

Content-Aware Steganography: About Lazy Prisoners and Narrow-Minded Wardens

Richard Bergmair and Stefan Katzenbeisser

December 2005

Abstract

We introduce content-aware steganography as a new paradigm of steganography stemming from a shift in perspectives towards the objects of steganography. In particular, we abandon the point of view that steganographic objects can be considered pieces of *data*, suggesting that they should rather be considered pieces of *information*. We provide some evidence to suggest that this shift in perspectives is in fact necessary, and pinpoint a *semantic problem* that has not received sufficient attention in the past. We also propose a solution to this problem, by putting forward a new kind of steganography that employs human interactive proofs as a security primitive.

Keywords content-aware, semantic, steganography, information hiding, cryptography, human interactive proof, HIP, CAPTCHA, human, machine, artificial intelligence, AI, AI-complete security primitive.

1 Introduction

In his 1984 landmark paper [25], Gustavus Simmons illustrated what is now widely known as steganography in terms of the *prisoners' problem*: Two accomplices in a crime, Alice and Bob, are arrested in separate cells. They want to coordinate an escape plan, but their only means of communication is by way of messages conveyed for them by Wendy the warden. Should Alice and Bob try to exchange messages that are not completely open to Wendy, or ones that seem suspicious to her, they will be put into a high security prison no one has ever escaped from.

We believe, that this communication setup is more widely deployed today than ever before. Alice's and Bob's problem resembles that of citizens trying to exercise their right to freedom of opinion and expression in a society that seeks disclosure of all private communication. Unfortunately cryptography provides only a partial solution to this problem, since its use is easily detected and is thus pointless under a legislation that does not tolerate it – as is in place in Russia, Belarus, Moldova, China, Pakistan, and South Africa [17]. Law enforcement in those countries, on the other hand, resembles Wendy's problem. The scenario is of increasing significance also for western nations, as advocates of anti-crypto-laws are gaining support in reaction to recent acts of terrorism. [19]

Simmons' solution to the prisoners' problem is based on the observation that Alice and Bob "will have to deceive the warden by finding a way of communicating secretly in the exchanges, i.e., of establishing a 'subliminal channel' between them in full view of the warden, even though the messages themselves contain no secret (to the warden) information" [25].

Considering that Alice is trying to convey a particular piece of information which is represented as a single datagram available to both Wendy and Bob, the idea that the message contains different information to Wendy than to Bob deserves some further elaboration. In particular note, that Simmons is using the term *information* in its strictest sense, as in the context of the well-known data-information-knowledge hierarchy. [1, 31]. The major idea is that data becomes information only under interpretation and information becomes knowledge only under understanding. Thus, one need not be a cryptologist to appreciate the fact that a piece of *data* which is purely symbolic in nature, is inherently ambiguous in the sense that it carries different information under different interpretation.

A subliminal channel is then simply one that transmits datagrams that have at least two possible interpretations. Each datagram is intentionally given an obvious interpretation (the cover) that is innocuous to Wendy, and a non-obvious interpretation (the secret) that is suspicious to Wendy, and thus could not be transmitted in plain sight. The security of the system then usually relies on some assumption of an advantage that Bob has over Wendy, when it comes to the interpretation of the message, so that Bob can interpret the message with regard to its secret meaning, and Wendy can only interpret the message as the cover. For example Bob might know when and where Alice is hiding data, or he might have exchanged a secret key with Alice before imprisonment.

In the past, many stegosystems have been constructed, most of them using digital audio or video as cover. A simple technique, often used for demonstrative purposes, is that of LSB-substitution in bitmap images. Bitmaps can represent digital images, say with 24 bits of color-depth, using three bytes to encode the color of a pixel, one for the strength of each a red, a green, and a blue light-source producing the color under additive synthesis. If we randomly toggle the least significant bit (LSB) of each of these bytes, it will result in the respective color-component of half of the pixels deviating in light strength by one in 2^8 units. As long as Alice and Bob do not expect Wendy to perceive the difference between a pixel's original color, and the same pixel after Alice has used it to encode a secret, thereby possibly making it one of 256 degrees more, say, reddish, they can in fact use this technique to establish a subliminal channel.

This goes well with our idea of steganography. A bitmap image acts as a cover, because it has an obvious interpretation, which is visual perception by a human user of the pattern that appears on screen when it is opened in his favourite image viewer. The image also has a non-obvious interpretation, which is to extract the LSBs and view their concatenation, say, in a hex-editor. Under the assumption that Alice constantly sends Bob bitmap images that Wendy is not willing to wade through with a hex-editor, the system might be attributed some kind of security.

2 Practical Lessons Learned in the Steganography vs. Steganalysis Arena

However, Wendy will not have to wade through all the images with a hex-editor, when a computer can automatically analyze them to gain knowledge of a subliminal channel. In this paper we will propose that Bob might be considered human, and Wendy a computer, which can also be a great advantage of Bob over Wendy.

To understand the way such an analyzer may work, it is important to bear in mind, that a bitmap image is not just a sequence of bytes, but rather a representation for specific semantic content. It could, for example, be a vector drawing consisting of uniformly colored geometric shapes. If a set of pixels can be identified as representing, say, an oval shape colored in a certain tone of blue, and half of these pixels deviate by their color in the LSB, this might give us some evidence of steganography taking place. A bitmap might also represent a paintbrush drawing. Since paintbrush has a color palette, we would expect the resulting image to use those colors most of the time that appear in the palette. If, however, all pixels happen to have paintbrush palette colors, with half of them deviating in the LSB, then some form of LSB substitution may have been involved. A 24-bit bitmap might also be a photograph taken by a digital camera with a CCD that leaves noise with special characteristics in the images. If these characteristics cannot be found in the image's LSBs, then, again, we have gained evidence to suspect that steganography is taking place. [20] These are some of the reasons why LSB substitution is considered completely insecure today. [12]

We believe the way in which LSB substitution has been compromised is stereotypical for how the steganography vs. steganalysis battle is usually fought, namely by steganalysis exploiting the false assumption made by steganography that a meaningful digital object can be specified solely in terms of syntactic properties. Stegosystems are usually broken by exploiting semantic inconsistencies introduced into the cover when hiding a secret. Thus, the battle between steganography and steganalysis is won (usually by steganalysis) not on grounds of mechanistic analysis of the syntactic properties of a steganogram, but on grounds of the ability to understand the content of a steganogram, not due to raw computational power, but by virtue of an accurate semantic model. This is why it is high time to bring the semantic dimension into the theoretic picture of steganography.

3 Theoretic Limits of the Old Paradigm

Up to now steganography has been a rather technical discipline. Much work has been devoted to the construction and breaking of different stegosystems (especially those used for media watermarking) but little has been done to develop a coherent theory underlying that practical work. In the recent past, however, we've seen some attempts at formalizing models for secure steganography. They all follow the same paradigm: Alice and Bob turn a suspicious secret into an innocuous-looking steganogram by way of some syntactic transformation. Wendy has strict formal criteria by which to mechanistically distinguish truly innocuous covers from the steganograms used by Alice and Bob. The

security of such a stegosystem is then defined in terms of the information theoretic uncertainty [6, 21, 7] or the computational complexity [15, 28, 2] that is involved in Wendy’s syntactic distinction problem.

Clearly, if, as a result of some mechanistic analysis of the syntax of a steganogram, Wendy can distinguish a steganogram from the noise she would usually expect in its place, she has gained suspicion of steganography going on, and has rendered the system insecure. So for a stegosystem to be secure, it is necessary for the syntactic distinction problem it poses to be hard. We agree with approaches in the tradition of the previous models, in that they can impose a necessary condition on the security of a system and therefore an upper bound on the level of security we can expect of it. However, we would like to raise the question whether, in the light of semantically meaningful steganograms, a formalistic model does (or even *can*) exist, that adequately imposes a sufficient condition in terms of a purely syntactic model, and therefore a lower limit on the security of a stegosystem relying on mechanistic processing only.

