The Deep Web: Implementation using Steganography

Youssef Bassil

Abstract*: The Deep Web is about the web content that is invisible to public and not indexed by search engines. The purpose of the Deep Web is to ensure the privacy and anonymity of web publishers who want to remain anonymous and untraceable. A popular method to create a Deep Web is to host web content on a private network that is secret and restricted. Tor short for The Onion Router is a private Deep Web network that is accessible only by using a special web browser called the Tor browser. It uses special non-standard communication protocols to provide anonymity between its users and websites. Although the Tor network delivers exceptional capabilities in protecting the privacy of data and their publishers, the fact that it is free, open-source and accessible can raise suspicions that confidential, sometimes illicit data exist. Moreover, the Tor traffic can be easily blocked and its nodes blacklisted. This paper proposes an innovative method for building Deep Web networks on the public World Wide Web using Steganography. In a nutshell, the method uses a steganography algorithm to hide secret web content into a benign carrier image that is hosted on a carrier website on the public domain. When using a regular browser, the carrier website displays the benign carrier image. 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. Likewise, results showed that the entire process is seamless and transparent as a particular web content can simultaneously be part of the Deep Web and the Surface Web while drawing no suspicions whatsoever regarding the existence of any secret data. As future work, more advanced steganography algorithms are to be studied and developed in an attempt to provide an irreversible yet reliable algorithm.

***Index Terms*: Deep Web, Dark Net, Steganography, Tor**

I. INTRODUCTION

Since the dawn of the World Wide Web, web content publishers have expressed an increasing caution about their data privacy and anonymity. Basically, Data Anonymization refers to the non-attribution of data to their original creators whose intent is privacy protection [1]. Data anonymization ensures that content publishers remain untraceable from the data presented about them on the Internet. In practice, online data anonymization was accomplished by sharing web content in a way that it is only accessible by those who know how to reach it; thus creating an alternative underground secret web domain only available to specific people using specific techniques and software. The Deep Web is one of these secret playgrounds whose content cannot be reached by search engines using conventional bots and web spiders, but by using special techniques and tools [2]. Technically, several methods have been exploited to implement the Deep Web, one of which is to host web content on the public World Wide Web and bypass search engine crawlers using unlinked resources, dynamic content, encrypted pages, and password-protected resources. Another method is to host web content on a proprietary private network only accessible using special proprietary protocols and browsers. For instance, the Darknet which is one of the most infamous and widely known Deep Web is an anonymous space describing the portion of the Internet purposefully not open to public view and not indexed by regular search engines such as Google and Bing [3]. The Darknet, which is part of the Deep Web [4], is implemented using non-HTML, non-HTTP languages possibly encrypted using proprietary protocols and can only be accessed using special software and browsers. Although, the Darknet sounds very secretive, the truth is not as it seems. In fact, the Darknet is not completely transparent as it is free, open-source, and available to anyone via special browsers. As a result, suspicious activities and illicit content can easily be observed and tracked down by governments. For instance, in October 2013, the Federal Bureau of Investigation (FBI) shut down many websites on the Darknet network and arrested many content publishers who were operating illegal businesses [5]. Furthermore, since the Darknet uses specific network ports and IP addresses, they can be easily blocked by network firewalls and other trivial security measures. This paper proposes a new method to implement the Deep Web using 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 image file called carrier file and then sent to the receiver. On the other side, the receiver deciphers the image 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 image 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 bitmap images 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 made up of bitmap images; 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 bitmap images 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.

Revised Version Manuscript Received on 25 January, 2019.

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

II. THE DEEP WEB

The Deep Web consists of web data content that are invisible to public and not indexed by search engines [7]. 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) [8]. 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 in Table 1 whose ultimate goal is to hide web resources from the swarm of search engines.

Method	How to Access
Unlinked content i.e. pages which are not linked to by other pages	Can be accessed using the full absolute URL - FQDM
Dynamically generated web content	Can be accessed by searching for something or submitting a query
Password protected web content using for instance HTTP Basic Access Authentication	Can be accessed by obtaining the right credentials
Private content hosted on private networks behind Firewall such as Intranet	Can be accessed by obtaining access rights allowing traffic in and out of the network
Web content built using Non-HTTP, Non-HTML, and Non-Web Standard protocols and ports	Can be accessed using special proprietary software, e.g. Tor

Table 1 – Methods to Implement a Deep Web

III. 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 [9]. 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 [10]. 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 [11].

