Beacons and Their Uses for Digital Forensics Purposes

An Investigation

Luke Lubbe
Department of Computer Science
University of Pretoria
u11156342@tuks.co.za

Prof. Martin Oliver
Department of Computer Science
University of Pretoria
molivier@cs.up.ac.za

***Abstract*— This article relates to the field of digital forensics with a particular focus on web (World Wide Web) beacons and how they can be utilized for digital forensic purposes. A web beacon or more commonly "web bug" is an example of a hidden resource reference in a webpage, which when the webpage is loaded, is requested from a third party source. The purpose of a web beacon is to track the browsing habits of a particular IP address. This paper proposes a novel technique that utilizes the presence of web beacons to create a unique ID for a website, to test this a practical investigation is performed. The practical investigation involves an automated scanning of web beacons on a number of websites, this scanning process involves identifying which beacons are present on a web page and recording the presence of those beacons, the results of this scanning process is then encoded into a table for human analyses. The result of the investigation show promise and incentivizes further research. Real world implications, future work and possible Improvements on the methods which were used in this study are finally discussed.**

Keywords-Web bugs;Web beacons; Digital forensics; Web analytics

I. INTRODUCTION

Web beacons have a ubiquitous presence on the internet [1]. Websites which aim to monetize content by utilizing serialized ad-networks contain web beacons as a means of tracking user usage patterns and behaviors. The subject matter of this research paper is the investigation of these web beacons. The investigation examines how web beacons function, their intended purpose and how their presence can be used for digital forensic purposes with a focus on leveraging the principles of DNA fingerprinting to create a similar proposed form of fingerprinting for websites.

Web beacons can predominantly be found on public websites and are an example of non-intrusive method of gathering information about a user's web activity. A web beacon commonly takes the form of a remotely included resource that is hidden to the web user, the resource can be a one by one pixel image, html object or a piece of JavaScript [2]. When a request is made for the resource information about the requester it is sent to the host of the resource, this information can include the source of the request i.e. the websites which hosted the web bug and more importantly the IP address of the requesting browsers computer. Examples of detected web beacons on a website are shown in Figure 1.0

The reason for the widespread presence of web beacons is that they have proven to be a very useful method for tracking user activity on the internet without impairing the users browsing experience. The ability to track a user's internet use and habits has become invaluable for web analysis and subsequently internet marketing. Web analysis is the act of studying internet user's behavior with the objective of identifying opportunities for improvements to the user experience or site performance. An interest in web analysis is predominantly being driven by commercial related interests.

Targeted marketing is the practice of delivering adverts to a web browser based on the browsing habits of the I.P. address of the browsers host computer, targeted marketing results in users only being shown adverts which pertain to their particular preferences. An attempt is made to define the users preferences by tracking the content of websites which are visited by the user [3].

To identify which websites a specific user has visited the IP addresses of visitors are stored. However storing the IP address of a visitor is not enough to define their particular preferences and in order to create a better assumption about a user's preferences their browsing habits across various websites have to be tracked. To this end networks are created that serialize the act of placing beacons on multiple websites and coordinating the placement of adverts based on the correlation of a user's browsing habits across web sites that are part of the network [4].

For research purposes an investigation of web beacons is justified in that the ubiquity of web beacons have led to them being encountered and interacted with on a day to day basis and as such the chances of a web beacon being involved in a digital forensic investigation or being encountered during an investigation are high. The novel approach of using the presence of web beacons as unique identifiers for websites, can result in a positive benefit to digital forensics as a unique identifier for websites would prove useful in an investigation that requires evidence that a specific website was visited or

978-1-4799-7755-0/15/$31.00 ©2015 IEEE

even existed. This paper aims to establish whether the proposed method is a viable solution to creating a unique identification system for websites. Furthermore the paper will introduce and contextualize web beacons by giving a brief overview of the history of web analytics, which is the field of interest that web beacons are a component of and then an explanation of Web beacons, both their purpose and mechanisms. The final section relates web beacons and the implications they can have on digital forensics which is followed by a practical research into the viability of the proposed method.