We will use Cachin’s information theoretic characterization of steganography [7] to illustrate our point. In Cachin’s model there is an alphabet \mathcal{Q} containing all symbols that can be exchanged over the channel established between Alice and Bob by Wendy. Random variables S and C are used to represent draws of a symbol q from \mathcal{Q} , according to the probability distributions P_S and P_C respectively, such that $P_S(q)$ is the probability for Alice to submit q in order to convey a hidden message and $P_C(q)$ is the probability for her to submit q as an open message. Once the alphabet and the distributions have been fixed, one can use the Kullback-Leibler distance (also called relative entropy or discrimination)

$$D(P_C||P_S) = \sum_{q \in \mathcal{Q}} P_C(q) \log \frac{P_C(q)}{P_S(q)}$$

to measure the security of the system. Intuitively $D(P_C||P_S)$ is a measure for the inefficiency of assuming that a distribution is P_S where the real distribution is P_C . More formally it is a convex function of P_C , always nonnegative, and zero only if $P_C = P_S$. Although it is not a distance in the mathematical sense, it is useful to think of it as one, and important to bear in mind that $D(P_C||P_S)$ measures the *insecurity* of a system, rather than the security, since a system is considered perfectly secure by Cachin if this measure becomes zero.

We will now use this model to track the security of a stegosystem through its lifecycle in the steganography vs. steganalysis battle. What we have in mind is to construct a stegosystem that encodes secrets into random English texts. We assume that Wendy generally tolerates English text as a medium of communication between Alice and Bob and neglect the possibility that a piece of English text generated at random may be suspicious to Wendy.

Let \mathcal{A} denote the well known English alphabet (so that $|\mathcal{A}| = 26$) and let \mathcal{A}^* be the set of finite sequences of symbols chosen from \mathcal{A} . Clearly every English language text (maybe after it has undergone some insignificant preprocessing) is in \mathcal{A}^* . Following the current paradigm in steganography we will also make the converse assumption that every element of \mathcal{A}^* is a meaningful English language text, in much the same spirit that may mislead many to make the false assumption that the set of $y \times x$ matrices of RGB-coded

color values is the set of all meaningful images of these dimensions. Obviously this is not the case since a $y \times x$ matrix of randomly chosen color values will most probably never end up in anyone’s digital photo album, and strings like *asdfasdf* are neither English nor inherently meaningful.

To construct a linguistic stegosystem S_1 , we observe that, thinking of a secret n as one of 26^k equiprobable choices for an integer with $0 \leq n < 26^k$, we can always encode the secret as a sequence of k symbols, each chosen from \mathcal{A} according to a uniform distribution. One simple coding scheme would do so, by numerically forming the base-26 expansion of n , denoting each base-26 digit by one alphabetic symbol. Clearly the resulting sequence of English alphabetic characters will be in \mathcal{A}^* .

How secure would a stegosystem constructed in such a way be? To quantify this within Cachin’s model, we need the probability distributions P_{S_1} and P_C . We already mentioned that S_1 would produce a string of characters, each of which is chosen from \mathcal{A} according to a uniform distribution, so $P_{S_1}(a) = \frac{1}{26} \approx 0.03846$ for every letter a . However English native speakers do not use the language as S_1 does. The letter ‘e’, for example, actually occurs much more often, at $P_C('e') \approx 0.13496$, and the letter ‘z’ occurs far less often, at $P_C('z') \approx 0.00027$. An adversary can use that knowledge to distinguish texts produced by S_1 from texts produced by a native speaker of English. Given the concrete numbers for P_{S_1} and P_C , we can quantify the insecurity of the system by $D(P_C||P_{S_1}) \approx 0.53648$.

We will now go on to refine S_1 , constructing a new stegosystem S_2 , by improving the model of the English language. Instead of the English alphabet \mathcal{A} , we will now use a dictionary \mathcal{D} of all words used in the English language¹ (thereby grossly oversimplifying things again), and assume that \mathcal{D}^* is the set of all meaningful English language texts. Our encoding technique works just like the one we used for S_1 . Since steganograms produced by S_2 are not random assortments of English letters, but random assortments of English words (each of which is an assortment of letters that is not random, but really used by native speakers), the stegotext appears far less suspicious, when evaluating the security under the same model we used for S_1 (i.e. in terms of the letter statistics). As opposed to the insecurity $D(P_C||P_{S_1}) \approx 0.53648$ we determined for S_1 , the insecurity of S_2 is now $D(P_C||P_{S_2}) \approx 0.077147$.

But now that Alice and Bob are using a more accurate semantic model of the English language, why should not Wendy? Let’s assume that Wendy’s linguistic model is at least as good as Alice’s and that she also has access to a dictionary \mathcal{D} . So instead of collecting statistics for the letter-frequencies P_{S_2} , and P_C over \mathcal{A} , she now observes the word-frequencies P'_{S_2} , and P'_C over \mathcal{D} . Due to the coding scheme, P'_{S_2} will be the uniform distribution again, but this time over words and not over letters, so $P'_{S_2}(w) = \frac{1}{6318} \approx 0.000158$ and again P'_C is determined by how native speakers use the English language. For example the word ‘the’ occurs much more often, at $P'_C('the') \approx 0.072185$, and the word ‘quantum’ occurs far less often, at $P'_C('quantum') \approx 0.000009$. We can see that these word-frequencies are much more significant evidence to distinguish steganograms

¹The dictionary our calculations are based on is a lexicalized word-frequency list of length $|\mathcal{D}| = 6318$ [16]. It lists the frequencies of the lemmatized tokens (words) occurring most often in the British National Corpus, a 100 Mio word long collection of text carefully hand-selected as a representative sample of modern English.

from covers. Putting this intuition into numbers, the insecurity of S_2 under Wendy’s new word-frequency model is $D(P'_C||P'_{S_2}) \approx 3.502983$, as opposed to $D(P_C||P_{S_2}) \approx 0.077147$ under Wendy’s old letter-frequency model.

We believe that the move from the alphabetic model of letter-semantics to the lexical model of word-semantics may be representative for improvements in the accuracy of a model for the semantics of the messages subject to steganalysis, for example as a result of additional efforts by Wendy or as a result of general progress in the state of the art of formulating such models. In our example, another step might be to improve the stegosystem by moving from the lexical model of word-semantics to a grammatical model, for example employing a context-free grammar instead of a flat dictionary. Note that we are using the linguistic term *semantics* in a very broad sense here, attributing semantics not only to linguistic expressions, but to every meaningful digital object. If we are not dealing with linguistic steganography, but, say, image steganography, we might dig into the semantic dimension of a bitmap-image by techniques of visual pattern perception or the like.

Turning back to our original discussion about the possibility of a model that adequately imposes a sufficient condition, and therefore a lower limit on the security of a given steganographic scheme in the light of semantically meaningful steganograms, we may remark that, at that point when we quantified the security of S_2 by $D(P_C||P_{S_2}) \approx 0.077147$, the belief that 0.077147 really was an upper limit on the insecurity of S_2 was somewhat illusionary, given that we later found $D(P'_C||P'_{S_2}) \approx 3.502983$, by varying the underlying semantic model. The problem is that the former result was obtained under a point of view that is treated, in Cachin’s model and under the old paradigm, as equally permissible, due to the fact that the semantic dimension is not taken into account.