IV. PITFALLS OF USING THE TOR NETWORK

Tor delivers unprecedented capabilities in protecting the privacy of data being published on its network while ensuring the total anonymity of their owners [12]. 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 [5]. 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 bitmap images that are 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 renders the benign bitmap images; 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 images 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. 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 [13]. 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, digital 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 itself. Essentially, digital steganography is governed by five key elements. They are as follows [14]:

1. *Covert Data*: Often known as the payload and refers to the overt data that need to be covertly communicated or stored. The covert data can be anything convertible to binary format, from simple text messages to executable files.
2. *Carrier File*: It is basically a file into which the covert data are concealed. The carrier file can be any computer-readable file such as image, audio, video, or text file.
3. *Stego File*: Sometimes called package, it is the resulting file which has the covert data embedded into it.
4. *Carrier Channel*: It denotes the file type of the carrier, for instance, BMP, JPG, MP3, PDF, etc.
5. *Capacity*: It denotes the amount of data that the carrier file can hide without being distorted.

VI. DESIGN SPECIFICATIONS

The proposed method employs a Steganography algorithm that is designed to work on 24-bit True Color uncompressed digital images such as BMP. Basically, the pixels of a 24-bit BMP image are each composed of three 8-bit color channels, namely R, G, and B channels [15]. The proposed algorithm conceals the secret data sequentially into the three LSBs (Least Significant Bit) of each of these color channels; thus, the hiding capacity is equal to 9 bits out of 24 bits or 37% of the total size of the carrier image (9/24=0.375=37%). The carrier image is manipulated as a matrix of a finite set of pixels each composed of a 24-bit color value divided into three 8-bit chunk color channels. 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 color channel in the carrier image. Figure 1 depicts the design behind the steganography algorithm used by our proposed method.

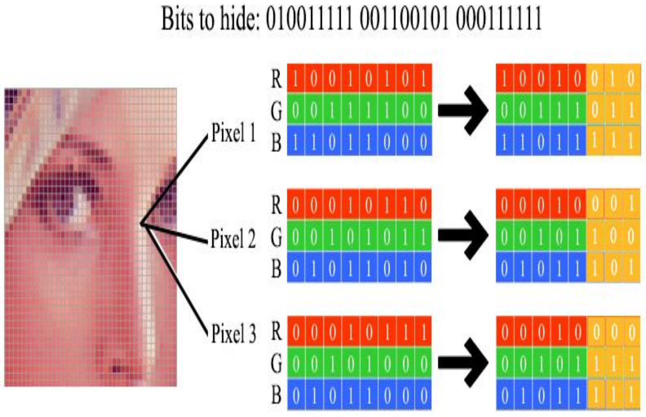

Bits to hide: 010011111 001100101 000111111
Pixel 1
R 1 0 0 1 0 1 0 1 → 1 0 0 1 0 0 1 0
G 0 0 1 1 1 1 0 0 → 0 0 1 1 1 0 1 1
B 1 1 0 1 1 0 0 0 → 1 1 0 1 1 1 1 1
Pixel 2
R 0 0 0 1 0 1 1 0 → 0 0 0 1 0 0 0 1
G 0 0 1 0 1 0 1 1 → 0 0 1 0 1 1 0 0
B 0 1 0 1 1 0 1 0 → 0 1 0 1 1 1 0 1
Pixel 3
R 0 0 0 1 0 1 1 1 → 0 0 0 1 0 0 0 0
G 0 0 1 0 1 0 0 0 → 0 0 1 0 1 1 1 1
B 0 1 0 1 1 0 0 0 → 0 1 0 1 1 1 1 1

Figure 1 – The Steganography Algorithm

Eventually, when the secret HTML page is completely concealed inside the carrier image, the image 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 image 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 image (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 displaying the benign carrier image along with 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 image.

