Forensic Science International: Digital Investigation 40 (2022) 301334

Contents lists available at ScienceDirect

Forensic Science International: Digital Investigation

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

DFRWS 2022 EU - Selected Papers of the Ninth Annual DFRWS Europe Conference

Prudent design principles for digital tampering experiments

Janine Schneider*, Linus Düsel, Benedikt Lorch, Julia Drafz, Felix Freiling

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

ARTICLE INFO

Article history:

ABSTRACT

We study the factors that lead to successful experiments in the field of digital evidence tampering, evaluating the studies conducted in the past and the mistakes that happened during the execution of our own experiments. We describe three lessons learned that arise from evaluating the experiments and provide advice on conducting future studies. We also report on qualitative results from our experiments and interviews with professional IT forensic experts.

1. Introduction

In recent years, there has been much discussion about the nature of digital evidence and how it differs from physical evidence (Dardick et al., 2014). Many believe that it is easier to tamper with digital rather than physical evidence (Caloyannides, 2003; Kumar et al., 2006) and that it is possible to create a perfect digital forgery. Accordingly, digital evidence is secured by implementing standard processes such as the creation of hash values. In addition to the creation of hash values, the integrity of digital evidence is also secured during the creation of the forensic images. The E01 format, for example, compresses the data and creates a checksum. Some also suggested to secure the integrity of digital evidence by using blockchain technology (Tian et al., 2019). However, all these mechanisms cannot protect against tampering with the evidence *before* the chain of custody has begun (i.e., a forensic image has been created and the first hash value has been computed). If the evidence was successfully manipulated by deleting or planting data, this can lead to wrong conclusions in further investigations.

Evidence tampering is a relevant issue since the legal system critically depends on evidence being authentic when weighing it in a criminal procedure. Casey (2011) states that

> [e]ven when there is a reasonable doubt regarding the reliability of digital evidence, this does not necessarily make it inadmissible, but will reduce the amount of weight it is given by the court. For instance, if there is concern that the evidence was tampered with prior to collection, this doubt may reduce the weight assigned to the evidence. In several cases, attorneys have argued that digital evidence was untrustworthy simply because there was a theoretical possibility that it could have been altered or fabricated.

Furthermore, there have also been cases of proven attempts to delete digital evidence. For example, in “U.S. v. Tucker” (United States Court of Appeals), the defendant attempted to manually delete pictures from the webcache, and in “State v. Mercer” (Wisconsin) the defendant attempted not only to delete the webcache but also to use software to erase other traces such as visiting certain websites. Accordingly, the influence of the possibility of tampering should not be underestimated.

There has been a substantial line of research work focusing on the analysis of manipulated multimedia evidence (Farid, 2016; Sencar and Memon, 2013). However, only little work has investigated the tampering in non-multimedia settings. For example, Casey (2002) examined the uncertainties that can be caused by data corruption and loss as well as tampering which leads to errors in the interpretation of digital evidence regarding network activities.

In 2018, Freiling and Hösch (2018) were the first to empirically investigate ways how manipulations within the browser evidence on disk images could be detected. In their study, 14 graduate level students with basic digital forensics education were asked to perform an evidence tampering task on a given hard disk image: the addition of a spurious browser history entry in Firefox on a standard Ubuntu Linux installation. Subsequently, the manipulated data was analyzed by other students to find out whether the tampering could be detected. Despite the task appearing rather simple, interestingly, all manipulations were correctly detected,

* Corresponding author.
E-mail addresses: janine.schneider@fau.de (J. Schneider), linus.duesel@fau.de (L. Düsel), benedikt.lorch@fau.de (B. Lorch), julia.drafz@fau.de (J. Drafz), felix.freiling@fau.de (F. Freiling).

https://doi.org/10.1016/j.fsidi.2022.301334

giving a first indication that tampering with digital evidence might not be as easy as expected. This result was corroborated by Schneider et al. (2020) who showed that post-mortem tampering of main memory images in the form of adding evidence of a spurious network connection was even harder in terms of tampering effort and similarly futile with respect to the probability of detection. Both studies were conducted under the assumption that understanding the creation of forgeries helps to detect forgeries. Accordingly, further experiments with different scenarios and pieces of evidence need to be conducted.

1.1. Research goal and contributions

We revisited the original study of Freiling and Hösch (2018) and attempted to investigate *why* the tampering task was so hard and the tampering results so easy to detect. We used a similar scenario, i.e., the tampering of browser evidence, and changed one parameter of the study. While Freiling and Hösch (2018) considered the *addition* of evidence to a hard disk image, our tampering task consisted of the *removal* of browser evidence. However, during the execution of our study we encountered some problems, which forced us to repeat the study. Again, we encountered problems during the retry, which prompted us to conduct a detailed evaluation of the various aspects of the failure of our experiments and to re-examine all the studies conducted in the past.

In our paper, we take a close look at all the studies conducted so far, including our own, and examine the various tasks, participants and the actual conduct of the studies. In addition, we report on the qualitative findings from our experiments and the insights we obtained through interviews with professional IT forensic experts. We also evaluate the individual aspects of the experiments, analyze the problems that have arisen and derive three lessons learned from that, namely:

1. Given that studies of digital evidence tampering require participants with a specialized skillset, only a limited number of participants can be expected. Instead of trying to increase the number of participants, the experiment should be designed in such a way that it can also be carried out with a small number of participants.
2. The description of a tampering task is particularly difficult because it includes the subjective interpretation of an analyst. This *relativity* implies special care regarding comprehensibility and executability.
3. Tampering is a complex task which is hard to control within a lab setting and therefore naturally involves a lot of noise. The experimental setup should therefore explicitly address ways to reduce noise, e.g., by designing for overcoverage or opting for qualitative evaluation rather than quantitative evaluation.

While some of these lessons may appear obvious for people experienced in empirical studies, reporting our insights still may be useful for digital forensics researchers who are usually not trained in such research methods. Furthermore, even after involving experts in empirical studies, we found that experiments in the domain of digital evidence tampering involve several domain-specific non-obvious twists, that may be helpful in the design of future experiments in this area.