A simplistic solution would be to enforce the condition that Alice’s and Wendy’s steganographic, respectively steganalytic activities are both based on the same semantic model, which is the best model either available to any one of them or any one of them can not rule out to be available to the other. Clearly $D(P_C||P_{S_2})$ can only be a measure for S_2 ’s security if Alice has a dictionary and Wendy does not, and $D(P_C||P_{S_1})$ can only be a measure for S_1 ’s security if Alice can be sure Wendy does not have a dictionary. Consequently $D(P'_C||P'_{S_2}) \approx 3.502983$ would be the only correct measure for S_2 ’s security, in a model refined in such a way that it enforces the dictionary to be public knowledge. Of course this idea is not new. It follows from Kerckoffs’ principle, because Alice’s and Wendy’s semantic model, could be viewed as part of the encoding and decoding techniques, which should, in the tradition of decades of cryptographic wisdom, always be considered public knowledge. In the case of a linguistic stegosystem, for example, it is obvious that the interpretation of a given piece of text must be considered public knowledge, as it is the very purpose of language to provide a representation of any semantic content that speakers might wish to communicate to each other, and must thus be known and agreed upon by many. Under a model refined in such a way, we could rely on Alice, Bob, and Wendy to share a single semantic model and can reduce the notion of security to a syntactic level.

Unfortunately the problem cannot be resolved that easily. Turning back to Cachin’s model, we would like to point out that $D(P'_C||P'_{S_2})$ can be a useful measure for the degree of security we can expect from S_2 , only if it is possible to fix a threshold ϵ , such

that the system is secure if $D(P'_C || P'_{S_2}) \leq \epsilon$, and insecure otherwise. However, as soon as we fix a threshold $\epsilon > 0$, we suppose that Wendy will not recognize messages as suspicious, that can be distinguished from covers in the semantic model used for the security analysis. If $\epsilon = 0$, on the other hand, then $P'_C = P'_{S_2}$. But from the information theory of cryptography [23] we know that such systems must be instances of one-time pads.

We believe that, if we carefully reviewed our formal models of steganography with the idea of semantically meaningful steganograms in mind, we would find peculiarities as the ones we have found in Cachin's model in all of them, amounting to the conclusion that, given a practical stegosystem that cannot be reduced to a one-time pad, we cannot fix a lower bound on its security, in any of them, without explicitly or implicitly making problematic assumptions about an underlying semantic model.

4 The New Paradigm

All of these limits, pointed out so far, do not seem surprising. After all, the idea that steganograms should be treated as meaningful digital objects in theory and practice is new. Traditionally, steganograms have been treated as meaningless objects, which is an assumption most probably stemming from cryptography, because in the context of cryptography it is the case that access to a cryptogram leaves an eavesdropper without any knowledge. By virtue of its definition, a cryptogram does not carry any meaning beyond that which must be deferred by means of the decryption routine. A steganogram, however, which has to resemble an innocuous cover in every respect, does carry such meaning. A steganogram can only be identified as innocuous or suspicious, after it has been interpreted and assigned meaning, which extends the cryptologic picture into a semantic dimension, as we move on from pure cryptography to steganography. Turning back to our intuitive picture of steganography, as introduced in Section 1 and the data-information-knowledge hierarchy, the essence of the new paradigm is that we are dealing with data in the context of cryptography, as opposed to steganography, which deals with information.

At this point, it may be necessary to elaborate a bit on the distinctions we commit to in this paper between the notions of data, information, and knowledge, because today it has become commonplace to use these terms somewhat interchangeably. The distinction we would like to make is based on the degree of understanding an observer has about a given observation he wishes to call a piece of data, information, or knowledge. In particular, we shall call an observation a piece of data, if we see it in a purely symbolic way, void of inherent meaning, but capable of being processed to make sense. Note in this context the latin origin of the word, *datum*, which literally translates to *that which is given*. We shall call an observation a piece of information, if the datum has been assigned meaning, and has thus been interpreted to be useful, and we shall call an observation a piece of knowledge if the observer has attained an even higher level of understanding, which there seems to be no broadly accepted consensus about, but which is outside the scope of this paper anyway. Such distinctions have been introduced by Russell Ackoff [1] on a technical level to the field of knowledge management, and are independently

treated in the literature by Milan Zeleny [31] in information science, Michael Cooley [11] in his discussion on common sense, Robert Lucky [18] in mathematical information theory and Harlan Cleveland [9, 10] in business information management, as well as in numerous pieces of secondary literature [24, 3, 13]. Some sources even credit T. S. Eliot for the idea, citing his 1934 poem *The Rock* as the first appearance of a data-information-knowledge hierarchy in the literature. Nevertheless Ackoff [1] is probably the most technically authoritative source.

Once we commit to this conception of data, information and knowledge, it becomes apparent, that the role of *understanding* as a means to elevate a given observation up the hierarchy from data to information and knowledge is quite crucial. Ackoff notes that understanding is by virtue of its nature a cognitive process, and can only be automated to the degree to which computers succeed in simulating this process. Thus, any claim attributing a human level of information- or even knowledge-processing capability to a fully computerized system must be presupposing a hypothesis whose confirmation has resisted decades of research in artificial intelligence and in cognitive science: that biological cognition is a computational process. Thus we feel driven to the point of view, that computers may not be regarded as directly operating on information or knowledge as such, in any way. Of course, the success of computerized systems in supporting human-controlled information- or knowledge-processing systems is undisputed. Yet, this does not contradict the view that computers are essentially limited in their domain of operation to simple data, since information- or knowledge-processing may still happen implicitly in a computerized system, if the data it operates on is elevated up and down the data-information-knowledge chain in the brains of its human users.

Of course, these ideas about data and information have a strong impact on data- and information-processing in the context of cryptology: In the new paradigm, we have in mind, a cryptosystem lies at the core of every stegosystem, in much the same way as a data-processing subsystem lies at the core of every other information-processing system. This cryptosystem is basically determined on a computational level by the encryption routine used by Alice's computer, the decryption routine used by Bob's computer, and the cryptanalytic attack used by Wendy's computer. The stegosystem is then an information-processing system that extends the data-processing cryptosystem by semantic aspects. This extension is determined on a cognitive and ontologic level, by the act of representation carried out by Alice, to make the encryption routine process the semantically meaningful information she wants to convey as a message-input, the act of interpretation carried out by Bob, to make sense of the message-output he gets from the decryption routine, and the act of steganalysis carried out by Wendy to make sense of the results she gets from the cryptanalysis routine. Figure 1 depicts this idea of content-aware steganography, and of how cryptography and steganography relate to each other in the context of the data-information-knowledge hierarchy.

The inner area of the figure depicts the cryptosystem: The message input to the encryption routine is treated as a piece of data. The encryption routine translates this message into a cryptogram which is another piece of data and the routines for decryption and cryptanalytic attack basically invert this mapping, so all the operations relevant to cryptography are closed within the data-domain. The encryption routine does not need to take into account any semantics, since it can always reinterpret its input as a random

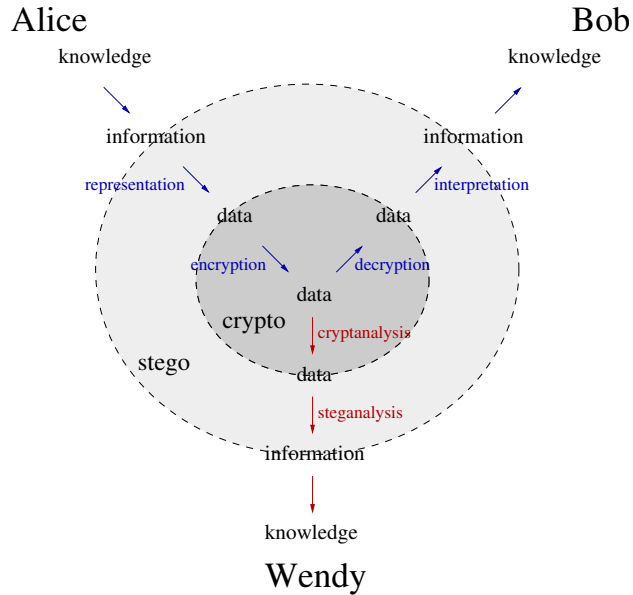


Figure 1: Cryptology and the DIKW hierarchy.

choice of one element from a finite message-space, regardless of whether this input is actually a representation for an image, a sound, or a text. The decryption routine and the cryptanalytic attack do not need to take into account any semantics, since they only need to reconstruct a random choice of one element from a finite message-space as its output, and the user will know how it is to be reinterpreted – whether the message represents an image, a sound, or a text.

