American Journal of Engineering Research (AJER)
e-ISSN: 2320-0847 p-ISSN : 2320-0936
Volume-8, Issue-5, pp-45-51
www.ajer.org
Research Paper Open Access

Audio Steganography Method for Building the Deep Web

Youssef Bassil
LACSC – Lebanese Association for Computational Sciences Registered under No. 957, 2011, Beirut, Lebanon

ABSTRACT:*The Deep Web is the portion of the Internet that cannot be crawled nor indexed by common web search engines. The idea behind the Deep Web is to host web content on a privatenetwork that is only accessible using proprietary software. For instance, the Tor network is a private network that hosts petabytes of data that are confidential by nature and not open to the public views. Additionally, the Tor uses non-standard protocols to ensure the anonymity of its users. Although the Tor network delivers unprecedented capabilities in protecting the privacy of data being published on its network while ensuring the total anonymity of their owners, the fact that it is free, open-source platform, and accessible via community tools, can raise suspicions that something illicit and secret exists, and as a result it can be easily shutdown and have its network nodes censored. This paper proposes a new method for implementing the Deep Web using Audio Steganography. In essence, the proposed method camouflages a secret web page into an audio carrier file that is hosted on a carrier website on the public domain. When users access the carrier website using a regular browser, they would only see the innocuous version of the website, mainly the website that plays the benign audio clip; whereas, when users access the carrier website using a proprietary browser that implements our algorithm, they would see a totally different website, mainly the secret website that was originally hidden inside the benign audio file using Steganography. Experiments proved that the proposed method is feasible to build and that it supportsimplementingthe Deep Web in plain sight without drawing any suspicions whatsoever regarding the existence of any secret data. As future work, file types other than audio are to be investigated and experimented including image files, video files, and text files.*

KEYWORDS:*Deep Web, Dark Net, Audio Steganography, Tor*

Date of Submission: 25-04-2019 Date of acceptance: 05-05-2019

I. INTRODUCTION

The Surface Web is the portion of the Internet that can be crawled and reached by web search engines. The opposite term for Surface Web is the Deep Web which is the unseen remaining portion of the Internet that cannot be indexed by search engines [1]. In fact, in order to make any given web resource as a part of the Deep Web, content publishers have to host it either on the World Wide Web and make it inaccessible to the public, or to host it on a private network that is only accessible using special proprietary software. The former method requires the use of unlinked pages, encrypted websites, dynamic resources, and password-protected content; while the latter method requires the use of an alternative secret private network such as the Tor network that is only accessible using special browsers and 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 content. Essentially, the purpose of the Deep Web is to provide data anonymity to web publishers in a way to keep their identity obscured 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 avoiding 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 can deliver anefficientway to ensure 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 to implement the Deep Web using Audio Steganography. Fundamentally, steganography is a technique for hiding data such as text into another form of data such as images or audio files [6]. For instance, a secret text message called covered text is concealed by the sender into an audio file called carrier file and then sent to the receiver. On the other side, the receiver deciphers the audio file and recovers the secret text that was hidden inside. If by any means, the carrier file is intercepted by eavesdroppers, it would simply appear like any regular audio file, and thereby avoiding raising suspicions and hiding the fact that a secret communication has taken place. The proposed method aims to implement the Deep Web on the public World Wide Web by hiding secret web content into regular WAV audio files hosted on a public carrier web page. Users who access the carrier web page using a regular browser would only see the innocent-looking version of the page mainly the page that is showingthe audio file; whereas, users who access the carrier web page using a proprietary browser that implements our algorithm would see a totally different page mainly the secret web content that was deciphered from the audio file using steganography. As the carrier web page looks benign on the public World Wide Web, it can reliably escapes traffic blocking and other security restrictions. Additionally, it does not raise uncertainties that some secret information is present and thereby reinforcing the data anonymization practices.

