Browser fingerprinting with HTML5

A closer look at the entropy provided by the <canvas> element

Bror Filip Bellander
Ulf Noring

Department of Computer and Systems Sciences

Degree project 7.5 HE credits
Degree subject: Cyber Systems Security
Degree project at the Pre-Bachelor level
Autumn 2013
Advisor: Danny Brash
Reviewer: Name of Reviewer
Swedish title: Webbläsar-fingeravtryck med HTML5

Stockholm
University

Browser fingerprinting with HTML5

A closer look at the entropy provided by the <canvas> element

Bror Filip Bellander
Ulf Noring

Abstract

As standard methods of tracking web users are becoming more and more difficult, new methods of tracking are surfacing. One of these is through creating a 'fingerprint' of a web browser. A method has been proposed to gather some fingerprinting information through the HTML5 `<canvas>` element. Only one study has been done to show how much entropy this method provides. This report therefore aims to provide more data to better answer to the question "How many bits of entropy does the `<canvas>` method provide?". The research question was answered by surveying 527 web browsers of participants for the results of six tests that utilize the HTML5 `<canvas>` element. The results of the tests were grouped ogether and entropy was calculated based on these groups. The result is a more accurate estimate of the entropy provided by the method than previously presented. The study shows that the `<canvas>` method offers 6.9 bits of entropy across the 527 participating web browsers. This new data could be used to motivate web developers to implement the methods where they were previously unsure of their helpfulness.

Keywords

Browser fingerprinting, <canvas>, HTML5, entropy, WebGL

Contents

List of Tables

List of Figures

1. Introduction

1.1 Background

It was estimated in 2011 that 35% of the world had access to an Internet connection(International Telecommunications Union, 2011:1), increasing to to 39% in 2013, or over 2.7 billion people(International Telecommunications Union, 2013:2). This is a lot of potential customers for any business with an online presence. As Tene and Polonetsky (2012) note there are many uses for tracking the activity of these Internet users, from customer analytics to information security. For instance, they state:

> "Online advertising is greatly enhanced by the ability to analyze and measure the effectiveness of ad campaigns and by online behavioral tracking, which tracks users' online activities in order to deliver tailored ads to that user."
>
> (Tene and Polonetsky, 2012:283)

But users might not want to be tracked for a number of reasons(Miyazaki, 2008). They might be whistleblowers who might fear prosecution if caught, or they might take offense when companies map their surfing habits. Most of the participants in Ur et al. (2012) found online behavioral advertising "smart but creepy". On the other hand, web users might also find the tracking to their benefit as they will see ads that are more relevant to their habits than they otherwise might be. Whoever the users might be, information security becomes essential in this matter, both for the tracker and also the one being tracked. Trackers want to be sure that they have correct information about users, and users might want to be sure that trackers get as little information as possible about them.

The field of information security is defined as preserving three things: confidentiality, integrity, and availability(ISO, 2013:1). Web users might want to protect their *confidentiality* by not giving out too much identifiable information about themselves. At the same time, websites might want to use user tracking to ensure *availability* and *integrity* of their services.

The most common way of performing user tracking is by using *cookies*(Eckersley, 2010:3). These are simple textfiles that are stored on the users device that the webpage can access and store information in. Since cookies are stored client-side, a webpage cannot (or should not) rely on them as a complete fact. A user is able to alter the content of the cookie, or remove it and block them completely. As more and more internet users are becoming aware of the privacy threats these cookies might imply they either disable them altogether or reject them from certain host(Eckersley, 2010:3)(Tene and Polonetsky, 2012:333).

In Eckersley (2010) and in Flood and Karlsson (2012), an alternative way of identifying and tracking users' browsers through browser fingerprinting were investigated. Eckersley (2010:1-2) explains the broader term of device fingerprinting in these words:

> "It has long been known that many kinds of technological devices possess subtle but measurable variations which allow them to be "fingerprinted". Cameras, typewriters, and quartz crystal clocks are among the devices that can be entirely or substantially identified by a remote attacker possessing only outputs or communications from the device."

This means that a slight difference in how something is configured makes it able to be identified, or *fingerprinted* if you will. For example, say that you have two pair of shoes

that are of the same brand, colour, and size etc. How do you tell them apart? One pair might have a slightly higher heel, or it might have a small smudge on the side; these are small differences that makes the shoes “fingerprintable”.

“*Browser* fingerprinting” is when you try to identify a browser instead of the device as a whole. Flood and Karlsson (2012:1) defines it as follows:

> "A browser fingerprint is a set of properties of a device and its software which is gathered from a web browser."

So instead of looking to extract the information from the device as is, the information is gathered via a web browser. Different fingerprints may be given by different browsers on the same device. These differences might be small and might not matter in the goal of identifying a device however. Eckersley (2010:3) states that

> "If there is enough entropy [1] in the distribution of a given fingerprinting algorithm to make a recognisable subset of users unique, that fingerprint may essentially be usable as a ‘Global Identifier’ for those users. Such a global identifier can be thought of as akin to a cookie that cannot be deleted except by a browser configuration change that is large enough to break the fingerprint."

A change of browser might be able to break this ‘global identfier’. It is however worth noting that since the identifying information is not stored in a textfile that might easily be deleted, as it is with cookies, it is much harder to get rid of it. The information that is used for identification is information that is often needed for the user to be able to see the content he or she requested. Removing these identifiable properties might therefore render the requested content useless to the user.

A method[2] of increasing the entropy of a browser fingerprint was proposed in Mowery and Shacham (2012). The method involvs the HTML5 `<canvas>` element which is a new element where content may be rendered. Because the image that is drawn onto the `<canvas>` is determined by the users computer, there might be small differences in how the end-picture looks. This could add further to the usefulness of fingerprinting techniques such as the ones proposed in Eckersley (2010) or Flood and Karlsson (2012). However, Mowery and Shacham (2012) were only able to give what they themselves called ”*a rough estimate*” of the entropy provided by their method.