II. WEB ANALYTICS

To better understand the current state of web analytics, it is prudent to consider the history of web analytics and in retrospect a little history of the open web. The first public use of the World Wide Web started in 1990 [5], as this was the first instance of large public access to the internet. At that time most companies simply used the internet as a global bill-board where information about their company could be displayed and accessed. To ascertain whether their company website was performing well, companies would interrogate the web server logs to access basic information.

Before the introduction of JavaScript the prime means of analyzing site usage was through log analysis [6]. This involved periodically monitoring server logs, which listed requests to a server for html content. Each of these visits was assumed to be a visit from a web browser to a website. Analyses of logs offered little in the way of insight into how a website was being used, i.e. how long a user stayed on a page or which part of a page was being focused on the most and if a website had a large amount of traffic, manual assessment of logs would become impractical.

A company called Webtrends, which was founded in 1993, provided the first commercial automated log analysis software [7]. Log analysis proved to be limited in the amount of information it could produce and the quality of information it could produce, as browsers would often cache pages and would not make a request to the server the second time a page was visited. This along with the rise in the trend of using proxies obscured information that could be collected from logs [7].

It was not until the introduction of JavaScript in 1996 [8] that browsing the internet became an interactive experience. JavaScript not only made it possible to create rich content on websites but it also created a means to send extra information about a visitor to the website's host webserver or to a remote third-party server. This information included the browser which the visitor was using or from which website link the visitor was directed from. This was achieved by employing a method called Page tagging.

Page tagging was introduced in 1997. Page tagging involves hosting an html object with JavaScript associated with it on a webpage. Page tagging can be viewed as the first instances of web beacons as the html object which was used for tagging, was most often hosted on a remote server with the specific purpose of serving requests and gathering associated information from the requester. Information would often be included in the requesting URL to the server hosting the requested content. Other methods of page tagging involved the contentious use of cookies, which involve creating a cookie for a visitor's web browser on their machines. These cookies could be accessed from repeat visitors to gather information about them.

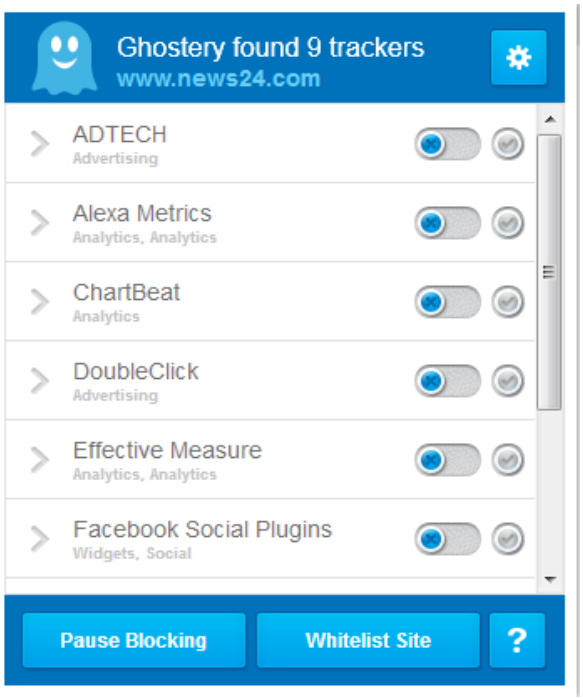

Figure1: An example of detected beacons on a website.

Growing concerns about privacy on the internet and legislative pressure lead to people not wanting any invasive data on their machines [1]; as such cookies have fallen out of favor and alternatives such as Web beacons have become the prominent mechanism for web analysis.

III. WEB BEACONS

Web beacons are examples of a non-intrusive means to gather information about a web browser without the user of the web browser knowing. A 1x1 bit image html object is commonly used and placed in an obscure manner that would not readily be perceivable to the human eye on a web page. This is not the only example of a web beacon. Other examples of web beacons are resources located on a web page that initiate requests to content hosted on third party servers, such as videos. This is the case with aggregate websites, such as news sites, which have content sourced from various other websites. In such cases it is still possible for the host of the original content to gain access to browsers information when a page is loaded on the aggregate site which contains remotely hosted content.