1.2. Paper outline

We describe the studies and experiments conducted in the past in Section 2 followed by a comparison between the detection and tampering effort values the studies described before. We then report on the three lessons in the following sections: small numbers of participants (Section 3), relativity of task description (Section 4), and data interpretation (Section 5). Section 7 concludes the paper.

2. Past studies and experiments

We now report on the studies conducted so far by other authors and by ourselves and the experiments performed within the studies. The individual experiments are listed in Table 1. Each experiment is given a tag to reference the individual experiments throughout the paper.

2.1. Study 1: Addition of browser activities (FH1 and FH2)

Freiling and Hösch (2018) describe a series of experiments they conducted with graduate level students to study the effort to perform an evidence manipulation task. In their study the students were handed out the image of a fully functional Ubuntu Linux virtual machine including the installation of a standard browser. The browser history on the virtual machine did not include any downloading activities.

In the first part of the study (FH1), the students were asked to manipulate the image of the virtual machine in such a way that a forensic investigator would reach the conclusion that a specific website has been visited and pictures have been downloaded from that website during a specified time period. For this, the students had permission to manipulate the image in any way and with any tools. Afterwards, the manipulated images (forgeries) were added to an image pool together with a set of non-manipulated images (originals) which contained untampered browser activities. The images were then randomly distributed to the students with equal probability to receive an original or a forgery. Each student got one image. The students were then asked to analyze the received image for traces of access to the website and downloading activities.

In the second part of the experiment (FH2), the first two phases of FH1 were repeated, but this time the students were only allowed to modify the image using a shell script they had created themselves for automated manipulation.

For each part of the study, the students had to document the time spent on the different tasks and to write a report on their proceedings and approaches. They also had to fill out a pre-study questionnaire about some demographic data, their experience and motivation.

Overall, 14 participants answered the pre-questionnaire, 11 completed FH1 and 6 completed FH2, whereby completed means that they participated in the tampering and the analysis part. The project diaries and the students' reports were used as the basis for the results. Accordingly, in the analysis part of FH1, 6 out of 6 forgeries were correctly identified as forgeries and 4 out of 5 originals were correctly identified as originals. It could also be observed that a higher manipulation effort leads to a higher analysis effort. In the analysis part of FH2, all images were correctly classified and it could be determined that an automatically generated forgery is easier to detect. Finally, the authors came to the conclusion that it is hard to produce a convincing forgery in their case.

2.2. Study 2: Addition of main memory traces (SWF1 and SWF2)

Based on the work by Freiling and Hösch (2018), another series of experiments was conducted in 2020. Schneider et al. (2020) studied the subject of tampering of main memory dumps. Similar to the previous work they conducted their study with graduate level students. This time the students received a main memory dump of a Kali Linux virtual machine to be tampered with.

Table 1
Key data of the past experiments on digital evidence tampering. Each experiment conducted with students consisted of a tampering and a subsequent analysis task. The number of participants therefore refers to the number of students performing the tampering and the analysis task. The experiments conducted with professionals consisted of an analysis task which could be solved in teams. The number of participants for these experiments is therefore divided into the number of teams and their members (teams/total participants). The number of data points refer to the number of analyzed images.

Tag	Authors	Subject	Participants	Number of Participants	Analyzed Images
FH1	Freiling and Hösch (Freiling and Hösch, 2018)	Browser evidence addition with full control	Students	11	11
FH2	Freiling and Hösch (Freiling and Hösch, 2018)	Browser evidence addition with partial control	Students	6	6
SWF1	Schneider, Wolf, Freiling (Schneider et al., 2020)	Main memory evidence addition	Students	22	22
SWF2	Schneider, Wolf, Freiling (Schneider et al., 2020)	Main memory evidence addition	Professionals	15/66	183
SFD1	Ourselves	Browser evidence removal	Students	22	43
SFD2	Ourselves	Browser evidence removal	Students	21	42
SFD3	Ourselves	Browser evidence removal	Professionals	14/42	101

The students were asked to tamper with the memory dumps such that an analyst would come to the conclusion that there was an active network connection to a specific server at a specified time and that administrative commands were executed on the server over this network connection. As in the previous study an image pool was created consisting of the tampered images (forgeries) and the non-manipulated main memory dumps (originals). Each student received one image from the image pool with an equal chance to receive a forgery or an original. Afterwards, the students were asked to examine the randomly chosen main memory dump and to decide whether it was an original or a forgery.

They also had to answer a pre-study questionnaire, log their efforts and write a short report for the tampering and the analysis part of SWF1.

Overall, 31 students answered the questionnaire, 23 participated in the tampering part and 22 participated in the analysis part of SWF1. All originals were successfully recognized as originals. Among the forgeries, 8 out of 10 were correctly identified as forgeries. Surprisingly the detection effort seemed to decrease with an increase of the tampering effort. The authors were also able to group and classify the different tampering approaches and associate them with the detection of forgeries. This made it possible to identify actions which in turn created traces of the actions themselves.

To obtain a larger amount of data and to expand the study to include professional participants, the authors decided to integrate the DFRWS EU 2019 Forensic Rodeo into the study as an experiment (SWF2). The conferees were instructed to analyze as many images as possible from a series of 40 images. In this capture-the-flag competition, the conferees were asked to work in teams to examine as many dumps as possible from a series of 40 main memory dumps which were drawn from the extended image pool of SWF1. For each memory dump they had to decide whether the found ssh connections were real connections or if the dump was tampered with. Since the competition was managed through a web platform, all data was logged automatically.

15 teams with 66 members participated in the Forensic Rodeo. The teams were able to correctly identify 84 out of 92 originals and 75 out of 91 forgeries. Furthermore, the professional participants were able to classify the images much faster than the students. The authors also investigated the effect of repetition, but could not find a positive effect.

2.3. Study 3: Removal of browser activities (SFD1)

Following up on (Schneider et al., 2020), we studied the case of malicious evidence removal. Like in the previous work the study was conducted with graduate level students.

