Digital Investigation 24 (2018) S83–S92

Contents lists available at ScienceDirect

Digital Investigation

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

DFRWS 2018 Europe — Proceedings of the Fifth Annual DFRWS Europe

Controlled experiments in digital evidence tampering

Felix Freiling*, Leonhard Hösch

Department of Computer Science, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany

Keywords:
Anti-forensics
Digital evidence
Forensic computing education

A B S T R A C T

We report on a sequence of experiments performed with graduate level students on the tampering of digital evidence. The task of the study participants was to manipulate a given disk image so that it looked as if a website had been accessed and images downloaded in the past. Later, the same students had to distinguish their forgeries from a set of originals in which the images actually had been downloaded. During all parts of the experiment, efforts were recorded in project diaries. Overall, the results show that the tampering task was difficult since none of the forgeries was taken as an original. Furthermore, the analysis effort to detect forgeries consistently was below the effort to create the forgery even in the worst case scenario where the manipulator had full control over the evidence. It also required generally less effort to correctly classify an original than to correctly classify a forgery. Additionally, we derived results confirming that the effort to construct consistently manipulated evidence increases with decreasing control, i.e., the ability to precisely act upon the evidence.

Introduction

In the history of forensic science, there is a well-established tradition to document the experiences in handling and interpreting evidence, which in the last 150 years over-whelmingly has been *physical* evidence (Kirk and von Thornton, 1974; Groß and Geerds, 1977; Lee and Harris, 2000). Today, however, an increasingly large portion of evidence in criminal cases is *digital* evidence, i.e., evidence that is stored on or transmitted over digital media (Casey, 2011). There has been much philosophical debate about the "nature" of digital evidence and how it differs from the characteristics of physical evidence (Dardick et al., 2014; Paul, 2009). One of the resulting issues is the volatile nature of the binding between support and information, which makes digital evidence — at least in principle — more susceptible to manipulation. In the community of digital forensic analysis, there appear to be ambivalent opinions whether digital evidence can be perfectly tampered. Many appear to believe this to be the case, as expressed by Caloyannides (2003), who states that digital data can be manipulated at will, and depending on the manipulator's sophistication, the alteration can be undetectable, regardless of digital forensics experts' competence and equipment.

Digital forensics processes have tried to cope with this problem and long since established standard processes that try to contain the dangers of manipulating digital evidence. The most prominent of theses methods is the use of cryptographic hash functions to document the state of a string of bits within the chain of custody. But even though cryptographic hash values are an established part of digital forensics processes, they cannot help to detect manipulations that have occurred *before* evidence collection, be it either direct tampering by the suspect or evidence planting by corrupt law enforcement officers.

Related work

Questions of manipulation appear to surface in a non-negligible number of practical court cases. Briefly spoken, there are often two opposing views:

1. One side, usually the prosecutor, claims that the state of collected evidence is consistent with a particular hypothesis H_1 of how the crime happened.
2. The other side, usually the defendant, claims that the collected evidence was manipulated such that it appears as if hypothesis H_1 were true, but in fact a different and opposing hypothesis H_2 is consistent with the evidence, which includes the manipulation and a different sequence of events implied by H_1.

In the literature, this phenomenon has been termed the "Trojan Horse Defense" (Brenner et al., Henninger) where the defendant claims that not he or she committed the offense (such as

* Corresponding author.
E-mail addresses: felix.freiling@cs.fau.de (F. Freiling), leonhard.hoesch@fau.de (L. Hösch).

https://doi.org/10.1016/j.diin.2018.01.011

performing a cyberattack or downloading illegal documents) but rather a Trojan horse installed on the computer on behalf of a third party. Unfortunately, there is not much concrete advice in the literature on how to technically deal with such cases apart from general statements such as the following [8, p. 49]:

The investigator should use all available resources to determine if a remote person could have used the application to commit the crime or to install additional software that could have committed the crime.

Obviously, in such cases also other forms of evidence are critical, e.g., the fact whether the suspect has sufficient knowledge to manipulate, hide or wipe evidence from the system. But little is known on the actual effort it takes to change digital evidence such that the "true" hypothesis H_2 might be mistaken for the "false" hypothesis H_1.

Moch, (2015) reports on some preliminary experiments with students that were instructed to manipulate evidence as follows: Students were given two disk images I_1 and I_2, where I_2 resulted from I_1 by executing a certain action (like sending a message on ICQ). The task of the students was to mount I_1 and manually change the evidence such that the same evidence as that in I_2 was present but without actually performing the action that originally created the evidence. The results showed that it was extremely difficult to perfectly manipulate the evidence because manual operations (like invoking vi or touch) create other artifacts (such as swap files or timestamps of value zero).

Another experiment by Moch, (2015) exhibited a problem to swap the content of two files in an ext4 file system because the metadata (especially the inode number) remains unchanged. Therefore, the set of differences between the manipulated image and the target image I_2 (calculated using `idifference` (Garfinkel, 2012)) often was not empty although students had perfect control over the image and the manipulation process. In conclusion, perfect manipulations (defined as an empty difference set) were possible but needed an extreme effort and care for detail. Unfortunately, this effort was not quantified.

Related to the problem of evidence manipulation is the area of *anti-forensics*, meaning "any attempts to compromise the availability or usefulness of evidence to the forensics process" (Harris, 2006). Interestingly, the literature on anti-forensics has mainly focused on rather obvious and aggressive techniques, such as hiding or encrypting evidence (Berghel, 2007; McDonald et al., 1999), overwriting/wiping evidence (Foster and Liu, 2005; Savoldi et al., 2012) or attacks against investigative tools (Wundram et al., 2013). Maybe the most advanced area in the analysis of manipulated (or counterfeit) evidence is multimedia security, e.g., where methods of blind image forensics can be used to detect manipulations (Johnson et al., 2006; Lin et al., 2009). However, we are not aware of any literature with a similar intention focusing on non-multimedia files.

When speaking to experienced investigators, many might agree that digital evidence can *theoretically* be manipulated perfectly, but *in practice* it is very hard to do this and not make mistakes. So while this indicates that every manipulation appears to also leave traces that can be detected, we are not aware of work that has systematically explored the effort to perform targeted evidence manipulation.

