Square Root Laws of Steganographic Capacity in the Context of Existing Models

Liam Fearnley

October 2009

Abstract

Steganography is the art and science of concealing information in other information in such a way as to render it both non-obvious and resistant to attack by malicious parties. An important summary of the limits of steganography was provided in two papers authored by Anderson and Petitcolas in the mid-late 1990s, but recent theoretical and practical results about secure steganographic capacity appear to differ from the behaviour predicted by the model proposed in this paper. We revisit the earlier work to show that there is in fact no contradiction and that the apparently contradictory results are in fact supportive of the Anderson-Petitcolas model of steganographic systems.

1 Introduction

Steganography can be loosely defined as the process of concealing information or a message in such a way that only authorised parties are aware that the information exists. The term was coined in its current form by the German abbot Johannes

Trithemius (who was also responsible for the first printed book about cryptography), in his *Steganographia*, making it slightly over 500 years old [13]. This belies a long and storied history dating back to at least 440BC, where steganographic methods are mentioned briefly in the works of the Greek historian Herodotus [6]. While in the ancient and medieval worlds, few, if any, would encounter any real need for steganographic systems (stegosystems), the advent of digital communications has brought steganographic systems into the mainstream. In the case of digital content, the rapid proliferation of data with covertly embedded steganographic data in the form of watermarks and other such identifying features has allowed rights holders to rapidly trace copyright infringement in several notable cases [10]. Steganography is also commonly used to hide data for other purposes, such as the hiding of communication or information transfer.

While the coming of this age has driven the need and adoption of steganographic techniques, it has also posed significant new problems for the field. The ability for individuals and groups to process large amounts of data quickly at minimal cost presents a major challenge, and the limits of steganographic systems, while . Ross Anderson (and Fabien Petitcolas in the later paper) present a survey of these limits in two key papers : *Stretching the limits of steganography* (1996) [3] and *On the limits of steganography* (1998)[1].

While excellent, these papers appear to avoid providing much detail on the subject of the *capacity* of steganographic systems, discussing it in general terms. Further work by Ker makes statements which appear at first to contradict these surveys, but in reality integrates nicely into Anderson's model and discussion of the limits of the art, strengthening the argument that that model is both valid and correct.

2 Definitions and General Model

2.1 Steganographic Terminology

Before proceeding it is necessary to provide definitions and our model for discussing steganographic systems that appear in the papers we survey.

Definition - Processes: A *steganographic system* or a *stegosystem* is defined as a system for the embedding of some secret information (the *embedded data*) in some other information (*cover data*) to yield *stegodata*. This process is referred to as *embedding*, and the reverse process (recovery of embedded data) as *extraction*. Additional information required to extract the embedded data is referred to as the *stegokey*. (We use terminology as defined in [12], which is standard for the discipline).

Definition - Applications: *Classical steganography* is defined as the application of a steganographic system for the purposes of obscuring the presence of some communication. *Steganographic watermarking* (or simply *watermarking* is defined as application of steganographic system for the purposes of marking content with some information.[1]

We define a base model for steganography derived from that presented in [3] and [1], and then modify it as appropriate to model both classical steganography and watermarking applications.

We have a party, Alice who wishes to place some data, E into other data, in such a way as to hide the presence of E from some adversary (or warden) Eve. To do so, Alice embeds E into some cover text C, using a stegosystem, yielding S, a set of stegodata. Eve is capable of intercepting S and we assume Eve has unlimited resources at her disposal to attack S.

Definition An attack on stegodata is referred to as *steganalysis*. Steganalysis can be either *passive*, which refers to detection of the presence of the embedded data, or *active* which refers to extraction, manipulation or destruction of the embedded data.[12]

In both the classical and the watermarking applications, passive steganalysis comprises an attempt to discern the presence of the embedded communication. In the watermarking case, where all parties may realise that there is a steganographic watermark embedded in the cover text (for example, a watermarked film), active steganalysis refers to an attack with the intent of modifying (destroying, altering, replacing) this watermark. In the classical sense, it refers to recovering the actual data embedded in the stegotext.

