Edustega: an Education-Centric Steganography methodology

Abdelrahman Desoky

Department of Computer Science and Electrical Engineering,
University of Maryland,
Baltimore County, USA
E-mail: abd1@umbc.edu

Abstract: This paper presents a novel Education-Centric Steganography Methodology (Edustega) that takes advantages of such text to conceal data. Edustega is based on Nostega paradigm, which implies that it neither hides data in a noise (errors) nor produces noise. Such materials, e.g., questions, answers, exams, examples, puzzles, competitions, etc., have an adequate room for concealing data where the observed bitrate by the experimental results of the current implementation is superior to all contemporary linguistic steganography approaches. Edustega can be applied to all languages. The presented implementation, validation, and experimental results confirmed that Edustega methodology is capable of achieving the steganographical goal.

Keywords: steganography; linguistic steganograpy; information security.

Reference to this paper should be made as follows: Desoky, A. (2011) 'Edustega: an Education-Centric Steganography methodology', *Int. J. Security and Networks*, Vol. 6, Nos. 2/3, pp.153–173.

Biographical notes: Abdelrahman Desoky is a Scientist and Computer Engineering Doctorate with over 18 years experience in the computer field. He is an experienced educator at both the graduate and undergraduate level. He is currently a CEO of The Academia Planet and an independent consultant, researchers, and instructor for both academia and practice sectors. He is an author and of Security book entitled "Noiseless Stenography: The key of Covert Communications". He received a PhD from the University of Maryland and a MSc from the George Washington University; both degrees are in Computer Engineering.

1 Introduction

Steganography is the science and art of concealing the existence of covert communications. In cryptography, the goal is to hinder the adversary from decoding a hidden message, which is the ciphertext. On the other hand, the steganographic goal is to prevent an adversary from suspecting the presence of covert communications (Desoky, 2012). If suspicion is raised when using any steganographic technique, the goal of steganography is defeated regardless of whether or not a plaintext is revealed (Kessler, 2004; Martin et al., 2005). The contemporary approaches hide data as noise in a cover that is assumed to look innocent. For instance, a message can be embedded by altering a digital media such as an image, text, graph, audio file, etc. (Martin et al., 2005; Desoky and Younis, 2008; Petitcolas, 1999). Detecting such noise can easily raise suspicion, which obviously defeats the steganographic goal (Petitcolas, 1999; Bennett, 2004). This is the major steganographic problem, which is still the persisting.

This paper introduces a promising novel technique, Education-Centric Steganography Methodology (Edustega). Edustega manipulates the popular educational documents, in academic and nonacademic environments, to camouflage both a message and its transmittal. Basically, Edustega exploits, mainly but not limited to, questions and answers to conceal data. For instance, the answer of multiple-choice, true-or-false, fill-in-the-space and matching questions can be the means to camouflage the data. These questions and answers can be for an exam, set of examples, puzzles, competitions, etc. The questions and answers can be fabricated in order to embed data without generating any type of suspicious pattern. To illustrate, true-or-false questions may conceal data in the sequence of answers with 'true' and 'false' indicating the binary bit '1' and '0' respectively. In addition, the answers of multiple-choice questions can conceal data where a choice, e.g., 'A', is mapped to a specific binary strings such '00' and so on. The steganographic code of a message can then be formed by concatenating the binary strings of the marked answers. Moreover, wrong answers can also be exploited to conceal data. For example, one can answer refer to water in a chemistry exam as 'N_2O' instead of 'H_2O'.

The main advantages of Edustega are as follows. First, the high demand for educational documents by a wide variety of people, in both the academic and nonacademic spheres, creates a high volume of traffic and averts suspicion in the presence of covert communication channels. Second, Edustega does not imply a particular

pattern (noise) that an adversary may look for. Third, the concealment process of Edustega has no effect on the linguistics of the generated cover (edu-cover). Therefore, an edu-cover is linguistically legitimate and is thus capable of passing both computer and human examinations. Fourth, Edustega can be applied to all languages. Fifth, as demonstrated later in the paper, educational materials have plenty of room for concealing data. The observed bitrate in the current implementation experiments is roughly 0.94–3.86%, which is superior to all contemporary linguistic steganography approaches found in the literature. Sixth, Edustega is resilient to popular attacks and the hidden message is anti-distortion. Since the reuse and alteration of educational documents are a common practice in academic, e.g., lectures, homework, exams, tests, quizzes, examples, and nonacademic setups, e.g., puzzles, competitions, etc., an edu-cover can pass comparison attacks. The implementation and steganalysis validation demonstrate that Edustega methodology is capable of achieving the steganographical goal.

The remainder of this paper is organised as follows. Section 2 discusses the related work and compares Edustega to the linguistic steganography techniques found in the literature. Section 3 explains the Edustega methodology in detail. Section 4 demonstrates the Edustega implementation. Section 5 presents the steganalysis validation of Edustega. Finally, Section 6 concludes the paper.

2 Related work

The aim of this section is to show the peculiarity of Edustega from the previous work on linguistic and nonlinguistic steganography. Contemporary steganographic approaches are often classified based on the steganographic cover type into image, audio, graph, or text (Desoky and Younis, 2008; Petitcolas, 1999). However, when linguistics is employed for hiding data, an approach is usually categorised as linguistic steganography (linguistic cover) to distinguish it from other steganographic technique types, e.g. image, audio, etc. Therefore, in this paper contemporary approaches are categorised to on linguistic (Section 2.1) and nonlinguistic steganography (Section 2.2) as follows.

2.1 Linguistic steganography

Linguistic steganography approaches conceal data in a linguistic-based textual cover. Linguistic steganography approaches can be categorised as follows: Series of Characters and Words, Statistical Based, Word Replacement, Noise Based, and Nostega-Based. *Series of Characters and Words* approach is also known as null-cipher (Kahn, 1996), which was used by the Germans during World War I (WWI). Null-cipher is a predetermined protocol of character and word sequence that is read according to a set of rules such as: read every seventh word or read every ninth character in a message. Apparently, suspicion is raised because the user is forced to fabricate a text-cover according to a predetermined protocol, which may introduce some peculiarity in the text that draws suspicion and defeats the steganographical goal. In addition, applying a brute force attack may reveal the entire message.

Statistical Based technique is known as mimic functions approach (Wayner, 1992, 2002). Mimic functions, as the name suggests, attempts to imitate the statistical profile of normal text. It employs the inverse of the Huffman Code by inputting a data stream of randomly distributed bits to produce text that obeys the statistical profile of a particular normal text. Therefore, the generated text by mimic functions is resilient against statistical attacks. In addition, mimic functions can employ the concept of both Context Free Grammars (CFG) and van Wijnaarden grammars to enhance the output. It is known that the output of regular mimic functions is gibberish rendering it extremely suspicious (Petitcolas, 1999; Bennett, 2004). However, the combination of mimic functions and CFG slightly improved the readability of the text (Wayner, 1992, 2002). Yet, the text-cover still contains numerous flaws such as incorrect syntax, lexicon, rhetoric, and grammar. Furthermore, the content of the text-cover is often meaningless and semantically incoherent. These shortcomings may raise suspicion in covert communications.

Word Replacement approach, is called NICETEXT or synonyms-based that, uses a big dictionary (Chapman and Davida, 1997, 2002; Chapman et al., 2001). NICETEXT employs a piece of text to manipulate the process of embedding a data in a form of synonym substitutions. This process preserves the meaning of text-cover as its original text that contains no hidden message every time it is used. The synonyms-based approach attracted the attention of numerous researchers in the last decade: Winstein (1999, 2008), Bolshakov et al. (2004), Bolshakov and Gelbukh, (2004), Calvo and Bolshakov (2004), Chand and Orgun (2006), Nakagawa et al. (2001), Niimi et al. (2003), Bergmair and Katzenbeisser (2004, 2007), Bergmair (2008), Topkara et al. (2006), Murphy and Vogel (2007) and Atallah et al. (2001, 2002). Although the text-cover of synonym-based approach may look legitimate from a linguistics point of view given the adequate accuracy of the chosen synonyms, reusing the same piece of text to hide a message is a steganographical concern. If an adversary intercepts the communications and oversees the same piece of text that has the same meaning over and over again with just different group of synonyms between communicating parties, he will question such use.

Noise Based technique is simply hides data in the linguistic errors (noise). There are few approaches that are Noise Based, which are as follows: Translation-based scheme, Confusing Approach, SMS-based scheme. The following details these approaches. Translation-based steganographic scheme (Grothoff et al., 2005; Stutsman et al., 2006) hides a message in errors (noise) that are encountered in a Machine Translation (MT). It embeds a message by substituting translated text using textual translation variations of multiple MT systems. In addition, it inserts popular errors of MT systems and also uses

synonym substitutions in order to increase the bitrate. The amount of linguistic flaws in noise-based approach is a concern because if the errors appear excessively, the steganographical goal will be defeated. In addition, Grothoff et al. stated that one of the concerns is that the continual improvement of machine translation may narrow the margin of hiding data (Grothoff et al., 2005; Stutsman et al., 2006). Conversely, an improvement in educational document generators, e.g., exam generators, is in fact beneficial to Edustega as demonstrated later in Sections 3 and 4. Furthermore, translation-based approach, as confirmed by Grothoff et al., cannot be used with all languages due to huge differences in the essential linguistics structures, e.g., English and Chines, English and Arabic, etc. This generates severely incoherent and unreadable text (Grothoff et al., 2005; Stutsman et al., 2006). Inversely, Edustega can be applied to all known languages without any exceptions while the generated edu-cover is scientifically and linguistically legitimate.

Confusing approach is a noise-based approach that employs typos and abbreviations in text of e-mails, blogs, forums, and any other similar type of noisy text in order to hide messages (Topkara et al., 2007). Finally, Shirali-Shahreza et al. (2007) presented an abbreviation-based scheme, which is known also as SMS-based approach that camouflages messages using short message service (SMS) of mobile phones. The SMS-based steganography approach takes advantages of SMS' size constraints and its use of phone keypad instead of the keyboard for hiding data in a noisy text. Nonetheless, these techniques are sensitive to the amount of noise (errors) that occurs in a human writing. Such shortcoming not only increases the vulnerability of the approach but also narrows the margin of hiding data. Conversely, Edustega neither employs errors nor uses noisy text to conceal data.

Recently, a novel paradigm in steganography research, namely Noiseless Steganography Paradigm (Nostega) has been introduced (Desoky, 2008, 2009a), in which a message is hidden in a cover as data rather than noise. A group of methodologies have been developed based on the Nostega paradigm. First one of these methodologies is the Summarisation-Based Steganography Methodology (Sumstega) (Desoky et al., 2008). Sumstega exploits automatic summarisation techniques to camouflage data in the auto-generated summary-cover (text-cover) that looks like an ordinary and legitimate summary. The second linguistic steganographic scheme that is also based on Nostega paradigm is the List-Based Steganography Methodology (Listega) (Desoky, 2009b). Listega manipulates itemised data to conceal messages in a form of textual list. The third linguistic steganography methodology, Notes-Based Steganography Methodology (Notestega) (Desoky, 2009c) that takes advantage of the recent advances in automatic notetaking techniques to generate a text-cover. Notestega pursues the variations among both human notes and the outputs of automatic notetaking techniques to conceal data. The fourth linguistic steganography methodology, Mature Linguistic Steganography Methodology (Matlist) (Desoky, 2011) employs random series of a domain specific subject along with NLG and template techniques to generate a text-cover that is naturally has a different legitimate meaning for concealing different messages while it remains semantically coherent and rhetorically sound. The fifth linguistic steganography methodology, Unlike all other approaches, the Normal Linguistic Steganography Methodology (NORMALS) neither generates noise nor uses noisy text to camouflage data. NORMALS employs Natural Language Generation (NLG) techniques to generate noiseless (flawless) and legitimate text-cover by manipulating the inputs' parameters of NLG system in order to camouflage data in the generated text (Desoky, 2010a, 2012). As a result, NORMALS is capable of fooling both human and machine examinations. Unlike Matlist, NORMALS is capable of handling non-random series domains. The implementation, validation, and experimental results demonstrate that these methodologies are capable of achieving the steganographical goal.

