Text Steganography:The Deep Web in Plain Sight

Youssef Bassil

Abstract: Essentially, the Deep Web also known as the Invisible Web is a hidden web whose content cannot be found by search engines and thus is inaccessible using conventional means. With the rise of activism, many has started using the Deep Web as a way to bypass regulations in order to distribute their ideologies while keeping their identity totally in secret. Tor short for The Onion Router is a Deep Web network that has been for many years used by many people from whistleblowers to cyber criminals to disguise their identities. However, as the Tor network is free and open to public, its inner workings and protocols can be seamlessly reverse-engineered. As a result, security experts were able to restrict the Tor traffic and block its network ports and IPs, making it prone to constant investigation by intelligence, security bodies, and law enforcement agencies. This paper proposes a novel method for implementing the Deep Web on the public Internet using Text Steganography. In short, the proposed method hides a secret page into another benign page called the carrier page using Cascading Style Sheets. When the carrier page is accessed using a regular browser, the benign page is rendered. Nonetheless, when the very same carrier page is accessed using a proprietary browser that implements the proposed algorithm, the hidden version of the page is rendered, mainly the secret web page that was originally concealed into the carrier page. The experiments conducted showed that the proposed method is plausible, seamless, and transparent as it allowed a single web page to exhibit two versions, one that is part of the Surface Web and another one that is part of the Deep Web. As future work, the proposed Text Steganography algorithm can be improved so much so to make it more robust and harder to reverse engineer.

Index Terms: Deep Web, Search Engine, Text Steganography, Tor

I. INTRODUCTION

The Surface Web is the part of the Internet that can be crawled and indexed by search engines. The opposite term for Surface Web is the Deep Web which is the hidden remaining part of the Internet that cannot be reached by search engines [1]. In practice, in order to make any particular web content as a part of the Deep Web, publishers have to host it either on the World Wide Web and make it unreachable to public views, or to host it on a private network that is only accessible using special proprietary software. The former approach requires the use of unlinked pages, dynamic content, encrypted websites, and password-protected resources; while the latter approach requires the use of an underground secret private network such as the Tor network that can only be accessed using special software [2]. Basically, Tor short for The Onion Router is a private network, part of the Deep Web, that can only be accessed using a special web browser that implements non-standard communication protocols and ports and provides anonymity to users and web resources. Interestingly, the purpose of the Deep Web in general and the Tor network

in particular is to provide data anonymity to web publishers in a way to keep their identity concealed and publicly unknown. For instance, with the rise of activism, many protestors, activists, journalists, whistleblowers among other oppressed people, in attempt to bypass regulations and laws, they started utilizing the Deep Web to publish and disseminate their ideologies while keeping their identity totally in the dark [3]. Alternatively, the Deep Web is also used to conduct several illicit activities ranging from exchanging secret documents, bypassing censorship and avoid the control of dictatorial regimes to cybercrime and selling illegal products such as drugs, weapons, child pornography, human organs, among other illegitimate stuff and services. Hence, the word Dark Web was coined for designating the darker, sinister, and criminal part of the Deep Web [4].

Although the Tor network is a very reliable solution to provide data anonymity to web publishers, the fact that it is free and open to public, its inner workings and protocols can be reverse-engineered. As a result, security experts were able to restrict traffic to the Tor network by blocking its network ports and blacklisting the IPs of its nodes. Besides, as the Tor is notoriously known for running illegal businesses and conducting criminal activities, it is prone to constant investigation by intelligence, security bodies, and law enforcement agencies [5].

This paper proposes a new method for hiding Deep Web content in plain sight using Text Steganography. Fundamentally, text steganography is an information hiding technology that covers data into text files. Its applications are diverse, including secret communication, copyright protection, digital watermarking, and tamper proofing. The proposed method transforms secret web content, mainly a website written using the HTML/JavaScript language, into another form of plain text, mainly Cascading Style Sheets (CSS) attributes that look benign to regular browsers and to the public World Wide Web. However, when this website is accessed using a proprietary browser that implements our algorithm, a totally different looking website is rendered, namely its original secret version.
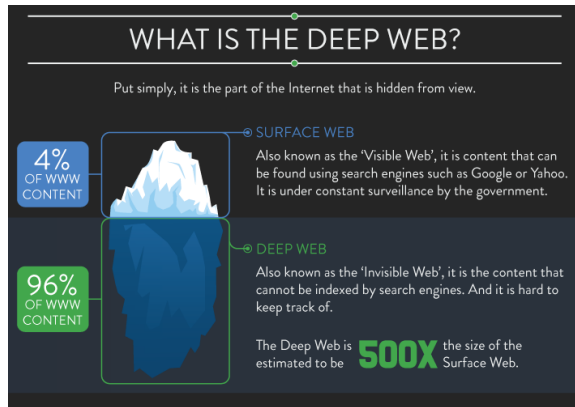
II. THE DEEP WEB AND THE ICEBERG ANALOGY