The model we use bears marked similarities to the information flow control model discussed in [11] which is perfectly adequate in the case of classical steganography. However, in the case of watermarking, our model does not necessarily require a second party (Bob), as Alice may be embedding data to retrieve later herself. One could model this within the constraints of Lampson's model by modifying the model such that Bob is in fact 'Future Alice' (ie, Alice at the time of retrieval), but it is simpler in this case to produce a specific model based on the intended use of the stegosystem, rather than forcing it into a general form.

In general, steganographic systems must provide both confidentiality and integrity to users [11] in order to be of use, but our primary concern here lies with the confidentiality of stegosystems. Note that while steganographic systems have notable similarities to cryptographic systems, there is a fundamental difference between the two. Informally, the purpose of a cryptosystem is to obscure the meaning of a message, as opposed to hiding the presence of such a message. The two schemes can be (and often are) combined by first encrypting E prior to embedding in C [7].

3 Steganographic Capacity in *Anderson and Petitcolas*

3.1 Image Least Significant Bit Stegosystems

A standard stegosystem used in these papers for experimental purposes is least significant bit embedding in images ([1], [8], [5], [9] all use or discuss such systems). Consider a steganographic system whose cover texts are images of various sizes specified in 24-bit colour. We wish to embed the message 'Message' into this image. We can do so by changing the least significant bits of the specification of each pixel, without perceptible change in the image using the following technique:

1. Take the binary encoding of the ASCII representation of a character from the string to be encoded. For example, $M = 77_{10} = 1001101_2$.

2. Take three pixels from the cover image. Each pixel is represented by some tuple of binary integers:

   Pixel 1: (01010101, 10001010, 11111111)
   Pixel 2: (01010101, 10001010, 11111111)
   Pixel 3: (01010101, 10001010, 11111111)

3. Change the least significant bits of the colour codes for the pixels to the bits 1001101 (embedded bits underlined):

   Pixel 1: (0101010$\underline{1}$, 1000101$\underline{0}$, 1111111$\underline{0}$)
   Pixel 2: (0101010$\underline{1}$, 1000101$\underline{1}$, 1111111$\underline{0}$)
   Pixel 3: (0101010$\underline{1}$, 10001010, 11111111)

The maximum amount of information embeddable by this system is then simply one ASCII character per three pixels - for a 1024x1024 image, we can encode 349,525 ASCII characters.

The least significant bit stegosystem described here is trivially easy to defeat in an active attack. Any form of image compression (even encoding as a JPEG or a GIF) will damage the embedded data [1]. If Alice and some other party, Bob were using this kind of image embedding as a channel for communication, all that Eve would need to do to disrupt the steganographic channel would be to intercept the stegotext, convert it into a JPEG, then back to its original format in order to largely erase the embedded data. Passive attacks on this kind of system are also relatively easy, requiring simple measurements.

As Anderson mentions, more sophisticated stegosystems do not encode data into every byte of data. An example of such a system would be a shared-key stegosystem, where Alice and Bob share some key, and use a keystream generator (as in a standard stream cipher) to generate a keystream which indicates which bytes contain data, or the order in which data is contained. This will reduce the capacity of such a scheme further, and while each type of this kind of embedding makes passive and recovery attacks more difficult, it is still vulnerable to active attacks by Eve, who simply needs to re-encode the image, or even crop it slightly, to disrupt the embedding. More complex stegosystems attempt to overcomes these problems by inserting the embedded data into slightly less obviously redundant parts of the cover data [1].

In general, these systems all have a general process in common - simply, Alice performs some transform such as compression, noise removal or transcoding on the cover text which renders some part of the cover text data redundant, in the sense that a subset of the cover text can be altered without being easily detectable by Eve.