Another purpose of web beacons is to act as a means to deliver cookies or locally shared objects to a targeted website visitor [2]. These cookies can also be used to track a user's activity across websites in real time. This negates the need for lookups made to the server which holds logs for resource requests and the unfeasible number of lookups that may be required and absence of ready access to web server logs.

Web beacons have also been used in emails to determine if a receiver of an email has viewed the email, this is done by inserting a beacon in the message which is hosted on a remote server and when the message is opened a request will be made to the server which hosts the beacon thus notifying the sender that the message has been opened [9]. Changes in the RFC for http have also lead to alternatives to using a gif. A webpage can simply make a request to a server, encoded in the can be information of interest about the requester, and the server can return an Http status Code 204, which indicates No content, this method will save a minuscule amount of data [10].

The predominant use of web beacons in all their forms has been that of collecting information for web analytics for targeted marketing [11]. With the rampart growth in marketing on the open web, web beacons have subsequently become ubiquitous on the open web. Web beacons have offered advance ways of monitoring and measuring how successful advertising is. This is primarily achieved by creating a cookie on a user's browser when an advertisement is shown. By monitoring that browser's cookie they can see if the advertisements referenced website page has been visited in order to measure the effectiveness of the advertisement.

So important has web beacons become to organizations, that collaboration through syndicated networks have been created with the purpose of managing the hosting of web beacons and their placement across a multitude of websites. The purposes of these networks are to analyze internet user behavior with the goal of servicing targeted adverts and services to internet users. This process involves the creation of a personal profile for a user's web browser which is linked to the user by means of a cookie placed on the user's machine when a browser makes a request for a web beacon which was part of the ad network [4]. By maintaining a network of web beacons across websites, which may all differ in content, a profile for a browser can be populated with browsers specific preferences. The profile for a browser is then used to place specific adverts that would pertain to the user's preferences.

This paper proposes that the various beacons which are found on a website can be used as a form of "unique" indicator which is similar to how indicators are used in DNA fingerprinting.DNA fingerprinting relies on various indicators that are found on DNA strands, these indicators known as Microsatellites are used as molecular markers for DNA fingerprinting. DNA fingerprinting is the identification and recording of unique indicators in DNA. A method for a similar mechanism of identifying websites based on using web beacons as markers isintroduced and discussed in this paper with relation to possible real world implications.

IV. WEB BEACONS AND DIGITAL FORENSICS

To determine what affect web beacons can have on digital forensics, it is prudent to analyze the value, significance and impact that the presence of web beacons can have on digital forensic investigations and how they will actually be handled and used by investigators.

Companies make use of web beacons to effectively track the browsing habits of a user [7] as discussed previously and these habits are meticulously stored and updated in an automated process. Typically the information which is stored includes:

- The last date and time of activity.
- The URLs of webpages visited by the browser on which the network has trackers.
- The IP address associated with the browser.
- The content/subject of websites that the browser has visited.

The above mentioned information can prove useful to a digital forensic investigator in a scenario where proof is needed that a specific browser has visited a specific site. By correlating information in a cookie with the information held by a web beacon's host that generated the cookie, it is possible to prove (assuming no advance spoofing method was employed) that a browser has visited a webpage. Furthermore without the act of correlation, the server has a record of the specific IP address of the browser which can be used to narrow down the area of a sites visitor which can also be used for proximity evidence.

The focus of this paper is the investigation of the variation of the type of trackers present on web site. Due to the different uses for web beacons it is common to encounter websites which have a multitude of embedded beacons in their home pages. A few examples of common types of web beacons which a browser can encounter are for:

- Analytics: These are web beacons that are commonly part of an analytics network which are in contrast to Local analytical beacons which are created and maintained by the first party website.
- Social: These are web beacons that are used for widgets that aim to incorporate social functionality into a website.
- Advertising: These beacons form part of a targeted ad-network.