Principally, the Internet that we already know is made up of two webs: The Surface Web and the Deep Web [6]. The Surface Web also known as the Visible Web is the public World Wide Web whose content is indexed by search engines and can be accessed using regular web browsers such as Google or Bing. Since the Surface Web dwells in the public domain, it can be censored by regulations and controlled by governmental agencies.

Youssef Bassil, Chief Science Officer of the Lebanese Association for Computational Sciences (LACSC), Beirut, Lebanon.

In contrast, the Deep Web also known as the Invisible Web is a hidden web whose content cannot be found by search engines and thus is inaccessible using conventional means. According to Michael Bergman [1] the Deep Web is nearly 500 times larger than the Surface Web having a size around 7,500 TB (7.5 PT). Figure 1 depicts an Iceberg whose larger underwater bottom represents the Deep Web; while its smaller tip on the top represents the Surface Web, it is a clear assertion that the Deep Web overhauls the Surface Web.

Figure 1 – The Iceberg Analogy

Aside from the Deep Web, there also exits the Dark Web which is the sinister portion of the Deep Web that exists on the Darknet. The Darknet is a private underground network within the Internet that can only be accessed using a specific software. The Darknet is made up of volunteer secret web servers that host websites and software clients. One of most renowned Darknet network is the Tor network, short for The Onion Router [7]. In essence, Tor uses proprietary and encrypted protocols to access web resources on the Tor network, they have distinct domain names that end with ".onion" such as "http://bdpuqvsqmphctrcs.onion". For this reason, the websites hosted on the Tor network cannot be accessed using standard browsers such IE or Firefox, they however require the use of the Tor browser. Historically, the Tor network is not the only Darknet available as part of the Deep Web, there also exist other Darknets including I2P [8], Freenet [9], and the alternative top-level domains TLDs [10], better known as rogue TLDs.

III. THE DEEP WEB AND THE SEARCH ENGINES

A web search engine uses a web crawler, also known as web spider, to automatically crawl the World Wide Web for the purpose of saving and indexing its content [11]. Web crawlers often start with a list of URLs to visit, called the seeds. As the crawler visits these URLs, it identifies all the embedded hyperlinks in the page and then recursively visit them so as in due course, every page that has a hyperlink on the web pointing to it, will get crawled and indexed. Alternatively, advanced crawlers can query domain registrars to index newly purchased domains that haven't yet been hosted. Any web content including pages, images, scripts, multimedia files, and downloadable resources that can be fetched by web crawlers is considered part of the Surface Web. On the other hand, web content that goes undetected by web crawlers is considered part of the Deep Web. In fact,

several techniques were devised to hide web content from search engines, thus automatically making them part of the Deep Web. These techniques can be classified as the following:

- *Unlinked Content:* They are web resources that do not have a link to by other pages.
- *Dynamic Content:* They are web resources generated in response to a submitted query or only through submitting a form.
- *Password-Protected Content:* They are web resources secured by a username and password, for instance by means of the HTTP Basic Access Authentication protocol.
- *Content on the Intranet:* They are web resources hosted on private IPs and thus cannot be reached from the Internet.
- *Blocked Content:* They are web resources that impose restrictions on search engines by using CAPTCHAs, pragma no-cache HTTP headers, and ROBOTS.TXT. This prevents web crawlers from indexing them.
- *Private Proprietary Network:* They are web resources built using incompatible content such as non-HTTP and non-HTML, and hosted using non-standard networking protocols and ports. The Tor network is an example of a private proprietary network.