The outer area of the figure depicts the stegosystem: The message that Alice actually wants to convey, is a piece of information. The act of representation degrades this information to data, so it can be run through the cryptosystem. The acts of interpretation or steganalysis, on the other hand reassign meaning to the data which is supposed to equal the original message, and therefore bring about information again, so the whole stegosystem essentially operates within the information-domain. Clearly, the act of representation must take into account semantics, since Alice has exactly one piece of semantic content in mind, when she represents it, and the acts of interpretation and steganalysis have to deal with semantics, since they have to reconstruct exactly that semantic content.

This brings the conception suggested by Figure 1 in line with the ideas of data and information in the context of cryptography and steganography as described before. It should be noted that this diagram can also be seen as depicting the data-flow in an actual construction, that we shall elaborate on a bit in the rest of this section.

This construction starts out with a message m like $m = \text{"Let's break out, Bob!"}$ as seen through Alice's or Bob's eyes. At this stage, we see m as a piece of information. The message originates from Alice's knowledge, and is supposed to become part of Bob's knowledge. Alice and Bob both assign a special meaning to it. They do not see it as

a string of ASCII-characters, and they would not be able to reinterpret this message as one of a finite number of equiprobable choices of a message from a given message space in any straightforward way. One cannot judge whether or not m contains “secret information to the warden”, in Simmons’ words, since m *is* secret information and does not *contain* any information whatsoever.

Only, as Alice (not her computer) fixes a representation, by typing m into her favourite editor, she degrades m from a piece of information to a piece of data \bar{m} (so we have $\bar{m} = R(m)$). What the computer gets to see of her message is only the syntactic representation, not a semantic interpretation. The only thing the encryption routine can work with may be a string of ASCII-characters “*Let’s break out, Bob!*” = $R(\text{“Let’s break out, Bob!”})$.² Next, Alice’s computer uses a traditional cryptosystem and applies encoder E to produce a cryptogram \bar{e} from the plaintext \bar{m} (so we have $\bar{e} = E(\bar{m})$). For example the encryption might output “*Good morning, Bob!*” = $E(\text{“Let’s break out, Bob!”})$. Then the computer uses an untrusted channel to transmit \bar{e} to both Bob and Wendy.

Bob’s computer will see \bar{e} and apply a decryption mechanism E^{-1} to reconstruct plain text $\bar{m} = E^{-1}(\bar{e})$ again (i.e. “*Let’s break out, Bob!*” = $E^{-1}(\text{“Good morning, Bob!”})$). Then Bob (not his computer) will find interpretation $m = I_B(\bar{m})$, and see the original message in its intended meaning $m = \text{“Let’s break out, Bob!”}$, thereby turning the data into information again.² Wendy’s computer will also see \bar{e} , and will apply cryptanalytic attack A to gain a suspicion $\bar{w} = A(\bar{e})$, such as “*The message probably just means ‘Good morning, Bob!’; It may also have some secret meaning, but I do not have any reason for suspecting so*” = $A(\text{“Good morning, Bob!”})$. Then Wendy (not her computer) will interpret this data as $w = I_W(\bar{w})$. Because she is unwilling to wait for a billion this interpretation will most probably be $w = \text{“Good morning, Bob!”}$.²

5 The Paradigm Shift

In the previous section we have introduced a new way of thinking about steganography, putting forward a new paradigm that puts special emphasis on the relations between data, information, and knowledge in cryptography and steganography. Where the traditional paradigm, implicit to current models of steganography, is widely ignorant of the semantic content of steganograms, the new paradigm views steganography as a phenomenon that resides at the interface between the representation of a steganogram and its content. This shift in perspectives is rich in implications on existing stegosystems some of which will be discussed in this section, when we try to fit content-unaware steganography into our new content-aware picture.

In order to discuss how content-aware steganography relates to content-unaware steganography, we shall first try to approach the highly related question of how content-aware steganography relates to cryptography. Let us, therefore, turn back to the con-

²Note that we use R , I_A and I_W only to denote the dataflow behind this complex cognitive process of representation, and ask the reader to keep in mind that R , for example, cannot be a function in the strict mathematical sense since the R -image of m on \bar{m} , will probably depend on additional common-sense and context-information not determined by m , or may be undefined, ambiguous, or inherently vague and indeterministic, etc.

struction explained in the previous section, in which Alice and Bob actually use a traditional cryptosystem as part of their content-aware stegosystem. We have often emphasized the fact that cryptography operates in a data-domain where steganography operates in an information-domain, so, in terms of our construction, the very purpose of the cryptosystem that is used, is to enforce a security condition of the form $Cryptosecure(\bar{e}, \bar{w}, \bar{m})$. Let's put aside, for a moment, the actual model of cryptographic security we have to employ to evaluate such a condition, and concentrate on the fact that cryptographic security is determined solely on the basis of the data of \bar{e} , \bar{w} , and \bar{m} .

Does the fact that Wendy gets *no data* out of a cryptographic message exchange really imply that she gets *no knowledge*? From the point of view of steganography, the answer is no. Petitcolas et al. [22] point out as an example the fact that a cryptogram has obvious implications if it is exchanged between a drug dealer and someone not yet under suspicion, or between a defence contractor and the embassy of a hostile power. Furthermore one might envision a situation in which Alice tries to encrypt the message *Let's have dinner, Bob!*, runs it through her PGP system and the ASCII-armored cryptogram spells out **LetsBreakOutBob** by coincidence.

Cryptographers are used to the idea that it is possible for Alice to communicate no knowledge to Eve the eavesdropper, and, possibly unknowingly, carried over this conception to steganography, assuming it was actually possible for Alice to communicate no knowledge to Wendy the warden. However, as we think about the original prisoners' dilemma and recall that Wendy functions as a conveyor, we see that Alice can certainly not communicate to Bob without communicating to Wendy first. This brings the actual motivations behind steganography in line with Paul Watzlawick's communication theory [29], which is built upon a "metacommunicative axiom" stating that, in any real life situation, it is *impossible* for a human *not to* communicate. This is exactly the kind of situation Alice finds herself in, as she faces Wendy and wants her to convey secret information to Bob. As Alice hands Wendy a piece of cipher text \bar{e} she can be sure that Wendy *will* give it an interpretation. Wendy not giving it an interpretation, Alice communicating *no* knowledge to Wendy, is not an option, however Wendy giving it a wrong interpretation, Alice intentionally communicating *misleading* knowledge to Wendy, is an option, and the very essence of steganography.

If \bar{e} is an ASCII-text that spells out '**Let's break out, Bob!**', Wendy will assume that Alice wants to break out, which is not a good thing for Alice. If \bar{e} appears to be a PGP-generated cryptogram, Wendy will assume that Alice has something to hide, which will also have serious consequences for Alice. Only if \bar{e} spells out '**Good morning!**', Wendy may find the message innocuous, unless, of course, it is communicated in the evening. This example even points out some context-sensitivity, and, once again, supports the point of view that steganalysis requires a sophisticated means to actually understand data. Clearly, what Wendy does, when she tries to judge such messages can hardly be reduced to simple comparisons, dictionary lookups, or syntax analyses. Steganalysis is all about making well-informed and intelligent judgements about the information e that can be inferred on the basis of the data \bar{e} , perhaps requiring analysis and synthesis of related contextual and common-sense knowledge.