II. AUDIO STEGANOGRAPHY

In practice, digital steganography is abouthiding a computer file, message, video, image, or audio within another file, message, video, image, or audio. The word steganography stems from the Greek words steganos, meaning "secret", and graphein, meaning "writing". In short, steganography refers to "secret writing" [7]. Audio steganography is yet a type of digital steganography that conceals data into either compressed audio files such as WAV or uncompressed audio files such MP3 or WMA files. Audio steganography exploits the Human Auditory System (HAS) which cannot detect the minor variations of high audio frequencies in the audible system.As a result, audio steganography can hide bits of secret data intothose high frequencies without destroying the quality or size of the original audio file [8]. Several methods were designed to perform Audio Steganography includingthe Least Significant Bit method [9], Silence Interval method [10], Echo hiding method [11], Phase and Amplitude Coding methods [12][13], Discrete Wave Transform method [14], and the Spread Spectrum method [15]. Basically, the LSB methodembedsevery secret bit from the data to hide into the rightmost bits of every audio sample of the original audio file. The LSB method allows hiding a fair amount of data without corrupting the quality of the audio file, in addition that it is relatively effective and easy to implement.Essentially, digital steganography is governed by four key elements. They are as follows [7]:

1. Covert Data: It is the secret payload that needs to be secretly communicated between two parties.
2. Carrier File: It is the main file into which the secret data are hidden. The carrier file can be an image file, an audio file, and even a text file.
3. Carrier Channel: It refers to the type of the carrier file, for instance, BMP, WAV, MP3, TXT, etc.
4. Capacity: It refers to the size of data that the carrier file can hold without gettingdamaged.

III. THE DEEP WEB

The Deep Web consists of web data content that are invisible to public and not indexed by search engines [16]. This content however can be accessed using direct URL or by using some sort of authentication mechanism; other content, are even deeper and require special tools and software to access. The Deep Web is nearly 500 times larger than the Surface Web with size around 7.5 Petabytes (the surface web is the opposite term of the deep web, that is the web that is visible to the public) [17]. The purpose of the Deep Web is to ensure the privacy and anonymity of web publishers who want to remain anonymous or create websites that cannot be traced back to a physical location or entity. The Deep Web also establishes covert communication channels between web content and web users who want to escape censorship, laws, and governmental regulations. From an implementation point of view, the Deep Web can be achieved using one of the outlined methods below whose ultimate goal is to hide web resources from the swarm of search engines.

- 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. TOR AND THE DARKNET

The Darknet is the notorious part of the Deep Web, and it is often operated by lawbreakers and convicts whether individuals or organizations. The Darknet is used for conducting illegal activities such as illicit trade, selling drugs, guns, counterfeit software, and human organs [18]. Moreover, the Darknet is used by activists who want to escape censorship and disseminate ideological, social, political, economic, and religious ideas as a way to guarantee freedom of speech. From a technical point of view, the Darknet is a private network which operates using specific software often using non-standard communication protocols and ports. The three most popular Darknet networks are Tor, Freenet, and I2P.

Tor short for The Onion Router is a private network that can only be accessed using a special web browser called the Tor browser [19]. It uses special non-standard communication protocols to provide anonymity between users and websites published on the Onion Router. The domain names hosted on the Tor network often end with ".onion" such as "http://bdpuqvsqmphctrcs.onion/". As a result, they cannot be accessed using standard browsers such IE or Firefox. Basically, the Tor network is composed of a worldwide volunteer network of servers where the traffic between them is randomly distributed through hops using the "onion routing" scheme in order to provide complete anonymity to all the communicating parties. Moreover, the Tor scheme employs cryptographic technologies to encrypt traffic between users and servers making it enormously difficult for eavesdroppers to unscramble the communication messages of the network [20].

In practice, Tor delivers unprecedented capabilities in protecting the privacy of data being published on its network while ensuring the total anonymity of their owners [21]. However, the Tor network is a free, open-source platform that anyone can have access to using the Tor browser. Consequently, illegal activities and counterfeit content can be monitored and censored by regulations. Several FBI investigations have led to the shutdown of numerous websites and to the arrest of several web publishers working on the Tor network. Likewise, web content present on the Tor network can raise suspicions as the network itself is known for its notoriety. Furthermore, security experts can find their way to block traffic in and out of the network by banning its network ports and blacklisting its node IPs. All in all, the Darknet and in particular the Tor network do not provide complete anonymity to web publishers and users as they do not obscure the fact that something secret is taking place.

V. PROPOSED METHOD

This paper proposes a novel method for implementing the Deep Web using Steganography over the public World Wide Web. The idea behind the proposed method is to hide secret web content, mainly a webpage written in HTML language, into benign audiofile that is hosted on a traditional website on the public domain that anyone can have access to using a regular web browser such as IE or Firefox. However, users who access the website using a regular browser would only be exposed to the innocent-looking version of it, mainly the website that plays the benign audio clip; whereas, users who access the website using a proprietary browser that implements our algorithm would be able to visualize the secret webpage that was originally hidden inside the benign audio file using Steganography. From the user perspective, the whole process is seamless and transparent, and therefore it does not raise any suspicions or hints regarding the existence of any secret data.

