2nd Workshop on Information Hiding: April 1998, Portland, LNCS 1525, Springer-Verlag, pp. 345-355

Modeling the security of steganographic systems[♣]

J.Zöllner*, H.Federrath**, H.Klimant**, A.Pfitzmann**, R.Piotraschke**, A.Westfeld**, G.Wicke**, G.Wolf*

Dresden University of Technology, 01062 Dresden, Germany
*Institute for Operating Systems, Databases and Computer Networks
**Institute for Theoretical Computer Science
{zoellner, federrath, pfitza, westfeld, wicke, g.wolf}@inf.tu-dresden.de
{klimant, pi}@tcs.inf.tu-dresden.de

Abstract. We present a model of steganographic systems which allows to evaluate their security. We especially want to establish an analogy to the known-plaintext-attack which is commonly used to rate cryptographic systems. This model´s main statement is that the embedding operation of a steganographic system should work **indeterministic** from the attacker´s point of view. This is proved by means of information theory.
Index Terms: Security and modeling of steganography, entropy, indeterminism, secret communication, hidden communication

1 A short introduction to steganography

Bruce Schneier characterizes steganography in the following way [1]: "Steganography serves to hide secret messages in other messages, such that the secret´s very existence is concealed." He also states some historic examples, such as "…invisible inks, tiny pin punctures on selected characters, minute differences between handwritten characters, pencil marks on typewritten characters, …".

These examples show that steganography itself is not a new technique. However, it experiences a renaissance due to the ubiquitious use of computers and multimedia; especially when graphical and audio data are involved. Consequently, most available implementations of steganographic algorithms work on graphics or sound.

In Figure 1 we illustrate the use of steganography on images.

♣ The work was sponsored by the German Ministry of Education, Science, Research and Technology and the Gottlieb-Daimler- and Karl-Benz-Foundation Ladenburg (Germany).

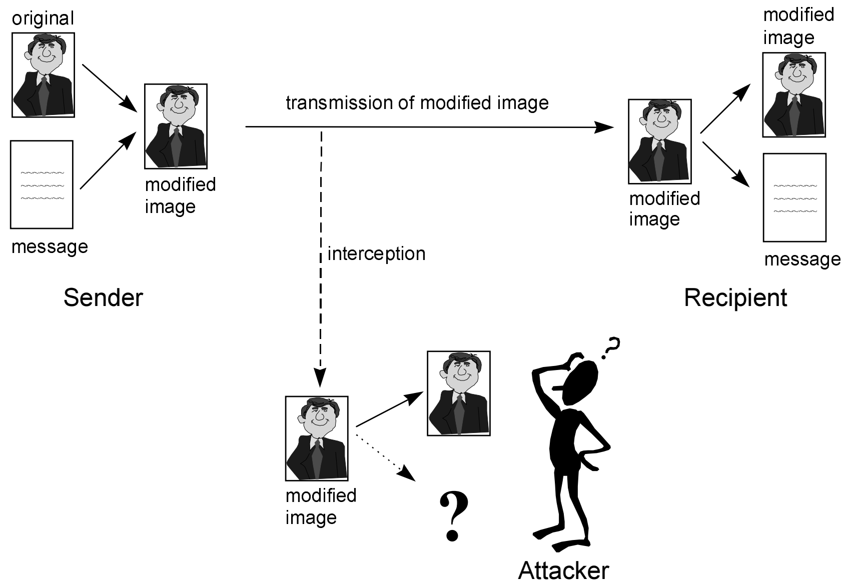


Fig. 1. Steganography with graphical data

On the left you can see the sender, who embeds the secret message into a graphic file (the "original"). She then transmits this modified file (here named "modified image") to the recipient shown on the right side. The attacker (at the bottom) intercepts this transmission. Only the recipient should be able to extract the message in the correct way. Of course this is possible only if there is a shared secret between the sender and the recipient. This could be for instance the algorithm for extraction itself or special parameters of the algorithm, e.g. keys.

2 Steganography vs. cryptography

How do steganography and cryptography compare? The purpose of both is to provide secret communication. Cryptography hides the contents of a secret message from an attacker, whereas steganography even conceals the existence of this message. Therefore the definition of breaking the system is different. In cryptography, the system is broken when the attacker can read the secret message (for the point under discussion it does not matter how he does this).