IV. THE DEEP WEB AND PRIVATE NETWORKS

As discussed in the previous section, web content that is not capable of being crawled by search engines is commonly considered as part of the Deep Web. Nonetheless, an alternative way to make this possible is using private networks such as the Darknet and the Tor network. Basically, the Tor network short for "The Onion Router" is a network within a network on the Internet that is private and cannot be accessed using regular web browsers but using a special proprietary browser called the Tor browser. Inherently, the Tor network is composed of Tor servers operated by a worldwide volunteer network of servers; and Tor browsers operated by anonymous user clients. The traffic between clients and servers are routed through hops using the "onion routing" scheme while providing complete anonymity and encryption to all the communicating parties. The Tor anonymity is based on the idea of distributing routing information during communication so that the original physical locations of the source and destination parties are kept unknown. In effect, the Tor connection starts with client X who requires to connect to server Y in order to retrieve a particular website on the Tor network. The client X using the Tor browser connects to a decentralized directory within the Tor network to retrieve addresses of the Tor nodes. The Tor browser picks randomly one of these addresses as an entry node and connects to it through an encrypted connection. Consequently, the entry node would start a chain of encrypted connections with random nodes until one of these nodes known as exit node identifies the destination of server Y.

The requested website is then returned in reverse order, back in chain of nodes, to the original sender, namely the client X. As aforementioned previously, this whole chain of connections is done using encryption protocols providing complete confidentiality to information being transmitted over the Tor network, while also providing complete anonymity to both clients and servers [12][13][14].

V. PROPOSED SOLUTION

This paper proposes a novel method for building a Deep Web platform using Text Steganography. It hides a secret web content, mainly a website written using HTML language into another HTML page using textual features and CSS (Cascading Style Sheets). The proposed algorithm works as follows: Every character in the secret web page is located randomly in the carrier page. This would result in an index pointing to the location of the character to hide in the carrier page itself. The size of this index is fixed to 4 decimal digits making it able to refer to 10,000 different character positions in the carrier HTML page, ranging from 0000 to 9999. Afterwards, the indexes produced by this process are used to generate numeric values to CSS attributes stored in the same carrier page. The carrier web page is then hosted on the public domain making it part of the Surface Web. When the carrier website is accessed using regular web browsers such as IE or Chrome, it is rendered normally showing only the innocent-looking version of the website. However, when this website is accessed using a proprietary browser that implements our algorithm, a totally different looking website is rendered, namely the original secret version.

VI. DIGITAL STEGANOGRAPHY

Fundamentally, steganography refers to "secret writing" in Greek, and is the art and science of hiding information inside innocuous files such as images, audio, and video files, in ways that avoid the detection of the hidden information [15]. The outcome of steganography is a covert channel of communication through which secret data can be transmitted in total secrecy avoiding drawing eavesdroppers' suspicions. In current practice, steganography is used to hide secret data such as text messages into carrier files such as images while maintaining the size and quality of the carrier file. As an algorithmic model, steganography can be thought as the "Prisoners' Problem" [16]. In this model, two prisoners put in jail, Alice and Bob, want to communicate about an escape plan. The challenge is that they can only communicate through the warden of the prison, Wendy, who prohibits both Alice and Bob to communicate in code using standard cryptography. As a result, Alice and Bob devise a new method for secret communication called Steganography. It is about hiding the message that needs to be communicated in an innocent-looking image using a computer. This image is then handed in by Bob to the warden Wendy to pass it along to Alice. Wendy, looking at the image, would not notice anything suspicious and would subsequently pass it along, not knowing that the pixels of the image encode the secret message. Alice, after receiving the image, would recover the secret message of Bob as she knows how the trick works. Formally, the steganography model can be mathematically