Research goal and contributions

With the increase of the amount of digital evidence in court, it must be expected that also the number of attempts to counterfeit, manipulate or forge such evidence will increase. Therefore, expert witnesses in digital forensics should be prepared to react to efforts by any of the opposing sides in the spirit of the Trojan horse defense. In this direction it is not only necessary to question the competence and motivation of a suspect to forge evidence, but also to

1. look for evidence of manipulation and
2. in case no such evidence can be found, to understand the effort necessary to perform such perfect manipulations.

In analogy to the handling of physical evidence one can then argue, that if the effort for evidence manipulation is very high and there is no evident motivation or competence of the suspect to forge evidence, then it is more probable that there has been no manipulation than the opposite.

In this paper we study the effort to perform an evidence manipulation task by running a controlled experiment within a graduate level course on digital forensics at Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) in Erlangen, Germany. Our focus was to study the class of manipulations that *make forgeries look like originals* and where no trace of tampering could be found. This is in contrast to many practical cases of tampering where evidence is blatantly overwritten or destroyed.

More specifically, the task involved to manipulate a given system disk image so that it appeared that files had been downloaded from a particular website in the past whereas in fact they had not been downloaded at that time. Independently we prepared a set of original evidence (i.e., evidence where the download action had actually happened). After performing the manipulation task, students had to investigate a randomly selected disk image and had to determine whether it was original or fake. During all of these activities, students were required to document their actions and log their effort in a project diary. The goal was to study the success probability of the manipulation attempts, the effort it takes and the factors that influence the quality of a forgery.

Overall, 14 students participated in the experiment and in this paper we report on the results of the analysis of the collected data. Since we were not aware of related work that performed similar experiments before, we could only state rough research questions instead of exact hypotheses to evaluate. Still, the following statements can be drawn from the data:

- All forgeries produced within our experiments were correctly classified as forgeries. This means that it appears to be generally hard to produce a convincing forgery in the given case.
- It required generally less effort to correctly classify an original than to correctly classify a forgery. This appears surprising since one would expect that forgeries can be classified as soon as obvious signs of tampering are found, whereas such signs are absent in originals.
- *Producing* a forgery generally requires more effort than *detecting* that a forged image is a forgery. This holds even in perfect manipulation environments where there are no restrictions on the tools and methods used to produce a forgery.
- Less control over the manipulation process (reduced toolset, more uncertainty over evidence) increases the effort to produce a forgery and reduces the effort to correctly detect a forgery as fake.

While our findings are limited and can be described as preliminary, we believe that the collected data, which is available online (Freiling and Hösch, 2018), will be helpful to shape further experiments in this relevant field in the future.

Paper outline

This paper is structured as follows: We first formulate the research questions along which the study was designed in Section 2. We then describe the experimental design of our study in Section 3. We report on the detailed quantitative (Section 4) and qualitative results (Section 5). We conclude in Section 6.

Research questions

Since nothing was known about the factors that influence the manipulation process, we are only able to formulate exploratory research questions instead of exact research hypotheses.

Scenario: disk image manipulation

The focus of our experiments is forgeries of disk images. Inspired by Brenner et al (Brenner et al., Henninger). the scenario was a case involving the download of illegal files from a website within a time period *T*. Before the beginning of *T*, we prepared a disk image *R*, a fully functional system disk including the installation of a standard browser. We ensured that previous usages of *R* did *not* include accesses to a particular website *W* before *T*. During the course of *T* and based on *R*, we generated a set of disk images $\{Q_1, Q_2, \dots\}$ that accessed website *W* and downloaded files from that website. These images are called *originals*.

We gathered a set of human participants for our experiment. At the end of *T*, these participants were given access to *R*. The manipulation task was to change *R* into *S* in such a way, that a forensic analyst investigating *S* comes to the conclusion that there had been accesses to *W* within *T* on *S*. In practice, such a scenario corresponds to cases where evidence of illegal activities is planted on a seized disk image by first responders performing evidence acquisition at the crime scene. Such an image *S* is called a *forgery*. Note that the construction of the manipulation task ensured that participants could not themselves produce originals.

Characterization of disk image tampering and tamper detection

The overall quantitative goal was to evaluate the success rate of participants performing the manipulation task. Intuitively, this was measured by relating false verdicts to true verdicts during the investigation phase (e.g., classifying a forgery as original or vice versa).

Independently, we were also interested to qualitatively investigate the different approaches that participants took to

1 perform tampering to produce forgeries in the given scenario, and to
2 determine whether a given disk image was an original or a forgery during the subsequent investigation.

With our experiment we hoped to find different classes of approaches that could be distinguished.

Factors influencing the quality of forgeries

Another research goal was to understand the factors that influence the quality of forgeries. Especially interesting were factors that could be used to predict the success of forgery production and forgery detection.

One possible factor we considered was the effort spent by participants both on producing a forgery and detecting a forgery. Among the numerous ways to define effort, we decided to measure the *effort* in the number of work minutes spent on performing the task because this was the simplest measure available.[1] We conjectured that the effort spent to create a forgery directly positively influences the effort to correctly classify it as a forgery, meaning that the more effort you spend to forge, the more difficult it is to correctly detect.

Another factor that we considered was the amount of "control" of the tampering process by participants. Here we distinguished *perfect* control, where participants had full access to the original image *R*, from *partial* control, where participants only knew the version of the installed operating system and browser and had to write a program that performed all modifications (see below). We conjectured that the more control you have over the image, the easier it is to create a good forgery, meaning that the manipulation effort in relation to the analysis effort is smaller.

We also considered other factors that might be able to predict the quality of a forgery, namely the prior experience of participants in forensic analysis and (more qualitatively) the tampering approach used by the participants.

Experimental design

We now describe in more detail the experimental design of our study.

General setting

The setting of the experiment was an advanced course on digital forensics at Friedrich-Alexander-Universitat Erlangen-Nurnberg (FAU) that was held between October 2016 and February 2017 with roughly 20 students. All students were computer science majors that had advanced technical skills and had successfully participated in an introductory course on digital forensics with an extensive file system forensic analysis part before October 2016. Participation in the experiment was mandatory, but the result was not graded and therefore not part of the grade of the course.