Breaking a steganographic system has two stages:

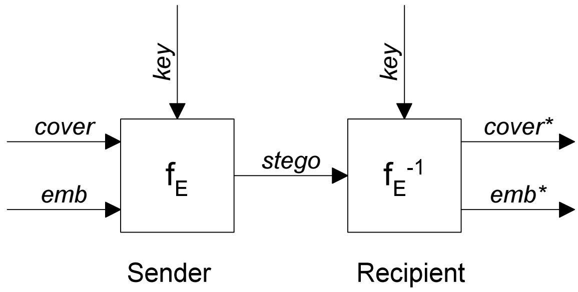
1. The attacker can detect that steganography has been used.
2. Additionally, he is able to read the embedded message.

In our definition a steganographic system is insecure already if the detection of steganography is possible (first stage).

3 Related work

3.1 The basic model of a steganographic system

The model in Figure 2 is based on the results of the discussions at the Information Hiding Workshop in Cambridge [2], which were continued in [3] and [4]. We will call this model the Embedding Model.



f_E: steganographic function "embedding"
f_E^{-1}: steganographic function "extracting"
cover: coverdata in which *emb* will be hidden
emb: message to be embedded
key: parameter of f_E
stego: coverdata with embedded message

Fig. 2. The Embedding Model

The input *cover* represents the untreated original data, *emb* the ones which will be embedded into *cover* by the function f_E. The resulting data called *stego* contain the message *emb*. The operation f_E^{-1} extracts the embedded data to *emb** and also produces an output *cover**. Naturally, *emb** should be equal to *emb* and in most cases *cover** is the same as *stego*. For concelation systems, *cover** is not of much interest anyway.

This model was not meant to be a model for evaluating the security of steganographic systems (or stegosystems for short) by the participants of the workshop. They merely tried – for a beginning – to put the ad-hoc knowledge of steganography into a more abstract form, for which purpose the figure shown serves quite well.

Therefore the above model is not of much use if you want to evaluate the security of a steganographic system. You can see the acting entities, the processing functions and their in- and output, as described in Chapter 1. But there are no comments on the behavior of the function f_E and the knowledge and capabilities of possible attackers.

3.2 Information theoretic setting

In the following we will evaluate the model from Chapter 3.1 by means of information theory. In "On the limits of steganography" [5] there is a chapter which addresses this approach. The authors argue with the entropy of *cover*, *emb* and *stego*, just like we will do, but don´t go further into detail. They had a different goal with their paper: while we would like to present a commonly valid model for steganographic systems (and prove its validity by means of information theory), they do concentrate on the practical issues of steganography. Consequently the mentioned chapter is rather short and does not contain an actual proof for the (of course reasonable) statements which are made.

Another interesting approach to information theoretic evaluation of steganography can be found in [6].

4 Using information theory to evaluate the security of steganographic systems

Borrowing from cryptology, we introduce two forms of attacks on steganographic systems:

1. stego-only-attack: The attacker knows only *stego*.
2. stego-cover-attack: The attacker knows both *stego* and *cover*.

We will concentrate on the second attack in analogy to the known-plaintext-attack on cryptographic systems where the attacker is allowed to know every in- and output except the key and still should not be able to break the system. In addition, the first attack is a special case of the second and thus included in our further considerations.

It seems obvious that the attacker can detect differences between *cover* and *stego* if he gets to know both of them. If the differences are caused only by f_E he can break the system easily. To avoid this, the first solution is: The attacker must not know *cover*. This can be proved by means of information theory:

The embedding process can be described as the function

$$stego = f_E(cover, emb, key).$$

We assume a *cover* of *m* bits in which we want to "hide" *n* bits and the following notation:

C	the set of all bitstrings
cover	actual bitstring of length *m* ($cover \in C$)
E	the set of all bitstrings
emb	actual bitstring of length *n* ($emb \in E$)
K	the set of all keys
key	actual key ($key \in \mathrm{K}$)
S	the set of all bitstrings ($S = C$)
stego	actual stego, i.e. bitstring that contains *emb* ($stego \in S$)

For a given alphabet X the entropy $H(X)$ describes the "uncertainty about X". That actually means the uncertainty about the occurrence of a certain element $x \in X$ [1]. The conditional entropy $H(X|Y)$ is the remaining uncertainty about X when knowing Y. The joint entropy $H(X,Y) = H(X) + H(Y|X)$ is the "union" of both entropies. The mutual information $I(X;Y)$ describes the amount of information about X you get if you know Y; $I(X;Y) = H(X) - H(X|Y)$ [7].