defined as follows [17]: the original file into which the secret message is to be concealed is denoted by A, the secret message to hide by M, and the carrier file by C. Actually, the carrier file C is visually identical to the original file A but with one difference is that it has the message M embedded inside it. The steganography encoding algorithm which is used to cover a secret message M into a file A, is denoted by S(A, M, Enc)=C where "Enc" denotes the encoding mode. On the other hand, the steganography decoding algorithm which is used to recover the secret message M out of the carrier file C, is denoted by S(C, Dec)=M where "Dec" denotes the decoding mode. Figure 2 depicts the basic mathematical model of steganography.
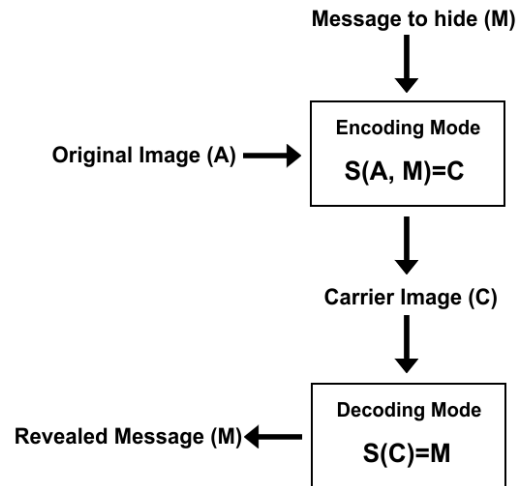

Figure 2 – Mathematical Model of Steganography

VII. THE PROPOSED ALGORITHM

The proposed algorithm entails several executional phases required to hide a secret HTML web page S into another HTML web page referred to as Carrier page and denoted by C. CSS attributes are used as part of the process and are denoted by CSS. Following is the process flow of the proposed algorithm:

1. An input secret HTML web page is fed to the algorithm. It is denoted by S=$\{s_i, s_{i+1}, s_{i+2}, \dots s_{n-1}\}$ where s is a single character in S, i is an index pointing to the ith character s and n is the total number of characters in S.

2. A benign HTML web page is created called the carrier web page whose purpose is to hide the secret web page S in it.

3. The carrier web page is denoted by C=$\{c_{k=0}, c_{k=1}, c_{k=2}, c_{k=3}, \dots, c_{k=9999}\}$ where c is a single character in C, k is an index pointing to the kth character c, and 9999 is the maximum number of characters that C can have.

4. A random number is chosen called random index denoted by INDEX such as INDEX<=9999. Its purpose is to point to a random character c in C that matches the corresponding character s in S. This operation is repeated until c_{index} matches s_i or until S is exhausted, such as WHILE $c_{index} \neq s_i$ → INDEX=$Rand(9999)$ THEN INDEX=$Rand(9999)$ and $i+1$

5. The list of output indexes are denoted by I={$index_0$, $index_1$, $index_2$, $index_{n-1}$} where n is the total number of indexes which is equal to the total number of characters in S. Every index has a fixed size of 4 decimal digits making it able to refer to 10,000 different character positions in the carrier HTML page C, ranging from 0000 to 9999. For instance, a generated sample of indexes would look like I={0320, 0043, 1645, 0005, 8733, 5543}, with further reduction to I={03, 20, 00, 43, 16, 45, 00, 05, 87, 33, 55, 43}

6. Working with the list of indexes I, a sequence of CSS attributes are generated whose values mimic the indexes in I such as CSS={att_0:$index_0$; att_1:$index_1$; att_2:$index_2$; att_{n-1}:$index_{n-1}$} where *att* is a CSS attribute, *index* replaces the attribute value, and *n* is the total number of indexes which is equal to the total number of CSS attributes which in turn is also equal to the total number of characters in the original secret web page S. For instance, a generated sample of CSS attributes would look like CSS={*font:3px ; line-height:20px ; height:0px ; border:43; margin-top:16px*}

In order to recover the secret web page S from C, a proprietary web browser that implements the proposed algorithm and method must be utilized. If a regular browser is used instead, the innocent-looking carrier page C would be rendered normally. The proprietary browser however would process the CSS attributes from a special predefined location in the carrier page, and uses their values to locate hidden characters in C that will eventually make up the original secret web page S.