It is worth noting that the presented Edustega methodology in this paper follows this new paradigm. However, it is unlike all other techniques, Edustega exploits educational documents such questions and answers of exams, examples, puzzles, competitions to camouflage data noiselessly.

2.2 Nonlinguistic steganography

Nonlinguistic steganography approaches can be categorised based on its file type such as text, image, audio, and graph. Textual steganography, which is based on nonlinguistic techniques, hides data by Textual Format Manipulation (TFM) (Petitcolas, 1999). TFM modifies an original text by employing spaces, misspellings, fonts, font size, font style, colours, and non-colour (as invisible ink) to embed an encoded message. However, comparing the original text to the modified text triggers suspicion and enables an adversary to detect where a message is hidden. In addition, TFM can be distorted and may be discerned by human eyes or detected by a computer (Petitcolas, 1999; Bennett, 2004).

On the other hand, image steganography is based on manipulating digital images to conceal a message. Such manipulation often renders the message as noise. In general, image steganography suffers from several issues such as the potential of distortion, the significant size limitation of the messages that can be embedded, and the increased vulnerability to detection through digital image processing techniques (Martin et al., 2005). Audio-covers have also been pursued. Example of audio steganography techniques include LSB (Desoky, 2010b; Cvejic and Seppanen, 2004), spread spectrum coding (Bender et al., 1996; Kirovski and Malvar, 2001), phase coding (Bender et al., 1996; Ansari et al., 2004), and echo hiding (Ansari et al., 2004; Gruhl et al., 1996). In general, these techniques are too complex, and like their image-based counterpart, are still subject to distortion and are vulnerable to detection

(Desoky, 2012; Martin et al., 2005; Petitcolas, 1999; Cvejic and Seppanen, 2004). The hidden message may become to a great extent a foreign body in the cover and thus makes those schemes vulnerable to detection. In addition, contemporary image/audio steganography schemes rely on private or restricted access to the original unaltered cover in order to avoid the potential of comparison attacks, which is considered a major threat to the covert communication. Basically, an adversary can detect the presence of a hidden message by comparing a particular image-cover or audio-cover to the original image or audio file and finding out that some alterations have been made.

Hiding information in an unused or reserved space in computer systems (Anderson et al., 1998; http://www.scramdisk.clara.net). For example, Windows 95 operating system has around 31 KB of unused hidden space, which can be used to hide data. Another example, unused space in file headers of image, audio, etc., can also be used to hide data. This depends on the size of the hard drive used. TCP/IP packets used to transport information across the Internet have unused space in the packet headers (Handel and Sandford, 1996). The TCP packet header has six unused (reserved) bits and the IP packet header has two reserved bits. There are tremendous packets transmitted over the Internet can convey and transmit a secret data. However, these techniques are vulnerable to distortion attacks (Desoky, 2012; Kessler, 2004; Petitcolas, 1999).

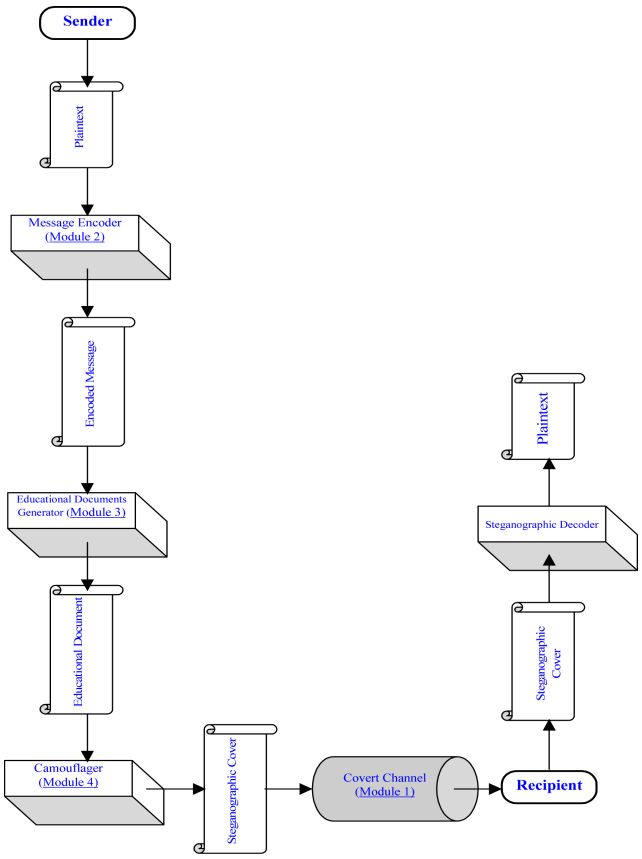
Recently, a Graph Steganography (Graphstega) methodology has been developed (Desoky and Younis, 2008). Unlike all other schemes, the message is naturally embedded in the cover by simply generating the cover based on the message. Graphstega camouflages a message as data points in a graph and thus the message would not be detectable as noise. The approach is shown to be resilient to a wide range of attacks, including a comparison attack by untraceable or authenticated data. Similarly, Chestega (Desoky and Younis, 2009) exploits popular games, like chess, checkers, crosswords, domino, etc., for concealing messages in an unaltered authenticated data. Graphstega and Chestega represent a new paradigm in steganography research in which the message is hidden in the cover as data rather than noise. Edustega follows this new paradigm by exploiting educational documents of the academic and nonacademic spheres to camouflage data by manipulating, mainly but not limited to, questions and answers in order to embed data without generating any suspicious pattern.

3 Edustega methodology

Educational documents are very popular and widely used all over the world by people of all ages. The excessive necessity for educational documents by a wide variety of people is not only by the academic communities but also by nonacademic spheres. This creates a high volume of traffic, which makes an adversary's job impractical to investigate all of them. Such huge traffic, due to the normal frequent exchange of educational documents in both electronic and printed formats, allows communicating parties to establish a covert channel to covertly transmit hidden messages. Yet, the text of educational documents not only is capable of concealing messages but also it has an adequate room for concealing data, rendering edu-cover to retain superior bitrate to all contemporary linguistic steganography approaches, as will be shown later. It is also can be applied to all languages. Consequently, the text of educational documents is an attractive steganographic carrier. Therefore, the novel Education-Centric Steganography Methodology (Edustega), which is presented in this paper, takes advantages of such text to securely conceal data. Edustega is based on Nostega paradigm (Desoky, 2008, 2009a), which implies that it neither hides data in a noise (errors) nor produces noise. Instead, it camouflages data in the text of educational documents by manipulating, mainly but not limited to, questions and answers of (e.g. multiple-choice, true-or-false, fill-in-the-space, matching, etc.) exams, examples, puzzles, competitions, etc., in order to embed data without generating any suspicious pattern.

The fundamental algorithm of Edustega system is consisted of four modules: the *Covert Channel*, *Encoder, Educational Documents Generator (EDG)*, and *Camouflager*. These modules are ultimate goal is to define an Edustega system configuration for the communicating parties to use. The four modules are highlighted as follows. First, the communicating parties opt to establish a *Covert Channel* (Module 1) which is the means for hidden delivery of steganographic carrier. Module 1 is only involved in the stage of constructing the configuration of Edustega system. Obviously, this module determines an appropriate educational topic(s) to be the context of the steganographic text-cover for achieving the steganographical goal. Second, *Steganographic Code Generator (SCG Module 2)* encodes a message in an appropriate form for the camouflaging process (Module 4). Encoding a message may be converted into binary representation of its ASCII code and slicing it to a particular length of digits e.g., 3, 4, 5, etc. For instance, a message after encoding it may look like the following: 00011, 10101, 00000, 11110, 11111, and so on. Third, *Educational Documents Generator* (EDG Module 3) that generates an original text, which the text that contains no hidden data because it is prior to embedding process. Fourth, the *Camouflager* (Module 4) embeds the steganographic code (encoded data), that is generated by the Steganographic Code Generator (SCG Module 2), in the generated text by the EDG (Module 3), in order to conceal a message. The output of this module is in a form of legitimate educational documents such as, mainly but not limited to, questions and answers. These modules are elaborated in the following sections.

Figure 1 An illustration of the interaction of the various Edustega modules and how the outputs of the individual modules are used for covert communications between two parties (see online version for colours)



3.1 Establishing covert channel (Module 1)

Edustega naturally camouflages the delivery of a hidden message in a way that makes it appear legitimate and innocent. To employ Edustega, the communicating parties first need to define and agree on the basic configuration of the covert channel. This step includes determining the following:

1 the topic of the educational documents that will be used as a cover

2 how the cover will be delivered from a sender to the recipient

3 how their interaction will be justified.

Selecting a suitable topic(s) can play an essential role for securing the steganographic communications by establishing an appropriate covert channel for delivering a hidden message. The chosen topic(s) must facilitate the process of embedding data without generating noise in order to achieve the steganographical goal. Since Edustega mainly, but not limited to, manipulates questions and answers to camouflage messages, any topic that allows the employing of questions and answers, such as examples, exercises, puzzles, tests, homework, etc., can be used. Although academic subjects such as mathematics and science are obvious choices, numerous nonacademic options can be pursued as well. Examples of nonacademic topics include training courses in industry, puzzle-based entertainment programs, competition, etc. In addition, the picked topic has to fit the communicating parties and provide some ground for justifying the communications, as elaborated below.

The second important configuration parameter is how the cover will be delivered to the recipient without raising suspicion. Covert transmittal of the steganographic cover is very crucial to the success of steganography. The fact that Edustega employs noiseless-based means for hiding data, enables great flexibility in delivering the steganographic cover to its recipient. Options may include web post and download, e-mail transmission, mailing hardcopies, specialised publications, TV broadcasting (http://en.wikipedia.org/wiki/Who_Wants_To_Be_A_Millionaire%3F), handed CD-ROM, videocassette, DVD, etc. Since the sender may mix an edu-cover among other legitimate documents, obviously, the basic configuration of the covert channel should include how a recipient can only decode the right covers. For instance, the communicating parties may agree on putting edu-covers among others similar documents by designating a particular sequence, such as odd number, even number, every other 3, etc., by placing edu-covers in a specific folder, or by specifying certain document contents such as homework.

The core of covert channels is how to prevent the association between a sender and recipient from drawing suspicion and to render it innocent communications. For example, exchanging e-mails would automatically imply a relationship between the communicating parties. Similarly, downloading files from a website indicates an interest in the accessed material. Due to the advances in monitoring tools for network and internet traffic, profiles of user's access pattern can be easily established. An adversary most probably will suspect the presence of a hidden message, even if the content does not look suspicious, because of the observed traffic pattern and the lack of a justification for the interest in the contents of the transmitted materials. For example, if a profession for one of the communicating party is an elementary English teacher and yet he sends or receives college level chemistry exams, then suspicion will likely be raised. Therefore, it is very important to rationalise the exchange of steganographic cover in order to avoid attracting any attention that may trigger an attack. The communicating parties need to agree on how to justify their interest in the education documents of the selected topic. This may include defining a role, such as mentoring or tutoring that a sender plays, a profession, or simple an interest that justify a peer relationship.