For the study a Ubuntu Linux virtual machine was created and cloned (to create untampered images). The virtual machine was then used to access a specific website and to download pictures from the website. Before and after the website access, the VM was used for usual browsing. Afterwards, the virtual machine image was handed out to the participating students which were asked to tamper with the image such that a forensic analyst would come to the conclusion that there is no evidence of an access to a specific website in a specified time period and that no pictures were downloaded from that website. The students were not allowed to just wipe or override the image or make the image unusable as the tampering should not be noticeable for the analysts. After the tampering, an image pool was created (similar to previous work) and each student received two images for analysis from that pool. The distribution of the images was done automatically with respect to a set of fixed conditions. The conditions ensured that, for example, no student received their own forgery or that each forgery was distributed only once.

All students had to log their effort in a project diary, they had to write a description of their tampering approach and a report for each analyzed image. They also had to complete a pre-study questionnaire.

Overall, 28 participants filled in the questionnaire, 23 participants took part in the tampering and the analysis task and 22 analyses were used for the evaluation. The study resulted in the correct identification of 12 out of 12 originals and 20 out of 31 forgeries. In general, the tampering effort was higher than the analysis effort and an increasing tampering effort did not seem to increase the analysis effort. The data also indicated that it requires more effort to overlook a forgery than to detect a forgery. We also observed a positive correlation between a high motivation and the tampering success and between good introductory course grades and the tampering success. We were also able to group and classify the tampering approaches and to relate them to the detection of forgeries.

2.4. Study 4: Removal of browser activities (SFD2 and SFD3)

After the evaluation of our study on the topic of evidence removal, we decided to repeat the study with modified parameters. The task description and the design of SFD1 were adopted with some changes to improve the comprehensibility of the task for SFD2.

Overall, 27 participants filled in the pre-study questionnaire, 23 students performed the tampering and the analysis part of SFD2 but only 21 participants submitted their project diary. Therefore, 2 participants were excluded from the evaluation. As a result 18 out of 21 originals were correctly classified as originals and 19 out of 21 forgeries were correctly classified as forgeries. In general, the effort to produce a forgery was higher than the effort to analyze the images. Furthermore, the detection effort seemed to decrease with an increasing tampering effort. However, this does not apply to the two misclassified forgeries.

Like in (Schneider et al., 2020), we decided to extend our study by adding the data of the DFRWS EU 2021 Forensic Rodeo.

Therefore, we used the task description and an extended image pool from SFD2. The general design was adopted from Schneider et al. (2020), but this time participants were asked to analyze a series of 10 images in succession, decide whether each image is a forgery or not and give a confidence level for the decision. Participants were also asked to complete a short questionnaire at the end of the event and the teams in the top two places were interviewed about their analysis approaches.

A total of 14 teams with 42 team members participated in the rodeo. Overall 30 out of 47 originals and 34 out of 54 forgeries were correctly classified. In contrast to the results of SWF2 the detection rate was significantly lower. The data also shows that about the same amount of time is needed to classify an original as an original as to classify a forgery as a forgery.

As already mentioned, we also interviewed the first and second placed teams. We asked the teams about how they prepared for the event, how they distinguished between original and forgery, what further investigations they would have done with more time, how they would act in a real life situation and whether they think it is possible to create a perfect forgery.

One team prepared by investigating the visited website and preparing all data needed from this website before the rodeo started. They also tried to replicate the image with the available information to generate a comparison baseline. Furthermore, they searched for all relevant locations where traces should probably be found and analyzed the structure of the relevant files. They distinguished between an original and a forgery by checking the prepared locations and looking for inconsistencies. If they had more time, they would have correlated the different inconsistencies and would have looked for more context knowledge. For real life cases they would focus on the crucial point of the case and check if the found traces are coherent. They would also again try to generate a baseline and recreate the traces. They stated that they do not believe that it is possible to create a perfect forgery, but they believe that it is easier to add data on a live system than to tamper with an image post mortem.

The other team prepared themselves by identifying the locations and the data that had to be analyzed, they also looked on their own machine how the data usually looks like and explored the suspicious website. They differentiated between originals and forgeries by checking specific locations and search for traces. With more time they would have performed a more detailed analysis of the images they classified as originals and they would have done a literature research on specific data structures. In a real case, they would focus on the correlation between traces and the correlation of information. In contrast to the other team, this team believes that it is possible to create a perfect forgery depending on how much effort somebody would put into the forgery. They also think that it is in general easier to add data to an image than to remove data and that it is easier to tamper with the image post mortem.

2.5. Data comparison

We now compare the detection and tampering effort values from the student experiment parts (FH1, FH2, SWF1, SFD1 and SFD2) of the studies described in Sec. 2. The different effort values are depicted in Fig. 1. All effort values are given in minutes. The figure shows that there are outliers in all experiments that may distort the overall picture. Nevertheless, a trend can be seen in Fig. 1a and Fig. 1c. Interestingly, the trend goes in two different directions, which may be related to the different tasks. Also in Fig. 1b a trend can be observed, unfortunately this trend is not representative because there are only two data points. In Fig. 1d and Fig. 1e, an attempt was made to increase the number of data points by giving students two images instead of one to analyze. No trend is apparent in either plot. This raises the question whether the trend in the previous experiments is only clearly visible because of the smaller number of data points.

Furthermore, Freiling and Hösch (2018) and Schneider et al. (2020) provided confusion matrices in their studies that contained the classification results and further analysis effort values. For a better overview we decided to transfer the data of the