As a mathematical model, the proposed steganography system can be summarized as having two processes. The encoding process is carried out such as $E(C, S)=C+CSS=C'$ where C is the carrier page into which the secret page S is hidden using the encoding algorithm E. The output is the carrier C in addition to a set of CSS attributes leading to a modified version of the carrier page denoted by C'. On the other hand, the decoding process is carried out such as $D(C')=D(C+CSS)=S$ where S is the secret page recovered out of the carrier page C'. Figure 3 depicts the mathematical model of the proposed method.
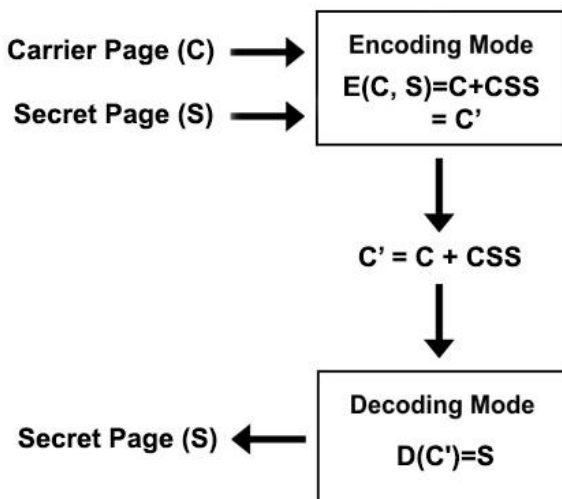

Figure 3 – Mathematical Model of the Proposed Method

VIII. EXPERIMENTS AND RESULTS

In the experiments, two web pages are created. The first one is an innocent-looking web page discussing how to travel to Hawaii during the summer season. It is a regular HTML page built using any web editor and meant to be part of the Surface Web (Picture 4). The second page is a sinister web page conveying some confidential information about a secret society (Picture 5). It is also written using HTML and meant to be part of the Deep Web. Using the proposed steganography algorithm, the secret society page is concealed in the Hawaii page which is referred to as the carrier page (Picture 6). Looking at the HTML source code of the carrier page, the CSS attributes in the header contain numeric values. They represent the text position of the secret page in the carrier page. As a final step, the carrier page containing the CSS styles is hosted on the World Wide Web under the public domain "http://www.hawaii-travel.com". Furthermore, a special proprietary web browser is built using C# and the .Net Framework [18]. It implements the proposed algorithm and it is used to access the carrier web page. When tested using a regular web browser such as IE, the domain "http://www.hawaii-travel.com" displayed the benign version of the website, mainly the one discussing holidays in Hawaii (Picture 7). However, when the same domain was accessed using our proprietary browser, the secret version of the website was rendered, mainly the one pertaining to the secret society (Picture 8).

```html
<!DOCTYPE html>

<html>
<head>
</head>
<body>

    <strong style="font-size:18pt">HAWAII</strong>

    <br/><br/>

    <div style="width:500px">
        Hawaii is the 50th and most recent state to have joined
        the United States, having received statehood on August 21, 1959.
        Hawaii is the only U.S. state located in Oceania, the only U.S.
        state located outside North America, and the only one composed
        entirely of islands.
    </div>

    <br/><br/>

    <div style="width:500px">
        A somewhat divisive political issue arose in 1978 when the Constitution
        of the State of Hawaii added Hawaiian as a second official state language.
        The title of the state constitution is The Constitution of the State of Hawaii.
        Article 5, Section 1 of the Constitution uses The State of Hawaii.
        Diacritics were not used because the document, drafted in 1949, predates
        the use of the kahako in modern Hawaiian orthography. The exact spelling of
        the state's name in the Hawaiian language is Hawaii.
    </div>

    <br/><br/><br/>

    <table>
      <tr>
        <td><img src="images\01.jpg" /></td>
        <td><img src="images\02.jpg" /></td>
        <td><img src="images\03.jpg" /></td>
      </tr>
    </table>

</body>
</html>
```

Figure 4 – The Innocent-Looking Page

```
<!DOCTYPE html>

<html lang="en">
<head>
    <meta charset="utf-8" />
    <title></title>
</head>
<body>
    <table width="800">
        <tr>
            <td width="400px"><img src="images\logo.jpg" /></td>
            <td>
                <h2>Welcome to our Secret Society on the Deep Web!</h2>
                <p>
                    Welcome to the informational pages of our secret
                    society, a Fraternity consecrated to Truth
                    and dedicated to the transmission of knowledge.
                    Our Venerable Order is a School of Light and
                    self-improvement in which the members, by means
                    of a progressive system of study and practical
                    application of what is learned, become not only
                    the Masters of their own lives, and Architects
                    of their own destiny, but also have the highest
                    existing ethical ideals into their personality.

                    If you are someone who truly aspires to be the
                    complete keeper of your own life, to ascend toward
                    perfection, to know the laws that govern the physical
                    realm and the mysteries of the mental and spiritual,
                    we invite you to accompany us on the Path of Knowledge.
                </p>
            </td>
        </tr>
    </table>
</body>
</html>
```