3.2 Detecting Stegosystems

In his 1996 paper (and in a similarly worded statement in [1]), Anderson makes the following statement:

> '...the more cover text we give the warden, the better he will be able to estimate its statistics, and so the smaller the rate at which Alice will be able to tweak bits safely. The rate might even tend to zero...'[3]

According to these papers, the best bound that can be given on capacity across all stegosystems is an upper limit dependent upon the attacker. Because secure steganographic concealment may be required for a considerable length of time (in the case of copyright watermarks and digital rights management, on the order of 70+ years), Eve may have access to orders of magnitude more computational power to attack the stegosystem than the designers had at the time of implementation. This does not necessarily pose an insurmountable problem for the designers of stegosystems - the hidden information in the third book of Trithemius's Steganographia was only detected after approximately 500 years after its writing [13].

The difficulty in developing secure stegosystems lies in the fact that it is exceptionally difficult to accurately model cover sources well enough to identify redundant bits, and that an adversary at some point in the future may have the ability to better model the cover text than the stegosystem's designers, and so identify the locations where redundant data capable of containing embedded data lies. As such, the authors in [1] provide their capacity limits with respect to the attacker's ability to measure some statistical aspect of the information for different transforms, such as in parity based embeddings, where the attack is based on entropy.

3.3 Entropy and Capacity

Informally, the Shannon entropy [14] provides a quantitative metric for the amount of information present in a message, given formally by the following calculation:

For a random variable $\mathcal{X}$ with n outcomes $\{x_1, x_2...x_n\}$:

$$\begin{aligned} \text{Entropy, } H(\mathcal{X}) &= E(I(X)) \\ H(\mathcal{X}) &= -\sum_{i=1}^{n} p(x_i) log_b p(x_i) \end{aligned} \tag{1}$$

Where $E(\mathcal{X})$ is the expected value function, $I(\mathcal{X})$ refers to the information content, and $p(x)$ is the probability mass function of $\mathcal{X}$ [14].

The logarithmic term in this calculation allows for the addition of two independent entropies directly to give their entropy after combination, a property of high significance for analysis of steganographic embeddings. If we first encrypt the data prior to embedding it, it will be indistinguishable from random data taken from the same alphabet, provided that the encryption process is a reasonable one. Therefore, for some stegotext S, a cover text C and information E being embedded in C [1]:

$$H(S) = H(C) + H(E) \tag{2}$$

Alice must attempt to maintain $H(E)$ at such a level that $H(E)$ is less than the variance in Eve's estimate of $H(C)$, which severely limits the information content of E. The only way to increase the information content that can be embedded in C is to preprocess C prior to embedding in order to reduce $H(C)$ by some amount (noise reduction is suggested as a method in [1]), allowing a corresponding increase in $H(E)$ without causing a change in $H(S)$.

This requirement can be generalised for most (if not all) statistical properties of the cover data, embedded data, and stegodata. It essentially states that for

a stegosystem to be secure, the changes made to the cover data to give the stegodata must be small enough that steganalysis produces results which cannot be distinguished from detector/estimation error [3], [1].

4 *Ker* - A Square Root Law Of Capacity?

Recent work primarily led by Andrew Ker has resulted in the claim of the existence of a square root law of secure steganographic capacity. Ker's (and coauthors') work ([8],[9],[5]) has led to make the following statements, which appear inconsistent with Anderson and Petitcolas at first reading:

> ...the square root law proved for batch steganography may also apply to the case of individual covers. There are suggestions...that the square root law should also hold in rather general circumstances for Markov chains: this would be powerful additional evidence for square root capacity in general...It is not widely known that the secure capacity of a cover is proportional only to the square root of its size [in the absence of perfect steganography]...[9]

Simply put, the capacity, $\chi = \sqrt{N}$ where N is the number of embedding locations in a single piece of stegodata. We outline how the idea of a square root law was arrived at, and show how it is, in fact, non-contradictory, and in fact integrates well with the work outlined in [1].