Thus, as opposed to cryptosystems that always enforce a security condition of the form $Cryptosecure(\bar{e}, \bar{w}, \bar{m})$ about a transmission \bar{e} and the related data, stegosystems

always enforce a security condition of the form $Stegosecure(\bar{e}, w, m)$ about a transmission \bar{e} and the related information. In the context of the construction from the previous section these two notions of security relate to each other in a quite obvious way, as we can substitute $I_W(\bar{w})$ for w , $A(\bar{e})$ for \bar{w} , $I_B(\bar{m})$, and $D(\bar{e})$ for \bar{m} . Thus the notion of cryptographic security of a given cryptosystem depends solely on the attack A and the encryption E , since it is secure only if $\forall \bar{e} : Cryptosecure(\bar{e}, A(\bar{e}), E^{-1}(\bar{e}))$, and the security of a given stegosystem depends on the composition $A \circ I_W$ of the attack A with Wendy's interpretation of the results of the attack I_W , and the composition $E^{-1} \circ I_B$ of the decryption E^{-1} with Bob's interpretation of the representation of secret messages I_B , since it is secure only if $\forall \bar{e} : Stegosecure(\bar{e}, I_W(A(\bar{e})), I_B(E^{-1}(\bar{e})))$.

This is where the content-unaware paradigm begins to fit into the content-aware picture. Clearly one of the essential characteristics of the content-unaware paradigm from the content-aware perspective, is that the old paradigm implicitly talks about the meaning of a piece of data, when it talks about data, and vice versa. In texts following the old paradigm a message can in one paragraph be regarded a statement about prison escape plans and in the next paragraph be regarded an ASCII-string or a number. Another important characteristic is that the old paradigm requires that the secret message, in the sense of a datagram representing a secret message, may never be seen by Wendy in the same form as seen by Bob, regardless of whether the representation is sensible to Wendy in the same way as to Bob.

These presumptions of content-unaware steganography can easily be fit into the our model: The essence of content-unaware steganography is that it handles the concept of interpretation by presuming a one-to-one mapping I that assigns a unique piece of information to each piece of data, and vice versa, and that this mapping is public knowledge, so that both Bob and Wendy use it and we have $I_W = I$, and $I_B = I$.

Here it is important to note, how content-unaware steganography which is secure only if $\forall \bar{e} : Stegosecure(\bar{e}, I(A(\bar{e})), I(E^{-1}(\bar{e})))$ becomes isomorphic to cryptography which is secure only if $\forall \bar{e} : Cryptosecure(\bar{e}, A(\bar{e}), E^{-1}(\bar{e}))$ under the content-aware point of view. This is due to the fact that, given a stegosystem, we can always view it as a cryptosystem where $Cryptosecure(\bar{e}, I^{-1}(w), I^{-1}(m))$ if $Stegosecure(\bar{e}, w, m)$. For example given a stegosystem like a fictional perfectly secure LSB-substitution system, we can always view it as a cryptosystem in which cryptograms happen to coincide in their representation with bitmap images. Conversely, given a cryptosystem, we can always view it as a stegosystem where $Stegosecure(\bar{e}, I(\bar{w}), I(\bar{m}))$, if $Cryptosecure(\bar{e}, \bar{w}, \bar{m})$. For example given a cryptosystem that employs a one-time-pad assigning ciphertext codes to cleartext messages, we can always view it as a stegosystem in which we assume that Wendy finds the ciphertext codes innocuous by themselves. This means that we would have to construct our cryptosystem in such a way that codewords that appear on the pad are publicly known to be regarded innocuous by Wendy.

We believe that the major weaknesses of content-unaware approaches to steganography as pointed out in sections 2 and 3 can widely be explained from a content-aware perspective by this isomorphism. Content-unaware stegosystems are constructed as simple syntactic wrappers around known cryptosystems, defining a one-to-one mapping I , as described above, between the cryptograms output by the cryptosystem and steganograms that are most often just *assumed* to appear innocuous to Wendy. These assumptions

are the ones we were talking about when we concluded section 2 by stating that, in practice, the steganography vs. steganalysis battle is won by steganalysis by exploiting false assumptions made by steganography, and the ones we were talking about when we concluded section 3 by stating that current theories of steganography tend to implicitly make problematic assumptions about an underlying semantic model. We hope that such mistakes are not made as easily with the isomorphism between content-unaware steganography and cryptography in mind, because it puts simple syntactic one-to-one wrappers into their theoretically trivial perspective, and it prevents theorists from unknowingly reinventing the cryptographic wheel, as they try to pin down what it means for a stegosystem to be secure without taking into account semantic aspects. Nonchalantly speaking, the lesson that content-aware steganography tries to teach us about content-unaware stegosystems, is that they fail to address the actual steganographic challenge.

6 HIP: A New Security Primitive for a New Kind of Steganography

So far, we have not given any clues on how to actually go about the construction of a content-aware stegosystem. To do so, we have to take a look into the black-boxes we denoted by the predicates $Cryptosecure(\bar{e}, \bar{w}, \bar{m})$ and $Stegosecure(\bar{e}, w, m)$ so far.

Today, the most common way of formalizing the notion of cryptographic security is based on the computational complexity involved in Wendy's attack. Here we presuppose a publicly known upper bound on the computational resources Wendy has and we fix a computational procedure to compute the suspicion \bar{w} . Then a definition of cryptographic security might require that $Cryptosecure(\bar{e}, \bar{w}, \bar{m})$ only if $\bar{w} = \bar{m}$ implies that, given only \bar{e} , the computation of \bar{w} requires computational resources beyond Wendy's reach. For example, we can verify such a statement about an RSA cryptosystem using moduli of size n , if we know that Wendy employs an attack that takes $\Theta(e^n)$ time, and choose n sufficiently large so that Wendy is not willing or able to wait for the attack to finish.

Another approach has been employed in information theoretic treatments of cryptographic security, such as that of Claude Shannon [23]. Here we presuppose a publicly known upper bound on the uncertainty Wendy is willing to accept to take the suspicions arising from her attacks for granted, and we fix a certain syntactic constraint on the appearance of messages and cryptograms. Then a definition of cryptographic security might require that $Cryptosecure(\bar{e}, \bar{w}, \bar{m})$ only if $\bar{w} = \bar{m}$ implies that, given only \bar{e} , random choice of \bar{w} from the set of all syntactically possible messages that may encipher to \bar{e} introduces a level of uncertainty into Wendy's suspicion that she is not willing to accept. For example, we can verify such a statement about an alphabetic substitution cipher with an alphabet of size n , if we know that Wendy has to guess which of the $n!$ equally possible codes was used, and we know that Wendy is not willing to act upon a suspicion \bar{w} that is correct only at probability $1/n!$ (as is the case for a perfectly secure system).

Once we've understood how these cryptosystems work, we can construct a stegosystem in much the same way. However, we have to keep in mind that a cryptosystem is

secure, only if $\forall \bar{e} : \text{Cryptosecure}(\bar{e}, A(\bar{e}), E^{-1}(\bar{e}))$, and a stegosystem is secure, only if $\forall \bar{e} : \text{Stegosecure}(\bar{e}, I_W(A(\bar{e})), I_B(E^{-1}(\bar{e})))$. Since we want to construct a content-aware stegosystem that is not isomorphic, in the sense described in the previous section, to any cryptosystem we also have to keep in mind that it is not necessarily the case that $I_W = I_B$, and that I_W and I_B are not simple mappings in the mathematical sense, but complex cognitive processes that are not easily simulated by computers.