Figure 5 – The Secret Page

```
<!DOCTYPE html>

<html>
<head>
  <style>
    .trsq{font-size:11px;top:3px;left:16px}.sform{height:65px}
    .sfbg{background:0214;box-shadow:0 1px 6px 0 rgba(32,33,36,0.28)}
    .sbibod{height:32px;margin:10px 0;border-radius:16px}
    .sbrqa{height:20px;width:20px} .searchform{width:100%}
    .sbico-c{height:32px;line-height:32px}.sbib_b{padding-top:0}
    .gsfs{font-size:14px;line-height:32px}.gsfs{font-size:14px}
    .logo{height:28px;width:86px} .mdiv {padding:17px 34px 0}
    .gb{top:11px} .sbse{padding:0px 0} .rdd{border-radius:16px}
    .vgt{border-bottom-right-radius:16px} .qs_st{line-height:32px}
    .a{top:32px} .gsri_a{background-size:20px 20px;height:20px;width:16px}
    .sbico{height:20px;width:20px}.hhform{top:311px} .srp{top:20px}
    .srrd{position:112;}.sfbg{height:69px;left:0;position:66;width:100%}
    .sfbtt{height:65px}. cnt{padding-top:20px;} .subctrl{min-height:0px}
  </style>
</head>
<body>

  <strong style="font-size:18pt">HAWAII</strong>

  <br/><br/>

  <div style="width:500px">
    Hawaii is the 50th and most recent state to have joined
    the United States, having received statehood on August 21, 1959.
    Hawaii is the only U.S. state located in Oceania, the only U.S.
    state located outside North America, and the only one composed
    entirely of islands.
  </div>

  <br/><br/>

  <div style="width:500px">
    A somewhat divisive political issue arose in 1978 when the Constitution
    of the State of Hawaii added Hawaiian as a second official state language.
    The title of the state constitution is The Constitution of the State of Hawaii.
    Article 5, Section 1 of the Constitution uses The State of Hawaii.
    Diacritics were not used because the document, drafted in 1949, predates
    the use of the kahako in modern Hawaiian orthography. The exact spelling of
    the state's name in the Hawaiian language is Hawaii.
  </div>

  <br/><br/><br/>

  <table>
    <tr>
      <td><img src="images\01.jpg" /></td>
      <td><img src="images\02.jpg" /></td>
      <td><img src="images\03.jpg" /></td>
    </tr>
  </table>

</body>
</html>
```
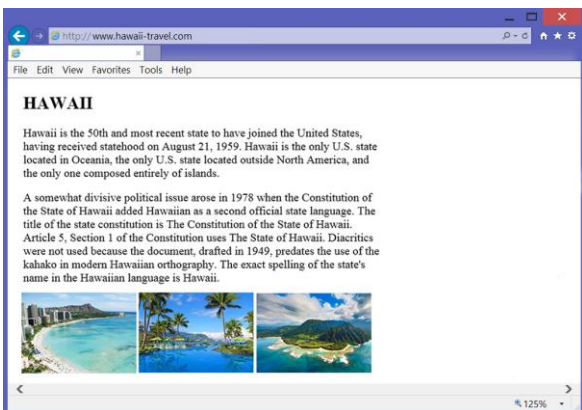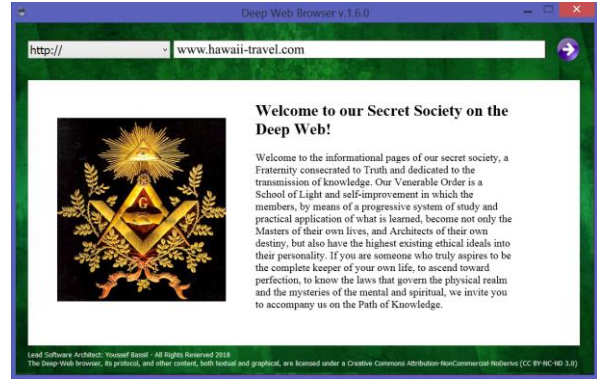
Figure 6 – The Carrier Page