The attacker does

- suppose that some *emb* is hidden in *stego*,
- know the steganographic functions,
- have the knowledge and abilities to perform an attack on the stegosystem,
- have unlimited time and resources.

If in spite of all his efforts the attacker can not confirm his hypothesis that *emb* is hidden we will call the system "information theoretically secure".

4.1 Why deterministic steganography can´t be secure

The stegosystem is information theoretically secure if the attacker cannot gain any information about *emb* or E by examining *stego* and *cover* (or S and C, respectively). Thus, the mutual information is zero:

$$I(E;(S,C)) = H(E) - H(E|(S,C)) = 0. \qquad (\mathbf{1})$$

That gives the fundamental security condition:

$$H(E|(S,C)) = H(E). \qquad (\mathbf{2})$$

That means that the uncertainty about E – the entropy $H(E)$ – must not be decreased by the knowledge of S and C. Conclusion: E has to be **independent** from S and C.

Is it possible to meet this condition? It seems logical that – with the given assumptions – an attacker gains knowledge about a hidden *emb* just by comparing the corresponding *cover* and *stego*. We can assume that not only the alphabets S and C but also their entropies $H(S)$ and $H(C)$ are equal. There are differences in the conditional entropies, however:

- without embedded information: $H(S|C) = H(C|S) = 0$,
- with embedded information: $H(S|C) = H(C|S) > 0$.

The connection of uncertainty and information allows us to say: the uncertainty about S, if we know C (or vice versa) corresponds to the information about E that you can get by looking at S and C. Therefore, by embedding $emb \in E$ into $cover \in C$ we have a mutual information

$$I(E;(S,C)) = H(E) - H(E|(S,C)) > 0. \qquad (\mathbf{3})$$

[1] Keep this relationship in mind when we partly look on only the alphabets in the following.

It follows:

$$H(E|(S,C)) < H(E). \tag{4}$$

This means that the security condition is not fulfilled. Therefore, the necessary and sufficient condition for secure steganography is:

$$H(S|C) = H(C|S) = 0. \tag{5}$$

This condition can be met only when

$\forall i \in \mathrm{N}$, $stego_i \in S$, $cover_i \in C$: $stego_i = cover_i$.

Thus the steganography is reduced to a practically irrelevant special case[2]. If we exclude this case, it follows:

The security condition (2) can not be fulfilled under the given assumptions. This basically means that secure steganography is impossible when both *cover* and *stego* are known to the attacker.

4.2 Indeterminism and steganography

An advanced solution to this problem is to have an indeterministic embedding operation. An indeterministic operation or process gives different results (within a certain range) every time it is computed. In other words, it contains randomness. Information theory supports this approach:

As stated above, it is impossible to provide information theoretically secure steganography if the attacker knows *cover* and *stego* (respectively C and S). Therefore we establish the following condition: When the attacker knows S, there remains an uncertainty about C, so that $H(C|S) > 0$. For that we introduce a new alphabet from which the actual *cover* is selected. We call this alphabet C_S or *Source*.

The effect of introducing C_S into the Embedding Model is shown in Figure 3. We assume that f_E, C_S and S (or *stego*) are publicly known, whereas K and C (respectively *key* and *cover*) are unknown to attackers.

Since the actual cover is selected from C_S, we assume $C \subseteq C_S$. In addtion, we assume $H(C_S) \geq H(C)$, which is both plausible for any selection and neccesary to achieve the intended indeterminism. It says that the uncertainty about the realisation of an actual *cover* from C_S must be greater than or equal to that about a realisation from C.

---

[2] This case is *cover* $\equiv$ *stego*: You have to find a *cover* that already contains *emb*.