Different categories of beacons are provided and maintained by different networks and service providers and as such can be found on a single website.[12].

This paper proposes that there could exist a unique combination of web beacons on a website and that these beacons can be used as a means of creating unique identification for a website similar to that of DNA finger printing. DNA forensics relies on DNA finger printing to create genetic profiles for a person so that any DNA samples which are found in the course of an investigation can be matched to a person of interest. It is often crucial in proving that a person's DNA was indeed present at a scene of a crime or if it was not.

A similar situation can arise in a digital forensic context, where the need to prove that a specific website was visited by a particular browser. If a specific set of web beacons were present on the website, it would be expected the beacons would generate a specific set of cookies for the browser that visited

	TechCrunch.com	TheNextweb.com	Wired.com	Tech2.com	Gizmodo.com
BlueKai	**1**	**1**	**0**	**0**	**0**
Chango	**1**	**0**	**0**	**0**	**0**
DoubleClick	**1**	**1**	**1**	**1**	**1**
Drawbridge	**1**	**0**	**0**	**0**	**0**
Facebook Connect	**1**	**1**	**0**	**1**	**1**
Facebook Social Graph	**1**	**0**	**0**	**0**	**0**
Google Adsense	**1**	**1**	**1**	**1**	**1**

Figure 2: An example of the tables which were created, with columns indicating websites and rows indicating detected beacons.

that site. If a data base existed which documented and stored the beacons found on various websites, it could be consulted to determine and identify which websites could have been visited by a browser using the cookies which were found on the browser. These cookies would have been generated by correlated beacons from the visited websites. Likewise the database could be consulted to determine which cookies should be present based on beacons found on a specific website of interest to assist in determining a forensic link.

V. DNA FORENSICS

DNA forensics has proven to be an invaluable tool in conventional forensics and more specifically DNA finger printing. This paper proposes that web beacons can assist in creating a similar mechanism for identification. DNA finger printing relies on specific points on human DNA to ascertain whether two samples are from the same person [13]. Ninety-nine percent of human DNA is identical but the one percent that is different is enough to individualize a person. Highly variable sequences in DNA, which are referred to as minisatellites, are arranged in unique patterns. These patterns are used as a means to identify a DNA match and similarly this research paper will rely on the beacons present on a website that can be used as unique identifiers for matching purposes. To test this the following practical investigation was performed.

VI. PRACTICAL INVESTIGATION

The objective of the practical investigation was to determine whether it is possible to uniquely identify a website based on the presence of particular/specific web beacons on it. To accomplish this, an automated method of scanning websites and the recording of the presence of web beacons was performed.

The investigation began with an assessment of the Ghostery add-on for Firefox which is an add-on which relies on a database of signatures of various web beacons which can be embedded into a web page. Ghostery can generate a report for each website which the browser visits, the web beacons which are on the web page and what their purpose is. However each website had to be physically be opened and the most efficient way to achieve this was to automate the process of opening, scanning, and encoding to leverage Ghostery's ability to detect web beacons and keep track of the generated results; to achieve this; an add-on was coded in JavaScript for Firefox. This add-on would extract the generated report from Ghostery and save it to a flat text file. The add-on relied on a timeout function to determine when to grab the results of the Ghostery report which was done as a means to prevent waiting indefinitely sometimes for a website to finish loading.

To automate the opening and closing of the web browser a java program was coded. This java program would accept a list of websites and open them sequentially, once the Firefox add-on had saved the ghostery report for the web page the java program would read the generated flat text file and save its results into a database table. The Ghostery results were encoded into a table in such a way that the rows would contain web trackers and the columns would contain the websites which were scanned. Any new web beacons which are encountered were added to the table. The cells in the table were filled with 1s and 0s which would represent the presence of a beacon on a web site or the absence of a beacon on a website respectively.

An assertion was then made that by considering each column as a binary number you would in effect have created a unique ID for each website.

The sample amounts of websites used were limited to three groups of ten websites as the current process used was not optimized for large data volumes, as the investigation only aimed at determine the validity of the proposed method. This small sample group yielded promising results which could warrant further research which would involve creating a more efficient process to scan websites.