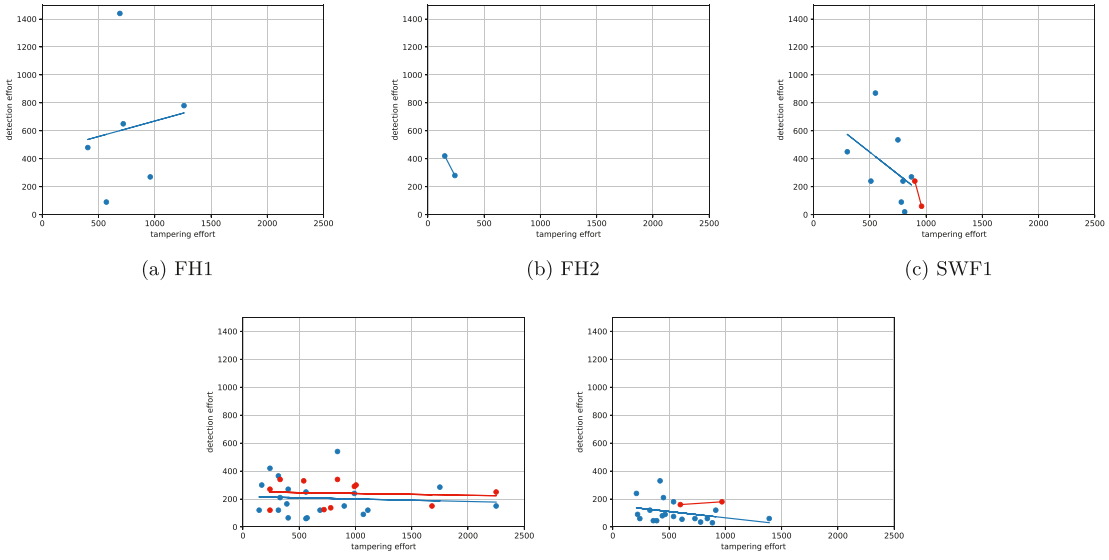


Fig. 1. Detection and tampering effort values for the student experiments of the four studies presented in Sec. 2. 1a shows the effort values of FH1 (browser evidence tampering with full control), 1b shows the values of FH2 (browser evidence tampering with partial control), 1c shows the effort values of SWF1 (main memory evidence tampering) and 1d and 1e show the effort values of SFD1 and SFD2 (browser evidence removal). Correctly classified images are depicted in blue, falsely classified images are depicted in red. The lines show a least squares polynomial fit. All values are given in minutes.

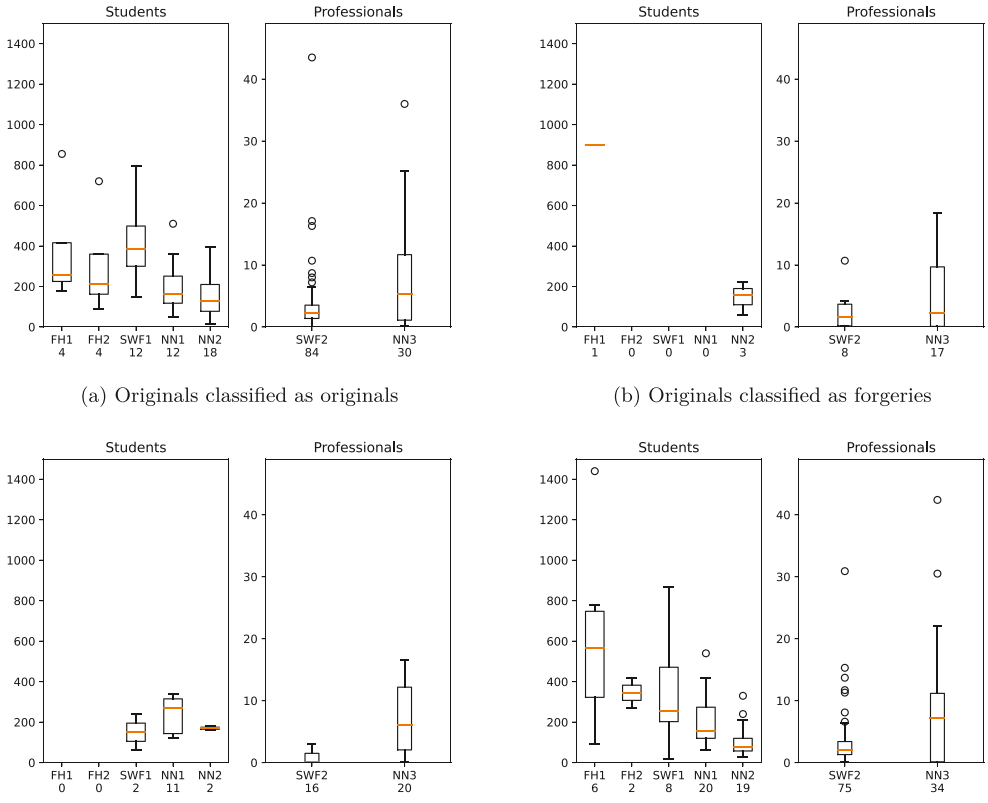
different experiments into boxplots and to split them according to the classification of the images. This allows the distribution and several different properties of the data to be displayed in one plot since it summarizes the minimum, first quartile, median, third quartile and the maximum. The boxplot also tells about the outliers and if the data is symmetrical, how the data is grouped and if and how the data is skewed.

In Fig. 2, the experiments with students and those with professionals were separated because the ranges of values are too far apart. The scales were adjusted for better comparison and the number of classified images is given below the experiment tag. The upper left subplot shows the effort values for all originals that were correctly classified as originals. The upper right subplot shows the values for all originals that were incorrectly classified as forgeries. The two lower subplots show the forgeries that were correctly classified as forgeries (left) and incorrectly classified as originals (right). This view corresponds to that of a confusion matrix.

As Fig. 2 shows, it seems to make no difference in regards to the analysis effort whether one examines an original or a forgery. Within the different experiments, the values for the detection of originals and forgeries are approximately the same, which seems odd since a forgery is mainly detected by identifying inconsistencies, which are not present in originals. Accordingly, it would be logical to need more time for the analysis of an original because it takes more time to look for inconsistencies in all relevant places than to conclude that it must be a forgery after the first inconsistency. However, according to the data, this does not seem to be the case. In addition, no major differences appear to exist between the different tasks of the experiments by Freiling and Hösch (2018) (FH1 and FH2) and our experiments (SFD1 and SFD2). Thus it makes no difference in analysis in regards to the effort if evidence is placed or removed. In Fig. 2b and Fig. 2c it can also be observed that there is a wide range of values and most of the experiments contained outliers that disturb the overall picture.