1.2 The Problem

The problem is that there is only an estimate of how much entropy the new method proposed in Mowery and Shacham (2012) provides. This is a problem because a website should be as efficient and responsive as possible. Many of the updates browser vendors make are for increasing performance in the browser and reduce load times(Mulazzani et al., 2013:2). With browser fingerprinting on the rise(Acar et al., 2013; Yen et al., 2012), developers will soon want to add these methods to their own websites. When adding different methods of browser fingerprinting to a website, logic dictates that total runtime of these tests increase with the number of tests added. Therefore, the fewer tests needed to gather sufficient entropy to identify a web browser, the better. However, to know if

[1] Flood and Karlsson (2012:14) define entropy as ”a measure of the level of disorder within the [a data] set, or more precisely the expected number of binary questions needed to classify a randomly picked instance from the data set.”

[2] Hereby known as the `<canvas>`-method

there are enough tests to gather a sufficient amount of entropy, it is necessary to know how much entropy each test provides. To get an exact number on the entropy provided by different tests over all web users today would be very impractical, but estimates can definitely be calculated on smaller sample sizes.

The sample size in Mowery and Shacham (2012), however, was 300 individuals, which can not be considered representative for the 2.7 billion people using the Internet[3] today. Mowery and Shacham state in their report:

> "We do not yet have the data necessary to estimate the entropy of our fingerprint over the entire population of the web, but given our preliminary findings, 10 bits is a (possibly very) conservative estimate."
>
> (Mowery and Shacham, 2012:10)

As well as:

> "However, since we do not believe that the hardware and software in our 300 samples is a representative sample of the internet as a whole, these metrics should be treated as rough guidelines at best."
>
> (Mowery and Shacham, 2012:7)

This indicates that more data is necessary in order to obtain a more accurate estimate of the entropy of the proposed method.

1.3 Research Question

The question is *"How many bits of entropy do the browser fingerprinting tests proposed in Mowery and Shacham (2012) provide?"*

The study would result in a more accurate estimate of the entropy provided by the `<canvas>`-method of browser fingerprinting proposed in Mowery and Shacham (2012) than the estimate given in the same study.

The `<canvas>` method consists of several tests that render text and a WebGL[4] scene to `<canvas>` elements on a website, then examining the pixels produced for variations among browsers. The tests cover the following areas:

- Rendering the Arial font which is one of the web safe fonts recommended by W3C (W3Schools, n.d.)
- A font loaded from the website (in this case Sirin Stencil)
- A case when the browser has to load the default font
- Rendering an image containing many areas of high detail

When creating systems for fingerprinting care should be taken to make sure that the fingerprint is accurate enough, while still keeping the process from becoming too slow, thereby hindering usability. A fingerprint with 33 bits of entropy is enough[5] to uniquely identify a person using a device. It is therefore important to know how much entropy

[3] As repoted by International Telecommunications Union (2013)

[4] WebGL is a web standard for a low-level 3D graphics API based on OpenGL ES 2.0 developed by The Khronos Group. http://www.khronos.org/webgl/

[5] Using the same formula for calculating enthropy as was used in Mowery and Shacham (2012:7) and also defined in Cover and Thomas (2005:14) $E = -\sum_{i=1}^{n} p(x_i) log_2 p(x_i)$. where $p(x_i)$ is the size of the ith group divided by the number of samples. we find that if we input p(x) for the probability of identifying a single person on earth, where current population is about 7 billion according to The United States Census Bureau (2013), we arrive at 32.6 bits. If 33 bits of entropy is input into the formula, it would be enough to identify one out of 2^{33}, or 8,589,934,592 people.

each browser fingerprinting method provides, so as to know whether enough entropy is gathered for the system to accurately identify a person.

2. Execution

2.1 Research Strategy

Three strategies have been considered for this study: Surveys and Sampling, Case Studies, and Experiments(Denscombe, 2010). Below follows an explanation of what implementing each strategy would mean for the study.

2.1.1 Surveys and Sampling

Denscombe (2010:11) states that the purpose of doing a survey generally is "to obtain data for mapping". Applying this to the research question, the surveys and sampling strategy could be used to map the browser fingerprints of the devices owned by the estimated 2.7 billion Internet users in 2013(International Telecommunications Union, 2013). By gathering configuration information and browser fingerprints from the devices and then analyzing this data to see how unique each fingerprint is, it would be possible to measure the amount of entropy provided by each test in the `<canvas>` method. To be specific, the tests of the `<canvas>` method, defined and described in Mowery and Shacham (2012), would be implemented on a website allowing visitors to submit a fingerprint by pushing a button, after having been presented with an explanation of what information would be gathered from their browser. Aside from the test data, information on browser version and operating system is also stored for control purposes. A cookie would also be stored on the respondents device to make sure the same device does not take the survey multiple times. As a cookie can not be fully trusted this is not a guarantee that participants will not participate multiple times, just as a locked door is no guarantee that a burglar will not enter your home. It is however a simple way of making sure most users will not take the survey multiple times. The alternative would be to store respondents IP addresses which carries with it ethical issues discussed further later on.

In order to use the surveys & sampling strategy, a number of questions must be answered(Denscombe, 2010):

- Type of survey
- Sampling technique
- Sample size

These questions will be answered in the sections below.

Type of Survey

Denscombe (2010) lists several types of surveys possible: Postal surveys, Internet surveys, telephone surveys, group-administered surveys, face-to-face surveys, observational surveys, and surveys of documents. As the study would be surveying Internet users' web browsers, it would necessarily have to be online, and therefore an Internet survey has been chosen.

Sampling Technique

The survey does not concern itself with *individuals* but with their *devices*, and intends to analyze the browser fingerprints of devices of Internet users. As the study intends to gather data on the real-world situation, it does not matter who the Internet users are, since any Internet user is part of the sample population. Since the authors of the study do not

have the resources to send out the survey to every Internet user, however, convenience sampling will be applied by spreading the link in various places on the Internet.