During the start of the experiment, participants were briefed about the nature and direction of the experiment. This included the instruction that participants work independently to minimize unwanted interference. Participants had to fill out a pre-questionnaire collecting data on their degree, year of study, experiences in working on forensic cases, planned effort and motivation for the experiment.

To decouple participation in the study from participation in the course, pseudonyms were used to link individual data. Pseudonyms of participants were deterministically constructed by taking defined parts of personal data like the first two letters of the place of birth and the last two digits of the date of birth of the mother, thereby allowing participants to "not forget" their pseudonym. After finishing the experiment, we exchanged pseudonyms by (random) identifiers in our data set.

The selection of participants ensured that all had prior experience in forensic analysis. Equally, because all participants had to construct a forgery before they were allowed to investigate a given disk image, all participants were aware of the details of the tampering task and the problems involved in creating a forgery.

The tampering task

The starting point *R* of our experiment was a current Ubuntu Linux disk image that has been used in a previous class on digital forensics which most participants had analyzed previously. It contained a user account `werner` and some minor browser and messaging activity from May 2016 that had been embedded in a "case" during the previous course. Participants also had access to a well-written and complete analysis report that documented and interpreted all evidence on R.

The tampering task was now specified as follows:

[1] A suggestion by a reviewer was to normalize effort based on the experience of the participant since good students usually work faster. Since participants had to report on their own efforts we believe that normalizing effort in this way could introduce more bias. This would be different if the effort was measured by an independent observer.

Manipulate *R* such that a forensic investigator will reach the conclusion that the website *W* has been visited and images (picture files) have been downloaded from that website during time period *T*.

Parts 1 and 2: perfect vs. partial control

We structured the experiment into two parts:

- Part 1: Tampering with perfect control over disk image.

  In this part of the experiment, participants had full access to the image R and were not restricted in any way regarding the tools that could be used to perform the tampering task. This means that they could manually change every bit in R using tools like hex editors or they could boot a copy of R within a virtual machine (VM) and in such a way perform experiments with the image

- Part 2: Tampering with partial control over disk image.

  In this part, participants only were given some basic knowledge of installed users, version number information on the installed operating system and applications on the target system *R*. Using this information, participants had to write a program that was run on the target machine with root privileges and should perform the tampering task.

Clearly, the manipulation task in Part 1 of the experiment was easier than in Part 2 because evidence manipulation could be performed purely from "outside" the image. In Part 2, evidence manipulation had to be performed from "within" the image, which also meant that the tampering approach had to hide its own traces in the resulting forgery.

Phases A and B: tampering and investigation

Each part had two phases:

- Phase A: Manipulation.

  In this phase, participants received information on the original disk image *R* and were instructed to produce the manipulation artifact together with a one-page description of their manipulation approach.

- Phase B: Analysis.

  In this phase, participants received the image of a hard disk and were instructed to answer two questions:

  1 Is there evidence of accesses to website *W* during time period *T*?
  2 Is there evidence of downloading images from *W* during time period *T*?

A separate notice in the investigation request mentioned possible tampering by the law enforcement officer that has acquired the disk image, and that participant should not fall for forged evidence.

Participants had to write a brief investigation report of at most 5 pages in which the results of their analysis were documented.

Sequence of experiments

We denote the overall sequence of experiments as 1.A, 1.B, 2.A and 2.B, where the number refers to either Part 1 or 2 and the letter refers to the Phase A or B within that part. The experiment started on October 21, 2016. Participants had about three weeks to perform each of the tampering phases 1.A and 2.A and about 10 days to perform the analysis phases 1.B and 2.B. The complete timeline is shown in Table 1. For each phase of the experiment the participants were required to log their effort in a project diary. To prevent a large number of fragmented entries, only continuous efforts of at least 30 min had to be reported.

Details on Part 1 (perfect control)

Fig. 1 depicts the design of Part 1. In Phase 1.A participants received *full* access to *R* and the desired artifact was a forged disk image *S*. The time span *T* in the tampering task was the period between October 2 and 17, 2016, and the website *W* was the German Wikipedia page on the term "rhinoceros", a tribute to the 2006 DFRWS Forensic Challenge (Richard, 2005). The period *T* was chosen such that it lay in the past when the course started. This ensured that participants could not themselves produce originals.

Next to a forged disk image, participants also were requested to write a one-page summary of the "design" of their manipulation and answer the following three questions:

1. How did you proceed?
2. What problems did you encounter?
3. How did you validate your result?

The resulting forgeries from Phase 1.A were added to a pool of original images produced by us (see Fig. 1). For every participant in Phase 1.B, we drew an image randomly from that pool and ensured that there was equal probability to receive an original and a forgery. Explaining this process was part of the initial briefing of participants so that it was clear that simply guessing whether an image was an original did not help. During that briefing, general advice on how to detect forgeries was given, e.g., looking for inconsistencies in log files and remains of manipulation or anti-forensic tools.

Details on Part 2 (partial control)

The design of Part 2 is visualized schematically in Fig. 2. As mentioned above, in Part 2 participants could not directly access image *R* but rather received information on the versions of software to be expected. The desired manipulation artifact was a program that performed the image manipulation within the described context.

In our case the system was a standard Linux installation of Ubuntu 16.04 x86-64 with at least two users (`werner, root`). This information implied the version of the installed browser. Additionally participants could assume the availability of Python 2.7 and GCC 5.4. Participants should assemble their manipulation program together with supplementary files in a tar-archive in which an executable program named run should be the entry point of the execution. As in Phase 1.A, participants were required to write a one-page "design document" about their solution.

For completeness, we note that the time period *T* that was relevant for the tampering task in Part 2 was the time between November 20 and December 1, 2016. The website *W* was the same page in Wikipedia as in Part 1.

Table 1
Timeline of experiments.

