftrightarrow c(t(e_i)) \leq c(t(e_j)).$$

Upon replacment, we obtain:

$$e_i \rightarrow e_j \Rightarrow c(t(e_i)) \leq c(t(e_j)).$$

Since $\tau_c(e_i) = c(t(e_i))$, we obtain the result:

$$e_i \rightarrow e_j \Rightarrow \tau_c(e_i) \leq \tau_c(e_j).$$

□

The monotonicity property of ideal clocks ensures that two causally connected events timestamped by the same ideal clock have timestamps

such that the timestamp of the latter event is never less than the timestamp of the former event.

3.6 Clock Hypothesis Formulation

In order to test if a certain theory holds for a clock, it is necessary to formulate a hypothesis about the clock function. The clock hypothesis, denoted by $c_h(t)$, is then tested against the set of observed timestamps.

Definition. A clock function $c(t)$ has two components, an ideal clock function $b(t)$ and a function $d(t)$ that represents the deviation from the ideal clock:

$$c(t) = b(t) + d(t).$$

The ideal clock $b(t)$ is called the base clock; $d(t)$ is the difference between the base clock and the clock of interest. Two clocks with a common base clock can be compared by examining their deviations. It is sometimes useful to express the time of an event in terms of the base clock. This is done by subtracting $d(t)$ as follows:

$$b(t) = c(t) - d(t). \tag{3}$$

3.7 Observed Event Sets and Correctness

During a digital investigation of a computer system, the investigator may observe a number of timestamped events that are based on the same clock. Some of these events will be causally connected. The set of observed timestamped events is called the "observation set."

Definition. An observation set O is a set of timestamped events that are related to one clock $c_o(t)$.

An observation set typically has a large number of timestamped events with a large number of causal connections. The data in an observation set is used to determine whether or not a clock hypothesis holds.

Definition. A clock hypothesis $c_h(t)$ for an observation set O is correct if $c_o(t) = c_h(t)$ for all t, i.e.,

$$c_o(t) = c_h(t) \Rightarrow \forall e_i(\tau_{c_o}(e_i) = c_h(t(e_i))).$$

If a clock hypothesis is correct, then all occurrences of timestamps must match the values predicted by the hypothesis. The correctness property can, therefore, be used to devise techniques for testing whether or not a clock hypothesis is correct.

Theorem 2. In a correct clock hypothesis $c_h(t)$ the timestamps of all causally connected events $e_i \rightarrow e_j$ in an observation set O must be such that the timestamp of the first event minus the deviation from a common base is not greater than the timestamp of the latter event minus the deviation from a common base, i.e.,

$$e_i \rightarrow e_j \Rightarrow \tau_{c_o}(e_i) - d_h(t(e_i)) \leq \tau_{c_o}(e_j) - d_h(t(e_j)).$$

Proof: Let $c_h(t)$ be a correct clock hypothesis. Let $b(t)$ be a common base for $c_h(t)$ and $c_o(t)$. Then,

$$b(t) = c_h(t) - d_h(t)$$

$$b(t) = c_o(t) - d_o(t).$$

Thus,

$$c_h(t) - d_h(t) = c_o(t) - d_o(t).$$

Also, since $c_h(t)$ is correct, we have $c_h(t) = c_o(t)$. Therefore,

$$\begin{aligned} d_h(t) &= d_o(t) \\ b(t) &= c_o(t) - d_h(t). \end{aligned}$$

Upon inserting the definition, we obtain:

$$b(t(e)) = \tau_{c_o}(e) - d_h(t(e)).$$

Note that $b(t)$ is an ideal clock. According to Theorem 1, ideal clocks satisfy the property:

$$e_i \rightarrow e_j \Rightarrow c(t(e_i)) \leq c(t(e_j)).$$

Inserting the expression for $b(t)$ yields the result:

$$\begin{aligned} e_i \rightarrow e_j &\Rightarrow b(t(e_i)) \leq b(t(e_j)) \\ e_i \rightarrow e_j &\Rightarrow \tau_{c_o}(e_i) - d_h(t(e_i)) \leq \tau_{c_o}(e_j) - d_h(t(e_j)). \end{aligned}$$

□

Conversely, if the property examined in Theorem 2 does not hold, the hypothesis is incorrect.