Sample Size

The primary strength of surveys is that they are good for getting information about a large amount of people.(Denscombe, 2010:12) As the survey in Mowery and Shacham (2012) had 300 respondents, this study would need more than 300 respondents to be able to get a better estimate of entropy of the `<canvas` method. Denscombe (2010:42) postulates that "when the population moves beyond 5,000,changes in the size of the population cease to have much impact on the sample size." Denscombe (2010) suggests using online sample size calculators to find out how many respondents a sample would need. Using one of these tools[1], we have arrived at a needed number of 1844 respondents if we would like a representative sample for the estimated Internet user population of 2.7 billion people, with a 3% error margin and 99% confidence rate. If we find that it is difficult to gather samples, we would only need 385 sample respondents to keep 95% confidence and a 5% error margin.

Conclusion

The surveys & sampling approach works the best when the information it is trying to gather is clear and well-defined.(Denscombe, 2010:12) As the suggested survey would only gather facts that are very well-defined in Mowery and Shacham (2012), such as image pixmap hashes and web browser user agent strings, this suggests the surveys & sampling approach would be well suited to answer the question of how much entropy the method provides.

If a respondent enters the survey multiple times with different devices and/or browsers, we would only have to find a way to make sure that the same browser data does not get counted multiple times. This could be accomplished by saving a cookie on the device and not allowing devices with that cookie to take the test again, or by saving the IP addresses of participating web browsers. A limitation with the cookie approach is that if the respondent deletes the cookie, they could partake in the survey again, thereby lowering reported entropy by having multiple identical signatures in the survey. The IP address method would work better in that sense, but if somebody brought their device to another network they could very easily enter the survey again. There would also be ethical aspects pertaining to saving IP addresses, see section 2.1.1 for a discussion on these. As long as it is ensured respondents do not enter multiple times, it does not matter who the respondents are and therefore all sampling techniques would be of equal value.

A problem with using the survey and sample research strategy would be that Internet surveys often have a low response rate(Denscombe, 2010). However, as there is a very large population to draw samples from - the devices of 2.7 billion Internet users - and the sampling can be convenience-based, the problem of a low response rate would be negated. The sample size minimum would be 385 respondents, allowing for a 95% confidence rate and 5% error margin. For a 3% margin of error 1068 respondents would be needed.

Ethical considerations

There has been some discussion over the ethics of tracking Internet users(Ball, 2013; Palmer, 2005). The survey would be gathering information that could potentially be used in identifying a unique device via its web browser. However, as shown in Tene and

[1]Found via a search at https://www.google.se for 'sample size calculator' and available at

Polonetsky (2012), identifying users can be used for many different purposes. It is therefore not the method in itself that is unethical but certain specific applications of it.

Storing an IP address together with the fingerprint information, to make sure that visitors cannot participate twice in the fingerprinting tests, would prove problematic as IP addresses are considered personal information in Sweden(Datainspektionen, n.d.) and there are rules and regulations that would have to be folllowed regarding storing such information(SFS 1998:204, 1998). If the survey strategy were to be followed then an anonymous cookie-based approach would be better.

Even if it was possible, however, to identify a web browser via these fingerprints, it would not be very useful information. Since no otherwise personally identifiable information of any kind would be stored by the test website - only the browser fingerprintsand an anonymous cookie, which, again, cannot be used to identify an individual, only a web browser - all the site could do is tell if a visiting web browser has taken the test previously or not. The privacy breach is therefore small enough to be negligible.

2.1.2 Case Studies

According to Denscombe (2010:52) the defining characteristic of the case study approach is the focus on one or perhaps two instances of the thing to be investigated. In this study, this could be done by preparing two different configurations of computer systems and then pointing web browsers on these systems to a series of tests implementing the method proposed in Mowery and Shacham (2012). It would then be possible to calculate the exact entropy provided by the fingerprint in these two cases. Great care would have to be taken to ensure that these systems are representative of a large population of web users; otherwise, the findings would not be useful as it would not be possible to generalize from them.

An advantage to this strategy would be the possibility of investigating the exact differences between two systems in rendering the HTML5 `<canvas>` element, and finding out *why* the entropy is of a certain amount.

An issue is that it is more suited to providing an explanation of *why* entropy is a certain amount, when the research question is *how much* entropy the method proposed by Mowery and Shacham (2012) provides(Denscombe, 2010).

Another issue is a lack of resources to perform the case study. The authors of this paper do not have the technical expertise needed to take such an in-depth look at the problem. It could also be difficult to get a hold of computer systems for the case study.

Thirdly, the matter of selecting two representable cases that would be easily generalized over would be difficult. Looking at the findings in Mowery and Shacham (2012), over 50 different rendering configurations were found. Generalization can be a problem in case studies(Denscombe, 2010:60) and it would be difficult to generalize the findings of such a study as proposed above due to the variation in configurations. This therefore discourages the use of the case study strategy to answer the research question.

Ethical considerations

As there would be no respondents involved in the above described way of using a case study to answer the research question, there would not be any ethical considerations to be made.

---

http://www.raosoft.com/samplesize.html

2.1.3 Experiments

Denscombe (2010:65) defines an experiment as ”an empirical investigation under controlled conditions designed to examine the properties of, and relationship between, specific factors.” Using the experiments strategy to answer the question of how much entropy the `<canvas>` method provides to a browser fingerprint would have us setting up a website which implements the `<canvas>` fingerprinting method and gathers the fingerprint of a visitor who has given his or her consent. The website would be a carefully controlled environment, only gathering the data necessary to answer the research question: hashes of the rendered `<canvas>` elements. This approach would be similar to the survey and sampling research strategy, in that subjects of the experiment would be directed to a website where different tests of the `<canvas>` method of fingerprinting would be performed.

An issue with this research strategy is that experiments are ”generally concerned with determining the *cause* of any changes that occur to the thing being studied.”(Denscombe, 2010:66) The research question in this case does not, however, involve any change to be studied and there would not be any cause or effect to investigate to answer the question. Therefore, the experiments strategy might be better suited for a different research problem, such as developing a browser fingerprinting system.