Oct. 16–31, 2016	Pre-questionnaire accessible
Oct. 18, 2016	Briefing
Oct. 21–Nov. 15, 2016	Phase 1.A
Nov. 22–Dec. 2, 2016	Phase 1.B
Dec. 2, 2016–Jan. 10, 2017	Phase 2.A
Jan. 17–31, 2017	Phase 2.B
Feb. 7, 2017	Debriefing

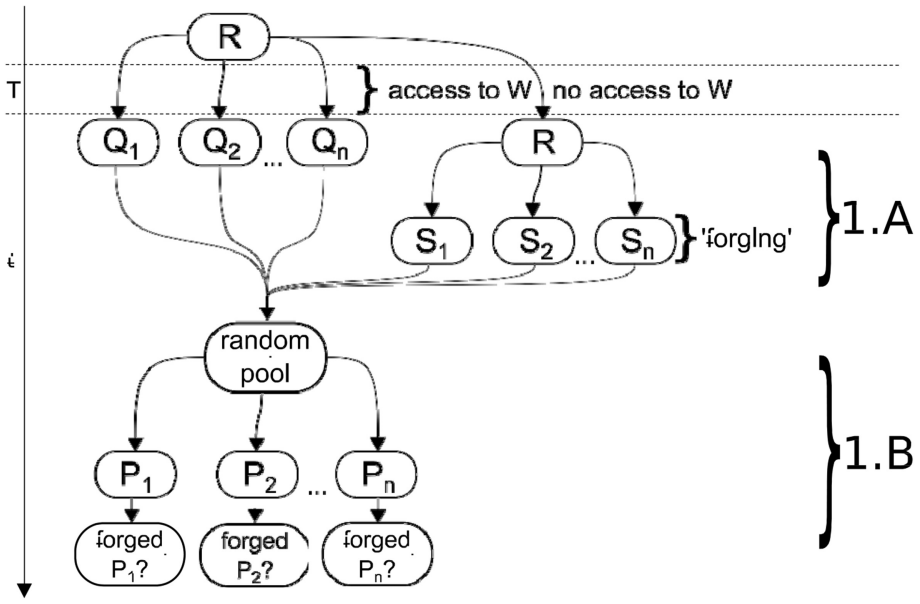

Fig. 1. Experimental design of Part 1 (perfect control).

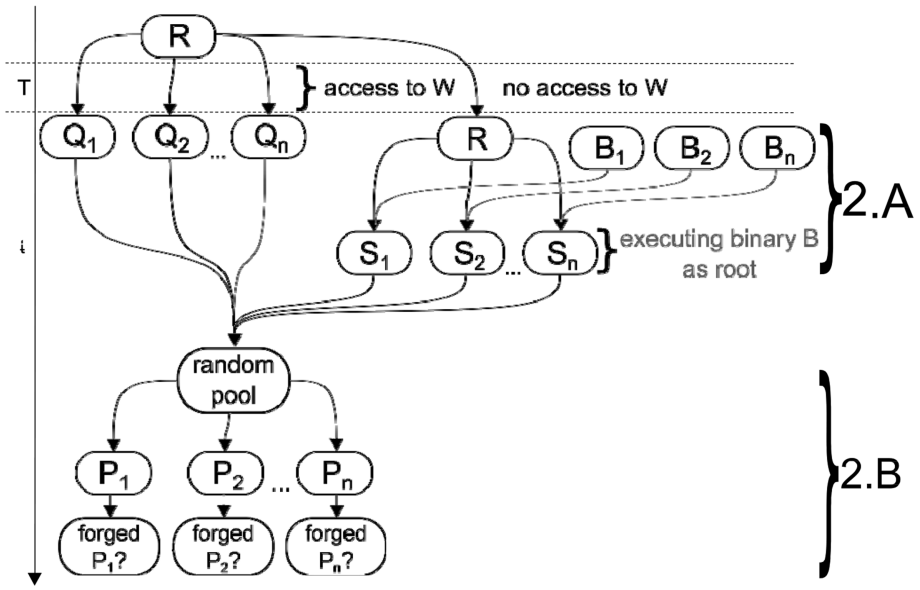

Fig. 2. Experimental design of Part 2 (partial control).

After receiving all programs from the participants, we automatically unpacked and executed them within the context of a virtual machine that was booted from the image that we had prepared (see Fig. 2). We waited for 20 min and then shut down the machine. The resulting forgery image was collected in a pool of forgeries.

Like in Part 1, we added a couple of originals to that pool and assigned elements from that pool to participants such that there was an equal probability to receive an original or a forgery.

Quantitative results

We now report on the qualitative and quantitative results of our experiments. We first give an overview of the participants from the pre-questionnaire and then discuss the findings related to the research questions formulated in Section 2.

Participants

Initially, 20 students had registered for the lecture. Of these, 14 participants were recruited and answered the questions from the pre-questionnaire. According to the prequestionnaire, all 14 participants were computer science students, all 14 were taking the course as part of a Master's degree although one participant was formally still enrolled for a Bachelor's degree. Most participants were still within their first year of their Master's studies.

Most participants agreed with the statement “I am motivated to take part in the experiment” (fully agree: 4, agree: 5, neutral: 2, disagree: 2, fully disagree: 1). Most participants (11) stated that this course was one of 5 which they were taking that semester, three participants were taking 3 other courses. Therefore the work load of all participants during the semester promised to be comparable.

Next to the motivation, we also asked how much of the weekly work time (in percent) were participants willing to invest in the course. Most participants stated that they were planning to invest 20–30% (8). Three were willing to spend more than 30% and three others less than 20%.

All participants stated that they had followed the introductory forensics course. Asked for their grade (range) in that course, 10 participants stated that they had received the marks 1.0 (best grade) or 1.3, three had received a grade between 1.7 and 2.3. One participant stated that she had received a grade between 3.7 and 4.0, which is the worst grade that still lets you pass the exam. All had similar experience in the analysis of disk images (between 2 and 6), mostly from the introductory forensics course.

As it is common in many courses, the number of participants declined throughout the semester. The number of participants taking part in the individual phases of the experiments is shown in Table 2. As mentioned above, participants were assigned a deterministic pseudonym that was used to link data items in the experiment. In this paper, we exchanged these pseudonyms with a sequence number so that individual participants are named "Participant 1" to "Participant 13".

Manipulation effort

All efforts were measured in minutes and the data was taken from the project diaries handed in by participants after every phase. The manipulation effort of the participants in Phases 1.A and 2.A is depicted in Figs. 3 and 4 and summarized in Table 3. It shows that the manipulation efforts in both parts were about the same on average (almost 9 h in total), however, the range of values (indicated by the standard deviation) in Phase 1.A was much larger than in Phase 2.A. This can possibly be explained by experience or exhaustion effects. The median indicates that the manipulation effort was in general higher in Phase 2.A than in Phase 1.A.