Theorem 3 (Test-A). *If a pair of causally connected events $e_i \rightarrow e_j$ exist in an observation set O for which the timestamp of e_i minus the hypothesis deviation from a common base is larger than the timestamp of e_j minus the hypothesis deviation from a common base, then the clock*

hypothesis is incorrect, i.e.,

$$\exists e_i \exists e_j ((e_i \rightarrow e_j) \wedge (\tau_{c_o}(e_i) - d_h(t(e_i)) > \tau_{c_o}(e_j) - d_h(t(e_j))))$$
$$\Rightarrow c_o(t) \neq c_h(t).$$

Proof: Let $c_h(t)$ be a clock hypothesis and O be an observation set with clock $c_o(t)$. Let (e_a, e_b) be a pair of events in O such that $e_a \rightarrow e_b$ and $\tau_{c_o}(e_a) - d_h(t(e_a)) > \tau_{c_o}(e_b) - d_h(t(e_b))$. Assume that $c_h(t)$ is correct, then $c_h(t) = c_o(t)$. Since $c_h(t)$ is correct, according to Theorem 3 we have:

$$e_i \rightarrow e_j \Rightarrow \tau_{c_o}(e_i) - d_h(t(e_i)) \leq \tau_{c_o}(e_j) - d_h(t(e_j)).$$

But for $i = a$ and $j = b$, we have assumed that:

$$(e_a \rightarrow e_b) \wedge (\tau_{c_o}(e_a) - d_h(t(e_a)) > \tau_{c_o}(e_b) - d_h(t(e_b))). \quad (4)$$

This contradicts the result from Theorem 2. Therefore, if Equation 4 holds, $c_h(t)$ cannot be correct. No assumptions or restrictions are imposed on events a and b; a and b could, therefore, be any event in the observation set O. For any event e_i and e_j, if Equation 4 holds, $c_h(t)$ cannot be correct. Consequently,

$$\exists e_i \exists e_j ((e_i \rightarrow e_j) \wedge (\tau_{c_o}(e_i) - d_h(t(e_i)) > \tau_{c_o}(e_j) - d_h(t(e_j))))$$
$$\Rightarrow c_o(t) \neq c_h(t).$$

□

Example 1. Consider the default clock hypothesis, which assumes that the clock of the target computer has always been equal to civil time, say UTC. Then $c_h(t) = b_h(t)$ and $d_h(t) = 0$. Let the observed set consist of timestamps for four events e_1 through e_4 where $e_1 \rightarrow e_2$ and $e_3 \rightarrow e_4$:

$$\begin{aligned} \tau_{c_o}(e_1) &= \text{Jan 12, 2003, 12:46:34} \\ \tau_{c_o}(e_2) &= \text{Apr 21, 2004, 10:22:38} \\ \tau_{c_o}(e_3) &= \text{Feb 9, 2003, 22:16:04} \\ \tau_{c_o}(e_4) &= \text{Dec 12, 2002, 02:46:32} \end{aligned}$$

If Test-A is applied for $i = 3$ and $j = 4$, we obtain:

$$(e_3 \rightarrow e_4) \wedge (\tau_{c_o}(e_3) > \tau_{c_o}(e_4)).$$

Since $d_h(t) = 0$, the test fails. Thus, the default hypothesis is incorrect for this observation set.

The result can be explained informally as follows: Since e_4 must have happened after e_3 and the timestamp of e_4 represents an earlier time than the timestamp of e_3, it cannot be the case that the clock was not adjusted between these two events.

Theorem 4 (Test-B). *In a clock hypothesis $c_h(t)$, for values c' of $c_h(t)$ for which $c_h(t) = c'$ has no solution, the existence of any timestamps in the observation set O with value $\tau_{c_o}(e_i) = c'$ implies that $c_h(t)$ is incorrect.*

Proof: Let $c_h(t)$ be a clock hypothesis and O an observation set with clock $c_o(t)$. Let e_a be an event in O and let $\tau_{c_o}(e_a) = c'$ be the timestamp of e_a. Furthermore, let c' have a value such that $c_h(t) = c'$ has no solution. If $c_h(t)$ is correct, $c_h(t) = c_o(t)$. Also,

$$\forall e_i(\tau_{c_o}(e_i) = c_h(t(e_i))).$$

This means that for $i = a$:

$$\tau_{c_o}(e_a) = c_h(t(e_a)).$$

This is a contradiction because $\tau_{c_o}(e_a) = c'$ and $c_h(t) = c'$ has no solution. Therefore, if $\tau_{c_o}(e_a) = c'$ and $c_h(t) = c'$ has no solution, then $c_h(t)$ cannot be correct.