We have already stressed quite heavily the fact that steganography is basically about information, and have also explained why we view information as something that never actually resides in a computer, but only emerges when a computer's human user interprets the data the computer confronts him with. At first sight this seems to complicate matters for a computerized steganography system, but in fact this is not necessarily an obstacle. Actually the fact that a computer will most often interpret a given piece of data quite differently from a human, may be the very thing that enables content-aware steganography. If we assume that Wendy is a computer, while Alice and Bob are human, then we will naturally have $I_W \neq I_B$. Furthermore, it is not unreasonable to assume that Wendy is a computer, while Bob is a human, in the case of online-steganography. Just imagine a human user monitoring mail-traffic on a large relay in realtime or assisting a web-crawler in finding suspicious web-content. Most often it is only natural to assume that the haystack of data that is out there, is way too large for humans to arbitrate in a cost-effective manner.

Once we admit that Wendy is a computer, and Bob is a human sitting in front of a computer, all we have to do is to make the solution to the problem of determining the secret interpretation m of the steganogram \bar{e} depend on the solution of a problem that only humans can solve correctly. More formally, we have to require that $\text{Stegosecure}(\bar{e}, w, m)$ only if $w = m$ implies that Wendy has to prove that she is human to determine w from \bar{e} .

This security primitive, known as human interactive proof (HIP) [14, 30, 26], better known under the more specific model of the CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) [27], has only recently gained attention in the computer security community, because of its usefulness in the fight against worms and spam, and the prevention of web-service abuse, denial-of-service, and dictionary attacks. Currently, the best-known HIPs are the OCR CAPTCHAs that display heavily distorted text for the user of a web-service to read and type into a text-field. This way the user of the web-service can prove that he is human. A computer, for example a robot trying to create a thousand e-mail accounts with a provider of free e-mail services to distribute worms and spam, can only do so, if he solves a problem of optical character recognition, that is far beyond the state of the art in artificial intelligence, in much the same way as the breaking of the RSA-cryptosystem requires the cryptanalyst to solve a problem that is far beyond the state of the art in seminumeric algorithms.

In general, a human interactive proof involves a set of tests $T = \{t_1, t_2, \dots\}$, and a set of solutions $S = \{s_1, s_2, \dots, s_{|S|}\}$, for $|S| \in \mathbb{N} \setminus \{0, 1\}$ and a procedure that produces a test/solution pair (t, s) where $t \in T$ and $s \in S$, such that one can assume that anybody who answers s to t is human. In theory, for an HIP to be perfectly secure, T must be countably infinite at least. In practice it is desirable, that $|T|$ is as large as possible. Furthermore we will assume that for each test $t \in T$ there is a set $C_t \subseteq T$ of candidate-

- 1 **for** $k \leftarrow 1$ **to** n
- do**
- 2 The tester constructs a test/solution pair (t_k, s_k)
 such that $t_k \in T$ and $s_k \in S$.
- 3 The tester sends the test t_k to the testee
 (and keeps its solution s_k private).
- 4 The testee makes a choice h_k for a solution of t_k .
- 5 The testee sends h_k to the tester.
- \triangleright The tester checks if Bob could be a computer.
- 6 **if** $h_k \neq s_k$
- 7 **then** Do not draw any conclusions and **stop**.
- 8 Conclude that the testee is human.

(a) n -step human interactive proof

- 1 **for** $k \leftarrow 1$ **to** n
- do**
- 2 Alice constructs a test/solution pair (t_k, s_k)
 such that $t_k \in T$ and $s_k \in C_{t_k}$.
- \triangleright Alice constructs a claim c_k :
- 3 $c_k \leftarrow I_{C_{t_k}}^{-1}((I_{C_{t_k}}(s_k) + \bar{m}_k) \bmod |C_{t_k}|)$
- 4 Alice sends the test/claim pair $\bar{e}_k = (t_k, c_k)$
 (and keeps the solution s_k private).
- 5 Bob makes a choice h_k for a solution of t_k .
- \triangleright Bob reconstructs the message character \bar{m}_k :
- 6 $\bar{m}'_k \leftarrow (I_{C_{t_k}}(c_k) - I_{C_{t_k}}(h_k)) \bmod |C_{t_k}|$
- 7 Wendy obtains a message hypothesis \bar{w}_k .
- \triangleright Wendy obtains the solution s'_k to t_k :
- 8 $s'_k \leftarrow I_{C_{t_k}}^{-1}((I_{C_{t_k}}(c_k) - \bar{w}_k) \bmod |C_{t_k}|)$

(b) n -character steganographic transmission

Figure 2: A generic construction of a stego-system from an HIP

solutions for t , and for each test there are at least 2 candidate solutions, i.e. $|C_t| \geq 2$ for all t . Let $I_{C_t} : C_t \mapsto \{0, 1, \dots, |C_t| - 1\}$ be a one-to-one mapping from the elements of a given set of candidate solutions to the smallest $|C_t|$ natural numbers.

For our example construction we will also simplify things by assuming that all tests $t \in T$ have the same number b of candidate solutions, i.e. $|C_t| = b$ for all C_t . Figure 2(a) shows how a human interactive proof is then handled, and Figure 2(b) shows how Alice can use the same security primitive to transmit a secret message string $\bar{m} = \bar{m}_1\bar{m}_2\dots\bar{m}_n$, of numbers $\bar{m}_i \in \{0, 1, \dots, b-1\}$. One can think of \bar{m} as the radix- b expansion of a number $0 \leq m < b^n$. Note that the construction can be straightforwardly generalized to the case of differing numbers of candidate-solutions $|C_t|$ by thinking of \bar{m} as a mixed-radix expansion.

Theorem 1 (decodability by humans) *After termination of the steganographic transmission outlined in Figure 2, the message $\bar{m}' = \bar{m}'_1\bar{m}'_2\dots\bar{m}'_n$ received by Bob will be equal to the original message \bar{m} submitted by Alice if Bob can be verified by the HIP as human.*

Proof Consider the k -th iteration of the HIP, respectively the stego-transmission. Since the HIP will verify Bob as human, we know that, in line 4 of the HIP, Bob must be able to choose h_k in such a way that $h_k = s_k$ (as he will otherwise fail the verification on line 6, and not be considered human). Thus he will also be able to choose h_k in such a way that $h_k = s_k$ in line 4 of the stego-transmission. Bob reconstructs the message-bit on line 6, by setting

$$\bar{m}'_k = (I_{C_{t_k}}(c_k) - I_{C_{t_k}}(h_k)) \bmod |C_{t_k}|.$$

We can substitute s_k for h_k . Furthermore, we can substitute c_k from the assignment on line 3 to obtain

$$\bar{m}'_k = (I_{C_{t_k}}(I_{C_{t_k}}^{-1}((I_{C_{t_k}}(s_k) + \bar{m}_k) \bmod |C_{t_k}|)) - I_{C_{t_k}}(s_k)) \bmod |C_{t_k}|.$$

Here $I_{C_{t_k}} \circ I_{C_{t_k}}^{-1}$ cancel out, and we can move the modulo operation to the outer scope:

$$\bar{m}'_k = ((I_{C_{t_k}}(s_k) + \bar{m}_k) - I_{C_{t_k}}(s_k)) \bmod |C_{t_k}|.$$

Obviously $I_{C_{t_k}}(s_k) - I_{C_{t_k}}(s_k)$ cancel out, so

$$\bar{m}'_k = \bar{m}_k \bmod |C_{t_k}|.$$

Since $\bar{m}_k < |C_{t_k}|$, we have

$$\bar{m}'_k = \bar{m}_k,$$

which means that Bob has correctly decoded the message. **qed.**

Theorem 2 (security against on-line non-human arbitrators) *After termination of the steganographic transmission outlined in Figure 2, Wendy's message hypothesis $\bar{w}' = \bar{w}'_1\bar{w}'_2\dots\bar{w}'_n$ will be equal to the original message \bar{m} submitted by Alice only if the HIP verifies Wendy as human.*