A.Design Specifications

The proposed method employs a Steganography algorithm that is designed to work on 8-bit uncompressed digital WAV audiofiles. Basically, an8-bit WAV audio file is composed of a set of audio samples each of which is of length 8 bits or 1 byte [22]. The proposed technique hides the secret data into the three LSBs of each of these audio samples; thus, the hiding capacity is equal to 3 bits out of 8 bits or 37% of the total size of the carrier audio file (3/8=0.37=37%). The carrier audio file is manipulated as a matrix of a finite set of audio samples S each composed of 8 bits. The secret data to hide, in our case a webpage or HTML script, is converted into binary format and substituted in the three LSBs of every 8-bitaudio sample in the carrier file. Figure 1 depicts the design behind the steganography algorithm used by our proposed method.

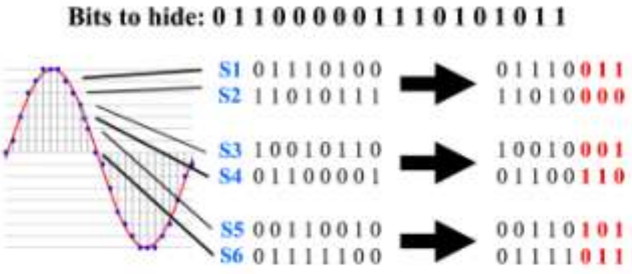


Fig. 1.The Steganography Algorithm

Eventually, when the secret HTML page is completely concealed inside the carrier audio file, the audio file itself is then hosted on a regular website (we will call it the carrier website) that the web publisher has created for the purpose of conveying his secret page. Alternatively, the web publisher can add any information or text along with the carrier audio to his carrier website just to make it sound more legit. Now, in order to access the secret page, a special browser that implements our method must be used. First, the browser opens the carrier website, then extracts the secret page from within the carrier audio file (already embedded in the carrier website) and renders it as if it was the original page that the user intended to open. In contrast, when the carrier website is accessed using a regular browser, its innocuous content would be rendered playing the benign carrier audio along with rendering any other information that the web publisher had originally included.

Below is a detailed process flow delineating how the proposed method and algorithm work:

1. The secret HTML page is converted into binary format so that it becomes compatible for storage inside the carrier audio file.

a. The secret HTML page, regardless of its content – whether HTML, JavaScript, CSS, or server scripts is converted into a binary form resulting into a string of bits denoted by D={b_0, b_1, b_2, b_3,…b_{n-1}} where b_i is a single bit composing the secret page and n is the total number of bits.

b. The string of bits D is organized into chunks of 3 bits, such as D={ C_0[b_0, b_1, b_2], C_1[b_3, b_4, b_5], C_2[b_6, b_7, b_8],…C_{m-1}[b_{n-3}, b_{n-2}, b_{n-1}] }, where $C_j$is a particular 3-bit chunk, m is the total number of chunks, and n is the total number of bits making up the secret page.

2. An8-bit WAV carrier audio file denoted by AUD is chosen to hide in it the string bits D.

3. The chunks of the secret page namely D that were created in step 1.b are embedded sequentially into the LSBs of every audio sample of the carrier audioAUD.

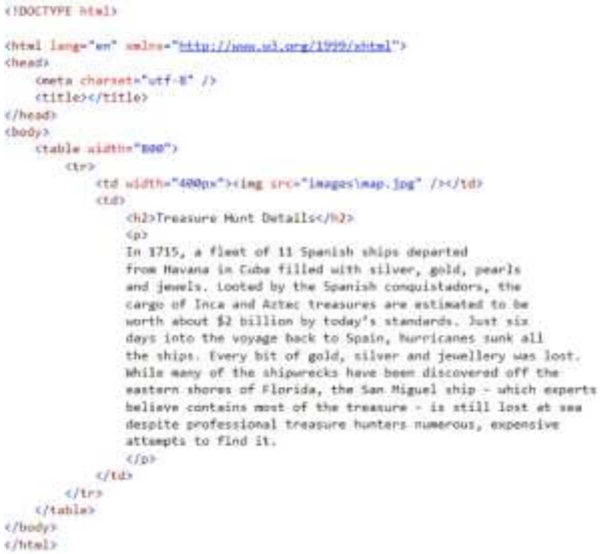
a. As the carrier audio file is of 8-bit format, meaning it is composed of multiple 8-bit audio samples, every chunk $C_j$is stored in the three LSBs of every audio sample in the carrier file such as S_t={fiveMSBs(S_t) + C_j }, where S is an audio sample belonging to carrier audio AUD, andt is the index of S. Furthermore, fiveMSBs(sample) is a function that returns the original five most significant bits of the audio sample in carrier audio AUD. The "+" operator concatenates the original five MSBs with 3 bits of a particular chunk from D, making the total number of bits in a given audio sample equals to 8 bits. In effect, the first 5 bits are the original five MSBs of the audio sample in AUD and the three LSBs are a particular chunk from the secret page in D.