3.2 Steganographic Code Generator (SCG Module 2)

Implementing the Steganographic Code Generator (SCG Module 2), the message encoder, follows a two-steps process: first, determining the encoding parameters in the topic picked by Module 1; second, defining a

steganographic coding based on these parameters. A parameter in this context means some aspects of educational document(s) that can be referred to steganographical values throughout an edu-cover. For educational documents, the order and style of questions as well as their answers can be exploited for concealing data. The definition of the steganographic code would depend on the selected parameters. For example, encoding a message using multiple-choice questions is different from encoding it using the order in which the various question styles appear and so on. The coding module of Edustega exploits these options and determines the parameter(s) that will be employed for concealing data. The selection criteria may be driven by the size of the message to be concealed, the popular question styles for the selected topic, and the availability of authenticated data (e.g., answers) that would match the encoded message. Concealing long messages is generally a challenge for most known steganography approaches. Edustega can hide long messages by simply employing more questions in an edu-cover or splitting the message over multiple documents, e.g., multiple homework, assignments, etc. Nonetheless, the popularity of certain question styles is an important factor in the selection since excessive appearance of certain style may draw suspicion. For example, having an exam that is mostly consists of true-or-false questions is not a usual practice in some disciplines. Finally, one would argue that the encoding parameters may actually influence the selection of a topic for the covert communication and should be done first. While this is a valid concern, the topic selection is crucial for justifying the interaction among the communicating parties and is thus more affected by the criteria for establishing a covert channel.

Edustega does not impose any constraint on the message encoding scheme (SCG Module 2), which is discussed in this section, as long as it generates a set of data values that can be embedded in an edu-cover. Given the availability of numerous encoding techniques in the literature that fit (Desoky, 2009a, 2010a, 2012), the balance of this section will focus on an example that will be used in Section 4 to demonstrate the applicability of Edustega. In the example, the encoding is done as follows. A message is first converted to a binary string. The string can be a binary of cipher text or a compressed representation. The binary string is then partitioned into groups of m bits. The value of m is determined based on the encoding parameters that Edustega exploits. For instance, if a message will be concealed in the answer of true-or-false questions, the value of m is 1 since each answer can conceal only 1 bit. On the other hand, if the edu-cover will be in a form of multiple-choice questions with four possible answers, A, B, C, and D, the binary message is partitioned it into groups of two bits, e.g., 00, 01, 10, and 11, corresponding to the possible choices. Again, this encoding scheme is just for illustration and many alternatives and more sophisticated schemes can be employed, as demonstrated in Section 4.

3.3 Educational documents generator and message camouflager (Modules 3 and 4)

The aim of this section is to discuss both the *Educational Documents Generator* (EDG Module 3) and the *Camouflager* (Module 4). The reason that these modules (Modules 3 and 4) are described in the same section is because the fact that both modules are highly interrelated to each other. The *Educational Documents Generator* (EDG Module 3) produces an original text that contains no hidden messages. Then, the *Camouflager* (Module 4) embeds the steganographic code (encoded message), that is generated by the Steganographic Code Generator (SCG Module 2), in the generated text by the EDG (Module 3), in order to conceal a message. The output of this procedure is in a form of legitimate educational documents such as, mainly but not limited to, questions and answers of multiple-choice, true-or-false, fill-in-the-space, matching, etc. Such text can be in a form of exams, examples, puzzles, competitions, etc., in order to embed data without generating any suspicious pattern.

From a steganography point of view, reusing or altering an existing text to hide data is not a recommended practice since an adversary can reference the original text and detect the differences. In addition, the reuse of same piece of text more than once may increase vulnerability of the covert communications. If an adversary intercepts the communications and oversees a similar piece of text being exchanged between communicating parties over and over again, suspicion may be raised because the adversary will wonder of such use. However, this is not a concern for Edustega because reusing and modifying educational documents are common practices. For example, an instructor may use and modify old documents such as lectures, examples, tests, exams, etc., for generating new versions. Such Edustega's strong feature eases the automation of an edu-cover. In addition, it is a trivial task that communicating parties to use contemporary educational document generators (exam generator) such as demonstrated in Section 4. The high demand for educational documents by a wide variety of people, in both the academic and nonacademic spheres, has motivated the development of numerous tools for automating the generation of text. Examples of exam generator system include:

- Bank of Chemistry Questions (Hoole et al., 2002)
- Exams and practice tests such as GRE, SAT, etc., (http://www.ets.org)
- Graduate Management Admission Test (GMAT) (http://www.mba.com/mba/TaketheGMAT)
- Exam Pro Software (http://www.exam-software.com)
- Test-generator (http://www.testshop.com/content.php?id=63)

- Chemistry Exam Generator at Department of Chemistry, Indiana University Northwest (http://www.iun.edu/~cpanhd/cgi-bin/generator/examgenerator.html)
- Chemistry Exam Generator at Department of Chemistry, Ohio State University (http://lrc-srvr.mps.ohio-state.edu/under/chemed/qbank/quiz/bank1.htm)
- Exam Generator (GRE Antonyms) (http://www.syvum.com/cgi/online/serve.cgi/gre/verbal/antonyms7.tdf?0)

The *Educational Documents Generator* (EDG Module 3) can be in a form of a data bank, which is simply a large database that contains a huge collection of educational text such as exam generator. Implementing such scheme is accomplished by collecting the required pieces of text, e.g., questions, or documents, e.g. lecture notes, that are initially developed by teachers, trainers and experts. In general, this kind of text is often linguistically legitimate given the rigor that the development of educational documents is subject to. For example, the wording of questions put on a test is often checked multiple times to ensure clarity and accuracy. In addition, the reuse of educational documents, which is a common practice as mentioned before, further strengthens them linguistically given the multiple review cycles that they go through. An example of such scheme is an exam generator of popular standardised tests like the GRE and SAT (http://www.ets.org). The educational document generator does not have to be centralised though. A distributed implementation as peer-to-peer system or web links can also be pursued. Updating such collection of educational document generator is continual and altering a question or document is not unusual and would not draw suspicion. This generates textual elements of educational documents that will form the edu-cover. The criteria of selection are based on the topic and the message encoding scheme. For example, if the topic is college-level calculus, the scope of the selection will be narrowed to such specific subject. On the other hand, if an edu-cover uses questions and answers, Edustega system will generate questions that form the edu-cover. The generated text has to enable the concealment process of the encoded messages. For example, if a message will be concealed by using correct answers of multiple-choice questions, a set of questions that matches the symbols (bit string) used in the encoded message have to be generated with the required order that will appear in generated text before embedding data.

On the other hand, the *Camouflager* (Module 4) is the cover generator scheme that is responsible for embedding the encoded message in the generated text by the EDG scheme (Module 3). The *Camouflager* (Module 4) embeds a message in the generated text by the EDG scheme (Module 3) in a cosmetic manner. To emphasise, the produced text by educational document generator (e.g., exam-generator systems (http://www.ets.org)) is fully legitimate because it is initially generated and reviewed numerous times by human and the embedding process has null effect on the linguistic proprieties of an edu-cover (text-cover) rendering it completely legitimate. For instance, concealing messages through embedding data in questions by changing the order of the choices in true-or-false, multiple-choice, matching, etc. will never generate any noise (linguistic flaws), as shown by implementation in Section 4. Since the sender may mix edu-cover among other legitimate documents, obviously, the basic configuration of the covert channel should include how a recipient can only decode the right covers. For instance, the *Camouflager* (Module 4) may put edu-covers among others similar, but null-coded, documents by following a particular sequence, such as odd number, even number, every other 3, etc., by placing edu-covers in a specific folder, or by specifying certain document contents such as homework. The next section will demonstrate the Edustega system through practical implementation.

4 Edustega implementation

The presented implementation in this section demonstrates the distinct feasibility of Edustega to achieve the steganographical goal with higher bitrate than all linguistic steganography approaches found in the literature. To illustrate Edustega, consider the following scenario. Bob and Alice are on a spy mission, which requires them to reside in two different countries. Before travelling, they plot a strategic plan and set the rules for communicating covertly while portraying themselves as a professor and a student. They basically agree on concealing messages in educational documents by manipulating questions and answers that naturally appear in lecture slides, exam samples, homework, examples, etc., in order to embed the secret data. The manipulated text document serves as a cover (edu-cover). Bob and Alice make sure that every time an edu-cover is generated it has different content and meaning in order to avert suspicion. To make this work, Professor Bob posts or e-mail an edu-cover, e.g., class notes, exam samples, homework, etc., for his students. Alice is one of Bob's students, which legitimises her interest in Bob's class web page and getting his e-mail announcements, etc. These covert transmissions will not look suspicious because the relationship between Bob and Alice is legitimate. Furthermore, Alice is not the sole recipient of Bob's messages; other non-spy students also receive their educational documents further warding off suspicion. When Alice decides to send Bob a message, she does it in the same manner as Bob, except she uses her role as a student to do so. She sends educational documents such as her homework solution, via e-mail to the professor. These educational documents conceal data. However, only Bob and Alice will be able unravel the hidden message because they know the rules of the game. In other words, nothing is suspicious about the communications traffic between Bob and Alice because they are using real data from their academic field to make their covert communications legitimate. Note that even after a class is over the relationship can still be exploited, e.g., when the student

become interested in a particular topic and pursue an independent study or a PhD.

The above scenario demonstrates how Edustega methodology can be used. Edustega methodology is detailed in the remainder of this section. It is worth noting that his section shows just few examples of possible implementations following the steps outlined in the previous section.

4.1 Edustega system

This section explains an implementation example of how Edustega modules are employed and configured to construct the overall Edustega system used in this section by the communicating parties.

The Covert Channel (Module 1): As indicated earlier the configuration of the covert channel includes the topic of the educational documents, the relationship between the sender and receiver and how an edu-cover can be delivered. In this section two topics are employed, namely, the Graduate Record Examination (GRE), and Chemistry. Obviously, these topics are just examples and any other topics may apply as stated in Section 3. The GRE is very popular worldwide among postgraduate students, both native and nonnative English speaking. Both topics, the GRE and Chemistry, offer numerous styles of questions that facilitate the process of camouflaging data. Tools are already available to enable the automation of the concealment process. Examples include the Exam Generator at Department of Chemistry, Indiana University Northwest (http://www.iun.edu/~cpanhd/cgi-bin/generator/examgenerator.html), or the Exam Generator at Department of Chemistry, Ohio State University (http://lrc-srvr.mps.ohio-state.edu/under/chemed/qbank/quiz/bank1.htm). In addition, both topics make it easy to legitimise the communications between sender and recipient. For instance, an instructor of a Chemistry class may post a homework assignment that conceals a message on the class web page. The student can conceal his message in the submitted assignment possibly through wrong answers. For example, the student may respond to a multiple-choice question with the correct answer if it matches the corresponding symbol in the message or intentionally marks the wrong answer that suits the symbol. Such relationship can justify the discernable association between the communicating parties to legitimise the transmittal of an edu-cover. Unlike NORMALS, Edustega is based on educational documents and it is not based on NLG techniques.