Success of manipulation

The success rate of the manipulation attempts by all participants in Part 1 is shown in Table 4. All investigated forgeries (6) and 4 out of 5 originals were correctly classified. Interestingly, one original was classified as a forgery. This is further summarized in Table 5. It shows that on average, it appears to be considerably easier to correctly classify an original (386 min) than a forgery (618 min).

The corresponding data for Part 2 is given in Tables 6 and 7. Since the task was to produce a program in Phase 2.A, we had the problem that one of these programs did not execute. Since this image was easy to identify as a forgery and no evidence was planted, we excluded this "forgery" from the data.

In Part 2 (partial control), *all* resulting images were correctly classified either as original or forgery. As in Part 1 (perfect control), it appears harder to correctly identify a forgery than to correctly identify an original. The average analysis efforts are both lower than in Part 1, which might indicate that it is easier to detect a forgery if the adversary has only partial control over the manipulation process. However, with only 4 investigated originals and 2 forgeries the statistical basis is very narrow.

Table 2
Number of participants in the individual phases of the experiments.

Phase	Number
Course registration	20
Pre-questionnaire	14
Phase 1.A	13
Phase 1.B	11
Phase 2.A	8
Phase 2.B	6

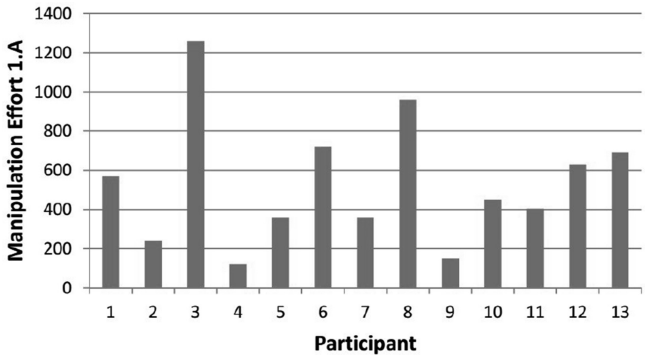

Fig. 3. Effort spent on creating a forgery in Part 1.A.

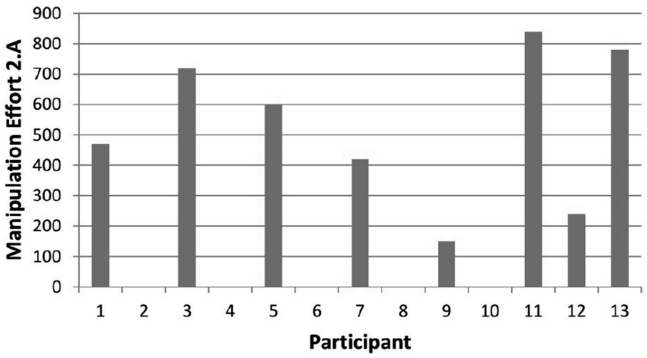

Fig. 4. Effort spent on creating a forgery in Part 2.A.

Table 3
Manipulation effort measured in minutes in Phases 1.A and 2.A.

	Phase 1.A ($n = 13$)	Phase 2.A ($n = 8$)
Min/max	120/1260	150/840
Average	532	528
Stddev.	325	252
Median	450	535

Relations between manipulation and analysis effort

Since (almost) all images were correctly classified, we unfortunately cannot determine the factors that influence wrong classifications. However, we can investigate the relation between

Table 4
Manipulation detection results of Part 1; "Rcvd" indicates either participant number who produced forgery or fact that participant received an original; "Effort 1.A" is the effort invested into forging the image that the participant analyzed (if it is not an original).

Participant	Rcvd	Effort 1.A	Effort 1.B	Correct?
1	orig	–	855	Y
2	orig	–	270	Y
3	11	405	480	Y
5	6	720	650	Y
6	orig	–	180	Y
7	orig	–	240	Y
8	13	690	1440	Y
9	8	960	270	Y
11	1	570	90	Y
12	3	1260	780	Y
13	orig	–	900	N

Table 5
Overview of classification results in Part 1 (with analysis effort ranges and average analysis effort).

	Classified as original	Classified as forgery
Original	4	1
Effort range	180–855	900
Effort average	386	900
Forgery	0	6
Effort range	–	90–1440
Effort average	–	618

manipulation effort and analysis effort in both parts of the experiment.

For the six forgeries that were investigated in Phase 1.B of the experiments, we plotted the manipulation effort of Phase 1.A in relation to the analysis effort in Phase 1.B. This plot together with an exponential trend line is shown in Fig. 5. In general it appears as if higher manipulation effort really result in higher analysis efforts and that the analysis effort was (apart from one outlier) always below the manipulation effort. The exponential trend line turns slightly upwards, indicating that with an increasing manipulation effort eventually it might cause an equal analysis effort.

For completeness purposes we plotted the same data for Phase 2.B (see Fig. 6). Interestingly, although it contains only two data points, in both the manipulation effort is less than the analysis effort.

Influence of control on effort

Now we turn to the manipulation efforts of individual participants and see how they developed during the experiment from Part 1 (perfect control) to Part 2 (partial control). Overall, eight participants created forgeries both in Phases 1.A and 2.A. We put these efforts in relation and the resulting plot is shown in Fig. 7. It shows generally that the more effort individual participants spent in Phase 1.A the more they also invested in Phase 2.A. Interestingly, the manipulation effort in Part 2 (partial control) of the experiment was generally higher with less control, i.e., it appears "easier" to produce forgeries if you have full control over the manipulated system.

Because increased control appears to reduce manipulation effort, we might expect that increased control will result also in an increased analysis effort, since it is less costly to produce a good forgery with increased control. For the six participants that performed image analysis in both parts of the experiments, we show the relation between the analysis efforts in Fig. 8. This confirms that increased control also increases the analysis effort.