Specific details of the experimental setup comprised of the following:

- A List of 30 websites: The websites were chosen in such a way that the subject matter would pertain to a similar topic for ten websites; this is done to investigate if websites with similar content would make use of similar web beacon service providers in an attempt to identify cases of websites which have similar or the same beacons present. These collisions, or more importantly the absence of them would indicate that the proposed technique for finger printing websites shows promise.

- Ghostery: A Firefox add-on which scans a website for web beacons based on an internal web beacon signature database and generates a report for found web beacons on a particular web site.
- A Firefox add-on developed for this experiment which stored the Ghostery reports into a flat text file.
- A Java program which read the list of websites and sequentially visits them. After their web beacons have been stored in the flat text file using the Firefox add-on, the program read the file and added the contents to a database in the manner described previously to determine the binary ID. The topics of the website were chosen in such a way as to minimalize the potential overlap in subject matter, the reasoning for this is twofold. Firstly: to determine if there are common web beacons that are related to each specific topic that would only be found on websites that have content matter relating to the topic and Secondly: to minimize the chances that web beacons would be repeated across the websites.
- The three topics which were used for the experiment were: technology, fishing and poetry.

VII. RESULTS

The results gathered from the study showed promise but revealed that the premise of using the presence of web beacons as an indicator without any restrictions is problematic. This was highlighted by the fact that there were many cases encountered were websites did not contain any web beacons. These sites leave no trace on a web browser except for the site being an entry in the browsers search history. This would pose a problem as there would be no digital finger print using this method and one would have to rely on getting access to web server logs to prove that a browser had visited such a site (which often requires subpoenas and other lengthy processes for investigators).

An example of the tables which were generated are shown in Figure 2. Websites which did have beacons did show promise as there was not many websites which shared a binary ID when compared, both within a group and when compared across the 3 groups except for 2 results. Results across all three groups are shown in Figure 3.

Number of unique sites	20
Number of sites which beacons were not found on	8
Number of sites which were not unique	2

Figure 3: Results of Scanning.

As shown there was a high number of sites which were unique when all the beacons were considered but it is important to note that across those sites there were a few select beacons which were extremely prominent. Examples of beacons like these are ones relating to Google analytics, which is expected, as its subsidiary DoubleClick is one of the largest ad networks. This presents is an issue when trying to use the beacons as unique identifiers, we wish to condense the amount of identifiers which are required to as small a sub-set as possible as this would provide logistic benefits but their exist websites which are outliers, these websites are made unique by the fact that they are missing a common beacon. As such if you were to remove the beacon from the set you would effectively lose a binary ID for a website.

A solution which is proposed for this involves separating websites of interest into groups. This is not an ideal approach but by separating websites into groups based on the prevalence of a certain beacon, it becomes possible to retain outlier websites. There was one instance of two websites having the same Binary ID, these two websites had very few beacons and the ones which they did share were examples of the "conventional" beacons mentioned above.

A benefit of using binary strings to represent the ID for each website is that the hamming distance between two ID's can indicate the similarity or dissimilarity between the IDs,with higher dissimilarity being favorable as it is an indication that a greater level of distinction would be possible. To this end in the Technology subgroup, a set was created of all the Binary IDs compared to one arbitrarily chosen ID. The average hamming distance was then calculated, which was 39.785, this value is favorable as the string length was 115 bits, where 115 would be the number of beacons detected across all technology related sites.

Upon inspection of sites which did not have any beacons present, it was found that Ghostery did not detect many beacons on sites which were not in English. The reason for this could be that the ghostery database of signatures or the scanning mechanism only caters for text in standard ascii characters and cannot read characters which are not part of the standard ascii set. This could be remedied by using a more sophisticated scanner.