Edustega Encoder (Module 2): It is worth noting that the focus of this paper is in the balance for showing how Edustega achieves the steganographical goal rather than making it difficult for an adversary to decode an encoded message. Employing a hard encoding system or cryptosystem to increase the protection of a message is obviously recommended and straightforward using any contemporary encoder or cryptosystem. Similarly, employing compression to boost the bitrate can easily be accomplished by using the contemporary techniques in the literature. Nonetheless, Edustega encodes a message in a form that suits the camouflaging process. The examples in this section conceal messages using multiple-choice questions. To increase the resilience to attacks, Edustega introduces some randomness to the steganographical coding through the use of a combinatorics operations to define the mapping of symbols to bit strings. The steganographical code in this Edustega configuration works as follows:

- Each correct answer (choice) conceals two or three bits according to the steganographic code in Table 1. For example, if a correct answer is the choice ‘A’, the steganographic value ‘00’ is assumed. On the other hand, if a correct answer is the choice ‘B’ will then carry the steganographic value ‘01’ instead, and so on as shown in Table 1.
- A wrong answer (choice) is also used to conceal data. The code is not dependent of the choice though. Instead, the first letter in a wrong answer is encoded according to Table 1. For example, when the incorrect answer starts with the letter ‘B’, it is concluded that the question conceals ‘0001’, as shown in Table 1. However, when the incorrect answer starts with the letter ‘C’, it is concluded that the question conceals ‘0010’, and son as shown in Table 1. In other words, a wrong choice is not encoded. Instead, if the answer of a question is incorrect, regardless which wrong choice is marked, the first letter of the picked answer is checked against the table to find out its code value.
- Based on a predetermined protocol, the presented implementation example of Edustega system in this section adds counter value like an index value ‘*i*’ to each steganographic bit string (e.g., 00+ *i*, 001+ *i*, 0101+ *i*, etc). To emphasise, it adds value of ‘0’ 1st time, ‘1’ 2nd time, ‘2’ 3rd time, and so on. This can either in an entire edu-cover or per an element (e.g., question) used. To illustrate, when using Table 1, which is for concealing data in correct choices, 1st time the choice ‘A’ caries the steganographic value of ‘00’ because the value of ‘00 + *i*’ while $i = 0$ first time used, the choice ‘A’ will then remain same which is ‘00’. The 2nd time, the choice ‘A’ caries the steganographic value of ‘01’ because the value of ‘00 + *i’* while $i = 1$ 2nd time used, the choice ‘A’ will then be equal to ‘01’. Similarly, when using Table 1, in order to conceal data in the wrong choices, if the key word answer starts by the letter ‘A’ implies ‘0000’ 1st time, ‘0001’ 2nd time used, and so on. In other words, when using Table 1 wrong choices are not encoded. Instead, the wrong choices are encoded according to the first letter of the wrong choices, which are selected to form the question according to the steganographic tables and the steganographic values of a particular message. The auto receiver to reveal the hidden message will check the correct answers and wrong choices against Table 1 to find out their steganographic code values. The use of these tables is illustrated later in this section.

Table 1 The steganographic code for camouflaging four bits in the word key answers in wrong choices of questions. In addition, the highlighted part shows the steganographic code for the choices (symbols) A, B, C, D, and E (see online version for colours)

Code values of Stega words		*Code values of Stega choices*	
Binary values	*First letter of Stega words*	*Choices of answers*	*Binary values*
0000	A	**A**	**00**
0001	B	**B**	**01**
0010	C	**C**	**10**
0011	D	**D**	**11**
0100	E	**E**	000
0101	F		
0110	G		
0111	H		
1000	I		
1001	J		
1010	K		
1011	L		
1100	M		
1101	N		
1110	O		
1111	P		
0000	Q		
0001	R		
0010	S		
0011	T		
0100	U		
0101	V		
0110	W		
0111	X		
1000	Y		
1001	Z		

Educational Documents Generator (EDG), and Camouflager (Modules 3 and 4): In this Edustega configuration example, edu-cover is mainly multiple-choice questions. These questions are grouped in a form of exams, examples, homework, etc. The Camouflager module employs online exam generator systems (http://www.iun.edu/~cpanhd/cgi-bin/generator/examgenerator.html, http://lrc-srvr.mps.ohio-state.edu/under/chemed/qbank/quiz/bank1.htm), online examples (http://www.greguide.com/verbal.html; http://greanalogies.blogspot.com/2008/04/analogies-91-95.html), online dictionaries (www.merriam-webster.com; www.online-dictionary.net, www.dictionary.cambridge.org), and Microsoft Thesaurus (built-in Microsoft Word 97) (http://www.microsoft.com/en/us/default.aspx) to embed the data and generate the edu-cover. The dictionaries and thesaurus are mainly exploited in order to pick appropriate vocabulary for the choices for a question. Since the chosen topics both the GRE and Chemistry use five options in the multiple-choice which are 'A' to 'E', the correct answers (choices) will be placed in an order that matches the bit string of an encoded message. According to Table 1 a correct answer may conceal either two or three bits. For instance, in the first time using Table 1 if the steganographic value that needs to be embedded is '00' then the correct answer (choice) will be placed in 'A', but if it is equal '11' then the correct answer (choice) is placed in 'D' and so on. Furthermore, Edustega system embeds data in wrong answers (choices) as well, either by selecting particular wrong choices or substituting them according to Table 1. As shown by examples in Section 4.2 that the use of first letters does not impose constraints on the employed vocabulary. Based on this Edustega configuration each wrong choice (incorrect answers) may conceal four to eight bits. This is not implying that the one who is supposed to answer the questions will mark the wrong choices. Instead, the questions are generated to fit the steganographic values, as detailed earlier and demonstrated next.

4.2 Edu-cover samples

This section shows few examples for how the Edustega configuration discussed above can be used by the

communicating parties to conceal messages. The following describes how a message is encoded and processed by the Edustega system prior to generating the edu-cover. A number of examples of edu-covers that conceal data are demonstrated afterward.

- The plaintext is: "Use my same security key"
- The Edustega Encoder converts the message to a concatenated binary string using the ASCII representation of the individual characters, as follows:

```
0101010101110011011001010010000001101101011110
0100100000011100110110000101101101011001010010
0000111001101100101011000110111010101110010011
0010111010001111001001000000110101101100101011
1001
```

- The encoder will then divide the above binary message into slices of sizes that matches those supported by the steganographic coding. The result is shown below. It should be noted that the binary string could have been encrypted or compressed prior to this step.

01	0101	0101	1100	1101	1001	0100	1000	0001
10	1101	0111	1001	0010	0000	0111	0011	0110
00	0101	1011	0101	1001	0100	1000	0001	1100
11	0110	0101	0110	0011	0111	0101	0111	0010
011	0100	1011	1010	0011	1100	1001	0000	0011
01	0110	1100	1010	1111	0001			

- The camouflage module considers the sliced bit string of the encoded message, generated by the encoder, and maps every slice to a question. The answer (choice) of each question will conceal a part of the message according to Tables 1 and 2. A slice of 2–3 bits will be assigned to a correct answer while 4 bits will be embedded in a wrong answer (choice). The results, shown in Table 2, finally used to generate an edu-cover.

Table 2 The sequence of answers that the Edustega Camouflage module will use to embed the encoded version of the message 'Use my same security key' in an edu-cover

Index	*Correct choice*	*Wrong choice*							
1	01	0101	0101	1100	1101	1001	0100	1000	0001
2	10	1101	0111	1001	0010	0000	0111	0011	0110
3	00	0101	1011	0101	1001	0100	1000	0001	1100
4	11	0110	0101	0110	0011	0111	0101	0111	0010
5	011	0100	1011	1010	0011	1100	1001	0000	0011
6	01	0110	1100	1010	1111	0001			

Samples of edu-cover: The following are sample edu-covers for the above message. Only part of the message is shown for space limitation. The samples demonstrate the effectiveness and efficiency of Edustega and are grouped based on the topic into GRE and Chemistry.

GRE antonyms

The following GRE question conceals the 18 bits "010101010111001101" and is generated by (http://www.syvum.com/cgi/online/serve.cgi/gre/verbal/antonyms7.tdf?0). Table 3 shows how the mapping of the individual slices of the bit string to the offered choices for the answer. The correct choice B, according to the steganographic code of Table 1, matches the first slice. The following 4 slices are embedded into the wrong choices according steganographic code Table 1. The letters of wrong choices are then matched to possible words, employing the dictionary as needed, and the picked words are sorted according to the order of the slices in the bit string. The question in the edu-cover is shown below.

PUTATIVE

a fruitful

b undisputed

c forceful

d modified

e noncommittal

GRE sentence completions

The edu-cover in this sample conceals the 34 bits "01010101011100110110010100100000001" using a GRE Sentence Completions question generated using (http://www.greguide.com/verbal.html). Table 4 shows how the bit string is mapped, again based on Table 1, similar to the previous sample. Note that it adds index value (counter value) to the values of Table 1 is used this time to embed the required data into the correct and wrong answers.

Table 3 Details encoding of a message (part of the message "Use my same security key") using a GRE question. Part of the message is embedded using correct and wrong answers based on Table 1 (see online version for colours)

	Correct answer	*Wrong answers camouflage data by 1st letter of key-words*			
Index ➔	1	2	3	4	5
Encoded message ➔	01	0101	0101	1100	1101
Camouflager uses this row ➔	B	F or V	F or V	M	N
	According to Table 1 and adding value of '0' first time	According to Table 1 and adding value of '0' first time			

Table 4 Details encoding of a message (only the first 34 bits) using a GRE Sentence Completions question style (see online version for colours)

	Correct answer	*Wrong answers camouflage data by 1st letter of key-words*							
Index ➔	1	2	3	4	5	6	7	8	9
Encoded Message ➔	01	0101	0101	1100	1101	1001	0100	1000	0001
Camouflager uses this row ➔	A	G or W	G or W	N	O	K or A	F or V	J or Z	C or S
	According to Table 1 and adding the index/counter value of '1' 2nd time used	According to Table 1 and adding the index/counter value of '1' 2nd time used							

The pressure of population on available resources is the key to understanding history, consequently any historical writing that does not take cognisance of_____ facts is _____flawed.

A demographic…intrinsically

B guard…weak

C national…object

D keen…feeling

E joint…congenial

GRE Analogies Question:

The following edu-cover, again, conceals the first 34 bits of the same bit strings using a GRE Analogies question formed using (http://greanalogies.blogspot.com/2008/04/analogies-91-95.html). Table 5 shows how the mapping of the individual slices of the bit string to the offered choices for the answer of the GRE analogies question.

DOSE : MEDICINE:

A hubris : hold

B oscillation : pulsation

C beat : groove

D alternating : disturbance

E sentence : punishment

Chemistry edu-cover samples:

The following two chemistry-based edu-covers conceal the 18 bits "010101010111001101", generated similar to the GRE questions above. They are generated using http://www.iun.edu/~cpanhd/cgi-bin/generator/examgenerator.html and http://lrc-srvr.mps.ohio-state.edu/under/chemed/qbank/quiz/bank1.htm, respectively. Note that Tables 6 and 7 shows how the mapping of the individual slices of the bit string to the offered choices for the answer of the following two Chemistry edu-cover samples respectively.