Batch steganography refers to a type of steganography wherein rather than considering a particular instance of a cover text and its steganographic capacity with respect to its size, a set of cover texts is analysed to determine their secure capacity. [8] This then raises an important question - how does one begin to define the idea of a secure stegosystem?

The attack in this context consists of the following - Eve has access to a set of stegodata, and has an idealised detector which provides an estimate of some statistic (Ker uses two specific detector systems for discovering embeddings via least-significant bit type mechanisms like those discussed earlier in his practical work) of the cover data, and a measurement for the same statistic in the stegodata. By pooling the estimates, Eve is able to perform a statistical test, which allows her to compare her estimated values to some null hypothesis, and thus claim the presence (or absence) of embedded data.

The proof relies on analysing the statistical tests available to Eve. The three tests of the results considered in [8] are:

1. Counting positive results obtained on a collection of stegodata;
2. Taking the average of the detection statistic across the collection of stegodata;
3. Testing likelihood ratios

Each of these methods provide Eve with a statistical tool which allows her to evaluate her observations in the context of the null and alternative hypotheses. A secure stegosystem, in this context, is one capable of embedding data in a set of cover texts without causing Eve to reject her null hypothesis.

Ker proves theoretically (for each of these statistical methods) that if Eve possesses n potential stegotexts, the capacity of these stegotexts to store information grows not directionally proportional to n, but rather in proportion with $\sqrt{n}$. This theoretical result is then validated experimentally using readily available detectors [8]. This result in and of itself dovetails nicely with Anderson's statement that an attacker's accuracy of estimation of such statistics improves as the attacker has more stegotext to analyse.

Intuitively and informally, this result implies that the more stegotext that you provide to an attacking steganalyst, the better their ability to assess the statistics related to the information contained within the cover text. Ker suggests that this refers not solely to the number of stegotexts that Eve has access to, but rather the amount of stegotext available, but does not prove this as a theoretical result, only to be the case in certain experimental situations [9].

Ker's definition of a secure system with respect to a single stegotext uses a set of scalar metrics which independently suggest the security of a system. Eve is assumed to have a binary classifier (either a message is embedded or it is not) for steganographic material, and the security metrics are based on the performance of this detector. The higher the value of these metrics, the better the performance of Eve's detector, and thus the less secure the stegosystem under analysis. Tests are run in accordance with the three methods outlined in the theoretical analysis, and secure capacity of a steganographic object in this context simply refers to the amount of information that can be embedded into a single cover text without detection [9].

As in the theoretical results, the capacity determined experimentally by Ker this capacity appears to be limited by a square root law.

Subsequent studies of the capacity problem under certain specialised conditions have provided further evidence to support Ker's suggestion that a square root law applies to secure capacity in a general classical (ie, non-batch) steganographic situation.

Work by Filler, Fridrich, and Ker (in [5] and [4]) sets out a situation where the cover data is defined as being a first order Markov Chain. By doing so, they are able to neatly sidestep the problem of having an adversary obtain a more accurate model of their cover data. They also specify a general type of embedding technique to be employed, and in doing so, are able to provide a theoretical proof

that the security of their system depends entirely upon the amount of information embedded, and that the most secure system possible under these constraints again sees data embedded at a rate lower than $\frac{1}{\sqrt{n}}$. These assumptions hold true for most practically used cover texts, and a significant proportion of embedding methods ([5]), so the theoretical result has significant practical utility.

While this result does not constitute a strong theoretical proof applicable over all steganographic systems or situations, the consistent appearance of square root laws of this type across multiple stegosystems of different types, with multiple statistical tools used by an attacking steganalyst strongly suggests that such a law is applicable to steganographic embeddings in general. How, then, does this reconcile with Anderson's claims?