Ethical considerations

Since the experiments strategy would be used in a way very similar to the way a survey strategy would be used, the same ethical considerations as in 2.1.1 would apply. That is to say, a cookie would have to be used to make sure that visitors do not enter multiple times, but also not personally identifiable.

2.1.4 Strategy choice

Out of the three strategies considered, the case study strategy is probably the weakest candidate for answering the research question due to the issues mentioned in 2.1.2. Of the other two, both the survey and sampling strategy and experiments strategy would lend themselves to similar implementations. Due to the experiments strategy being more suited for finding cause and effect, and the surveys & sampling strategy for surveying a large population, however, we have chosen to use the survey & samples strategy in this study as it is the one best suited for answering the research question.

2.2 Method

In Denscombe (2010), four different research methods are presented: Questionnaires, interviews, observation and documents. To answer the research question on how much entropy the `<canvas>` methods provide, using the research strategy of surveys & sampling, data needs to be gathered on how uniquely different web browser configurations render the different elements of the `<canvas>` method. Explained below is how each of the four methods could be used to gather the data needed.

2.2.1 Questionnaires

Denscombe (2010:156) puts up three criteria that questionnaires should fulfill to qualify as a questionnaire(Denscombe, 2010:155-156). A questionnaire should:

- Be designed to collect information which can be used subsequently as data for analysis.
- Consist of a written list of questions.

- Gather information by asking people directly about the points concerned with the research.

A questionnaire could be used to ask people to visit a web site that would show them the pixmap hashes of the rendered `<canvas>` elements and ask them to submit them. This data gathering could be automated by a website, however, making it redundant to introduce questionnaires as an extra (unnecesary) step to the data gathering process.

2.2.2 Interviews

Interviews would not work to gather the kind of data needed for this survey. To answer the research question, data provided by users' web browsers is needed, not data provided by questioning the users themselves.

2.2.3 Observation

Denscombe (2010) defines two possible methods of observation: systematic observation and participant observation. The difference is that with the former, the researcher tries to merely observe the subject and gather data from the behavior recorded, while with participant observation the researcher, ackording to Becker and Geer (1957), "...participates in the daily life of the people under study, either openly in the role of researcher or covertly in some disguised role, observing things that happen, listening to what is said, and questioning people, over some length of time.' As our research question requires data from web browsers, there is no obvious reason to participate in the daily life of web users. Therefore, *systematic observation* would be used.

With systematic observation, it would be possible to observe how a web browser renders the different elements of the `<canvas>` method and to store these results in a database. The results could then be compared to each other to calculate the entropy of each test.

2.2.4 Documents

Since the only study performed so far to explore the `<canvas>` methods is the one in Mowery and Shacham (2012), and the research question is aimed to provide an answer that requires more data than provided in that study, the research question would not be able to be answered through document studies.

2.2.5 Method choice

Observation has been chosen as the method to use for the survey that is to be performed in this study. Document studies and interviews are not able to answer the research question. As observation would use the same process as questionnaires, only without the added extra step introduced by the latter method, observation will be chosen as it is more efficient.

2.3 Analysis Method

Denscombe (2010:204) states that the systematic observation method "produces quantitative data which are pre-coded and ready for analysis." The subject in this study is entropy of the `<canvas>` methods, that is to say, how many bits of information the methods provide. The data that would be gathered by the survey would be `<canvas>` element

pixmap hashes, and user operating system and browser version. We have therefore chosen a quantitative data analysis approach as the data is nominal.

Analysis would be performed on grouped frequency distributions of the different <canvas> element pixmap hashes for each test in the <canvas> method, with each unique pixmap hash making up a group. For example, browser A produces pixmap hash X for the test **text_arial**, browser B produces pixmap hash Y for the same test, and browser C produces pixmap hash X for the test. These hashes would then be grouped, with X and Y being the groups. Browsers A and C would belong to the X group and browser B would belong to the Y group. An entropy calculation would then be performed on these grouped frequency distributions to find out how much entropy is supplied by each test in the <canvas> method, that is, how uniquely identifiable are the individual pixmap hashes of each browser? The formula used to calculate entropy was provided in Mowery and Shacham (2012:7) and is as follows:
$E = -\sum_{i=1}^{n} p(x_i) log_2 p(x_i)$. where $p(x_i)$ is the size of the ith group divided by the number of samples, to calculate the amount of entropy provided.

2.4 Methods of investigation

The study was conducted by presenting a website to web users where they were able to read about the purpose of the tests, as well as what kind of data that was collected. These ”letters to the respondents” can be read in Appendix A and B, with A being what the users saw when they first entered the web site, and B being the ”About” page where more information could be found. After the initial text that briefly explained the tests, the users were presented with an orange button labeled ”Start tests”. This button would then start the tests defined in Mowery and Shacham (2012):section 3.1.1-3.1.4: **text_arial**, **text_arial_px**, **text_webfont**, **text_webfont_px**, **text_nonsense**, and **webgl**. These tests draw either text or an image onto the <canvas> element. The results of our own tests can be seen in figures 2.2-2.5. The tests would then gather the pixeldata from the users image and compare it to all the other samples that had been collected, giving the user a result like *“1 / 497”* which would mean that the user was the only one out of 497 that had an image that looked exactly like that. The users User Agent String was also gathered for a sanity check to see that real browsers were used, this data served no other purpose.

Mowery and Shacham (2012) describes their, and so following, our, tests as follows:

> "**3.1.1 Arial Text**
> In our first two tests, we render a short sentence in Arial, a font known for its ubiquity on the web. To exercise each letterform, we use the pangram “How quickly daft jumping zebras vex.”, along with some added punctuation.
>
> For text_arial, the text is rendered to the canvas in 18pt Arial. In text_arial_px, we change the font specification to 20px Arial.
>
> **3.1.2 WebFont Text**
> These two tests are extremely similar to the Arial tests, with the added complexity of loading a new font from a web server. In a more sophisticated or targeted fingerprint, the delivered font could be carefully tuned by the fingerprinter to exercise corner cases in font loading.
>
> In our case, however, we use the WebFont Loader to load “Sirin Stencil” from the Google Web Fonts server . Once it loads, we render the same pangram as in our Arial tests. For text_webfont, the text is set in 12pt Sirin Stencil, while text_webfont_px uses 15px Sirin Stencil."