Proof We assume that, somehow, Wendy has managed to guess \bar{w}_k in such a way that $\bar{w}_k = \bar{m}_k$ in the k -th iteration. In line 8, Wendy uses that message hypothesis to obtain a solution to the HIP

$$s'_k = I_{C_{t_k}}^{-1}((I_{C_{t_k}}(c_k) - \bar{w}_k) \bmod |C_{t_k}|)$$

Now we can substitute \bar{m}_k for \bar{w}_k , and c_k from the assignment on line 3 to obtain

$$s'_k = I_{C_{t_k}}^{-1}((I_{C_{t_k}}(I_{C_{t_k}}^{-1}((I_{C_{t_k}}(s_k) + \bar{m}_k) \bmod |C_{t_k}|)) - \bar{m}_k) \bmod |C_{t_k}|).$$

Again, $I_{C_{t_k}} \circ I_{C_{t_k}}^{-1}$ cancel out, and we can move the modulo operation to the outer scope:

$$s'_k = I_{C_{t_k}}^{-1}((I_{C_{t_k}}(s_k) + \bar{m}_k - \bar{m}_k) \bmod |C_{t_k}|).$$

Obviously $\bar{m}_k - \bar{m}_k$ cancel out:

$$s'_k = I_{C_{t_k}}^{-1}(I_{C_{t_k}}(s_k) \bmod |C_{t_k}|).$$

Since $I_{C_{t_k}}(s_k) < |C_{t_k}|$ and $I_{C_{t_k}}^{-1} \circ I_{C_{t_k}}$ cancel out,

$$s'_k = s_k.$$

This means that Wendy can solve the HIP. **qed.**

If we turn back, for a moment, and think about the nature of the security we can expect from this stegosystem, we will see that, in addition to a property of the form *Cryptosecure*($\bar{e}, \bar{w}, \bar{m}$) about the data that is involved, it also exhibits a property of the form *Stegosecure*(\bar{e}, w, m), about the information based on it. In theorem 2 we have established that $\bar{w} = \bar{m}$ implies that Wendy can prove that she is human if she can determine \bar{w} given only $\bar{e} = (t_1, c_1)(t_2, c_2)\dots(t_3, c_3)$.

However, since both Wendy and Bob are human now, they will both be able to interpret the message in the same way, i.e. $I_B = I_W$. So we will naturally have *Stegosecure*($\bar{e}, I_W(\bar{w}), I_B(\bar{m})$), which is a claim about information of the form *Stegosecure*(\bar{e}, w, m). Of course $w = m$ entails $\bar{w} = \bar{m}$, so we conclude that *Stegosecure*(\bar{e}, w, m) only if $w = m$ implies that Wendy can prove that she is human if she can determine w from \bar{e} . Here we assume that Wendy's interpretation of the message meaning only depends on the content of \bar{e} , not on its pure existence. In other words, we assume that Wendy will not find strings of test/claim pairs of the form $\bar{e} = (t_1, c_1)(t_2, c_2)\dots(t_3, c_3)$ suspicious as such.

In practice we could, for example, assume that Wendy generally tolerates English language text being exchanged between Alice and Bob. We can then set up a stegosystem on the basis of a text-domain HIP (such as the word-sense disambiguation HIP [5]) Alternatively we could assume that Wendy tolerates images being exchanged. We would then have to use an image HIP (such as the famous OCR CAPTCHA [27] or image recognition CAPTCHAs [8]).

7 An Image Recognition Stegosystem

In order to show how our generic construction can be applied to a particular domain, we will consider an image recognition CAPTCHA [8], and will show how to turn it into a stegosystem in this section.

The original set-up of the image recognition CAPTCHA is as follows: The tester is a computer that has available a dictionary of words that can be easily illustrated, and a database of images labelled according to their content with one of the dictionary words. This knowledge is assumed to be private to the tester in some way (the data-sources might, for example, be of a dynamic nature). The tester then generates a test/solution-pair (t, s) by picking two dictionary words w_s and w_e , five pictures $\{p_0, p_1, \dots, p_4\}$ labelled w_e , and one picture p_5 labelled w_s . These six pictures make up the set of candidate solutions $C_t = \{p_0, p_1, \dots, p_5\}$. The test t itself, as it is transmitted over the network and as it is presented to the testee is a random permutation of $\langle p_0, p_1, \dots, p_5 \rangle$. For a human testee it will be trivial to pick the solution $s = p_5$ to the problem “which of the pictures does not belong here?”. For a computer this task will amount to a highly non-trivial problem, way beyond the state of the art in automatic image recognition.

Now suppose Alice wants to submit a message m which can be reinterpreted as a number $0 \leq \bar{m} < 6$. She can now create a test/solution pair (t, s) as described before, with $t = \langle p_0, p_1, \dots, p_5 \rangle$, and $s = p_i$ for some $0 \leq i < 6$. Alice can now construct a claim $c = (i + \bar{m}) \bmod 6$. She can exploit a channel that transmits images by transmitting the pictures in t in order, and transmitting $c + 6 * k$ images after that to encode for c , where k is picked randomly between 0 and some arbitrarily large K . That is to say, the transmission \bar{e} will always be of the form $\bar{e} = \langle p_0, p_1, \dots, p_5, \dots, p_n \rangle$. Wendy and Bob will see $n = 6 + c + 6 * k$ images.

Now Bob, being a human, can look at the first 6 images and determine the image p_i that does not belong into the set. Furthermore he can reconstruct the claim, since $c = n \bmod 6$. He can now subtract from the claim the solution to the test to get $\bar{m}' = (c - i) \bmod 6$. As proved before $\bar{m}' = \bar{m}$.

Wendy, on the other hand, cannot do so, since she cannot find the solution p_i , without also solving the image recognition CAPTCHA. Furthermore she can not formulate any suspicions based on the transmission as such, since images may be transmitted, and the length of the transmissions, i.e. the number of images appear nearly random. More specifically, the distribution $P(N = n)$ on a random variable N that holds the number of images submitted in total will be uniform between 6 and $6 * (K + 2) - 1$.

8 A Word-Sense Disambiguation Stegosystem

We can also think of scenarios that do not relate directly to the generic construction outlined before, but also amount to an implicit human interactive proof being carried out in the course of a stegosystem. More particularly, we would like to introduce a linguistic stegosystem we studied more extensively before [4], that can be broken only by an arbitrator who has managed to solve the quite non-trivial task of word-sense disambiguation, a problem that can also be used as the basis of a word-sense disambiguation HIP [5].

Again, Alice wants to submit a message m , which we will regard as a number $0 \leq \bar{m} < 4$. She can embed this number into an innocuous piece of text, like the sentence

They built a new depot near the docks.

by substituting words for synonyms.

For example the word *depot* can be replaced, *in this context*, by any other word from the set $\{\textit{depot}, \textit{storage}, \textit{store}, \textit{storehouse}\}$. Here we can set each word to encode for a different message, for example we could assign codewords in alphabetic order and encode messages as follows:

$$\textit{They built some new} \left\{ \begin{array}{l} 0 \textit{ depot} \\ 1 \textit{ storage} \\ \mathbf{2} \textit{ store} \\ 3 \textit{ storehouse} \end{array} \right\} \textit{ near the docks.}$$

Say Alice wants to submit the message with $\bar{m} = 2$, then she will transmit

They built a new store near the docks.

When Wendy is confronted with the word *store*, she can look it up in a synonymy-dictionary, and will find that *store* can be substituted from the synonymy set $\{\textit{depot}, \textit{storage}, \textit{store}, \textit{storehouse}\}$ as in “*They built a new store near the docks*”, or from the synonymy set $\{\textit{stock}, \textit{store}\}$ as in “*He has an impressive store of wine*”. Thus the synonymy of a particular word with another word is determined by the context, which selects for a word by what is ultimately its meaning.