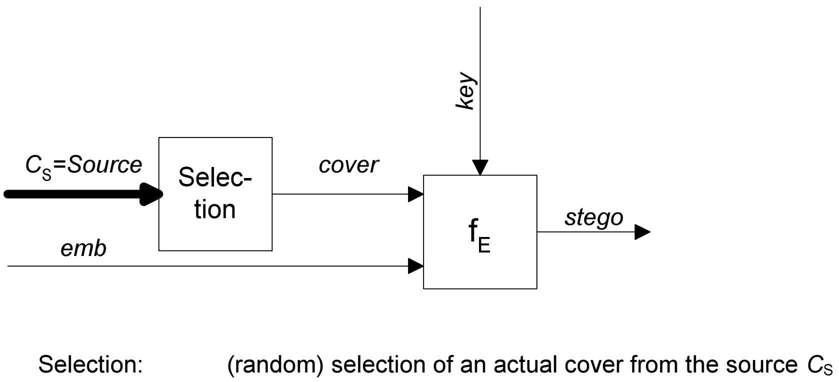


Fig. 3. Selection of covers from a source

The necessary uncertainty about C is then achieved by selecting every *cover* in a truly random process and keeping it secret afterwards. One example for such a process is the sampling of analog input, e.g. speech or images. The inaccuracy of the quantization provides the needed uncertainty. If the changes of *cover* during the embedding process remain within this range, the manipulations cannot be detected.

In analogy to the proof given before the fundamental security condition is:

$$H(E|(S,C_S)) = H(E), \qquad (\mathbf{6})$$

what means that the uncertainty about E – the entropy $H(E)$ – must not be decreased by the knowledge of S and C_S, in other words: E has to be independent from S and C_S.

How can this condition be fulfilled? The attacker should not be able to detect changes in *cover* which are due to the embedding process by examining *stego*. Therefore we need a certain amount of uncertainty about *cover*, what means $H(C|S) > 0$. The necessary amount results from the relation between conditional entropy and mutual information:

$$H(C|S) \geq I(E;(S,C)) = H(E) - H(E|(S,C)). \qquad (\mathbf{7})$$

If we assume the worst case that the attacker can determine E completely from S and C, we get:

$H(E|(S,C)) = 0.$

It follows that

$$H(C|S) \geq H(E). \qquad (\mathbf{8})$$

This can be interpreted in the following way: Because the mutual information can be at most the size of $H(E)$, the necessary uncertainty about C must be at least the same size to make an attack on S impossible.

The same applies of course to attacks on C_S. Therefore we assume

$$H(C|C_S) = H(C|S) \qquad (\mathbf{9})$$

and

$$H(C|C_s) \geq H(E). \tag{10}$$

With these conditions we need a joint entropy

$$H_o = H(C,C_s) = H(C) + H(C_s|C). \tag{11}$$

Because $C \subseteq C_s$ and $H(C_s) \geq H(C)$, it follows

$H(C_s|C) \geq H(C|C_s)$.

When considering these relations we get a lower bound for the necessary joint entropy:

$H_o \geq H(C) + H(C|C_s)$.

Using (10), we get

$$H_o \geq H(C) + H(E). \tag{12}$$

Since $H(C_s) \geq H(C)$, we assume $H(C_s,S) \geq H(C,S)$. From this follows:

$$H(C_s|S) \geq H(C|S). \tag{13}$$

According to Equation (8) it follows that the security-relevant bound

$$H(C_s|S) \geq H(E) \tag{14}$$

can be met. We may draw the following conclusion: When you observe the lower bounds for $H(C|S)$ (Equation (8)) and $H(C|C_s)$ (Equation (9)), attacks with knowledge of S and C_s (stego-source-attack) to prove the existence of E in S are not successful: The fundamental security condition (6) can be fulfilled.

Additionally we look at the conditions under which the stegosystem is secure when it is attacked via K. Therefore we require that an attacker (who knows S and C_s) can not obtain any information about (K, E)[3]. This can be expressed as follows:

$$\begin{aligned} I((K,E);(S,C_s)) &= H(K,E) - H((K,E)|(S,C_s)) = 0 \qquad (15)\\ &= H(K,E) - H(K|(S,C_s)) - H(E|(S,C_s,K)) = 0. \end{aligned}$$

When taking into account that $H(E|(S,C_s,K)) = 0$, we get:

$H(K|(S,C_s)) = H(K,E)$

or

$$H(K|(S,C_s)) = H(E) + H(K|E) \geq H(E), \tag{16}$$

respectively.