4. The secret HTML page is now fully concealed inside the carrier audio AUD, the audio file itself is then hosted on a website called carrier website that the web publisher has created for the purpose of conveying his secret page.

5. The final output is a one medium page, namely the carrier website, made up of two components. The first component is the carrier audio file which embeds the secret page into its audio samples such as AUD={S_0,S_1,S_2,S_{t-1}}, where S is a carrier audio sample and t is the total number of carrier samples; while, the second component is the carrier website itself displaying the carrier audio file AUD in addition to other innocent-looking layout and text. The carrier website is then hosted on any public domain on the World Wide Web under any Top-Level Domain.

VI. EXPERIMENTS & RESULTS

For experimentation purposes, a simulation web browser is built using C#.Net and MS Visual Studio 2015 under the MS .Net Framework 4.5 [23]. The web browser implements the proposed method along with the steganography algorithm. In the experimentation, a secret web page is built using HTML containing undisclosed information about some treasure hunts in remote locations around the world. The page is meant to be part of the Deep Web as it is confidential and contains secret information. Moreover, an 8-bit WAV audio file is created to act as the carrier file in which the secret web page is concealed using steganography. The audio file is a musical extract from the famous song "Strangers in the Night" originally performed by Frank Sinatra in 1966. Likewise, another web page is built using HTML discussing the biography of the late American singer Frank Sinatra. It represents the innocent carrier web page in which the carrier audio file is embedded. Figure 2 depicts the HTML source code of the secret web page; while, Figure 3 depicts the carrier audio file in which the secret web page is concealed using steganography.



```
<!DOCTYPE html>

<html lang="en" xmlns="http://www.w3.org/1999/xhtml">
<head>
    <meta charset="utf-8" />
    <title></title>
</head>
<body>
    <table width="800">
        <tr>
            <td width="400px"><img src="images\map.jpg" /></td>
            <td>
                <h2>Treasure Hunt Details</h2>
                <p>
                In 1715, a fleet of 11 Spanish ships departed
                from Havana in Cuba filled with silver, gold, pearls
                and jewels. Looted by the Spanish conquistadors, the
                cargo of Inca and Aztec treasures are estimated to be
                worth about $2 billion by today's standards. Just six
                days into the voyage back to Spain, hurricanes sunk all
                the ships. Every bit of gold, silver and jewellery was lost.
                While many of the shipwrecks have been discovered off the
                eastern shores of Florida, the San Miguel ship - which experts
                believe contains most of the treasure - is still lost at sea
                despite professional treasure hunters numerous, expensive
                attempts to find it.
                </p>
            </td>
        </tr>
    </table>
</body>
</html>
```

Fig. 2.The HTML code of the Secret Web Page

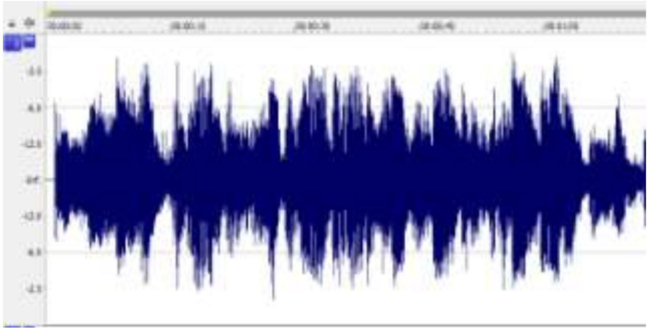
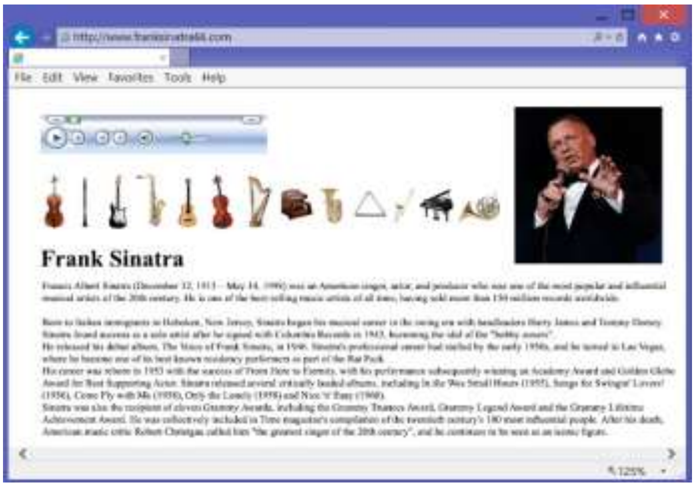