Furthermore, the effort values for SWF2 and SFD3 range between zero and several minutes for both experiments. This is probably because of the time pressure during the competition, which led to guessing. Even though the median of SFD3 and the values of SFD3 in general are higher, guessing seems also to be a problem. The figure also shows that the subplots of the experiments with professionals contain a lot of outliers and the data is very widely distributed.

Fig. 3 shows the distribution of the classfication of the images. The view is again based on that of a confusion matrix. The dark areas show a high number of classified images. In general the rate of misclassified images is very low for FH1, FH2 and SWF1. This rate increases in SFD1 and SFD2 and is highest in SFD1. The figure shows that the diagonal from the upper left to the lower right is strongly occupied, so there is a lot of data for correctly classified images. This could be an indication that tampering digital evidence is difficult. The rate of misclassified images seems to be generally higher for the experiments with professional participants (SWF2 and SFD3). This is not surprising considering the setup and the time pressure of these experiments.



(a) Originals classified as originals

(b) Originals classified as forgeries

(c) Forgeries classified as originals

(d) Forgeries classified as forgeries

Fig. 2. Analysis effort values for the different image classifications over all experiments. In each subplot the student experiments are shown on the left hand side while the experiments with professionals are shown on the right hand side: 2a shows the analysis effort values for the originals correctly classified as originals, 2b shows the effort for originals wrongly classified as forgeries, 2c shows the values for forgeries wrongly classified as originals, and 2d shows the analysis effort values for forgeries correctly classified as forgeries. In some cases there is no value since none of the images in this experiment was classified as specified. All values are given in minutes. The values below the experiment tag represent the number of images classified as specified.

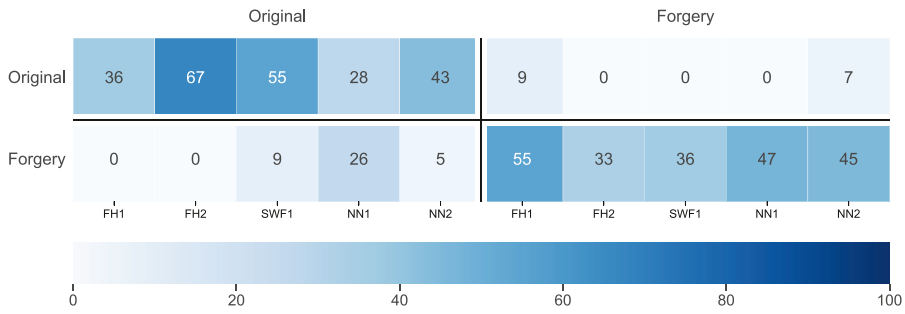


Fig. 3. The heatmap shows the relative amount of originals and forgeries correctly or incorrectly classified in each case, normalized to 100 percent. The normalization respectively refers to the sum of all images within an experiment. The darker a field is colored, the more images were assigned to the respective category.

So far, we have presented the experiments conducted in the past and compared the results of these. Next, we will discuss the various problems that arose during the experiments and see what we can learn from them for future experiments.

3. Lessons learned 1: Accept small numbers

One of the biggest problems in conducting experiments on digital evidence tampering is the fact that the participants must have specialized knowledge. Therefore, such an experiment is ideally conducted with a larger group of professional experts. Since it is very difficult to recruit professional experts as participants for such experiments, students were recruited in the past to replace the experts. Unfortunately, this leads to a number of problems.

3.1. Problems occurring in FH1 and FH2

In their study, Freiling and Hösch (2018) recruited students from a graduate level course on digital forensics as participants. In FH1, 11 students participated and each of them analyzed one image. In total, 5 of the analyzed images were originals. In FH2, 7 students participated but one result had to be excluded from the data set because the manipulation program failed to execute (Freiling and Hösch, 2018). Each student analyzed one image, whereby 4 images were originals. Thus, the number of participants was relatively low from the beginning and decreased further in the second part of the study. Furthermore, the students not only analyzed forgeries but also originals. This is necessary in order to establish a control group by which false criteria for a forgery can be identified. The distribution of the originals is also necessary to guarantee the blinding of the experiment. Furthermore, none of the forgeries created were successful in either FH1 or FH2.

As a consequence, no quantitative evaluation of the success rate of participants performing the tampering tasks could be performed. Although a quantitative evaluation of the unsuccessful forgeries was carried out, no statistically significant results could be obtained due to the very small number of participants. Nevertheless, the authors were able to perform a qualitative analysis of the tampering approaches and the detection of forgeries, but due to the lack of a successful forgery they were not able to analyze which factors influence the quality of the forgery and the detection or non-detection.

3.2. Problems occurring in SWF1

In SWF1, participants were also recruited from a graduate level course on digital forensics. This time 22 students participated of which 12 students received originals and 8 students received forgeries. Accordingly, no statistically significant results could be observed. In contrast to FH1 and FH2, however, 2 forgeries were classified as originals, which at least opened the possibility to evaluate the factors for a successful forgery.

In addition to the problems already discussed that arise from the substitution of professional participants with students, there are other problems that have not yet been addressed. First of all, it is quite difficult to recruit students with the necessary skillset and thus the number of participants will always be comparatively small. In SFD1 and SFD2, we therefore decided to hand out two images to each student. This generated more measurements and made it possible to hand out each forgery at least once and thus collect more data about the forgeries.

Furthermore, it cannot be assumed that the participating students are all at the same level in terms of prior knowledge, experience and motivation. This leads to outliers that can distort the overall picture. This problem could be addressed by having each image examined at least twice. This reduces the data noise and results in more reliable measurements.

Aware of the problems that arise with replacing professionals by students, Schneider et al. (2020) attempted to expand the pool of participants while acquiring a new group of attendees. For this purpose, the authors used the DFRWS EU 2019 Forensics Rodeo. Through the rodeo it was possible to recruit a larger number of experts as participants. However, this created new difficulties.

3.3. Problems occurring in SWF2

Through the rodeo, the authors were able to recruit 15 teams with a total of 66 team members for the experiment. During the rodeo, each team had to analyze as many images as possible from a sequence of 40 randomly drawn images and decide whether they were forgeries or originals. This way, the professional teams analyzed significantly more images than the students in SWF1. The participants had 105 min for the analysis.