Dalton revitalised the concept of the atom, which had been dormant for close to 2000 years, in the l9th century. Which ancient philosopher coined the term 'atomos', or atom?

a Farnsworth

b Aristotle

c Vonnegut

d Mazdak

e None of the above

An atom with a positive charge is known as

a dogion

b WC12

c Ga

d neutrons

e O

4.3 Performance results

The goal of this section is to show the bitrate performance of contemporary linguistic steganography approaches vs. the achieved by Edustega. The bitrate is defined as the size of the hidden message relative to the size of the cover. Table 8 shows the bitrate achieved in the sample edu-covers above. It is worth noting that the bitrate differs from one question to another, from one topic to another, and from one implementation to another as indicated in Table 8.

Table 5 Details encoding of the 34 bits "0101010101110011011001010010000001" (see online version for colours)

	Correct answer	*Wrong answers camouflage data by 1st letter of key-words*							
Index ➔	1	2	3	4	5	6	7	8	9
Encoded Message ➔	01	0101	0101	1100	1101	1001	0100	1000	0001
Camouflager uses this row ➔	E	H or X	H or X	O	P	L or B	G or W	K or A	D or T
	According to Table 1 and adding the index/counter value of '2' 3rd time used	According to Table 1 and adding the index/counter value of '2' 3rd time used							

Table 6 Details encoding of the first 18 bits of the 'Use my same security key' message to generate Chemistry edu-cover (see online version for colours)

	Correct answer	*Wrong answers camouflage data by 1st letter of key-words*			
Index ➔	1	2	3	4	5
Encoded Message ➔	01	0101	0101	1100	1101
Camouflager uses this row ➔	B	F or V	F or V	M	N
	According to Table 1 and adding the index/counter value of '0' 1st time used	According to Table 1 and adding the index/counter value of '0' 1st time used			

Table 7 Encoding the 18 bits '0101010101111001101' in the second Chemistry edu-cover (see online version for colours)

	Correct answer	*Wrong answers camouflage data by 1st letter of key-words*			
Index ➔	1	2	3	4	5
Encoded Message ➔	01	0101	0101	1100	1101
Camouflager uses this row ➔	A	G or W	G or W	N	O
	According to Table 1 and adding the index/counter value of '1' 2nd time used	According to Table 1 and adding the index/counter value of '1' 2nd time used			

Table 8 The bitrate of the presented Edustega examples

Index sample	*Topic*	*Edustega Bitrate* (%)
1	GRE Antonyms	3.26
2	GRE Sentence Completions	1.46
3	GRE Analogies	3.86
4	Chemistry (1st Sample)	0.94
5	Chemistry (2nd Sample)	2.81

To put these bitrate figures in perspective, the bitrate of contemporary linguistic steganography approaches has been investigated. The following reports on the findings, categorising them based on the pursued approaches. Table 9 provides a concise summary of these findings.

1 The statistical-based approach, namely mimic functions: An experiment has been conducted using 30 samples generated using Spam Mimic (http://www.spammimic.com). An average bitrate of 0.90% is observed.

2 *Synonym-based approaches*:

- For the NICETEXT scheme, the samples in Chapman and Davida (1997, 2002) are used to estimate the bitrate, which is found to be approximately 0.29%.
- The Winstein's scheme (Winstein, 1999, 2008) roughly hides about 6 bits per sentence, which yields a bitrate of approximately 0.5% based on the sentences listed in these publications. However, this rate cannot be generalised since not every sentence in the text-cover conceals data. In addition, the size of sentences will affect the bitrate because there are short and long sentences. Nonetheless, the 0.5% figure is assumed given that it is based on the samples developed by the authors.
- The capability of the scheme of Murphy and Vogel (2007) again is reported as the number of bits per sentence. Based on the samples provided in their publication, the achievable bitrate is roughly 0.30% per sentence.

- Nakagawa et al. (2001) have provided two samples for their scheme. The samples achieve bitrate of 0.06% and 0.12% respectively. However, it has been noted that when tried in a real application, only a bitrate of 0.034% could be reached.

3 *Noise-based approaches*

- The bitrate for the translation-based scheme reported in Stutsman et al. (2006) is roughly 0.33%.
- Based on the examples in Topkara et al. (2007), the confusing scheme approximately achieves a bitrate of 0.35%.
- The linguistic technique of the SMS-based approach (Shirali-Shahreza et al., 2007) is claimed to be capable of hiding few bits in a file of several kilobytes, which yields an extremely low bitrate.

Comparing Tables 8 and 9, it is obvious that Edustega achieves much more superior bitrate than all comparable approaches, making it a very effective steganography approach. The high bitrate also enables the use of reasonable cover sizes, a major concern for all steganography approaches linguistic and nonlinguistic.

In the other hand, the following shows a brief comparison between Edustega and NORMALS, as shown in Table 10.

Table 9 The bitrate of contemporary linguistic steganography approaches

Approach	*Bitrate* (%)	*Comment*
Mimic functions (Wayner, 1992, 2002)	0.90	Based on 30 samples generated at www.spamimc.com
NICETEXT (Chapman and Davida (1997, 2002)	0.29	Based on the samples in the cited papers
Winstein (1999, 2008)	0.5	Based on the samples in the cited papers, and also confirmed in Murphy and Vogel (2007)
Murphy and Vogel (2007)	0.30	Average per sentence (as reported in Murphy and Vogel (2007))
Nakagawa et al. (2001)	0.12	As reported in Nakagawa et al. (2001), Bitrate achieved in real application is only 0.034%
Translation-based (Stutsman et al., 2006)	0.33	Noted by the authors in the cited papers
Confusing (Topkara et al., 2007)	0.35	Based on the samples in the cited papers

Table 10 A brief comparison between Edustega and NORMALS

	Edustega	*NORMALS*
Domains	*Educational*	*Domain Specific Subjects* (*DSS*)
Steganographic cover form	Generally, but not limited to, questions and answers of exams, examples, puzzles, competitions, etc	Reports, letters, etc., which is not a set of questions. The questioner in the implemented NORMALS was used only to collect data inputs, which is not the steganographic cover itself
System	This implementation is based on Exam Generators	NLG-based
Bitrate	According to this implementation roughly may reach 0.94–3.86%	According to current NORMALS implementation the bitrate may reach 0.29%
Note	The questions entirely made and fully reviewed by human before used in the system, e.g., Exam Generator that selects a set of questions that are already mad by human. Edustega employs such system to embed a message in the educational documents	NLG system generates the text-cover

5 Steganalysis validation

The aim of this section is to show the resilience of Edustega to possible attacks. Again the success of steganography is qualified with its ability for avoiding an adversary's suspicion of the presence of a hidden message. It is assumed that an adversary will perform all possible investigations. In addition, the adversary is also aware of Edustega, as a public methodology, but he does not know the Edustega configuration that the sender and recipient employ for their covert communication.

5.1 Traffic attack

One of the possible attacks an adversary may pursue is to analyse the communications traffic and the access patterns to publicly available or exchanged documents, images, graphs, files, etc. For example, the intelligence community has a number of tools at their disposal for analysing traffic on the internet, tracking access to websites, monitoring checked out literature from public libraries, etc. The main goal of a traffic attack is to detect unusual or questionable association between a sender and a recipient. Traffic analysis intuitively can identify who communicates with whom. The relationship between the communicating parties will be then qualified based on the contents of the message. Traffic attacks can be a threat for most contemporary steganographic techniques regardless of the steganographic cover type (e.g., image, graph, audio file, text, etc). In the context of Edustega, the subject of the cover is checked rather than its validity and the consistency of its contents.

If someone sends, receives, and accesses some materials without a legitimate reason for doing so, e.g., a Math teacher sending a Chemistry assignment to one of his students, suspicion can be raised and further investigation may be warranted. The additional investigations will involve a thorough analysis of a steganographic cover, as detailed in the next subsections.

Traffic analysis is deemed ineffective with Edustega. Edustega camouflages the transmittal of a hidden message (edu-cover) to appear legitimate and thus suspicion is averted. Basically, Edustega ensures that the involved parties establish a covert channel by having a well-plotted relationship with each other. Analysing the traffic between them will not reveal any questionable association and will not trigger any further investigation. In addition, the high demand for educational documents by a wide variety of people, in both the academic and nonacademic spheres, creates a high volume of traffic that makes it impractical for an adversary to investigate all traffics. The voluminous traffic allows the communicating parties to establish a covert channel in order to transmit an edu-cover without drawing attention, rendering educational documents an attractive steganographic carrier. Finally, it is noted that if further investigations on an edu-cover are triggered by traffic analysis, they would not be successful, as elaborated next. In Edustega, differentiating between an edu-cover that contains a hidden message and another peer educational document without a hidden message is extremely difficult.

5.2 Contrast and comparison attacks

One of the intuitive sources of noise that may alert an adversary is the presence of contradictions in an edu-cover. Examples of these contradictions include finding repeated questions in an exam, errors in an answer sheet, naïve problems for an advanced class, etc. Also, if an edu-cover in a form of exam, it is not expected to contain errors, and if there is, it is not expected to be numerous. Such contradictions would surely raise suspicion about the existence of a hidden message, especially when they are present in the same document. The vulnerability of Edustega to contrast attacks is generally very limited and depends on how the cover is generated. Automating the generation of an edu-cover through the use of contemporary question banks and exam generation tools makes the cover very resilient to this type of attacks. As demonstrated in Section 4, the use of a tool like (http://www.iun.edu/~cpanhd/cgi-bin/generator/exam generator.html; http://lrc-srvr.mps.ohio-state.edu/under/chemed/qbank/quiz/bank1.htm; http://www.syvum.com/cgi/online/serve.cgi/gre/verbal/antonyms7.tdf?0; http://www.greguide.com/verbal.html; http://greanalogies.blogspot.com/2008/04/analogies-91-95.html) allows the selection of appropriate questions that not only match the encoded messages but also ensure the scientific validity of the questions, the clarity of the wording and the suitability of the scope.

From a steganography point of view, reusing or altering an existing text to hide data is not a recommended practice since an adversary can reference the original text and detect the differences. In addition, the reuse of same piece of text more than once may increase vulnerability of the covert communications. If an adversary intercepts the communications and oversees a similar piece of text being exchanged between communicating parties over and over again, suspicion may be raised because the adversary will wonder of such use. However, this is not a concern for Edustega because reusing and modifying educational documents are common practices. For example, an instructor may use and modify old documents such as lectures, examples, tests, exams, etc., for generating new versions. Such Edustega's strong feature eases the automation of an edu-cover. In addition, it is a trivial task that communicating parties to use contemporary educational document generators such as the tools that are similar to the systems mentioned (exam generators) in Section 3.3 and as demonstrated in Section 4. Meanwhile, noise in the context of comparison attacks reflects alteration of authenticated or previously used documents. The goal of the adversary is to find any incorrect and inconsistent data that may imply the manipulation of contents of an edu-cover in order to embed a hidden message. As stated above, since reusing and modifying educational documents are common practices in academic, e.g., homework, exams, etc. and nonacademic setups, e.g., puzzles, competitions, etc., comparison and contrast attacks are deemed ineffective.

5.3 Linguistics attacks

Linguistics examination distinguishes the text that is under attack from normal human language. Distinguishing the text from normal human language can be done through the examination of meaning, syntax, lexicon, rhetoric, semantic, coherence, and any other feature that can help in detecting or suspecting the existence of a hidden message. These examinations are used to determine whether or not the text that is under attack is abnormal. Generally, the text used in educational documents is normal. In addition, the produced text by exam-generator systems usually meets the expected properties of a normal human language because it is initially generated by human and any alteration is done is more of cosmetic. For example, changing the order of the choices in multiple-choice questions will not generate any noise (linguistic flaws). As a result, the generated cover as demonstrated in the implementation section is normal text. Furthermore, if there are errors in the exam generator engine, it should not be a concern for two reasons; first, it applies to all the generated text with and without a hidden message; second, nothing is concealed in errors. In addition, an engine error of an exam generator is most likely fixable. Therefore, Edustega is capable of passing any linguistic attack by both human and machine examinations.