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. A 24-bit color BMP carrier image denoted by IMG 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 color channel of every pixel of the carrier image IMG.

a. As the carrier image is a 24-bit colored image, every pixel would have three color channels R, G, B, each of length 8 bits. For this reason, every chunk C_j is stored in the three LSBs of each of the color channels of every selected carrier pixel such as P_t={ *fiveMSBs*(R_t) + C_j ; *fiveMSBs*(G_t) + C_{j+1} ; *fiveMSBs*(B_t) + C_{j+2} }, where P is a pixel belonging to carrier image IMG, t is the index of P, and R, G, and B are the three Red-Green-Blue color channels of pixel P_t. Furthermore, *fiveMSBs*(*channel*) is a function that returns the original five most significant bits of the color channel in carrier image IMG. 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 color channel equals to 8 bits. In effect, the first 5 bits are the original five MSBs of the color channel in IMG and the three LSBs are a particular chunk from the secret page in D.

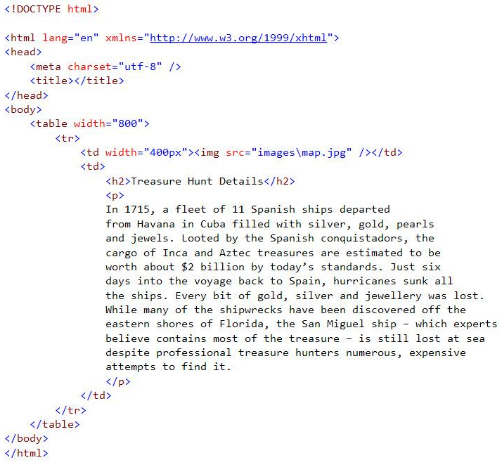
4. The secret HTML page is now fully concealed inside the carrier image IMG, the image itself is then hosted on a website called carrier website that the web publisher has created for the purpose of conveying his secret page.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

5. The final output is a one medium page, namely the carrier website, made up of two components. The first component is the carrier image which embeds the secret page into its pixels such as IMG={P_0,P_1,P_2,P_{t-1}}, where P is a carrier pixel and *t* is the total number of carrier pixels; while, the second component is the carrier website itself displaying the carrier image IMG 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.