To estimate the effort required to analyze each image during the rodeo, various timestamps, such as the time when an image is displayed or the time when a solution is submitted, were automatically captured. Thereby, comparable effort values should be generated in order to compare them with the values from SWF1. However, this was not possible due to various factors. First of all, the CTF web platform did not display the same image every time the page was refreshed. This allowed the participating teams to work on different images in parallel. In addition, the database did not record the first time that the image was shown to the team. This made it possible to “reset” the effort needed to decide on a specific image. Furthermore, it could be observed that many of the teams were guessing instead of really analyzing, especially at the end but also during the rodeo. This is

certainly due to the character of the rodeo and the time pressure. Due to the teamwork, the effort values could not be compared with the individual participant values from SWF1. Because of all these factors, comparing the effort values of SWF1 and SWF2 or obtaining other meaningful results from SWF2 was not feasible.

3.4. Problems occurring in SFD3

Despite the problems encountered in SWF2, we tried to use the DFRWS EU 2021 Forensic Rodeo in SFD3 to increase the number of participants and add experts to the participant pool. For this we introduced some changes that should improve the experiment. The task description and the image pool were adopted from SFD2. The general design was adopted from SWF2, but this time participants were asked to analyze a series of 10 images in succession, decide whether it was a forgery or not and give a confidence level for their decision. Improvements to the web platform prevented parallel work and improved the automatic recording of data. By penalizing wrong answers with a point deduction, guessing as a (successful) strategy was discouraged. Participants were also asked to complete a short questionnaire at the end of the rodeo and the teams in the top two places were interviewed about their analysis approaches.

Unfortunately, these improvements could only eliminate some of the weaknesses of SWF2. The introduced confidence level served not only as a measure of decision certainty but also to determine a point multiplier. Thus, confident decisions were rewarded with a higher multiplier and unconfident decisions were disadvantaged by a lower multiplier. The confidence level itself led some teams to make their decisions more deliberately, but due to the competitive nature of the event, in the majority of cases the teams simply chose to give a high level, as this meant a higher point multiplier.

Due to the improvements to the CTF web platform, even after refreshing the web page, the same image was always displayed until the displayed task was solved. However, this only apparently prevented parallel work. With the change, it was no longer possible to work on multiple images at the same time, but the single image could of course still be examined simultaneously by multiple people with different tasks. Thus the effort value remains a team value which is not comparable with the individual value from SFD2.

In contrast to SWF2, an enormous effort was made in SFD3 to blind the experiment as much as possible. For this purpose, the CTF web platform was adapted so that each team received a different random sequence of images. Furthermore, the teams had no indication of whether their decision was correct or not. Also during the rodeo, the participants were only shown progress in terms of the images they had already processed. The scoreboard was only displayed halfway through the rodeo to prevent any conclusions being drawn about the answer based on the points scored. This prevented teams from working together to achieve a high score. In addition, deducting points for incorrect answers was intended to prevent teams from guessing the answers. Unfortunately, none of these measures could prevent some teams from guessing after all.

However, the results obtained in this way are more robust than the results from SWF2, but unfortunately they are still not suitable for comparison with the results from SFD2. Also a separate examination of the data did not lead to any conclusions, since the data contained too much noise to be able to make statements due to the structure of the event and the time pressure.

Nevertheless, through the questionnaire at the end of the rodeo and the interviews with the first and second placed team, interesting insights could be obtained. Through the statements of the teams, the procedure for the analysis of the images could be evaluated and the most important criteria for the decision could be identified. Furthermore, the view of the participants helped to better understand the data and to explain the results.

3.5. Lesson 1

Due to the specialized knowledge needed for such experiments, the number of participants will always be low. CTF contests are not necessarily suitable to increase the number of participants.

4. Lessons learned 2: Take relativity of task definition into account

Another important aspect is the description of the tampering task. The usual advice, namely that the experimental task must be formulated clearly such that each participant knows exactly what needs to be done and why it needs to be done, is particularly difficult for tampering tasks because they involve the interpretation of another human. The task definition is therefore naturally relative to subjective interpretation. What happens when this is not observed is what we experienced in SFD1.

4.1. Problems occurring in SFD1

The change from “addition” to “removal” made the formulation of the task surprisingly difficult. This was due to the fact that we had no experience with this type of task and were not prepared for the different interpretations of the task. While the task description of the initial study by Freiling and Hösch (2018) seemed rather straightforward (“add evidence of website accesses”), the converse task description (“remove evidence of website accesses”) was surprisingly ambiguous when trying to narrow it down. Is it a successful forgery if, for example, the entire browser history is deleted, or if the entire disk is wiped? In SFD1 we decided to answer these questions negatively. Removal of the entire browser history of a browser that is known to have been used in the past necessarily creates suspicion. A wiped hard disk does this even more. Our goal was to investigate manipulations that change only a minimal amount of data on the disk and therefore should be much harder to detect. In the end, the ambiguity of the tampering task led to the background story having to be changed in the middle of the experiment in order to address the students’ interpretation of the task. This in turn led to different levels of knowledge among the students and an unclear definition of manipulation traces which made the evaluation of the reports very difficult and created a lot of additional work in the data analysis phase since all forensic reports had to be analyzed in detail for mentionings of particular traces. Future work should take this problem into account from the beginning. Therefore, it is not recommended to reuse an already existing tasks from an other experiment. Instead, for each new task, it is important to consider exactly what the goal of the task is and how the task should be formulated in order to leave no room for interpretation.

Because of this we decided to repeat SFD1 and therefore conducted SFD2. The task description and the design of the previous study were adopted. However, in order to solve the problems of SFD1, some major changes were made.

First of all, the task description was adapted and the students were told in detail what their task entailed and what not. This included for example that the students were told that the virtual machine they received must not be started to perform the manipulation task. Students were also told in detail what exactly traces are and what artifacts are not considered as traces of the crime in terms of the experiment. For the analysis task, the students were reminded that they should examine their received images for traces of tampering and clearly decide whether an image is a forgery or not. By the introduction of a confidence level for the decision we tried to ensure that the students consider their decision and weigh all arguments in favor or against forgery. The analysis report was also replaced by a questionnaire to force the

students to answer certain questions and to collect the information in a more structured way.