We can conclude from the proof above that cover must contain an uncertainty for the attacker to allow secure steganography between sender and recipient.

Furthermore the proof shows that information theoretically secure steganography is possible, if two conditions are met:

---

[3] Although this requirement is actually too strong, it can be chosen for the theoretic approach, because we weaken it in the result (see Equation (14)).

1. Knowledge of C_S and S must not decrease the uncertainty of an attacker about E and *emb* (see Equation (6)):
   $H(E|(S,C_S)) = H(E|S) = H(E)$.
   To achieve this, the following constraints (compare Equations (12), (8) and (14)) apply:
   $H_O = H(C,C_S) \geq H(C) + H(E)$,
   $H(C|S) \geq H(E)$,
   $H(C_S|S) \geq H(E)$.
2. The conditional entropy of the key must be greater or equal to $H(E)$ to prevent an attack via K (see Equation (16)):
   $H(K|(S,C_S)) \geq H(E)$.

A third condition can be established (receiver condition): For the receiver (who knows $key \in K$) there must not be any uncertainty about $emb \in E$:

$H(E|(S,C_S,K)) = H(E|(S,K)) = 0$.

5 Introducing indeterminism into the steganographic function

As we have seen in the previous chapter, the embedding has to be indeterministic to provide security against attackers who get to know the in- and output of the system.

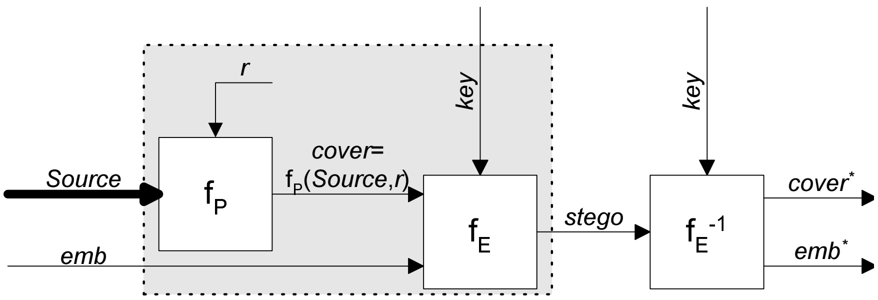
The solution is to split the steganographic process "embedding" into an indeterministic and a deterministic part. These parts must not be distinguishable by the attacker and therefore have to take place in a trusted domain of the sender.

When, for example, the sender takes a digital image with an electronic camera, an attacker may know the scene which is depicted and the camera exactly. But the attacker – and even the sender – does not know the position of the camera and the direction it was pointed in a sufficiently exact manner (turning the camera even by a fraction of a degree results in a different image [8]). Thus, if the attacker gets hold of a digital image he is unable to decide whether the picture is an original one or has been treated by a steganographic system.

Another example is the sampling of analog data we already mentioned in Chapter 4. When the sender takes samples of one analog waveform several times, he is extremely unlikely to get exactly the same digital data each time. This phenomenon is due to the inaccuracy of analog-digital-converters and the characteristics of quantization. If the function f_E mimics the process of sampling in a sufficiently exact manner, nobody without the key is able to decide whether a given digital sample contains steganographic data or not.

In both cases the **preprocessing** f_P (e.g. camera positioning, sampling) introduces randomness into the cover data. f_P can even be a product of multiple operations. The first of the above cases gives an example: It contains a sampling step, too.

The considerations so far lead us to a refined model of steganography which features a preprocessing f_P in addition to the embedding function f_E:



f_P: preprocessing

r: random part of *cover* introduced by f_P (may be parameter of f_P, but does not have to)

Fig. 4. Model with preprocessing

The difference to the Embedding Model shown in Figure 2 is the enhanced view on steganography. We no longer concentrate on only the steganographic core function f_E but look at the whole steganographic system. In the Embedding Model the attacker simply must not know *cover* for secure steganography. This is true for the model above as well, but with the parameter *Source* we model the uncertainty of the attacker about *cover*. This allows a more exact evaluation of steganographic systems.