For the text_webfont and text_webfont_px tests, we did not use the Google Web Fonts server to load the font but instead chose to host the font on the server the website was hosted from. This is because the original study seemed to encounter bugs with loading the web font from Google sometimes:

> "While running the experiment, six users experienced failures in the text_webfont and text_webfont_px tests, returning a blank PNG instead of one containing text. Upon investigation, we attribute these failures to a known race condition in the WebFont Loader library."
>
> (Mowery and Shacham, 2012:5)

Continuing through the test definitions, the next test is the Nonsense Text test:

> "**3.1.3 Nonsense Text**
> Code-wise, this test is nearly identical to the two Arial tests. However, instead of a valid font specification, we set the 2d font specification to "not even a font spec in the slightest". This exercises the fallback handling mechanisms in the browser: what does it do with an invalid font request? The browser's choice of fallback font, as well as its positioning and spacing, can be quite telling."

Examples of text_arial, text_arial_px, text_webfont, text_webfont_px, and text_nonsense, can be found in Figures 2.3-2.5. Code-wise these tests are all very similar. The only difference is the font declaration[2].

```
<script type="text/javascript">
    var canvas = document.getElementById("arial18pt");
    var context = canvas.getContext("2d");
    context.font = "18pt Arial";
    context.textBaseline = "top";
    context.fillText("Hello there!", 2, 2);
</script>
```

Figure 2.1: Example code of javascript drawing to a canvas element

The text-tests are all quite small, both code-wise and resource-wise. The one test that does not write text to the canvas is the WebGL test. This test instead renders a hyperbolic paraboloid with ambient and directional light with the ISO-standard 12233 test image for digital cameras. Mowery and Shacham (2012) describes it as follows:

> "**3.1.4 WebGL**
> webgl is our only test whose code spans more than a few lines. As WebGL scenes go, however, this scene is almost minimal. We create 200 polygons, approximating the hyperbolic paraboloid $z = y2-x$, with $-3 \leq y \leq 3$ and $3-3 \leq x \leq 3$. Over this surface, we apply a single texture: a 512 by 512 pixel rasterized version of ISO 12233, the ISO standard for measuring lens resolution. Designed for measuring sharpness and resolution in electronic still-picture cameras, this texture contains many areas with high detail. We then add an ambient light with color (0.1, 0.1, 0.1) and a directional light of color (0.8, 0.8, 0) and direction (2,4,9). Placing our surface at $z = -10$, we render this simple tableau."

An example render of the webgl test can be seen in figure 2.2.

[2]The code for these tests can be found at https://github.com/qwelyt/Browser-Fingerprinting

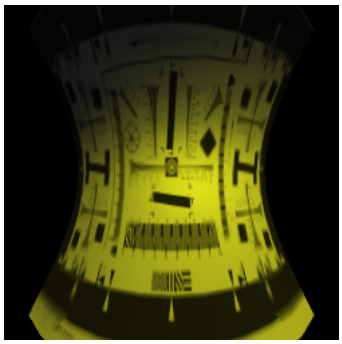
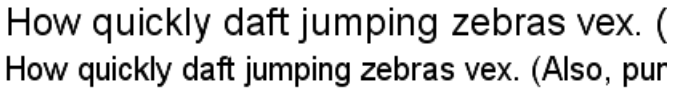


Figure 2.2: Example of the **webgl** test in our survey.



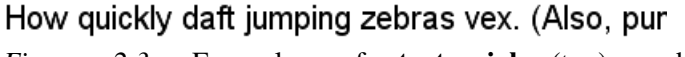
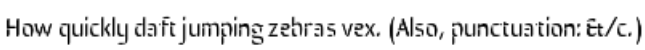


Figure 2.3: Examples of **text_arial** (top) and **text_arial_px** (bottom) from our survey.



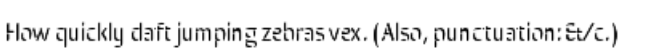


Figure 2.4: Examples of **text_webfont** (top) and **text_webfont_px** (bottom) from our survey.

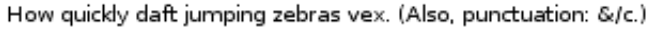


Figure 2.5: Example of **text_nonsense** from our survey.

After the tests had been run, the pixmap hashcode[3] of each resulting canvas element was saved in a database. Browser version and operating system version information was also saved to make sure that the tests were performed from an actual browser and not with software like wget[4].

2.4.1 Respondent selection

Since it does not matter who the respondents are, as explained in section 2.1.1, convenience sampling was used which consisted of posting the link to the website on various sites on the Internet and asking for respondents to spread the link to their friends, thereby creating a kind of ”snowball effect”. Some examples of how the link got spread are through IRC[5] chat rooms, Facebook[6] posts and various forums which the authors are familiar with.

2.4.2 Infrastructure

The tests were implemented using pure JavaScript. This is a deviation from how Mowery and Shacham (2012) did it as they wrote their tests with the help of Ruby. How the tests worked is never the less the same. After the tests had been run the results were sent with an AJAX-call to a PHP-script that saved the results in a MySQL database. The PHP-script thereafter returned the uniqeness-value of the tester which was then presented in a HTML-table.

2.4.3 Ethical considerations

Denscombe (2010:Appendix 1) lists four key principles which underlie codes of research ethics. These key principles are:

1. Participants’ interests should be protected
2. Participation should be voluntary and based on informed consent

[3] A pixmap is a map of pixels, i.e. an image. A pixmap hashcode is the textual representation of a pixmap.
[4] GNU Wget is a free software package for retrieving files using HTTP, HTTPS and FTP. http://www.gnu.org/software/wget/
[5] Internet Relay Chat
[6] A large social media site: https://www.facebook.com/

3. Researchers should operate in an open and honest manner with respect to the investigation
4. Researchers should comply with the laws of the land

Below follows an explanation on how these three principles were followed in this study.