Figure 7 – IE rendering the Innocent-Looking page

Figure 8 – Proprietary Browser rendering the secret page

IX. CONCLUSIONS

This paper proposed a novel method for implementing the Deep Web on the public World Wide Web using Text Steganography. In short, the method hides a secret page into another benign page called the carrier page. The algorithm uses Text Steganography and CSS in order to encode the hidden data. When the carrier page is accessed using a regular browser, the benign page is displayed. Nevertheless, when the very same carrier page is accessed using a proprietary browser that implements the proposed algorithm, the darker version of the page is rendered, mainly the secret web page that was originally concealed into the carrier page. The experiments conducted showed that the proposed method is plausible, seamless, and transparent. In other words, the proposed method allows a single web page to exhibit two versions at the same time, one that is part of the Surface Web - displayed when the page is accessed using a regular browser; and one that is part of the Deep Web - displayed when the page is accessed using a proprietary browser. Besides, as the proposed method uses HTTP and HTML standards in addition to the de-facto Internet protocols, it can prove to be difficult to be discovered, monitored, and restricted, thereby ensuring the anonymity of the data published on the Deep Web.

X. FUTURE WORK

As future work, more complicated types of content can be exploited besides the textual data. This includes audio and video streaming, flash objects, and client-side scripts. Furthermore, the presented Text Steganography algorithm can be improved so much so to make it more robust, more compact, and harder on steganalyst to reverse engineer it and recover its inner workings.

ACKNOWLEDGMENT

REFERENCES

1. Bergman, Michael K, "The Deep Web: Surfacing Hidden Value", the Journal of Electronic Publishing, vol. 7, no. 1, 2001.
2. Lin, K. and Chen, H., "Automatic Information Discovery from the Invisible Web", in Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'02), 2002.
3. Senker, Cath, "Cybercrime & the Dark Net: Revealing the hidden underworld of the internet", London Arcturus Publishing, 2016, ISBN 9781784285555
4. Janis Dalins, Campbell Wilson, Mark Carman, "Criminal motivation on the dark web: A categorization model for law enforcement", Elsevier Digital Investigation Journal, vol 4, pp. 62-71, 2018
5. "Sealed Complaint 13 MAG 2328: United States of America v. Ross William Ulbricht", p. 6, 2014.
6. Devine Jane, Egger-Sider Francine, "Beyond Google: the invisible web in the academic library", The Journal of Academic Librarianship, vol. 30, no. 4, pp. 265–269, 2004
7. "Tor Project: FAQ", www.torproject.org, retrieved 25 Dec 2018.
8. "The Invisible Internet Project", https://geti2p.net/en/, retrieved 25 Dec 2018.
9. "The Free Net Project", https://freenetproject.org/, retrieved 25 Dec 2018.
10. Bastick, Zach, "Our Internet and Freedom of Speech Hobbled by History: Introducing Plural Control Structures Needed to Redress a Decade of Linear Policy", European Commission: European Journal of ePractice, vol. 15, pp. 97–111, 2012
11. Ross Nancy, Wolfram Dietmar, "End user searching on the Internet: An analysis of term pair topics submitted to the Excite search engine", Journal of the American Society for Information Science, vol. 51, no. 10, pp. 949–958, 2000
12. M.G. Reed, P.F. Syverson, D.M. Goldschlag, "Anonymous connections and onion routing", IEEE Journal on Selected Areas in Communications, vol. 16, no. 4, 1998
13. TOR - Onion routing online documentation, https://www.torproject.org/, retrieved on June 2014
14. P. Syvneron, G. Tsudik, M. Reed, C. Landwehr, "Towards Analysis of Onion Routing Security", Designing Privacy Enhancing Technologies: Workshop on Design Issue in Anonymity and Unobservability, vol. 1, no. 1, pp. 96-114, 2000.
15. R.J. Anderson, F.A.P. Petitcolas, "On the Limits of Steganography", IEEE Journal Selected Areas Communnication, vol. 16, no. 4, pp.474-481, 1998
16. G. J. Simmons, "The prisoners' problem and the subliminal channel," in Advances in Cryptology: Proceedings of Crypto 83, pp. 51–67, 1984.
17. K. Bailey, K. Curran, "Steganography the Art of Hiding Information", BookSurge Publishing, 2005.
18. Charles Petzold, "Programming Microsoft Windows with C#", Microsoft Press, 2002.