□

3.8 Clock Hypothesis Consistency

Theorems 3 and 4 can be used to refute a clock hypothesis for an observation set O based on the timestamps of events in O. In the case of Test-A (Theorem 3), a clock hypothesis is incorrect when observations of timestamps for two causally connected events are not ordered correctly by the clock hypothesis being tested. On the other hand, Test-B (Theorem 4) stipulates that a clock hypothesis is incorrect when timestamps are observed that cannot be produced by the clock hypothesis because it is a discontinuous function. By iterating over all events and event pairs, every timestamp can be checked for consistency using Test-A and Test-B.

The tests can refute a clock hypothesis, but they cannot prove that it is correct. This leads to the following definition of a consistent clock hypothesis.

Definition. Given a set of tests Z, a clock hypothesis is consistent under Z with an observation set O if no test $z \in Z$ shows that the

hypothesis is incorrect for O. A clock hypothesis is inconsistent under Z with an observation set O if it is not consistent under Z with O.

The distinction between the definitions of a correct hypothesis and a consistent hypothesis is useful in the context of digital investigations. In a correct clock hypothesis, all possible time values are always based on the clock of interest. Such a hypothesis can only be verified if the clock has been observed at every moment in its history. This is inconceivable for the clock on a target machine in a digital investigation. Therefore, at best, the investigator can attempt to establish a consistent clock hypothesis. In such a hypothesis, none of the timestamps of the events in O used in the tests in Z are able to show that the hypothesis is incorrect. Nevertheless, the presence of large numbers of timestamps and causally connected events in O impose strict constraints on a consistent hypothesis, which can be used to justify the hypothesis. The more data available in O that is supplied to the tests in Z, the greater the justification provided to the consistent clock hypothesis.

3.9 Clock Hypothesis as a Scientific Hypothesis

In Carrier's hypothesis based investigation model [2], a digital investigation is a process that formulates and tests hypotheses to answer questions about digital events and/or the state of digital data. According to Carrier, an investigative process is scientific if the hypothesis is scientific and is tested by conducting experiments. Carrier cites Popper in that the "criterion of the scientific status of a theory is its falsifiability or refutability or testability."

The question here is whether or not the methods for clock hypothesis formulation and testing adhere to these criteria. From the previous discussion, a clock hypothesis is a theory that is falsifiable and therefore testable. The clock hypothesis theory thus meets the requirements of a scientific theory. The hypothesis forbids certain things from happening, i.e., the occurrence of timestamp configurations described in Test-A and Test-B. The two tests examine the evidence to refute hypotheses. They do not look for confirmation; instead, they seek to detect inconsistencies. Even when a test does not refute a hypothesis, the testing has value as a serious but unsuccessful attempt to falsify the hypothesis, which can be viewed as offering a certain amount of confirming evidence.

4. Conclusions

Timestamps of computer and network events are routinely used for incident reconstruction in digital forensic investigations. However, their

evidentiary value can be questioned because they are easily manipulated and the clocks used to create them could have been erroneous or improperly adjusted. The proposed formalism enables digital forensic investigators to define clock hypotheses based on historical adjustments to clocks and to test the consistency of the hypotheses with respect to stored timestamps. When the number of timestamps is large and many of the timestamped events are causally related, the consistency tests place clock hypotheses under close scrutiny. Even when a test does not refute a hypothesis, its mere application provides important confirming evidence. Clock hypothesis specification and testing is readily implemented in a software tool. Such a tool would enable investigators to verify the evidentiary value of timestamped data. Also, it could be used to investigate alternative hypotheses related to incident reconstruction, such as those postulated by prosecutors and defense attorneys.

References

[1] C. Boyd and P. Forster, Time and date issues in forensic computing – A case study, *Digital Investigation*, vol. 1(1), pp. 18–23, 2004.

[2] B. Carrier, A hypothesis-based approach to digital forensic investigations, Technical Report 2006-06, Center for Education and Research in Information Assurance and Security, Purdue University, West Lafayette, Indiana, 2006.

[3] C. Fidge, Logical time in distributed computing systems, *IEEE Computer*, vol. 24(8), pp. 28–33, 1991.

[4] P. Gladyshev and A. Patel, Formalizing event time bounding in digital investigations, *International Journal of Digital Evidence*, vol. 4(2), 2005.

[5] L. Lamport, Time, clocks and the ordering of events in a distributed system, *Communications of the ACM*, vol. 21(7), pp. 558–565, 1978.

[6] B. Schatz, G. Mohay and A. Clark, A correlation method for establishing the provenance of timestamps in digital evidence, *Digital Investigation*, vol. 3(S1), 98–107, 2006.

[7] M. Weil, Dynamic time and date stamp analysis, *International Journal of Digital Evidence*, vol. 1(2), 2002.