Factors from pre-questionnaire

We also investigated the influence of items we asked in the pre-questionnaire on the effort invested by participants. There was a slightly positive correlation between motivation, grade of prior course and experience (year of 9 study) to manipulation and analysis efforts in Part 1. This means that students that were more motivated, had better grades and had more study experience spent more effort in manipulation and analysis. Since (almost) all analysis results were correct, we could not study the influence of these factors on the quality of the analysis.

As might have been expected, there was no obvious correlation between the effort (work hours per week) planned by participants and real effort spent.

Table 6
Manipulation detection results of Part 2 (partial control); "Rcvd" indicates either participant number who produced forgery or fact that participant received an original; "Effort 2.A" is the effort invested into forging the analyzed image (if it is not an original).

Participant	Rcvd	Effort 2.A	Effort 2.B	Correct?
1	orig	–	186	Y
3	11	840	270	Y[a]
5	9	150	420	Y
7	orig	–	240	Y
11	orig	–	90	Y
12	orig	–	720	Y
13	12	240	280	Y

[a] Indicates a manipulation program that failed to execute, therefore this row was excluded from the results.

Table 7
Overview of classification results in Part 2 (with analysis effort ranges and average analysis effort).

	Classified as original	Classified as forgery
Original	4	0
Effort range	90–720	–
Effort average	309	–
Forgery	0	2
Effort range	–	280–420
Effort average	–	350

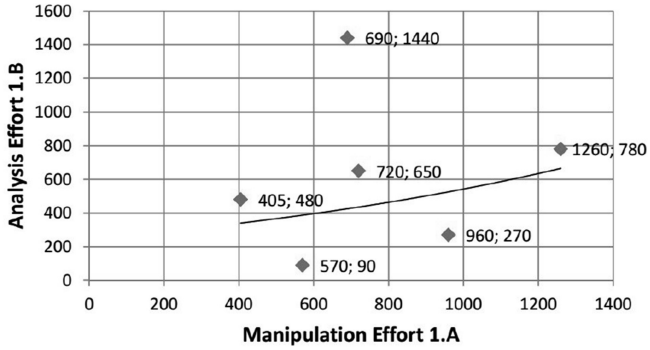

Fig. 5. Effort on detection vs. effort on manipulation in Part 1 (perfect control).

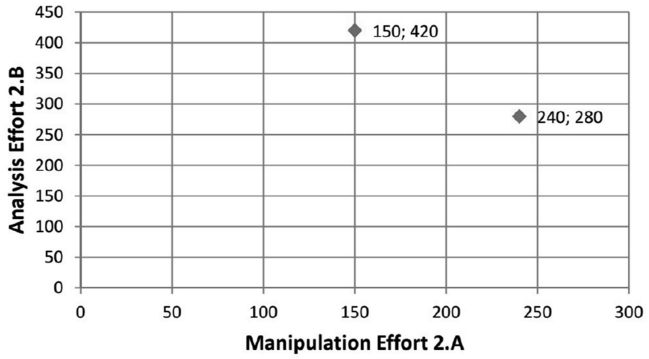

Fig. 6. Effort on detection vs. effort on manipulation in Part 2 (partial control).

Qualitative results

We now turn to more qualitative considerations by looking partly at the concrete tampering approaches taken by the participants and why they were correctly classified as forgeries by other participants. We concentrate on Part 1 of the experiments since the insights from having full control over the evidence shed light also on scenarios where there is less control and since we had so little data from Part 2. We also look at the single case where one participant falsely classified an original as forgery.

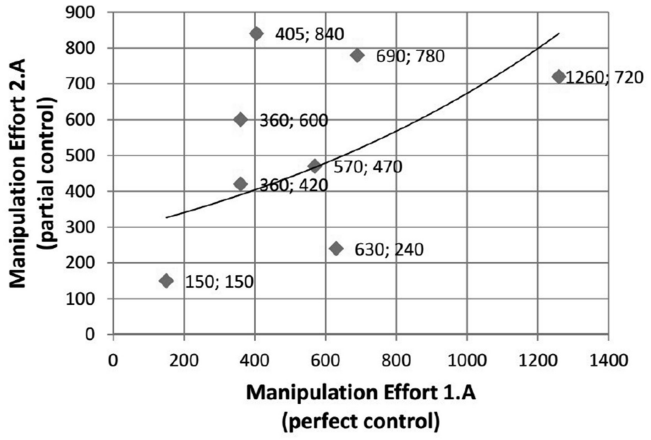

Fig. 7. Manipulation effort in Phase 1.A and 2.A of individual participants ($n = 8$).

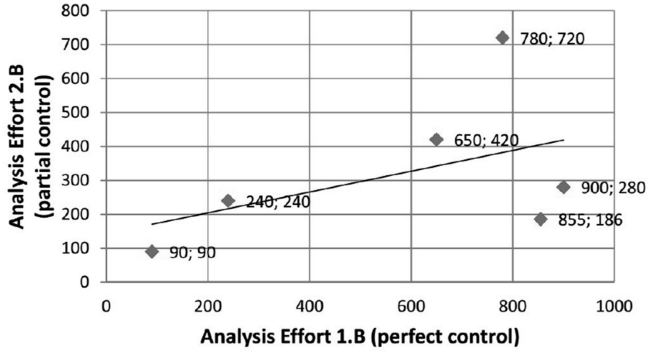

Fig. 8. Analysis effort in Phase 1.B and Phase 2.B of individual participants ($n = 6$).

Forgeries by participants

Participant 1

Participant 1 booted the original image within a VM, performed the web accesses and the image downloads, and then manipulated all timestamps (MAC times of downloaded images and browser history (`places.sqlite`). Three relevant log files (`syslog, auth.log, bash history`) were also manipulated regarding content and timestamps. The comment that she could not switch off the NTP time synchronization service and the evidence found by the analyst indicate that all manipulations were being performed from within the running system.

The image was analyzed by Participant 11, who found many indications of manipulation. The clearest indications were timestamps after time *T* in many relevant logfiles, e.g., of the browser (`places.sqlite, cookies.sqlite`), the Zeitgeist activity manager (`activity.sqlite`), and recently used files in the Ubuntu dash. The file auth.log showed three successful logins after time T, and an installation of an sqlite editor was documented after time T, which in this context was taken as the installation of a manipulation tool. The local `bash_history` clearly documented manipulation attempts, e.g., setting the hardware clock, opening log files like `syslog` with `gedit`, or invocations of `sqlite3` on `places.sqlite`. The analyst concludes that the evidence "speaks a clear language of manipulation".