VII. 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 [16]. 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, a bitmap image of "Leonardo Da Vinci" is created to act as the carrier image in which the secret web page is concealed using steganography. Likewise, another web page is built using HTML. It represents the innocent carrier web page in which the carrier image is embedded. Figure 2 depicts the HTML source code of the secret web page; while, Figure 3 depicts the carrier image of Leonardo Da Vinci 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>
```

Figure 2 – The HTML code of the Secret Web Page

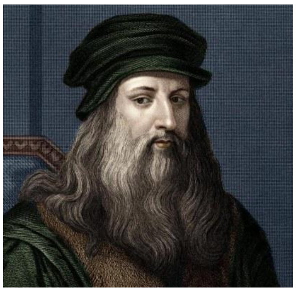
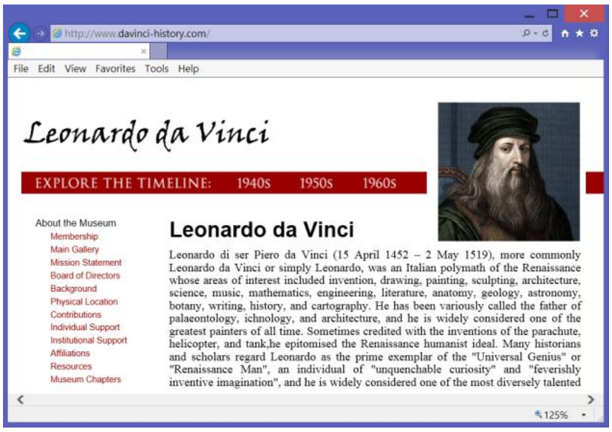


Figure 3 – The Carrier Image

The carrier web page along with the carrier image are hosted on a public domain on the Internet namely "davinci-history.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.

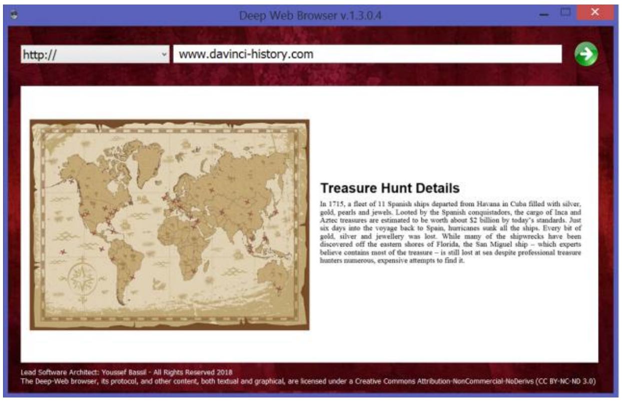


Leonardo da Vinci

Leonardo di ser Piero da Vinci (15 April 1452 – 2 May 1519), more commonly Leonardo da Vinci or simply Leonardo, was an Italian polymath of the Renaissance whose areas of interest included invention, drawing, painting, sculpting, architecture, science, music, mathematics, engineering, literature, anatomy, geology, astronomy, botany, writing, history, and cartography. He has been variously called the father of palaeontology, ichnology, and architecture, and he is widely considered one of the greatest painters of all time. Sometimes credited with the inventions of the parachute, helicopter, and tank,he epitomised the Renaissance humanist ideal. Many historians and scholars regard Leonardo as the prime exemplar of the "Universal Genius" or "Renaissance Man", an individual of "unquenchable curiosity" and "feverishly inventive imagination", and he is widely considered one of the most diversely talented

Figure 4 – IE rendering the Carrier page

Obviously, the Internet Explorer rendered the carrier web page normally. Interestingly, the content looks genuine and no evident artifacts can be detected. However, when the same domain "davinci-history.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 image of Leonardo Da Vinci. Figure 5 depicts the results.



Treasure Hunt Details

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.

Lead Software Architect: Youssef Bassil - All Rights Reserved 2018
The Deep-Web browser, its protocol, and other content, both textual and graphical, are licensed under a Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND 3.0)

Figure 5 –Deep Web browser rendering the Carrier page

VIII. 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 image file that is hosted on a benign carrier website on the public domain. When using a regular browser, the benign carrier website displays the carrier image. 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. Likewise, 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.

IX. FUTURE WORK

As future work, more types of web content are to be investigated and experimented including audio 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. Bin Zhou, Jian Pei, WoShun Luk, "A brief survey on anonymization techniques for privacy preserving publishing of social network data", ACM SIGKDD Explorations Newsletter, vol. 10, no. 2, pp. 12–22, 2008.
2. 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.
3. Gayard, Laurent, "Darknet: Geopolitics and Uses. Hoboken, NJ: John Wiley & Sons", 2018, ISBN 9781786302021.
4. Senker, Cath, "Cybercrime & the Dark Net: Revealing the hidden underworld of the internet", London: Arcturus Publishing, 2016, ISBN 9781784285555
5. "Sealed Complaint 13 MAG 2328: United States of America v. Ross William Ulbricht", p. 6, retrieved January 2014.
6. Peter Wayner, “Disappearing cryptography: information hiding: steganography & watermarking”, 3rd Edition, Morgan Kaufmann Publishers, 2009.
7. 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.
8. Michael Bergman, "The Deep Web: Surfacing Hidden Value", The Journal of Electronic Publishing (JEP), vol 7, no. 1, 2000.
9. Wood, Jessica, "The Darknet: A Digital Copyright Revolution", Richmond Journal of Law and Technology, vol. 16, no. 4, pp. 15–17, 2010.
10. "Tor Project: FAQ", www.torproject.org, retrieved 25 Dec 2018.
11. Oppliger, Rolf, "Privacy protection and anonymity services for the World Wide Web". Future Generation Computer Systems, vol. 16, no. 4, pp. 379–391, 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, n. 4, 1998.
13. Youssef Bassil, "Steganography & the Art of Deception: A Comprehensive Survey", Int. J Latest Trends Computing, vol. 4, no. 3, 2013.
14. 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.
15. A. K. Jain, “Fundamentals of Digital Image Processing”, Englewood Cliffs publications, Prentice-Hall, 1989.
16. Charles Petzold, “Programming Microsoft Windows with C#”, Microsoft Press, 2002.