Principle 1: Participants' interests should be protected

Denscombe (2010:331) lists three things that should be considered to uphold this principle.

The first is that participants should not come to any physical harm. Due to the survey being online, this was not possible in this case.

Participants should also not come to any psychological harm. This was also not possible in the case of this study. All that was sent out was a survey website explaining the different tests.

The third item is that participants should suffer no personal harm from the disclosure of information collected during the research. The only information gathered in this study is on what web browser and operating system the participant uses, as well as how it renders the different `<canvas>` elements. Since the cookie used to make sure participants did not enter several times was anonymous it was not possible to identify specific participants with the help of the cookie. Therefore, no personal harm should arise from the divulging of this information.

Principle 2: Participation should be voluntary and based on informed consent

According to Denscombe (2010:333), participants should be given enough information so that they can give their informed consent to participating. Their participation should also be voluntary. In the case of this study, informed consent was established by the way of two explanatory texts on the survey website - see Appendix A and B for these texts. It was explained what information the tests would gather, how they work and what they do. It was also made clear that all participation would be anonymous. Users were also given a button to click to initiate the tests themselves, after being given the opportunity to read about them first. Links were given to the original research in Mowery and Shacham (2012) for those who wanted to know more. Participants were also informed that the study was being performed at Stockholm University for a course requiring a small-scale thesis to be written. No contact information was given to the researchers as this information could easily have been abused on the Internet.

Principle 3: Researchers should operate in an open and honest manner with respect to the investigation

A full explanation of the aims of the research and the data that were to be collected was given to participants. See Appendix A and B for the explanations given to participants.

Principle 4: Researchers should comply with the laws of the land

To avoid any potential breaking of Swedish law, no personally identifiable information (such as IP addresses) was gathered. All data gathered was kept on a private web server used only for the purpose of the survey website, thus making sure that it could not fall into the wrong hands.

3. Results & Analysis

In this chapter the results of the survey will be presented, as well as an analysis of the collected data.

3.1 Results

In total our web survey received 527 samples.

These samples were grouped together according to the pixmap hash data of respondents so that all members of a group had the same pixmap hash for that specific test. The resulting groups are presented below for each test as a number of groups of different sizes.

3.1.1 text_arial

For **text_arial** a total of 102 groups were discovered.

Size	# of groups with size
1	59
2	15
3	4
4	7
6	4
7	3
8	1
9	1
12	1
13	2
18	2
23	1
36	1
203	1

3.1.2 text_arial_px

For **text_arial_px** a total of 97 groups were discovered.

Size	# of groups with size
1	53
2	14
3	5
4	8
6	4
7	3
8	1
9	1
12	1
13	2
18	1
19	1
23	1
36	1
203	1

3.1.3 text_webfont

For **text_webfont** a total of 82 groups were discovered.

Size	# of groups with size
1	40
2	14
3	6
4	2
5	4
6	4
7	1
8	1
9	2
12	1
15	1
16	1
18	1
19	1
23	1
53	1
202	1

3.1.4 text_webfont_px

For **text_webfont_px** a total of 79 groups were discovered.

Size	# of groups with size
1	37
2	13
3	6
4	3
5	4
6	4
7	1
8	1
9	2
12	1
15	1
16	1
18	1
19	1
23	1
51	1
203	1

3.1.5 nonsense_text

For **nonsense_text** a total of 95 groups were discovered.

Size	# of groups with size
1	54
2	15
3	5
4	5
5	1
6	2
7	3
8	1
9	1
10	1
12	1
14	1
18	1
19	1
27	1
53	1
200	1

3.1.6 webgl

For **webgl** a total of 108 groups were discovered.

Size	# of groups with size
1	56
2	13
3	12
4	8
5	1
6	2
7	2
8	1
9	2
10	1
14	3
16	1
18	1
24	1
32	1
38	1
126	1

3.1.7 All the tests together

It is not only possible to calculate entropy on each of the tests, as has been done above, but the tests can also be combined (together they are known in this study as ”the `<canvas>` method”). When combining all the tests the result is a total of 227 different groups.

Size	# of groups with size
1	161
2	27
3	9
4	8
5	5
6	2
7	2
8	3
9	1
10	2
12	1
13	1
16	1
18	1
23	1
31	1
36	1

3.2 Analysis

From the results presented in the previous sections,the entropy of the tests was calculated with the following formula:

$E = -\sum_{i=1}^{n} p(x_i) log_2 p(x_i)$, where $p(x_i)$ is the size of the ith group divided by the total number of samples (527 in this survey). In PHP code it would translate to the something like following

```
$entropy = 0;
for($i=0; $i<$numberOfGroups; $i++){
  $entropy += -(($samplesInGroup[i] / $totalNumberOfSamples) *
                log($samplesInGroup[i] / $totalNumberOfSamples), 2));
}
```

With the above code a result with 11 decimals was reached. For brevity and the sake of comparison with the numbers in Mowery and Shacham (2012), these results have been rounded off to two decimals. The results are presented below in Table 3.1.

Test name	Calculated entropy
Arial	4.61
Arial Px	4.43
Webfont	4.22
Webfont Px	4.19
Nonsense	4.31
WebGL	5.16
All tests	6.90

Table 3.1: *The calculated entropy for all the tests*

4. Conclusion & discussion

4.1 Conclusion

We have shown that when released 'into the wild' onto the Internet, the <canvas> method tests perform even better in identifying unique users than previously thought, providing more entropy for each test than measured in Mowery and Shacham (2012).