Another aspect is the definition of inconsistencies. So far, all participants were told to look for inconsistencies when analyzing the images without defining what inconsistencies are. Defining this term is not trivial, but necessary to avoid misunderstandings. Is a changed timestamp an inconsistency or does the timestamp change for other reasons? Is the absence of a particular file an inconsistency or part of a garbage collection mechanism? The question of what an inconsistency actually is should therefore be clarified for each experiment and should be addressed in future research.

Accordingly, SFD1 could be considered a preliminary study to SFD2. Since experiments like this are particularly susceptible to such problems, a preliminary study should be carried out in any case, since such problems become apparent in the preliminary study. However, we noticed this methodological error only after performing and repeating SFD1. Therefore, for future experiments, we strongly recommend following the good practices for carrying out studies (known from social sciences) and conducting at least a small preliminary study.

4.2. Lesson 2

The task description is one of the most important aspects and should be carefully considered and, in any case, verified with a preliminary study. The task description needs to take into account its relativity to the subjective interpretation of the analyst.

5. Lessons learned 3: Reduce noise in data collection

The previous lessons referred to the problems that were encountered when generating measurements. Now we discuss how the generated data can lead to problems in the evaluation.

As already illustrated in Sec. 2, the interpretation of data can be very difficult depending on the quality and amount of the data. For example, the data of FH1 and FH2 are qualitatively interesting but not quantitatively evaluable. The same holds for the data of SWF2. SFD1 was difficult to evaluate because of the task definition that led to confusion among the participants during the experiments. It is thus questionable if anything can be drawn from this except for the lesson learned. Furthermore, the character of SWF2 and SFD3 makes the results incomparable to the other experiments but also difficult to evaluate because of the high data noise created through the competitive situation. But what about SFD2? In SFD2 we had already learned about the task definition problem and were aware of the problems arising through the small number of participants. Thus we handed out two images to each student, made sure that the number of originals and forgeries was equally distributed and that each forgery was examined at least once. Unfortunately, we had not considered a problem that had already been mentioned namely the different levels of knowledge and motivation of the students.

5.1. Problems occurring in SFD2

Although the experiment was significantly improved, no quantitative analysis could be performed or conclusions drawn from the data. This is due to the fact that the knowledge, the motivation and the workload of the different students were not considered during the distribution of the images. For example, a student with little forensic expertise and motivation may create a forgery that is then randomly given to an experienced student who recognizes the forgery rather quickly. On the other hand, a very motivated student can create a forgery that is only roughly analyzed by a student with a lot of other courses and little time, and is therefore incorrectly classified. These factors came together in a very unfortunate way in SFD2. Carrying over the intention that a manipulation was successful if it was not noticed by the analyst, we stumbled across the question of what it means that an analyst “does not notice” the manipulation. While it appears easy to simply remove textual references and file content from an image by overwriting it with data, doing this carelessly will raise suspicion. Furthermore, it is not so clear how to treat a forgery where the analyst does no detect the first-order signs of tampering (e.g., directly in the browser history in the context of the entries with the illegal access) or second-order signs (like collateral manipulations of other browser history entries that not directly affect the traces of illegal access). To avoid this problem, a selection of forgeries could instead be distributed to the students allowing each of these forgeries to be examined twice by two different students. This ensures that very obvious forgeries do not distort the results and that outliers can be eliminated by the fact that each image is analyzed twice. In conclusion, it is more important to collect high quality data than to try to increase the amount of data.

In FH1, FH2 and SWF1 the quality of the data of the student experiments was relatively high but the amount was unfortunately quite small. In SFD1 and SFD2 the amount of data was high, but the quality was rather low. For SWF1 and SFD3 the amount of data was very high, but the quality of the data was poor. Accordingly, it is not surprising that it was still possible to evaluate the data adequately and even identify trends in the data of FH1, FH2 and SWF1, but not in SWF2 and SFD3 or SFD1 and SFD2. This leads to the conclusion that in future experiments the focus should be on the quality of the data and not on the increase of the participant numbers. Therefore, qualitative research should be conducted instead of quantitative research. Our goal is not to create the perfect forgery, but to understand how forgeries are created, how they can be detected and which factors lead to them not being detected. For this, the evaluation of the tampering approaches, the inconsistencies created through the tampering and the analysis procedure is much more important than quantitative results.

It has also been found that a questionnaire is more suitable for collecting the results than a freely formulated report, as this makes it possible to force certain questions to be answered, all students have to provide the same information, and information can be collected in a structured manner. This also prevents that formulations in the reports have to be interpreted or the question whether it is a forgery is not answered clearly. It can also be useful to conduct a structured interview in which specific questions can be asked again and misunderstandings can be cleared up. An interview also allows the study participant to share information that seems important but was not asked for, thus providing new perspectives.

An interview can also help avoid another issue, which is the lack of objective criteria for detecting a forgery. Since the decision about whether something could be a forgery or not is very subjective, it makes no sense to consider and evaluate the decision detached from the participants and their arguments. Instead of evaluating only the decision and the time taken, in the future each participant could be interviewed one or several times to find out how the participants proceed (during the tampering and the analysis) and how the participants come to one or the other conclusion during the analysis. Maybe some participants change their mind during the analysis or change their approach after a certain insight. This could also clarify the question of what exactly it means when an analyst does not recognize a forgery. Is that because the forgery was so good or was the analyst not skilled enough to spot the forgery? Did the analyst miss to search in a specific location? Or did the analyst misinterpret a trace? Or did the analyst simply not have enough time or motivation?

5.2. Lesson 3

The main problem in analyzing the experiment data is data noise which can be avoided by selecting a set of forgeries (only convincing ones) and then having them analyzed at least twice by different participants. Another option would be to assign students to groups with similar skills and knowledge levels and share the images within the group.