5 Conclusions

Recall that the capacity of a steganographic object is dependent entirely upon Eve's ability to estimate the underlying statistical properties of some cover text used to produce a stegotext (as outlined in the example of Shannon entropy). Eve is capable of measuring the statistical properties of a stegotext directly, so any statistically significant difference between the estimate of cover text properties and the stegotext measurements indicates a discrepancy that could indicate the presence of steganographically embedded information.

Informally, Ker's square root law that the secure capacity of a stegosystem grows proportionally with $\sqrt{N}$ attempts to maintain the amount of embedded data within the limits of Eve's ability to accurately estimate the statistics of the underlying coverdata.

To return to the example of entropy that we take from [1], the variance in Eve's estimate or measurement of $H(C)$ decreases as the amount of stegodata

available to Eve increases. The square root law for secure capacity attempts to ensure that the amount of data embedded in the system is within levels that would be considered statistically insignificant, ensuring the security of the stegosystem against passive attack.

Viewed in this context, the square root law for secure steganographic capacity is entirely consistent with the model laid out in [1] and [3]. This finding strengthens the model proposed by [1] rather than contradicting it, by providing further evidence for its validity through the ability of this result, which is novel in the context of the original paper, being able to integrate well into the prior work.

References

[1] Ross Anderson and Fabien Petitcolas. On the limits of steganography. *IEEE Journal of Selected Areas in Communications*, 16:474–481, 1998.

[2] Ross J. Anderson, editor. *Information Hiding, First International Workshop, Cambridge, U.K., May 30 - June 1, 1996, Proceedings*, volume 1174 of *Lecture Notes in Computer Science*. Springer, 1996.

[3] Ross J. Anderson. Stretching the limits of steganography. In *Information Hiding* [2], pages 39–48.

[4] Tomás Filler and Jessica J. Fridrich. Fisher information determines capacity of -secure steganography. In Stefan Katzenbeisser and Ahmad-Reza Sadeghi, editors, *Information Hiding*, volume 5806 of *Lecture Notes in Computer Science*, pages 31–47. Springer, 2009.

[5] Tomáš Filler, Andrew D. Ker, and Jessica Fridrich. The square root law of steganographic capacity for markov covers. In Edward J. Delp III, Jana

Dittmann, Nasir D. Memon, and Ping Wah Wong, editors, *SPIE Media Forensics and Security*. SPIE, 2009.

[6] Herodotus. *The History of Herodotus*. MacMillan and Co, 1890. `http://www.gutenberg.org/dirs/2/7/0/2707/2707.txt`.

[7] Stefan Katzenbeisser and Fabien A. Petitcolas, editors. *Information Hiding Techniques for Steganography and Digital Watermarking*. Artech House, Inc., Norwood, MA, USA, 2000.

[8] Andrew D. Ker. Batch steganography and pooled steganalysis. In Jan Camenisch, Christian S. Collberg, Neil F. Johnson, and Phil Sallee, editors, *Information Hiding*, volume 4437 of *Lecture Notes in Computer Science*, pages 265–281. Springer, 2006.

[9] Andrew D. Ker, Tomás Pevný, Jan Kodovský, and Jessica J. Fridrich. The square root law of steganographic capacity. In Andrew D. Ker, Jana Dittmann, and Jessica J. Fridrich, editors, *MM&Sec*, pages 107–116. ACM, 2008.

[10] D Kravets. Watermarking could lead to x-men uploader, April 2009. `http://www.wired.com/threatlevel/2009/04/watermarking-co/`.

[11] Butler W. Lampson. Computer security in the real world. *Computer*, 37(6):37–46, 2004.

[12] Birgit Pfitzmann. Information hiding terminology - results of an informal plenary meeting and additional proposals. In Anderson [2], pages 347–350.

[13] J Reeds. Solved: The ciphers in book III of Trithemius' Steganographia. *Cryptologia*, XXII(4):291–318, October 1998.

[14] C. E. Shannon. Prediction and entropy of printed english. *Bell Systems Technical Journal*, 30:50–64, 1951.