Participant 3

Participant 3 also attempted to manipulate the original image by booting it in a VM and performing manipulations from within. First she made backup copies of all relevant log files. Then she determined a download time that was consistent with the login times of the user and performed the relevant web accesses. Originally she had planned to set back the hardware clock and then perform web accesses, but this caused many certificate checks to fail, so web accesses were performed in real time. Afterwards, she stopped time synchronization, set back the system clock and performed logfile manipulations using the originally backuped files. For these steps she had written programs that performed the manipulation automatically. Special care was taken to insert the manipulated web accesses into the browser history `places.sqlite` in such a way that created consistent primary key (id) entries. These entries are created automatically by Sqlite and act like logical timestamps since they use Sqlite's autoincrement feature. Finally, the superfluous backups were deleted, original `bash_history` was reinserted and MAC times of all files were set to the value of their backuped counterparts.

The analysis, which was performed by Participant 12, found browser files to be consistent, but many of the logfiles (such as `wtmp` and `auth.log`) to be empty. It appears that re-installing the backup files had not worked. The MAC times of the downloaded images files were also set to a time after T. The analyst also found remains of the `bash_history` containing commands to manipulate the hardware clock (`sudo date`).

Participant 6

From looking at the `/var/log/wtmp login` database on the original image, Participant 6 first identified a suitable time in the past at which the user had been logged in and the web accesses therefore could have taken place. She then booted the image inside a VM and performed the web accesses and downloaded two images while keeping track of all files that were changed or added by these actions. These files were analyzed and changed outside of the VM. She then wrote a program that could manipulate timestamps within the complex Cache2 format of Firefox. Timestamps within sqlite files were manipulated manually (using SQL commands). Finally, these files were injected into a mounted copy of the original image. Subsequently, the `inode` timestamps of these files were dated back using a script and the `debugfs` tool.

The analyst of this image (Participant 5) found the two rhino images and did not find any obvious inconsistencies within the cache and history files of Firefox. However, she observed that the download of one of the two images was missing in the browser history and that the MAC times of the images were inconsistent with the timestamps in the browser history (they showed an earlier timestamp than the download).

Participant 8

Participant 8 performed a similar strategy as Participant 3, only that file manipulations were performed from the outside (i.e., after mounting the image). She took backup copies of all relevant files, then booted the image within a VM and performed the web accesses. After shutting down the VM, she systematically compared the differences between the old and the new log files. These observations were used to "merge" the new data into the old log files using a suitable time for the downloads which was found issuing last while the VM was running. Particular effort was spent on a consistent manipulation of the browser history (`places.sqlite`) and setting the MAC times of the downloaded files.

Participant 9, who analyzed this image, found that the image contained logins after the time period *T*. Obviously, Participant 8 had performed the file manipulations on the copy of the image which she had booted in the VM and not on the original image. This, together with an inconsistent timestamp in the `places.sqlite` was taken as an indication of manipulation.

Participant 11

Participant 11 originally planned to boot the original image in a VM, set back the hardware clock within the VM, and then download the images onto the system. Setting back the time, however, was the cause of failed certificate checks on the server of Wikipedia, so this approach was aborted. In the end, Participant 11 ended up with manual manipulations of all relevant files on the image: Firefox history and cache, Firefox cookies, the Zeitgeist activity scanner file (`activity.sqlite`), recently used files in the Ubuntu Dash, also MAC times of the file system. The participant commented that it was "a lot of work" and that she was still not sure whether she had overlooked anything.

The analysis by Participant 3 found that the manipulated files has MAC timestamps that were inconsistent with the download history, i.e., their creation time was after time *T* and the last modified and access times were *before* the downloads. This was taken as an anomaly indicating manipulation.

Participant 13

Participant 13 took a different approach to manipulation in that she downloaded the images and then deleted all relevant files, i.e., `bash_ history` and the downloaded images. The browser history was deleted from within Firefox to avoid a deleted version being found using carving. These actions were embedded into a story of user werner talking to a "dealer" of rhino images who gives instructions on how to delete evidence. Additionally, the story contained another chat user getting external access to the system who planted a malicious ZIP bomb in the file system and inserted the command `rm -rf` into the `.bashrc` file of user `werner`.

Participant 8, the analyst of this image, had a hard time figuring out what had happened since she could only find the names of the deleted files and not their content. However, she was able to recover the chat protocol of the user, found evidence of external accesses by another party, and found indications of changes of system times in auth.log.1. She also found the ZIP bomb and the malicious command in `.bashrc`. All this and especially timestamps after time T supported the conclusion that evidence was inconsistent and system manipulation probable.

On the effect of manipulation approaches

Overall it is very interesting to follow the manipulation and analysis approaches of all participants. Interestingly, the highest analysis effort was caused by the image of Participant 13. This indicates that the removal of evidence complicates analysis severely, but obviously makes the evidence inconclusive (in favor or against the suspect).

Two of those manipulation attempts where participants made obvious errors, i.e., Participant 1 and Participant 8, were identified as forgeries very quickly, i.e., after 90 and 270 min, respectively (see Table 4). It is not clear why Participant 12, who analyzed the forgery of Participant 3, took in total 780 min to perform the analysis despite obvious forging mistakes.

The manipulation attempts that were closest to being taken as successful forgeries, i.e., those of Participants 6 and 11, needed an analysis time of 480 and 650 min, respectively. This is still more than the average analysis time for originals, which was 386 min (see Table 5). Overall, it appears that manipulations from outside the image are more convincing and less error prone than manipulations from within the image. This re-inforces the issue of *control* over the evidence. Furthermore, the necessity to manipulate evidence of network interactions complicated the manipulation task considerably. Such interactions contain much non-deterministic and non-replayable evidence, e.g., involving cookies and certificates. Constructing this in a consistent manner often involves some form of guessing, which increases the probability to make mistakes.

The false positive

Participant 13 caused the only false positive in our experiments, identifying an original as a forgery. Reading her analysis report, the verdict in favor of a forgery is very weak ("tend to believe that the image is a forgery"). The main evidence on which this verdict is based are two deactivations of the logging service and indications of time manipulations in syslog and `bash_history` (`sudo hwclock`). All other evidence (browser files, MAC times), however, are consistent, as the analyst reports.