Furthermore, a questionnaire should be distributed to participants to collect data, rather than having participants prepare a report. In this context, it is extremely important to ensure that the questions are formulated appropriately because vague or ambiguous questions can increase the measurement error (Lenzner, 2012). Furthermore, the cognitive effort required to comprehend the questions should be minimized (Lenzner, 2012). For further clarification, an additional interview can be conducted, for example in the form of a group discussion at the end of the study (Flick et al., p. 214–221). Together with the expected small numbers of participants, all this suggests to prefer qualitative studies over quantitative ones.

6. Discussion

Despite the rather negative findings regarding necessary improvements, the experiments conducted have nevertheless yielded a qualitative (and probably also subjective) insight into what makes manipulations so difficult. It refers to the difference between data and metadata. If metadata is involved in the tampering task, it appears to be much more likely that tampering is detected through a measurable inconsistency in data structures than if only (content) data is involved. In fact, it might even be impossible to create a perfect forgery once metadata is involved because of the difficulty in getting *all* metadata right, because often enough, tampering resembles a cat-and-mouse game where a seemingly never-ending series of timestamps, reference pointers or similar needs to be modified, each one causing another inconsistency that would need to be fixed. This not only reminds us of distinctions made in multimedia forensics (Ho and Li, 2015), but it also corresponds to the difference between syntactical (internal) and semantical (external) consistency notions that permeate the common notions of *integrity* from computer security and cryptography (see for example Biskup (Biskup, 2009, Chapter 2.2) and Gollmann (Gollmann, 2011, Chapter 3)). Understanding the nature of tampering of digital evidence after all might be much closer to established ways of looking at data integrity than expected.

7. Conclusion

To summarize, we conducted three experiments on the removal of digital evidence based on the scenario by Freiling and Hösch (2018). Because of problems encountered during the first experiment we decided to repeat the experiment and extend it. The second attempt also failed, which led us to the evaluation of our own and past experiments in order to perform better experiments in the future.

Therefore, we revisited the studies conducted by Freiling and Hösch (2018) and Schneider et al. (2020), compared the results and elaborated the problems occurring in the different experiments.

From that we derived three lessons learned. The first lesson we learned is that experiments in digital evidence tampering are too dedicated to be conducted with participants without expertise and that it is difficult to recruit suitable participants. So far, such experiments have been conducted with students with prior knowledge, but the number of participants has been limited. All attempts to increase the number of participants through CTF competitions failed. Handing out several tasks to each participant is therefore better suited to improve the data set. Furthermore, we observed that the description of the task is non-trivial and very important as this is the basis for the success of the experiment. A preliminary study can ensure that the task is clear and understandable. The last lesson we learned concerns the analyzability and interpretation of the data. Three factors play a role here: the quantity of data, the quality, and the method of collection. As mentioned earlier, the quantity of data can be increased by handing out multiple tasks to each participant. To increase the quality, overlapping should be ensured. The collection of data should be realized by a questionnaire because questionnaires are easier to evaluate than free text.

CRediT authorship contribution statement

Janine Schneider: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization. **Linus Düsel:** Conceptualization, Methodology, Software, Investigation, Data Curation, Writing - Review & Editing. **Benedikt Lorch:** Formal Analysis, Writing - Review & Editing. **Julia Drafz:** Methodology, Writing - Review & Editing. **Felix Freiling:** Conceptualization, Methodology, Investigation, Resources, Writing - Review & Editing, Supervision, Project Administration, Funding Acquisition.

Acknowledgments

We wish to thank all participants of the experiments, especially the students from the course on “Advanced Forensic Computing” at FAU and the participants of the DFRWS EU 2021 Forensic Rodeo. We also thank the anonymous reviewers for their helpful comments on previous versions of the paper. Work was supported by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of the Research and Training Group 2475 “Cybercrime and Forensic Computing” (grant number 393 541 319/GRK2475/1-2019).

References

Biskup, J., 2009. Security in Computing Systems - Challenges, Approaches and Solutions. Springer.
Caloyannides, M.A., 2003. Digital “evidence” and reasonable doubt. IEEE Security & Privacy 1 (6), 89–91.
Casey, E., 2002. Error, uncertainty and loss in digital evidence. Int. J. Digit. EVid. 1 (2).
Casey, E., 2011. Computers and the Internet. In: Digital Evidence and Computer Crime - Forensic Science, third ed. Academic Press.
Dardick, G.S., Endicott-Popovsky, B., Gladyshev, P., Kemmerich, T., Rudolph, C., 2014. Digital evidence and forensic readiness (dagstuhl seminar 14092). Dagstuhl Reports 4 (2), 150–190.
Farid, H., 2016. Photo Forensics. MIT Press.
Flick, U., von Kardorff, E. and Steinke, I. [n.d.], A Companion to Qualitative Research, SAGE Publ.
Freiling, F.C., Hösch, L., 2018. Controlled experiments in digital evidence tampering. Digit. Invest. 24, S83–S92.
Gollmann, D., 2011. Computer Security (3. Wiley.
Ho, A.T.S., Li, S. (Eds.), 2015. Handbook of Digital Forensics of Multimedia Data and Devices. Wiley.
Kumar, V., Srivastava, J., Lazarevic, A., 2006. *Managing Cyber Threats: Issues, Approaches, and Challenges*, Massive Computing. Springer US.
Lenzner, T., 2012. Effects of survey question comprehensibility on response quality. Field Methods 24 (4), 409–428. https://doi.org/10.1177/1525822X12448166.
of Appeals of Wisconsin, C.. State v. mercer. n.d. https://casetext.com/case/state-v-mercer-57.
Schneider, J., Wolf, J., Freiling, F., 2020. Tampering with digital evidence is hard: the case of main memory images. Forensic Sci. Int.: Digit. Invest. 32, 300924.
Sencar, H.T., Memon, N. (Eds.), 2013. Digital Image Forensics: There Is More to a Picture than Meets the Eye. Springer.
Tian, Z., Li, M., Qiu, M., Sun, Y., Su, S., 2019. Block-def: a secure digital evidence framework using blockchain. Inf. Sci. 491, 151–165.
United States Court of Appeals, T.C.. ‘U.S. v. Tucker’. https://casetext.com/case/us-v-tucker-41 n.d.