On the other hand, a statistical attack refers to tracking the profile of the used text. A statistical signature (profile) of a text refers to the frequency of words and characters used. An adversary may use the statistical profile of a particular topic of educational documents that contains no

hidden message and compare it to a statistical profile of the suspected edu-cover to detect any differences. An alteration in the statistical signature of a particular topic of educational documents may be a possible way of detecting a noise that an adversary would watch for. Unlike image steganography, tracking statistical signatures is an ineffective means for attacking linguistic steganography (Grothoff et al., 2005a, 2005b; Stutsman et al., 2006). Nonetheless, Edustega is resistant to statistical attacks because it is simply opt to use legitimate text that is generated naturally by human. In addition, the generated textual cover (edu-cover) by Edustega retains the same profile of its other peer educational documents that contains no hidden message. Basically, most alterations introduced by Edustega are nonlinguistic and do not produce any flaws (noise), as demonstrated in the implementation section, deeming statistical attacks on edu-cover very ineffective.

5.4 Statistical signature

In this paper, the statistical signature (profile) of a text refers to the frequency distribution of words, and characters. An adversary may use the statistical profile of normal text that contains no hidden message and compare it to the profile of the suspected text in order to detect any differences. Tracking statistical signatures may be an effective means for attacking text-cover because it can be automated. However, Edustega is resistant to statistical attacks as demonstrated by the experimental results in this section.

Three main schemes to capture the statistical profile are pursued for validating Edustega. Two of validation schemes are based on some fundamental concepts of Natural Language Processing (NLP), namely the Words Frequency Distribution (WFD) discussed in Section 5.4.1 and the Letters Frequency Distribution (LFD) detailed in Section 5.4.2. Finally, Section 5.4.3 demonstrates the results when using Kullback–Leibler Divergence (KLD) to assess the level of similarity between Edustega-based Covers and normal text with respect to WFD and LFD. Note that the following are concise summaries and for more information refer to Desoky (2009a, 2012).

5.4.1 Zipfian signature

Human language in general and the English language in particular, have been statistically investigated to discover its statistical properties. The most notable study on the frequency of words was done by Zipf (1968) and Li (1992). Zipf investigated the statistical occurrences of words in the human language and in particular the English language. Based on the statistical experimental research, Zipf concluded his observation, which is known as Zipf's law. Zipf's law states that the word frequency is inversely proportional to its rank in an overall words frequency table, which lists all words used in a text sorted in a descending order of their number of appearances. Mathematically, Zipf's law implies that $W_n \sim 1/n^a$, where W_n is the frequency of occurrence of the nth ranked word and 'a' is a constant that is close to 1. Based on such a mathematical relationship, a logarithmic scale plot of the number of words' appearance and their ranks will yield a straight line with a slope '$-a$' that is close to –1. The value of 'a' is found to depend on the sample size and mix. Unlike the presented experiment in this paper, Zipf's law was originally observed on a huge bundle of textual collections containing numerous different Domain Specific Subjects (DSS) by different authors, different writing-styles, different writing-fingerprints, etc. Consequently, this huge bundle of textual collections is fairly blended which causes the occurrence of approaching or reaching Zipfian of –1. The following is the first experiment, which applies Zipf's law and for more information about this experiment refer to Desoky (2009a).

Experiment # 1

The aim of this experiment is inspecting Zipfian signature of Edustega Cover. Therefore, Experiment # 1 investigates Zipfian signature for the following texts:

- Unaltered educational documents that are generated by auto-exam (http://www.iun.edu/~cpanhd/cgi-bin/generator/examgenerator.html; http://lrc-srvr.mps.ohio-state.edu/under/chemed/qbank/quiz/bank1.htm). Generally, the size of this text is about 10 questions each.
- Edustega Covers that retain similar size to the unaltered versions.

Basically, Zipf's law is applied to edu-cover and its original text that contains no hidden data. Table 11 shows the results. It is observed that edu-cover and its peer, which contains no hidden data, hold the same Zipfian values for 23 samples. These Zipfian values fluctuate from –0.7195 up to –0.9758 and hold an average Zipfian value of –0.82582. Obviously, edu-cover can fully fool such an attack.

5.4.2 Letter frequency distribution

The human language can be defined as a set of characters. In modern languages, this set of characters represents the letters of a particular language. Generally, in any language letters may have different frequency of usage. Thus, the following experiment has examined such phenomena, opting to identify distinct patterns.

Experiment # 2

The aim of the Experiment # 2 is to inspect the LFD of Edustega Covers. Therefore, this experiment investigates the following:

- Educational documents generated by auto-exam (http://www.iun.edu/~cpanhd/cgi-bin/generator/examgenerator.tml, http://lrc-srvr.mps.ohio-state.edu/under/chemed/qbank/quiz/bank1.htm) that contain no hidden data. Generally, the size of these texts is about 10 questions each. These documents are the original text before hiding data (the peers of edu-covers).

- Edustega Covers. The size of edu-cover is similar to the previous one.

The following observations can be made about the results of experiment # 2, as shown in Table 12.

Since the LFD for both edu-cover and its original text before hiding data are exactly same and therefore LFD curves will be same for both. Therefore, it is concluded based on the experimental results, as shown in this section, that edu-cover is capable of fooling such an examination. Due to the size constraint of this paper, for more information about this experiment refer to Desoky (2009a).

5.4.3 Kullback–Leibler divergence

In probability theory and information theory, the Kullback–Leibler Divergence (KLD) (Solanki et al., 2006; Kullback and Leibler, 1951; Kullback, 1959, 1987) is considered a non-commutative measure. It measure difference between two probability distributions P and Q; where P represents the true distribution of data, observations, or a precise calculated theoretical distribution, and Q represents a theory, model, approximation of P, or description. Therefore, KLD is employed in this paper in order to assess the similarities of Edustega-based Covers to normal text with respect to WFD and LFD. In this paper the KLD values may be considered as a divergence between Edustega-based Cover and other normal text.

To illustrate the interpretation of the KLD values, consider the following:

- Zero KLD value means no divergence between a cover and its comparable normal text.
- Small KLD value which can not be as an evidence of hidden data.
- Large KLD value, which defiantly raise suspicion.

For probability distributions P and Q of a discrete random variable the measure of K–L divergence of Q from P is defined to be:

$$D_{KL}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}.$$

Experiment # 3:

The aim of this experiment is to investigate the KLD between Edustega Covers and their peers' texts. Therefore, Experiment # 3 investigates the following:

- Educational documents generated by auto-exam (http://www.iun.edu/~cpanhd/cgi-bin/generator/examgenerator.html, http://lrc-srvr.mps.ohio-state.edu/under/chemed/qbank/quiz/bank1.htm) that contain no hidden data. Generally, the size of the considered text is about 10 questions each.
- Edustega Covers. The size of edu-cover is similar to the above one.

The Experimental Result Retains Zero KLD:

The experimental results for all of these samples retain zero KLD on Letter Frequency Distribution (LFD) and word Letter Frequency Distribution (WFD). As shown in Table 11 (results of Experiment # 1) and in Table 12 (results of Experiment # 2), the LFD and WFD for edu-covers verses the original text (the text that contain no hidden data) retain same values; and therefore, the KLD values is zero. This result is very much expected for both LFD and WFD since $D_{KL}(P\|Q)$ zero if and only if $P = Q$ (Solanki et al., 2006). For more information about this experiment refer to Desoky (2009a).

Table 11 The Zipfian distribution (logarithmic scale) of the original text that contain no hidden data and edu-cover hold same Zipfian values. The equation is a linear curve fitting of the results (Slope [$-a$]). R^2 is the squared error (see online version for colours)

	Text without hidden data			*Edu-cover*		
Text #	*Equation*	R^2	*Slope*($-a$)	*Equation*	R^2	*Slope*($-a$)
1	$-0.7948x + 1.4646$	0.9111	–0.7948	$-0.7948x + 1.4646$	0.9111	–0.7948
2	$-0.7436x + 1.4305$	0.8933	–0.7436	$-0.7436x + 1.4305$	0.8933	–0.7436
3	$-0.8931x + 1.7267$	0.9549	–0.8931	$-0.8931x + 1.7267$	0.9549	–0.8931
4	$-0.8201x + 1.4454$	0.9285	–0.8201	$-0.8201x + 1.4454$	0.9285	–0.8201
5	$-0.8894x + 1.5987$	0.9052	–0.8894	$-0.8894x + 1.5987$	0.9052	–0.8894
6	$-0.9089x + 1.8336$	0.9419	–0.9089	$-0.9089x + 1.8336$	0.9419	–0.9089
7	$-0.7436x + 1.4128$	0.9371	–0.7436	$-0.7436x + 1.4128$	0.9371	–0.7436
8	$-0.7975x + 1.5543$	0.9234	–0.7975	$-0.7975x + 1.5543$	0.9234	–0.7975
9	$-0.7661x + 1.541$	0.9145	–0.7661	$-0.7661x + 1.541$	0.9145	–0.7661
10	$-0.7195x + 1.391$	0.9056	–0.7195	$-0.7195x + 1.391$	0.9056	–0.7195
11	$-0.7314x + 1.386$	0.8847	–0.7314	$-0.7314x + 1.386$	0.8847	–0.7314
12	$-0.8147x + 1.658$	0.9398	–0.8147	$-0.8147x + 1.658$	0.9398	–0.8147

Table 11 The Zipfian distribution (logarithmic scale) of the original text that contain no hidden data and edu-cover hold same Zipfian values. The equation is a linear curve fitting of the results (Slope [$-a$]). R^2 is the squared error (see online version for colours) (continued)

	Text without hidden data			*Edu-cover*		
13	$-0.8338x + 1.7116$	0.9542	–0.8338	$-0.8338x + 1.7116$	0.9542	–0.8338
14	$-0.7926x + 1.6032$	0.9219	–0.7926	$-0.7926x + 1.6032$	0.9219	–0.7926
15	$-0.9103x + 1.7424$	0.9434	–0.9103	$-0.9103x + 1.7424$	0.9434	–0.9103
16	$-0.8818x + 1.7571$	0.935	–0.8818	$-0.8818x + 1.7571$	0.935	–0.8818
17	$-0.7597x + 1.479$	0.9148	–0.7597	$-0.7597x + 1.479$	0.9148	–0.7597
18	$-0.7396x + 1.3786$	0.9201	–0.7396	$-0.7396x + 1.3786$	0.9201	–0.7396
19	$-0.9685x + 1.8276$	0.937	–0.9685	$-0.9685x + 1.8276$	0.937	–0.9685
20	$-0.7462x + 1.4402$	0.9206	–0.7462	$-0.7462x + 1.4402$	0.9206	–0.7462
21	–0.9758x + 1.7782	0.9286	–0.9758	–0.9758x + 1.7782	0.9286	–0.9758
22	–0.8586x + 1.542	0.9199	–0.8586	–0.8586x + 1.542	0.9199	–0.8586
23	–0.9043x + 1.4876	0.889	–0.9043	–0.9043x + 1.4876	0.889	–0.9043
	Average		–0.82582	Average		–0.82582