This word-sense disambiguation problem has been of considerable interest to computational linguists ever since the first attempts at automatic machine translation were made in the 1950s. To this day, the performance of machines in word-sense disambiguation is nowhere near the performance of humans, and the problem is now widely considered “AI-complete” in the sense that a solution to this problem presupposes a solution to the “strong AI-problem” of the synthesis of a human-level intelligence.

The problem Wendy faces when she does not know the correct set of synonyms that Alice used during encoding is that each of the possible synonymy-sets establishes a different code. She might just as well try to decode the steganogram as follows:

$$\textit{They built some new} \left\{ \begin{array}{l} 0 \textit{ stock} \\ 1 \textit{ store} \end{array} \right\} \textit{ near the docks.}$$

In order to determine whether the steganogram decodes to $\bar{m} = 2$ or to $\bar{m} = 1$ Wendy has to solve this word-sense disambiguation problem and thereby prove that she is human.

9 Conclusion

In this paper we have introduced the concept of content-aware steganography as a new paradigm of steganography, stemming from a shift in perspectives towards the objects

of steganography. We pointed out that, in the predominant paradigm of steganography, the nature of these objects is that of data.

We departed from the observation that systems relying on this paradigm are eventually broken on grounds of attacks that exploit the fact that the digital objects we encounter in everyday life are more than data – that they are meaningful and can be interpreted to give us information. Then we considered a theoretic model of steganography and showed how to derive different predictions from it about the security of one stegosystem, depending on the semantic model implicitly (most often unknowingly) employed when any such a prediction is made.

This lead us to abandon the point of view that steganographic objects can be characterized in terms of the data that represent them, and to take the new point of view that steganographic objects should be considered pieces of information as such. Then we introduced a new kind of steganography that relies on human interactive proofs as a security primitive, and showed how this kind of steganography fits the picture established by the new paradigm.

We presented a generic construction of a stegosystem from a human interactive proof, that is secure against a non-human arbitrator if the human interactive proof is secure. We considered an image recognition CAPTCHA, and showed how our construction could be used in practice to turn it into a stegosystem. Furthermore we introduced a linguistic stegosystem that does not directly follow our construction, but can also be viewed to run an implicit human interactive proof as a security primitive.

References

- [1] Russell L. Ackoff. From data to wisdom. *Journal of Applied Systems Analysis*, 16:3–9, 1989.
- [2] Michael Backes and Christian Cachin. Public-key steganography with active attacks. In Joe Kilian, editor, *Theory of Cryptography: Second Theory of Cryptography Conference, TCC 2005*, volume 3378 of *Lecture Notes in Computer Science*, page 210ff. Springer, February 2005.
- [3] Gene Bellinger, Durval Castro, and Anthony Mills. Data, information, knowledge, and wisdom. website, 2004. accessed 2005-08-31.
- [4] Richard Bergmair. Towards linguistic steganography: A systematic investigation of approaches, systems, and issues. final year project, April 2004. submitted in partial fulfillment of the degree requirements for “B.Sc. (Hons.)” to the University of Derby.
- [5] Richard Bergmair and Stefan Katzenbeisser. Towards human interactive proofs in the text-domain. In *Proceedings of the 7th Information Security Conference (ISC '04)*, Springer Lecture Notes in Computer Science, September 2004.
- [6] Christian Cachin. An information-theoretic model for steganography. In *Information Hiding, 2nd International Workshop*, volume 1525 of *Lecture Notes in Computer Science*, pages 306–318. Springer, 1998.

- [7] Christian Cachin. An information-theoretic model for steganography. *Inf. Comput.*, 192(1):41–56, 2004.
- [8] Monica Chew and J. D. Tygar. Image recognition CAPTCHAs. In *Proceedings of the 7th Information Security Conference (ISC '04)*, Springer Lecture Notes in Computer Science, September 2004.
- [9] H. Cleveland. Information as resource. *The Futurist*, pages 34–39, December 1982.
- [10] H. Cleveland. *The Knowledge Executive: Leadership in an Information Society*. Truman Talley Books, New York, 1985.
- [11] M. Cooley. *Architecture or Bee?* The Hogarth Press, London, 1987.
- [12] Jessica Fridrich, Miroslav Goljan, Dorin Hoge, and David Soukal. Quantitative steganalysis of digital images: estimating the secret message length. *Multimedia Systems*, 9:298–302, 2003.
- [13] Jonathan Hey. The data, information, knowledge, wisdom chain: The metaphorical link. Technical report, University of Berkeley, 2004. accessed 2005-08-31.
- [14] Nicholas J. Hopper and Manuel Blum. Secure human identification protocols. In *Advances in Cryptology, Proceedings of Asiacrypt '01*, 2001.
- [15] Nicholas J. Hopper, John Langford, and Luis von Ahn. Provably secure steganography. In M. Yung, editor, *Advances in Cryptology - CRYPTO 2002: 22nd Annual International Cryptology Conference*, volume 2442 of *Lecture Notes in Computer Science*, pages 77–92. Springer, August 2002.
- [16] Adam Kilgarriff. Bnc database and word frequency lists. website, March 1996. accessed 2005-04-10.
- [17] Bert-Jaap Koops. Crypto law survey. website, January 2005.
- [18] Robert W. Lucky. *Silicon Dreams: Information, Man and Machine*. St. Martin's Press, New York, 1989.
- [19] Declan McCullagh. Congress mulls stiff crypto laws. WIRED news online, September 2001.
- [20] Nasir Memon Mehdi Kharrazi, Husrev T. Sencar. Blind source camera identification. In *Proceedings of the National Conference on Image Processing (ICIP '04)*, 2004.
- [21] Thomas Mittelholzer. An information-theoretic approach to steganography and watermarking. In *IH '99: Proceedings of the Third International Workshop on Information Hiding*, pages 1–16. Springer-Verlag, 2000.
- [22] Fabien A. P. Petitcolas, Ross J. Anderson, and Markus G. Kuhn. Information hiding – a survey. *Proceedings of the IEEE*, 87(7):1062–1078, July 1999.

- [23] Claude E. Shannon. Communication theory of secrecy systems. *Bell System Technical Journal*, 28(4):656–715, 1949.
- [24] Nikhil Sharma. The origin of the data information knowledge wisdom hierachy. website, February 2005. accessed 2005-08-31.
- [25] G. J. Simmons. The prisoners’ problem and the subliminal channel. In *Advances in Cryptology, Proceedings of CRYPTO ’83*, pages 51–67, 1984.
- [26] Luis von Ahn, Manuel Blum, Nicholas J. Hopper, and John Langford. HIPs. <http://www.aladdin.cs.cmu.edu/hips/>.
- [27] Luis von Ahn, Manuel Blum, Nicholas J. Hopper, and John Langford. CAPTCHA: using hard ai problems for security. In *Advances in Cryptology, Eurocrypt 2003*, volume 2656 of *Springer Lecture Notes in Computer Science*, pages 294–311, May 2003.
- [28] Luis von Ahn and Nicholas J. Hopper. Public-key steganography. In Christian Cachin and Jan Camenisch, editors, *Advances in Cryptology - EUROCRYPT 2004: International Conference on the Theory and Applications of Cryptographic Techniques*, volume 3027 of *Lecture Notes in Computer Science*, pages 323–341. Springer, May 2004.
- [29] Paul Watzlawick, Janet H. Beavin, and Don D. Jackson. *Menschliche Kommunikation. Formen, Stoerungen, Paradoxien*. Huber, 2000.
- [30] Xerox PARC. *First Workshop on Human Interactive Proofs*, January 2002.
- [31] M. Zeleny. Management support systems: Towards integrated knowledge management. *Human Systems Management*, 7:59–70, 1987.