Summary and Conclusions

We performed a sequence of experiments within a graduate level course on digital forensics with the aim to understand the effort and factors influencing the tampering of digital evidence. The focus was on the production of *perfect* forgeries, i.e., evidence where a forensic analyst comes to the conclusion that the evidence is not manipulated and actions logged in the evidence have actually happened.

Interestingly, all forgeries produced within our experiments were correctly classified as forgeries. One original was even classified as a forgery. This means that it appears to be generally hard to produce a convincing forgery in the given case. Our data also indicates that producing a good forgery generally requires more effort than the one it takes to detect that a forged image is a forgery. This is good news for investigators, because they can allocate their resources according to the level of manipulation competence of the suspect. As might be expected, we also found that less control over the manipulation process (in our case: automated external control though a program) increases the effort to produce a forgery and reduces the effort to correctly detect a forgery as fake. This is also good news for practicioners since their analysis effort can be reduced if the manipulation means of the suspect are technically restricted.

The results of our controlled experiments may not be statistically significant, but they can help to shape further research in the direction to better understand manipulation of digital evidence. If we were to repeat it, we would add a post–experiment questionnaire to check the workload of students during the semester caused by other courses. We would probably also choose the sequence of Part 1 and Part 2 randomly per participant to control experience and exhaustion effects. We would also consider having participants analyze two images, since in our setting we still have 7 forgeries produced by participants that have not yet been analyzed. To pursue this line of research, more participants are needed to obtain statistically meaningful results. Ideally, such an experiment would be done with 50 experienced digital investigators in a controlled environment. Given the restrictions of resources and number of skilled personnel in practice, however, this will be hard to achieve.

Acknowledgments

We wish to thank all participants of the experiment for their patience and participation. We also thank Robert Frank for comments on a prior version of this paper and the anonymous reviewers for their helpful feedback.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.diin.2018.01.011.

References

Berghel, H., 2007. Hiding data, forensics, and anti-forensics. Commun. ACM 50 (4), 15–20. https://doi.org/10.1145/1232743.1232761.

Brenner, S.W., Carrier, B., Henninger, J., 2004. The trojan horse defense in cybercrime cases. Santa Clara High Technol. Law J. 21 (1). URL: http://digitalcommons.law.scu.edu/chtlj/vol21/iss1/1.

Caloyannides, M.A., 2003. Digital "evidence" and reasonable doubt. IEEE Secur. Priv. 1 (6), 89–91. https://doi.org/10.1109/MSECP. 2003.1266366.

Casey, E., 2011. In: Digital Evidence and Computer Crime – Forensic Science, Computers and the Internet, third ed. Academic Press.

Crime Investigation Kirk, P.L., 1974. In: von Thornton, John I. (Ed.), 2. Auflage. John Wiley & Sons.

Dardick, G.S., Endicott-Popovsky, B., Gladyshev, P., Kem-merich, T., Rudolph, C., 2014. Digital Evidence and Forensic Readiness (Dagstuhl Seminar 14092). Dagstuhl Reports 4 (2), 150–190. https://doi.org/10.4230/DagRep.4.2.150. http://drops.dagstuhl.de/opus/volltexte/2014/4549.

Foster, J.C., Liu, V., 2005. Catch Me, if You Can, Presentation at Blackhat Briefings. http://www.blackhat.com/presentations/bh-usa-05/bh-us-05-foster-liu-update.pdf.

Freiling, F., Hösch, L., 2018. DFRWS EU 2018 Data Set. https://faui1-files.cs.fau.de/public/publications/freiling-hosch-dfrws-eu-2018-data.xlsx.

Garfinkel, S.L., 2012. Digital forensics XML and the DFXML toolset. Digit. Invest. 8 (3–4), 161–174.

Groß, H., Geerds, F., 1977. Handbuch der Kriminalistik, Vol. 1. Ver-lagsgesellschaft Manfred Pawlak.

Harris, R., 2006. Arriving at an anti-forensics consensus: examining how to define and control the anti-forensics problem. Digit. Invest. 3 (Suppl. 1), 44–49.

Johnson, M.K., Farid, H., 2006. Exposing digital forgeries through chromatic aberration. In: Voloshynovskiy, S., Dittmann, J., Fridrich, J.J. (Eds.), MM&Sec, ACM, pp. 48–55. https://doi.org/10.1145/1161366.1161376.

Lee, H.C., Harris, H.A., 2000. Physical Evidence in Forensic Science. Lawyers and Judges Publishing.

Lin, Z.C., He, J.F., Tang, X., Tang, C.K., 2009. Fast, automatic and fine-grained tampered JPEG image detection via DCT coefficient analysis. Pattern Recogn. 42 (11), 2492–2501. https://doi.org/10.1016/j.patcog.2009.03.019.

McDonald, A.D., Kuhn, M.G., Steg, F.S., 1999. A steganographic file system for linux. In: Pfitzmann, A. (Ed.), Information Hiding, Vol. 1768 of Lecture Notes in Computer Science. Springer, pp. 462–477.

Moch, C., 2015. Automatisierte Erstellung von U bungsaufgaben in der digitalen Forensik (Ph.D. thesis). Friedrich-Alexander-Universität Erlangen-Nürnberg, Technische Fakultät.

Paul, G.L., 2009. Foundations of Digital Evidence. Amer Bar Assn.

Richard III, G.G., 2005. DFRWS Forensic Rodeo Challenge. https://www.cfreds.nist.gov/dfrws/Rhino_Hunt.html.

Savoldi, A., Piccinelli, M., Gubian, P., 2012. A statistical method for detecting on-disk wiped areas. Digit. Invest. 8 (3–4), 194–214.

Wundram, M., Freiling, F.C., Moch, C., 2013. Anti-forensics: the next step in digital forensics tool testing. In: Seventh International Conference on IT Security Incident Management and IT Forensics, IMF 2013, Nuremberg, Germany, March 12–14, 2013, pp. 83–97. http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=6568360. In: http://www.computer.org/csdl/proceedings/imf/2013/4955/00/index.html.