Table 12 The letter frequency distribution (LFD) values of both edu-cover and its original text that contain no hidden data. The LFD's values are exactly same for both edu-cover and its original text because the LFD values are same (see online version for colours)

Letter	*LFD of Edu-Cover # and its original text number that contain no hidden data*											
	1 (%)	*2* (%)	*3* (%)	*4* (%)	*5* (%)	*6* (%)	*7* (%)	*8* (%)	*9* (%)	*10* (%)	*11* (%)	*12* (%)
A	7.65	7.32	7.76	8.66	7.98	7.98	8.73	8.97	9.00	8.74	8.04	8.63
B	2.02	2.53	2.56	2.84	2.55	2.50	1.95	2.62	3.00	2.26	2.18	2.21
C	4.27	5.23	4.56	7.61	6.91	6.00	5.63	4.99	4.81	5.30	4.17	5.28
D	3.04	2.79	2.78	4.03	2.45	3.36	2.41	3.55	3.08	4.03	2.78	2.81
E	12.15	11.42	9.68	10.45	10.64	9.43	11.71	10.58	12.00	12.38	12.21	10.57
F	2.92	2.62	2.70	4.33	3.40	3.43	2.87	3.30	1.97	3.05	2.78	2.54
G	3.15	2.88	2.14	0.90	1.49	2.77	1.84	3.64	2.76	1.96	2.68	4.28
H	4.61	4.18	4.06	5.67	5.74	5.67	4.94	4.06	3.71	4.13	5.16	3.75
I	7.99	7.67	6.69	6.27	7.98	7.45	7.69	7.36	7.42	7.66	7.35	7.42
J	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.13
K	0.56	0.09	0.14	0.00	0.00	0.40	0.46	0.34	0.32	0.39	0.30	0.13
L	3.60	5.06	2.85	5.22	5.21	4.61	6.43	5.08	5.84	5.21	5.36	4.82
M	4.27	4.27	3.06	3.43	2.66	5.01	3.79	4.65	4.42	3.44	4.27	4.62
N	6.75	5.75	8.68	5.22	8.19	4.55	6.54	6.26	6.47	5.70	6.85	6.29
O	6.30	7.93	10.04	10.00	9.89	9.36	7.69	7.45	6.63	7.27	6.95	8.36
P	1.57	2.53	2.35	1.79	1.81	2.31	1.72	1.18	0.71	1.96	1.69	1.94
Q	0.11	0.17	0.00	0.00	0.64	0.07	0.11	0.17	0.47	0.39	0.30	0.07
R	4.84	4.53	5.62	6.42	3.40	4.61	5.28	5.50	4.66	5.30	5.06	4.01
S	8.21	7.24	7.54	6.27	5.32	5.60	5.86	5.33	6.79	5.89	6.16	6.22
T	9.56	8.81	9.47	5.97	8.19	6.92	7.58	8.38	8.60	8.45	7.85	8.56
U	2.47	3.05	3.63	2.54	2.55	2.83	2.18	2.28	2.53	1.87	2.88	2.88
V	0.45	0.52	0.28	0.60	0.32	0.46	1.15	0.68	0.32	0.39	0.70	0.40
W	1.80	1.48	1.57	1.34	1.60	2.24	1.61	1.86	1.89	1.67	2.38	1.67
X	0.90	0.17	1.07	0.00	0.21	1.05	0.69	1.18	1.26	1.08	0.70	0.60
Y	0.79	1.48	0.78	0.45	0.53	1.25	1.03	0.51	1.03	1.38	0.99	1.61
Z	0.00	0.09	0.00	0.00	0.32	0.13	0.11	0.08	0.32	0.10	0.20	0.20
Total	100	100	100	100	100	100	100	100	100	100	100	100

Table 12 The letter frequency distribution (LFD) values of both edu-cover and its original text that contain no hidden data. The LFD's values are exactly same for both edu-cover and its original text because the LFD values are same (see online version for colours) (continued)

Letter	*LFD of Edu-Cover # and its original text number that contain no hidden data.*										
	13 (%)	*14* (%)	*15* (%)	*16* (%)	*17* (%)	*18* (%)	*19* (%)	*20* (%)	*21* (%)	*22* (%)	*23* (%)
A	8.75	6.98	9.65	8.43	8.12	8.43	8.12	6.86	8.35	7.56	8.20
B	1.98	2.54	2.04	1.87	1.97	1.87	1.97	2.55	2.22	2.52	4.21
C	4.60	5.14	4.25	5.04	4.68	5.04	4.68	4.05	8.67	6.10	8.65
D	3.70	2.28	3.89	3.39	3.12	3.39	3.12	2.46	4.12	3.05	3.33
E	9.71	10.91	7.35	9.37	10.99	9.37	10.99	11.43	7.93	10.08	7.76
F	2.43	2.79	3.10	2.95	2.13	2.95	2.13	3.17	5.29	4.51	4.21
G	3.26	3.23	2.74	4.03	2.46	4.03	2.46	1.41	1.27	1.46	1.11
H	4.53	4.25	4.96	5.26	3.77	5.26	3.77	4.40	5.07	5.44	7.54
I	8.68	9.13	9.03	8.86	8.20	8.86	8.20	8.18	5.60	6.10	4.21
J	0.13	0.00	0.00	0.14	0.08	0.14	0.08	0.00	0.00	0.00	0.00
K	0.13	0.13	0.18	0.36	0.00	0.36	0.00	0.09	0.63	0.00	0.00
L	4.47	4.19	7.43	5.19	5.09	5.19	5.09	3.61	5.18	5.17	3.99
M	4.02	4.25	4.16	3.89	4.68	3.89	4.68	3.87	3.28	4.64	3.33
N	6.64	6.40	4.60	5.98	5.25	5.98	5.25	8.53	6.13	4.91	4.43
O	8.30	7.61	8.32	7.71	8.45	7.71	8.45	9.67	10.57	10.08	13.30
P	1.66	1.90	1.86	1.59	1.97	1.59	1.97	2.02	1.06	1.46	4.21
Q	0.19	0.06	1.42	0.22	0.41	0.22	0.41	0.09	0.00	0.00	0.00
R	3.96	5.14	3.36	3.46	5.50	3.46	5.50	5.45	6.13	7.16	6.43
S	7.28	6.79	6.73	8.00	7.14	8.00	7.14	7.04	5.60	6.63	3.33
T	8.75	9.58	7.61	6.84	9.35	6.84	9.35	8.62	5.39	6.23	5.76
U	3.00	2.92	3.36	2.23	3.12	2.23	3.12	2.55	3.07	3.05	2.44
V	0.32	0.13	0.62	0.50	0.41	0.50	0.41	0.88	0.11	0.40	0.00
W	2.11	1.59	1.50	2.02	1.39	2.02	1.39	1.41	1.80	1.46	2.00
X	0.32	0.51	0.35	1.01	0.25	1.01	0.25	0.18	0.32	0.27	0.22
Y	1.02	1.52	1.50	1.59	1.31	1.59	1.31	1.23	0.74	0.66	1.33
Z	0.06	0.06	0.00	0.07	0.16	0.07	0.16	0.26	1.48	1.06	0.00
Total	100	100	100	100	100	100	100	100	100	100	100

6 Conclusion

This paper has presented a novel Educational-Centric Steganography Methodology (Edustega) that conceals data in educational documents. The high demand for educational documents by a wide variety of people, in both the academic and nonacademic spheres, allows the communicating parties to establish a covert channel to transmit hidden messages rendering educational documents an attractive steganographic carrier. Edustega neither hides data in a noise (errors) nor produces noise. Instead, it camouflages data in educational documents by manipulating, mainly but not limited to, questions and answers (e.g., multiple-choice, true-or-false, fill-in-the-space, matching, etc.) of exams, examples, puzzles, competitions, etc., in order to embed data without generating any suspicious pattern. It has been shown that Edustega can conceal data in both correct and incorrect answers of questions. An implementation example has demonstrated that a bitrate of up to 3.86% can be achieved. Such bitrate is superior to contemporary linguistic steganography approaches found in the literature, confirming the effectiveness of Edustega and the high capacity that educational documents provide for concealing data. Furthermore, Edustega can be applied to all languages. The steganalysis validation has shown Edustega methodology is capable of achieving the steganographic goal.

References

Anderson, R.J., Needham, R. and Shamir, A. (1998) 'The steganographic file system', *Proceedings of the 2nd International Workshop on Information Hiding*, Vol. 1525 of *Lecture Notes in Computer Science*, Springer, pp.73–82.

Ansari, R., Malik, H. and Khokhar, A. (2004) 'Data-hiding in audio using frequency-selective phase alteration', *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, Vol. 5, 17–21 May, pp.389–392.

Atallah, M.J. *et al.* (2001) 'Natural language watermarking: design, analysis, and a proof-of-concept implementation', *Proceedings of the 4th International Workshop* on *Information Hiding*, Lecture Notes in Computer Science, Vol. 2137, April, pp.185–199.

Atallah, M.J. *et al.* (2002) 'Natural language watermarking and tamperproofing', *Proceedings of the 5th International Workshop on Information Hiding*, Lecture Notes in Computer Science, Vol. 2578, October, pp.196–212.

Bender, W. *et al.* (1996) 'Techniques for data hiding', *IBM Systems J.*, Vol. 35, Nos. 3–4, pp.313–336.

Bennett, K. (2004) *Linguistic Steganography: Survey, Analysis, and Robustness Concerns for Hiding Information in Text*, Technical Report CERIAS Tech. Report 2004-13, Purdue University.

Bergmair, R. (2008) *Towards Linguistic Steganography: A Systematic Investigation of Approaches, Systems, and Issues*, Technical Report, The University of Derby, Austria, April 2004. Available: http://richard.bergmair.eu/pub/towlingsteg-rep-inoff-b5.pdf, Accessed on 18 September, 2008.

Bergmair, R. and Katzenbeisser, S. (2004) 'Towards human interactive proofs in the text-domain', *Proceedings of the 7th Information Security Conference (ISC'04), Lecture Notes in Computer Science*, Vol. 3225, September, pp.257–267.

Bergmair, R. and Katzenbeisser, S. (2007) 'Content-aware steganography: about lazy prisoners and narrow-minded wardens', Proceedings of the 8th Information Hiding Workshop*, Lecture Notes in Computer Science*, Vol. 4437, September, pp.109–123.

Bolshakov, I.A. (2004) 'A method of linguistic steganography based on collocationally-verified synonymy', *Proceedings of* 6th International Workshop on Information Hiding, Lecture Notes in Computer Science, Vol. 3200, May, pp.180–191.

Bolshakov, I.A. and Gelbukh, A. (2004) 'Synonymous paraphrasing using wordnet and internet', *Proceedings of the* Natural 9th International Conference on Applications of Natural Language to Information Systems, Lecture Notes in Computer Science, Vol. 3136, June, pp.312–323.

Calvo, H. and Bolshakov, I.A. (2004) 'Using selectional preferences for extending a synonymous paraphrasing method in steganography', in Sossa Azuela, J.H. (Ed.): Avances en Ciencias de la Computacion e Ingenieria de Computo – CIC'2004: XIII Congreso Internacional de Computacion, October, pp.231–242.

Chand, V. and Orgun, C.O. (2006) 'Exploiting linguistic features in lexical steganography: design and proof-of-concept implementation', Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS '06), Vol. 6, January.

Chapman, M. and Davida, G. (1997) 'Hiding the hidden: a software system for concealing ciphertext as innocuous text', *Proceedings of the International Conference on Information and Communications Security*, Vol. 1334 of *Lecture Notes in Computer Science*, Springer, Beijing, PR China, November, pp.335–345.