Fig. 3.The Carrier Audio File

The carrier web page along with the carrier audio file are hosted on a public domain on the Internet namely "franksinatra66.com"; thus making it exposed to search engines and part of the Surface Web. Figure 4 shows the carrier web page when accessed using a regular browser such as Internet Explorer.



http://www.franksinatra66.com
File Edit View Favorites Tools Help
125%

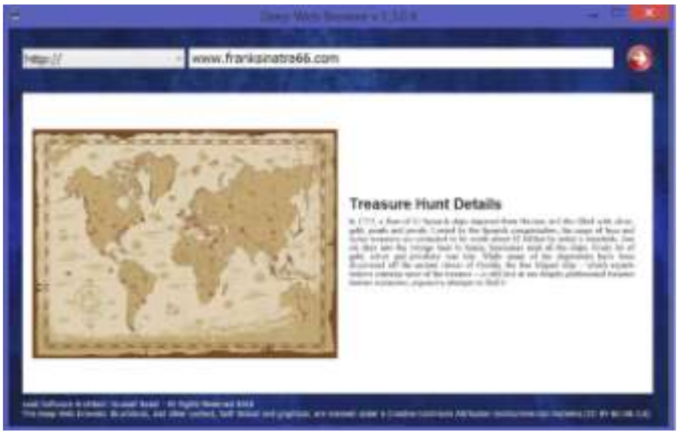
Frank Sinatra

Francis Albert Sinatra (December 12, 1915 – May 14, 1998) was an American singer, actor, and producer who was one of the most popular and influential musical artists of the 20th century. He is one of the best-selling music artists of all time, having sold more than 150 million records worldwide.

Born to Italian immigrants in Hoboken, New Jersey, Sinatra began his musical career in the swing era with bandleaders Harry James and Tommy Dorsey. Sinatra found success as a solo artist after he signed with Columbia Records in 1943, becoming the idol of the "bobby soxers".
He released his debut album, The Voice of Frank Sinatra, in 1946. Sinatra's professional career had stalled by the early 1950s, and he turned to Las Vegas, where he became one of its best known residency performers as part of the Rat Pack.
His career was reborn in 1953 with the success of From Here to Eternity, with his performance subsequently winning an Academy Award and Golden Globe Award for Best Supporting Actor. Sinatra released several critically lauded albums, including In the Wee Small Hours (1955), Songs for Swingin' Lovers! (1956), Come Fly with Me (1958), Only the Lonely (1958) and Nice 'n' Easy (1960).
Sinatra was also the recipient of eleven Grammy Awards, including the Grammy Trustees Award, Grammy Legend Award and the Grammy Lifetime Achievement Award. He was collectively included in Time magazine's compilation of the twentieth century's 100 most influential people. After his death, American music critic Robert Christgau called him "the greatest singer of the 20th century", and he continues to be seen as an iconic figure.

Fig. 4. IE rendering the Carrier page

Obviously, the Internet Explorer rendered the carrier web page normally and played the audio file normally as well. Interestingly, the content looks genuine and no evident audible artifacts can be detected. However, when the same domain "franksinatra66.com" is accessed using our proprietary browser, a different web page is displayed. It is actually the original secret web page that was covered using steganography in the audio clip. Figure 5 depicts the results.



http://

www.franksinatra66.com

Treasure Hunt Details

Fig. 5.Deep Web browser rendering the Carrier page

VII. CONCLUSIONS

This paper proposed an innovative method for building Deep Web network on the public World Wide Web using Steganography. In a nutshell, the method uses a steganography algorithm to hide secret web content into a carrier audio file that is hosted on a benign carrier website on the public domain. When using a regular browser, the benign carrier website displays and plays the carrier audio file. However, when a special proprietary browser is used, the secret web page is displayed. Experiments proved that the proposed method is plausible and can be implemented. Equally, results showed that the entire process is seamless and transparent as a particular web content can be simultaneously part of the Deep Web and part of the Surface Web while drawing no suspicions whatsoever regarding the existence of any secret data. Furthermore, as the proposed method uses HTTP and HTML standards in addition to the de-facto Internet protocols, it can be difficult to be detected, monitored, and restricted, thereby ensuring the anonymity of data published on the Deep Web.