VIII. CONCLUSION

As the results have shown it could potentially be possible to use the presence of web beacons as a means to uniquely identify a website, but recording and cataloguing their presence , as was done in the experiment alone will not be a viable solution. Further research will be needed to determine whether a large enough sample of the open web would contain similar results as was found in this research paper. As was found in this research even though a small amount of the websites shared the exact same binary ID, when extrapolated to the total amount of websites on the open web, could result in many sites with shared IDs that stem from only having a few beacons on their webpages. Examples of these types of websites which only utilize a small common set of beacons could potentially not be of interest as there would be little in the way of differentiating websites such as these except for content.

The more beacons there are present on a website the higher the chances of diversity, but for efficiency purposes research can be performed to determine whether an optimum amountof beacons need to be present to be able to uniquely identify a website. Another aspect which would need to be considered is that of stability, this refers to how constant these IDs are. A website owner could change service providers which would host different beacons on the website. Future work could include recording how stable the IDs are by scanning for IDs over a time period and tracking changes. This process can be used to determine what the optimum frequency is to re-scan websites to keep the IDs current.

A combination of various identification methods should be used which could include the Binary ID as described in this paper for a viable approach to finger printing websites. Another aspect to consider is that of the open web and the deep web, the deep web can refer to websites which are not accessible by conventional means or networks. Web pages on these sorts of networks will most likely not make use of serialized services which provide beacons and would therefore not be suitable for this method.

Future research will include the collection of more data; this data will include beacons on websites as well as data pertaining to the temporal nature of the beacons, as changes to the beacons would require the re-generation of new IDs. To facilitate the collection of beacons a more scalable approach to the gathering of web beacons has to be devised. The detection method which was used in the research for this paper, namely the browser add-on Ghostery, will also need to be investigated to better understand how it affects the results gathered.

REFERENCES

[1] M. Y. Eltoweissy, A. Rezgui, and A. Bouguettaya, "Privacy on the Web: Facts, challenges, and solutions," IEEE Secur. Priv., vol. 1, no. 6, pp. 0040–49, 2003.

[2] A. Hamed, H. Kaffel-Ben Ayed, M. A. Kaafar, and A. Kharraz, "Evaluation of third party tracking on the web," in Information Science and Technology (ICIST), 2013 International Conference on, 2013, pp. 471–477.

[3] "Ad Servers and Advertising Networks." [Online]. Available: https://saylordotorg.github.io/text_emarketing-the-essential-guide-to-online-marketing/s06-05-ad-servers-and-advertising-net.html. [Accessed: 04-May-2015].

[4] C. A. Dwyer, "Behavioral targeting: A case study of consumer tracking on levis. com," Available SSRN 1508496, 2009.

[5] L. Kleinrock, "History of the Internet and its flexible future," Wirel. Commun. IEEE, vol. 15, no. 1, pp. 8–18, 2008.

[6] D. S. Sisodia and S. Verma, "Web usage pattern analysis through web logs: A review," in Computer Science and Software Engineering (JCSSE), 2012 International Joint Conference on, 2012, pp. 49–53.

[7] H. Singal, S. Kohli, and A. K. Sharma, "Web analytics: State-of-art & literature assessment," in Confluence The Next Generation Information Technology Summit (Confluence), 2014 5th International Conference-, 2014, pp. 24–29.

[8] "A Brief History of JavaScript," in DOM Scripting, Apress, 2005, pp. 3–10.

[9] "Big Brother is Watching: An Update on Web Bugs - big-brother-watching-update-web-bugs-445." .

[10] "Mobile Analysis in PageSpeed Insights - PageSpeed Insights — Google Developers." [Online]. Available: https://developers.google.com/speed/docs/insights/mobile?csw=1. [Accessed: 04-May-2015].

[11] A. Goldfarb and C. E. Tucker, "Privacy Regulation and Online Advertising," Manag. Sci., vol. 57, no. 1, pp. 57–71, 2011.

[12] "KnowYourElements.com - Presented by Ghostery." [Online]. Available: http://www.knowyourelements.com/#tab=list-view&date=2014-09-04. [Accessed: 04-May-2015].

[13] A. Goyal, "DNA Fingerprinting," Harv. Sci. Rev., 2006.