Chapman, M. and Davida, G.I. (2002) 'Plausible deniability using automated linguistic steganography', *Proceedings of the International Conference on Infrastructure Security (InfraSec'02)*, *Lecture Notes in Computer Science*, Springer, Vol. 2437, pp.276-287.

Chapman, M. *et al.* (2001) 'A practical and effective approach to large-scale automated linguistic steganography', *Proceedings of the Information Security Conference (ISC '01)*, Vol. 2200 of *Lecture Notes in Computer Science*, Springer, Malaga, Spain, pp.156–165.

Cvejic, N. and Seppanen, T. (2004) 'Reduced distortion bit-modification for LSB audio steganography '04', *Proceedings of the 7th International Conference on Signal Processing (ICSP 04)*, Vol. 3, August, Beijing, China, pp.2318–2321.

Desoky A. and Younis, M. (2009) 'Chestega: chess steganography methodology', *Journal of Security and Communication Networks*, March.

Desoky, A. (2008) 'Nostega: a novel noiseless steganography paradigm', *Journal of Digital Forensic Practice*, Vol. 2, No. 3, March, pp.132–139.

Desoky, A. (2009a) *Nostega: A Novel Noiseless Steganography Paradigm*, PhD Dissertation, University of Maryland, Baltimore County, USA, May.

Desoky, A. (2009b) 'Listega: list-based steganography methodology', *International Journal of Information Security*, Springer-Verlag, April.

Desoky, A. (2009c) 'Notestega: notes-based steganography methodology', *Information Security Journal: A Global Perspective*, Vol. 18, No. 4, January, pp.178–193.

Desoky, A. (2010a) 'NORMALS: normal linguistic steganography methodology', *Journal of Information Hiding and Multimedia Signal Processing*, Vol. 1, No. 3, July, pp.145–171.

Desoky, A. (2010b) 'Comprehensive linguistic steganography survey', *Int. J. Information and Computer Security*, Vol. 4, No. 2.

Desoky, A. (2011) 'Matlist: mature linguistic steganography methodology', *Security and Communication Networks*, Vol. 4, No. 6, June, pp.697–718.

Desoky, A. (2012) *Noiseless Steganography: The Key to Covert Communications*, Information Security Publisher/Taylor & Francis ISBN: 1439846219/ ISBN: 9781439846216 (In Process).

Desoky, A. and Younis, M. (2008) 'Graphstega: graph steganography methodology', *Journal of Digital Forensic Practice*, Vol. 2, No. 1, January, pp.27–36.

Desoky, A. *et al.* (2008) 'Auto-summarization-based steganography', *Proceedings of the 5th IEEE International Conference on Innovations in Information Technology*, Al-Ain, UAE, December.

Grosvald, M. and Orhan Orgun, C. (2011) 'Free from the cover text: a human-generated natural language approach to text-based steganography', *Journal of Information Hiding and Multimedia Signal Processing*, Vol. 2, No. 2, April, pp.133–141.

Grothoff C. *et al.* (2005b) 'Translation-based steganography', *Proceedings of Information Hiding Workshop (IH 2005)*, Springer-Verlag, Barcelona, Spain, June, pp.213–233.

Grothoff, C. *et al.* (2005a) *Translation-Based Steganography*, Technical Report CSD TR# 05-009, Purdue University (CERIAS Tech Report 2005-39).

Gruhl, D., Lu, A. and Bender, W. (1996) 'Echo hiding', *Proceedings of First International Workshop on Information Hiding*, Lecture Notes in Computer Science, Vol. 1174, May, Springer, Cambridge, UK, pp.295–316.

Handel, T.G. and Sandford, M.T. (1996) 'Data hiding in the OSI network model', *Proceedings* of *First International Workshop on Information Hiding*, Vol. 1174 of *Lecture Notes in Computer Science*, Springer, pp.23–38.

Hoole, D. *et al.* (2002) 'A bank of chemistry questions on an on-line server', *Journal of Science Education and Technology*, Vol. 11, No. 1, March, pp.9–13.

Kahn, D. (1996) *The Codebreakers: The Story of Secret Writing*, Revised Ed., Scribner, December.

Kessler, G.C. (2004) 'An overview of steganography for the computer forensics examiner', *Forensic Science Communications*, Vol. 6, No. 3, July.

Kirovski, D. and Malvar, H. (2001) 'Spread-spectrum audio watermarking: requirements, applications, and limitations', *Proceedings of the 4th IEEE Workshop on Multimedia Signal Processing*, October, Cannes, France, pp.219–224.

Kullback, S. (1959) *Information Theory and Statistics*, John Wiley and Sons, NY.

Kullback, S. (1987) 'The Kullback-Leibler distance', *The American Statistician*, Vol. 41, pp.340–341.

Kullback, S. and Leibler, R.A. (1951) 'On information and sufficiency', *The Annals of Mathematical Statistics*, Vol. 22, No. 1, pp.79–86, doi:10.1214/aoms/1177729694. MR39968.

Li, W. (1992) 'Random texts exhibit Zipf's-law-like word frequency distribution', *IEEE Transactions on Information Theory*, Vol. 38, No. 6, pp.1842–1845.

Martin, A., Sapiro, G. and Seroussi, G. (2005) 'Is image steganography natural?', *IEEE Transactions on Image Processing*, Vol. 14, No. 12, December, pp.2040–2050.

Murphy, B. and Vogel, C. (2007) 'The syntax of concealment: reliable methods for plain text information hiding', *Proceedings of the SPIE International Conference on Security, Steganography, and Watermarking of Multimedia Contents*, January.

Nakagawa, H. *et al.* (2001) 'Text information hiding with preserved meaning – Japanese text case', *Transaction of Information Processing Society of Japan*, Vol. 42, No. 9, pp.2339–2350.

Niimi, M. *et al.* (2003) 'A framework of text-based steganography using sd-form semantics model', *Transaction of Information Processing Society of Japan*, Vol. 44, No. 8, August.

Petitcolas, F.A.P. (1999) 'Information hiding – a survey', *Proceedings of the IEEE*, Vol. 87, No. 7, July, pp.1062–1078.

Shirali-Shahreza, M. *et al.* (2007) 'Text steganography in SMS', *Proceedings of the International Conference on Convergence Information Technology*, November, pp.2260–2265.

Shirali-Shahreza, M.H. and Shirali-Shahreza, M. (2006) 'A new approach to persian/arabic text steganography', *Proceedings of 5th IEEE/ACIS International Conference on Computer and Information Science (ICIS-COMSAR 2006)*, 10–12, July, Honolulu, Hawaii, pp.310–315,

Solanki, K., Sullivan, K., Madhow, U., Manjunath, B.S. and Chandrasekaran, S. (2006) 'Provably secure steganography: Achieving zero K-L divergence using statistical restoration', *Proceedings of ICIP*, October, Atlanta, Georgia, USA.

Stutsman, R. *et al.* (2006) 'Lost in just the translation', *Proceedings of the 21st Annual ACM Symposium on Applied Computing (SAC'06)*, Dijon, April, France.

Topkara, M., Topkara, U. and Atallah, M.J. (2007) 'Information hiding through errors: a confusing approach', *Proceedings of the SPIE International Conference on Security, Steganography, and Watermarking of Multimedia Contents*, January.

Topkara, U., Topkara, M. and Atallah, M.J. (2006) 'The hiding virtues of ambiguity: quantifiably resilient watermarking of natural language text through synonym substitutions', *Proceeding of the 8th workshop on Multimedia and security (MM&Sec '06)*, pp.164–174.

Wayner, P. (1992) 'Mimic functions', *Cryptologia*, Vol. XVI/3, pp.193–214.

Wayner, P. (2002) *Disappearing Cryptography*, 2nd ed., Morgan Kaufmann, pp.81–128.

Winstein, K. (1999) *Lexical Steganography Through Adaptive Modulation of the Word Choice Hash*, Secondary education at the Illinois Mathematics and Science Academy, January, Available: http://alumni.imsa.edu/~keithw/tlex/lsteg.ps, Accessed on 15 April, 2008.

Winstein, K. (2008) *Lexical Steganography*, Available: http://alumni.imsa.edu/~keithw/tlex, Accessed on 03 August, 2008.

Zipf, G.K. (1968) (Introduction by Miller, G.A.) *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. MIT Press, Cambridge, MA.

Websites

An Online Exam Generator (GRE Antonyms), Available: http://www.syvum.com/cgi/online/serve.cgi/gre/verbal/antonyms7.tdf?0, Accessed on 31 July, 2008.

An Online Exam Generator at Department of Chemistry, Indiana University Northwest, Available: http://www.iun.edu/~cpanhd/cgi-bin/generator/examgenerator.html, Accessed on 27 July, 2008.

An Online Exam Generator at Department of Chemistry, Ohio State University, Available: http://lrc-srvr.mps.ohio-state.edu/under/chemed/qbank/quiz/bank1.htm, Accessed on 27 July, 2008.

An Online Exam Generator: Exams and tests such as GRE, SAT, etc., Available: http://www.ets.org, Accessed on 27 July, 2008.

An Online Exam Generator: Graduate Management Admission Test (GMAT), Available: http://www.mba.com/mba/TaketheGMAT, Accessed on 27 July, 2008.

Cambridge Dictionaries Online – Cambridge University Press, Available: www.dictionary.cambridge.org, Accessed on 31 July, 2008.

Dictionary and Thesaurus – Merriam-Webster Online, Available: www.merriam-webster.com, Accessed on July 31, 2008.

Exam Pro Software, Available: http://www.exam-software.com, Accessed on 27 July, 2008.

GRE Analogies, Available: http://greanalogies.blogspot.com/2008/04/analogies-91-95.html, Accessed on 30 July, 2008.

GRE Sentence Completions, Available: http://www.greguide.com/verbal.html, Accessed on 31 July, 2008.

Microsoft Word 97, Available: http://www.microsoft.com/en/us/default.aspx, Accessed on 31 July, 2008.

Online Dictionary Net, Available: www.online-dictionary.net, Accessed on 31 July, 2008.

ScramDisk: Free Hard Drive Encryption For Windows 95 & 98, Available: http://www.scramdisk.clara.net, Accessed on 03 August 2008.

Spam Mimic, Available: http://www.spammimic.com, Accessed on 31 July, 2007.

Testgenerator, Available: http://www.testshop.com/content.php?id=63, Accessed on 27 July, 2008.

TV show Who Wants to Be a Millionaire, Available: http://en.wikipedia.org/wiki/Who_Wants_To_Be_A_Millionaire%3F, Accessed on 27 July 2008.

Bibliography

Cvejic, N. and Seppanen, T. (2004) 'Increasing robustness of LSB audio steganography using a novel embedding method', *Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'04)*, Las Vegas, April, Nevada.

Dénes, J. and Keedwell, A.D. (1991) *Latin Squares (Annals of Discrete Mathematics)*, Vol. 46, Elsevier Science Publishing Company Inc., North-Holland, Amsterdam, January.

Laywine, C.F. and Mullen, G.L. (1998) *Discrete Mathematics Using Latin Squares*, 1st ed., Wiley-Interscience, 3 September, 1998.

Li, B., He, J., Huang, J. and Shi, Y.Q. (2011) 'A survey on image steganography and steganalysis', *Journal of Information Hiding and Multimedia Signal Processing*, Vol. 2, No. 2, April, pp.142–172.