VIII. FUTURE WORK

As future work, more types of web content are to be investigated and experimented including image files, video files, and digital streaming. Moreover, more advanced steganography algorithms are to be studied and developed in an attempt to provide a more robust, a more complicated, and a hard-to-break algorithm.

ACKNOWLEDGMENT

This research was funded by the Lebanese Association for Computational Sciences (LACSC), Beirut, Lebanon, under the “Deep Web Research Project – DWRP2019”.

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]. Peter Wayner, “Disappearing cryptography: information hiding: steganography & watermarking”, 3rd Edition, Morgan Kaufmann Publishers, 2009.

[7]. Youssef Bassil, "Steganography & the Art of Deception: A Comprehensive Survey", Int. J Latest Trends Computing, vol. 4, no. 3, 2013.

[8]. Youssef Bassil, "An Image Steganography Scheme using Randomized Algorithm and Context-Free Grammar", Journal of Advanced Computer Science and Technology, vol. 1, issue. 4, pp. 291-305, 2012.

[9]. K. Gopalan, “Audio steganography using bit modification”, Proceedings of International Conference on Multimedia, vol. 1, pp.629-632, 2003.

[10]. SajadShirali-Shahreza, Mohammad Shirali-Shahreza, “Steganography in Silence Intervals of Speech”, IEEE International Conference on Intelligent Information Hiding and Multimedia Signal Processing, 2008

[11]. Yunlu Wang, et al, "Steganalysis on positive and negative echo hiding based on skewness and kurtosis", 9th IEEE Conference on Industrial Electronics and Applications, 2014.

[12]. Yin-Cheng Qi, Liang Ye, Chong Liu, “Wavelet Domain Audio Steganalysis for Multiplicative Embedding Model”, Proceedings of the 2009 International Conference on Wavelet Analysis and Pattern Recognition, 2009.

[13]. F. Djebbar, B. Ayad, K. Abed-Meraim, H. Habib, “Unified phase and magnitude speech spectra data hiding algorithm”, Journal of Security and Communication Networks, John Wiley and Sons, 2012.

[14]. Khan, K., "Cryptology and the origins of spread spectrum", IEEE Spectrum, vol. 21, pp. 70-80, 1984.

[15]. N. Cvejic, T. Seppanen, "A wavelet domain LSB insertion algorithm for high capacity audio steganography", Proc. 10th IEEE Digital Signal Processing Workshop and 2nd Signal Processing Education Workshop, pp. 5355, 2002.

[16]. Dong, Y., Li, Q., "A deep Web crawling approach based on query harvest model", Journal of Computer Information System, vol. 8, no. 3, pp.973-981, 2012.

[17]. Michael Bergman, "The Deep Web: Surfacing Hidden Value", The Journal of Electronic Publishing (JEP), vol 7, no. 1, 2000.

[18]. Wood, Jessica, "The Darknet: A Digital Copyright Revolution", Richmond Journal of Law and Technology, vol. 16, no. 4, pp. 15–17, 2010.

[19]. "Tor Project: FAQ", www.torproject.org, retrieved 25 Dec 2018.

[20]. Oppliger, Rolf, "Privacy protection and anonymity services for the World Wide Web". Future Generation Computer Systems, vol. 16, no. 4, pp. 379–391, 2000.

[21]. M.G. Reed, P.F. Syverson, D.M. Goldschlag, "Anonymous connections and onion routing", IEEE Journal on Selected Areas in Communications, vol. 16, n. 4, 1998.

[22]. Ifeachor, Emmanuel C., and Jervis, Barrie W., "Digital Signal Processing: A Practical Approach", Pearson Education Limited, 2002.

[23]. Charles Petzold, "Programming Microsoft Windows with C#", Microsoft Press, 2002.

Youssef Bassil" Audio Steganography Method for Building the Deep Web" American Journal of Engineering Research (AJER), vol.8, no.05, 2019, pp.45-51