Test name	Our calculated entropy	Entropy from Mowery and Shacham (2012)
Arial	4.61	3.05
Arial Px	4.43	2.86
Webfont	4.22	2.93
Webfont Px	4.19	2.95
Nonsense	4.31	N/A
WebGL	5.16	4.3
All tests	6.90	5.73

Table 4.1: *Comparison of entropy for the different tests.*

4.2 Discussion

It is to be expected that our results gain a higher entropy. Mowery and Shacham (2012) state in their report that *"...the user population in our experiments exhibits little variation in browser and OS."* The little variation is could be due to the research being conducted on Amazon's Mechanical Turk website[1]. Ours is a more scattered distribution which translates into higher entropy for the tests as they look for differences in hardware and software. Thus, the results in Mowery and Shacham (2012) should probably be seen as a lower bound estimate for the possibly entropy that can be gathered by the <canvas> method.

As can be seen in 4.1, WebGL is the method generating the best fingerprint out of the six. This is probably due to it rendering a much more complicated canvas than the simpler font tests which makes for more points where differences might occure. One should not forget that this test also made a bigger <canvas> element in size, meaning that there were more pixels to draw on. This might have an implication on how much the picture can differ and might be a task for future studies.

Where the font tests usually had one large group that rendered fonts identically, that large group in WebGL instead consisted of browsers that did not support WebGL. This comes as no surprise to us as WebGL is still in development and support for it is not implemented for every graphics card in every browser. One interesting point with the WebGL results was how big the differences were between some of the results. Some respondents' browsers even drew the picture upside down, which begs the question of how WebGL was implemented in the browser.

[1] https://www.mturk.com/mturk/

Unfortunately we were not able to compare the value of the text_nonsense test as Mowery and Shacham (2012) fail to provide their findings for this test. However, since a higher degree of entropy was measured for all other tests, it can be assumed that this would have been the case for text_nonsense as well.

In Mowery and Shacham (2012:10) it was suggested that the entropy of the `<canvas>` method fingerprint over the entire population of the web would be around 10 bits. As we can see in table 4.1, the measured entropy over 300 respondents was a little more than half of that. In this study, it was measured to 6.9 bits. It is possible that we would have gotten even higher entropy values over a larger sample size. Our findings therefore point in the direction of the claim in Mowery and Shacham (2012) that the method could provide 10 bits of information when used in real-world situations.

With 6.9 bits of entropy, we could safely identify one in 119[2] people. With 10 bits of entropy this would be 1024. With 2.7 billion internetusers, this is not a very reliable number. With 6.9 bits of entropy you would get groups containing 23 tousand samples. Combining the `<canvas>` method with other ways, say looking for plugins and fontlists, an even higher entropy could be reached. The `<canvas>` method on it's own however is not enough to identify all users of the internet. If an entropy of 22 bits could be reached however, this would be possible.

4.2.1 Limitations of the study

One issue with the study is the repeatability of the tests. In this study, a cookie was used that, if present, prevented that browser from participating again in the survey. However, if a user manually deleted the cookie from their device, they were easily able to re-enter the survey. This would have lowered the entropy of the final results. As it is still higher than the results in Mowery and Shacham (2012) points to the fact that most participants probably only participated in the survey once.

Representativeness

We found that getting a sample base that is representative of the Internet is difficult. The sample of 527 respondents would be more representative if it was larger. There is also the issue of most of the survey participants being friends and family of the researchers, sharing their interest in technology. This could have affected the results - for instance, it could be that technology enthusiasts use a broader range of different devices than the average web user. This would have raised the amount of entropy in the results as there would have been more unique configurations participating in the survey.

Things that could be improved

For a better estimate of the amount of entropy provided by the `<canvas>` method, an expanded study would be necessary. An example would be the ongoing EFF Panopticlick[3] study on different browser fingerprinting techniques, which has over 3.5 million samples as of 2013(Electronic Frontier Foundation, 2013). Such a study would have better odds at getting a representative sample.

[2] With 6.9 bits of entropy we can use $2^{6.9}$ to see how unique a person, device or webbrowser is. With $2^{6.9}$ we arrive at around 119.43, meaning 1 in 119.43 can be identified.
[3] http://panopticlick.eff.org/

4.2.2 Uses

As more and more websites begin to use browser fingerprinting(Acar et al., 2013; Yen et al., 2012), and with more evidence of the amount of entropy contributed by the `<canvas>` method, web developers may turn to include the `<canvas>` method as they now have a stronger indication of its validity.

4.3 Future research

Since this study was very limited, a larger study should be conducted with respondents in the hundreds of thousands, such as The EFF Panopticlick study(Electronic Frontier Foundation, 2013). That way, a more representative sample could be gathered and it would be possible to get an even more accurate estimation. It would also be interesting to see how much more information each test could provide if the `<canvas>` element was made bigger than the size used in this study, as it would then render more text leading to a more unique render.

A. Appendix

This text originally appeared on the front page of the survey website.

A.1 How unique is your HTML5 <canvas> fingerprint?

A.1.1 A study of how your browser renders the HTML5 <canvas> element.

One of the new elements introduced with HTML5 is the <canvas> element, which is used to render text or images in the browser. The text and images differ slightly in how they are rendered depending on your software and hardware configuration. The tests on this page are part of a study that aims to answer the question about how uniquely a browser renders the <canvas> element. The tests create six pictures and then compares their differences to other users' results.

After the tests have finished, you will be presented with the results which will show you the ratio of devices that share your browser's <canvas> fingerprint, to the total number of browsers that have participated in the study. By participating, you will help us estimate how well these tests work to identify a browser.

The test data cannot be used to identify individual visitors to the webpage. For more information on who we are, what data is saved and how the tests work, see the About page.

Now then - how unique is *your* configuration?

Start tests

This page uses cookies. These cookies can not be used to identify individual visitors. For more information, see the About page.

B. Appendix

This text originally appeared on a special section of the website titled ”About this page”.

B.1 About this page

B.1.1 About us

We are two students at the Computer Systems and Science department of Stockholm University in Sweden currently undertaking a preparatory course for writing our bachelor's thesis. As part of this course, we are supposed to write a 'mini-thesis'. This survey is intended to help gather data for that paper.