The gray area in Figure 4 marks a trusted domain which the attacker cannot intrude. You can see that he may know *Source* and *emb* and still should be unable to detect the steganography if he does not know *key*.

We would like to illustrate the concept of the trusted domain with an example: the ISDN telephony network. ISDN gives bit-transparent transport of digital data, which is crucial for secret communication with steganography. Imagine an ISDN telephone with analog-digital conversion and steganography integrated into one chip: Naturally the analog sound is *Source,* the digitized speech is *cover* and the voice-sampling is the preprocessing f_P. Whether the output of the integrated chip is *stego* or *cover* cannot be determined from the outside. An attacker may know *Source*, *emb* and even the characteristics of the sampling chip, but it does not help him to decide whether the output contains *emb* or not.

Now imagine the sampling and the steganography on different chips. If the attacker is able to eavesdrop the in- and output of the steganography chip alone, he is of course able to detect the use of steganography because he knows the actually used *cover*. Therefore, both chips have to be inside a trusted domain of the sender. In the first case the integrated chip serves as this trusted domain.

It seems obvious that the embedding function f_E should be implemented according to the characteristics of the preprocessing. If, for instance, f_P introduces white noise into the least significant bits of *cover* (as most analog-digital-converters do), then the embedding should spread *emb* over these bits of *cover* in a way that resembles white noise. Other processes may require completely different embedding techniques.

6 Conclusions

We can name two necessary conditions for secure steganography:
1. *Key* remains unknown to the attacker.
2. The attacker does not know the actual *cover*.

How can we guarantee this? The concealment of *key* corresponds to the one of symmetric cryptosystems. The second point is at first sight simply a condition to be met. As an alternative we can assume a set of input data named *Source*, from which the stegosystem chooses the actual *cover*. The attacker knows only *Source*. This model is well suited for the implementation of actual steganographic systems because the embedding can be tuned to f_P. The embedding exploits the randomness introduced by f_P and thus provides secure steganography. To implement a good steganographic function you naturally have to have as much knowledge about the preprocessing as possible.

References

1. B. Schneier, *Applied Cryptography*, 2nd ed. New York: John Wiley & Sons, 1996, p. 9.
2. B. Pfitzmann, "Information Hiding Terminology". In R. Anderson, *Information Hiding: first international workshop, Proceedings (Lecture notes in computer science; Vol. 1147)*, Berlin: Springer, 1996.
3. J. Zöllner, H. Federrath, A. Pfitzmann, A. Westfeld, G. Wicke, G. Wolf, "Über die Modellierung steganographischer Systeme". In G. Müller, K. Rannenberg, M. Reitenspieß, H. Stiegler, *Verläßliche IT-Systeme. Zwischen Key-Escrow und elektronischem Geld*, Friedr. Vieweg & Sohn Verlagsgesellschaft mbH, Braunschweig/Wiesbaden, 1997, pp. 211-223.
4. H. Klimant, R. Piotraschke, "Informationstheoretische Bewertung steganographischer Konzelationssysteme". In G. Müller, K. Rannenberg, M. Reitenspieß, H. Stiegler: *Verläßliche IT-Systeme. Zwischen Key-Escrow und elektronischem Geld*, Friedr. Vieweg & Sohn Verlagsgesellschaft mbH, Braunschweig/Wiesbaden, 1997, pp. 225-232.
5. R. Anderson, F. Petitcolas, "On the limits of steganography". To be published in *IEEE Journal on Selected Areas in Communications, Special Issue on copyright and privacy protection.* Available at: http://www.cl.cam.ac.uk/ftp/users/rja14/steganjsac2.ps.gz
6. C. Cachin, "An Information-Theoretic Model for Steganography". In *Information Hiding: second international workshop, Preproceedings*; 15-17 April 1998; Portland, Oregon.
7. R. G. Gallagher, *Information Theory and Reliable Communication*. John Wiley & Sons, 1968.
8. A. Westfeld, G. Wolf, "Steganography in a Video Conferencing System". In *Information Hiding: second international workshop, Preproceedings*; 15-17 April 1998; Portland, Oregon.