B.1.2 About the tests

The tests performed create a fingerprint of your browser by checking how it uses the HTML5 `<canvas>` element to render different fonts and, if possible, a WebGL model. It then compares your rendered elements to other people's elements and sees how unique they are. How unique they are defines how easily identifiable your browser fingerprint is.

The six tests are further defined and described in section 3.1 of the report ”Pixel Perfect: Fingerprinting Canvas in HTML5” by Keaton Mowery and Hovav Shacham at the Department of Computer Science and Engineering of the University of California in San Diego. The report can be found here: http:cseweb.ucsd.edu/ hovav/dist/canvas.pdf

For research purposes, we store the result data from the tests, as well as a cookie to help identify returning devices. For control purposes, browser version and operating system version information is also saved. None of the stored data can on its own be used to identify any individual visitors to the site. The data will only be used to calculate an estimate of the entropy (uniqueness) of the fingerprints generated by the different tests. For an explanation on entropy and browser fingerprinting, see https:www.eff.org/deeplinks/2010/01/primer-information-theory-and-privacy

B.1.3 Technologies

The website is written in HTML, with the tests in JavaScript using PHP to save the data to a MySQL database.

Bibliography

Acar, G., Juarez, M., Nikiforakis, N., Diaz, C., Gürses, S., Piessens, F. and Preneel, B. (2013). FPDetective: Dusting the web for fingerprinters, *Proceedings of the 20th ACM Conference on Computer and Communications Security (CCS 2013),*, ACM. Accepted.
URL: *https://lirias.kuleuven.be/handle/123456789/417523*

Ball, A. (2013). The cookie is still crumbling: the challenges facing cookie tracking research, *International Journal of Market Research* **55**(1): 34 – 41.

Becker, H. and Geer, B. (1957). Participant observation and interviewing: a comparison, *Human Organization* **16**(3): 28–35.

Cover, T. M. and Thomas, J. A. (2005). *Entropy, Relative Entropy, and Mutual Information*, John Wiley & Sons, Inc.

Datainspektionen (n.d.). Vad är en personuppgift? [Electronic] Accessed on 2013-10-31.
URL: *http://www.datainspektionen.se/fragor-och-svar/personuppgiftslagen/vad-ar-en-personuppgift/*

Denscombe, M. (2010). *The good research guide for small-scale social research projects*, fourth edn, Open University Press, Milton Keynes.

Eckersley, P. (2010). How unique is your web browser?, *Proceedings of the 10th international conference on Privacy enhancing technologies*, PETS'10, pp. 1–18.

Electronic Frontier Foundation (2013). Panopticlick. [Electronic] Accessed on 2013-11-01. Take the test to find out how many have participated.
URL: *http://panopticlick.eff.org/*

Flood, E. and Karlsson, J. (2012). *Browser fingerprinting*, Master's thesis, Department of Computer Science and Engineering, Chalmers University of Technology.

International Telecommunications Union (2011). The World in 2011: ICT Facts and Figures, ITU Telecom World 2011. Accessed on 2013-10-31.
URL: *http://www.itu.int/ITU-D/ict/facts/2011/material/ICTFactsFigures2011.pdf*

International Telecommunications Union (2013). The World in 2013: ICT Facts and Figures. Accessed on 2013-10-31.
URL: *http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2013.pdf*

ISO (2013). ISO/IEC 27002:2013 - Information technology – Security techniques – Code of practice for information security management.
URL: *http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=54533*

Miyazaki, A. D. (2008). Online privacy and the disclosure of cookie use: Effects on consumer trust and anticipated patronage., *Journal of Public Policy & Marketing* **27**(1): 19–33.

Mowery, K. and Shacham, H. (2012). Pixel perfect: Fingerprinting canvas in HTML5, *in* M. Fredrikson (ed.), *Proceedings of W2SP 2012*, IEEE Computer Society.

Mulazzani, M., Reschl, P., Huber, M., Leithner, M., Schrittwieser, S. and Weippl, E. (2013). Fast and reliable browser identification with javascript engine fingerprinting, *Web 2.0 Workshop on Security and Privacy (W2SP)*.

Palmer, D. E. (2005). Pop-ups, cookies, and spam: Toward a deeper analysis of the ethical significance of internet marketing practices, *Journal of Business Ethics* **58**(1/3): pp. 271–280.

SFS 1998:204 (1998). Personuppgiftslag. Justitiedepartementet.

Tene, O. and Polonetsky, J. (2012). To Track or "Do Not Track": Advancing Transparency and Individual Control in Online Behavioral Advertising, *Minnesota Journal of Law, Science & Technology* **13**(1): 63–69.

The United States Census Bureau (2013). Population clock. [Electronic] Accessed 2013-09-26.
URL: *http://www.census.gov/popclock/*

Ur, B., Leon, P. G., Cranor, L. F., Shay, R. and Wang, Y. (2012). Smart, useful, scary, creepy: perceptions of online behavioral advertising, *Proceedings of the Eighth Symposium on Usable Privacy and Security*, SOUPS '12, ACM, New York, NY, USA, pp. 4:1–4:15.
URL: *http://doi.acm.org/10.1145/2335356.2335362*

W3Schools (n.d.). CSS Web Safe Fonts. [Electronic] Accessed on 2013-10-17.
URL: *http://www.w3schools.com/cssref/css_websafe_fonts.asp*

Yen, T.-F., Xie, Y., Yu, F., Yu, R. P. and Abadi, M. (2012). Host fingerprinting and tracking on the web: Privacy and security implications., *NDSS Symposium 2012*, The Internet Society.
URL: *http://dblp.uni-trier.de/db/conf/ndss/ndss2012.html#YenXYYA12*

Stockholm University
Department of Computer and Systems Sciences
Forum 100
SE-164 40 Kista
Phone: +46 8 16 20 00
www.dsv.su.se

Stockholm
University