RC 20931 (07/23/97)
Computer Sciences/Mathematics

IBM Research Report

On Public-key Steganography in the Presence of an Active Warden

Scott Craver

IBM Research Division
T.J. Watson Research Center
Yorktown Heights, New York
email: sac@watson.ibm.com

IBM Research Division
Almaden · T.J. Watson · Tokyo · Zurich

On Public-key Steganography in the Presence of an Active Warden

Scott Craver
IBM T.J. Watson Research Center
P.O. Box 704
Yorktown Heights, N.Y. 10598
Department of Mathematics
Northern Illinois University
DeKalb, IL 60115
email: caj@niu.edu, sac@watson.ibm.com

Abstract

The so-called **prisoners' problem,** in which two individuals attempt to communicate covertly without alerting a "warden" who controls the communications channel, has taken a number of forms, adorned with various assumptions or requirements which make the problem more or less difficult. One assumption which makes the problem considerably more managable is that the participants are allowed to share some secret information (such as an encryption key) prior to imprisonment. Another assumption, which makes the problem much more difficult, is that the warden be allowed to modify messages sent between the prisoners as well as read them. This paper describes techniques for *pure* steganography, in which no secret information needs to be shared before imprisonment. First, a modification of an existing protocol will be shown to admit pure steganography if the warden is not allowed to modify the contents of the channel. Then, a technique will be described which in certain scenarios allows pure steganography between two prisoners in the presence of an active (content-modifying) warden.

This technique is possible through the use of two distinct channels rather than one: the subliminal channel for steganographic communication is augmented by a *supraliminal* channel, one in which information is not hidden from the warden but cannot be modified.

1 The prisoners' problem

The prisoners' problem was first posed by G.J. Simmons in [7], and is generally considered to be the *de facto* model of covert communication. In this problem, two people, usually named Alice and Bob[1], are thrown in prison and intend to co-author an escape plan. The problem is that all communication between them is arbitrated by a warden, here named Wendy, who will place both parties in solitary confinement at the first sign of any suspicious communication. Alice and Bob must trade inconspicuous-seeming transmissions which contain hidden information that, they hope, Wendy will not notice.

Using terminology agreed upon in [5], the inconspicuous data that is used to hide the real message is usually referred to as cover-data or *cover-objects:* a letter is often called a cover-text, for instance, while an image may be called a cover-image. The hidden or *embedded* message is placed therein, turning the cover-object into a *stego-object*. Alice's and Bob's goal, then, is to trade stego-objects without Wendy realizing that they are in fact stego-objects. This scenario is depicted in figure 1.

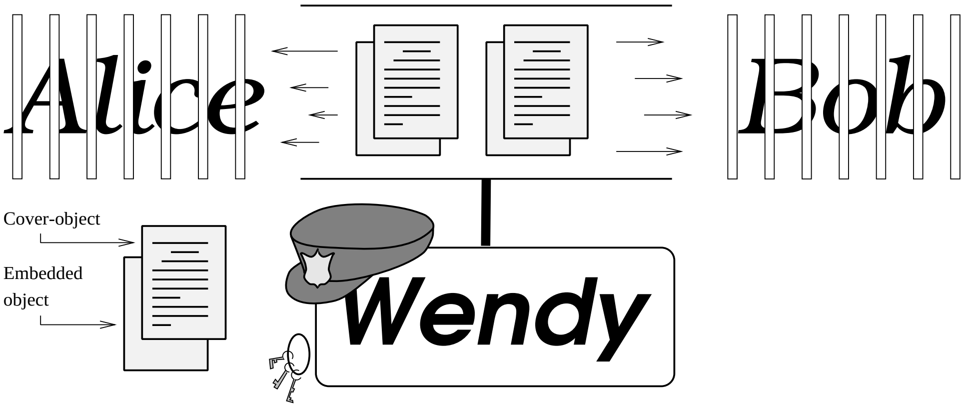

Figure 1: The prisoners' problem, illustrated

Further complications may hinder Alice's and Bob's escape. Wendy the warden may, in certain situations, be allowed to slightly modify messages as they are passed between the prisoners, to foil any hidden codes depending on the exact wording of the

[1]In the field of Cryptography, communications protocols usually involve two fictional characters named Alice and Bob. The standard convention is to name the participants in a protocol either alphabetically (Carol and Dave often succeed Alice and Bob in multi-person protocols), or with a name whose first letter matches the first letter of their role, such as Wendy the Warden, above.

communication between them. In this case we call Wendy an *active* warden; without this ability she is considered a *passive* warden. In the worst case, Wendy may be truly malicious, sending Alice and Bob bogus messages, in an attempt to expose their plans directly. One real-world example of an active warden is the censoring of telegrams by the United States government during World War II: the semantic content of telegrams could not be changed, but censors would slightly alter their exact wording, replacing words with close synonyms to foil possible secret codes [1].

In the field of steganography, the prisoners' problem serves as a metaphor for covert communication in general: Alice and Bob represent two people who wish to exchange secret information, giving no sign that communication of any kind is taking place. Wendy represents an eavesdropper, who may be able to both snoop and manipulate the communications path between the two, searching for evidence of any form of secret communication, and possibly attempting to prevent it.

Certain assumptions, perhaps unrealistic for an actual prison, are given to the problem to make it applicable to the real world. First, the two prisoners and the warden have computers at their disposal, and can trade data in a wide variety of formats (images, sound clips, text, etc.). Also, it is assumed that if either of the prisoners uses a specific algorithm for embedding a hidden message in a transmission, then the warden also knows of that same algorithm, and may suspect its use. In cryptography, the secrecy of an algorithm is rarely considered secrecy at all—it is often assumed that no matter which algorithm one uses to send covert communication, attackers will ultimately deduce which one is being used; hence security must rest entirely with the secrecy of the keys used.

It may be beneficial at this point to describe some common variations of the prisoners' problem. First, the warden's power to alter the transmissions between the prisoners affects the difficulty of the problem. This paper recognizes three distinct categories of warden:

- A **passive warden** can do nothing but spy on the communications channel between the prisoners.

- An **active warden** is allowed to modify (slightly) the data being sent between the prisoners. Mild modification of text which does not alter its semantic content (say, replacing words with close synonyms) is an example of an active warden being active. The active warden must not modify data so much that innocent communication would be foiled.

- The case of a **malicious warden** is the most difficult. A malicious warden

> would be one who may alter the prisoners' messages with impunity, perhaps composing entire messages for one prisoner while pretending to be the other. This was the kind of warden described in the original statement of the problem, which focused on the problem of authenticating messages sent by a warden attempting to deceive the prisoners [7].
>
> Fortunately, most real-world situations prevent a warden from grossly altering the content of messages. Imagine the confusion if a large number of telegrams sent during World War II were altered in meaning, suppressed, or entirely fabricated by crafty censors on the lookout for spies! In this paper, the term "active" denotes a warden who can modify content, but is not malicious.

As for the prisoners themselves, it should be pointed out that in the best case, they would not have to communicate prior to imprisonment, so as to (say) trade an encryption key. This best-case scenario, here called **pure steganography,** is very difficult to engineer. One would suspect that it is not even possible in the presence of a truly malicious warden, for without any prior information Alice has no way to distinguish between Bob and Wendy. Steganographic protocols typically assume that some information must be shared between the prisoners prior to imprisonment. If this assumption was not allowed, little progress in information hiding could have been made to date.

The remainder of the paper is organized in the following fashion. In section 2 it will be shown how steganographers have managed to send information covertly in the presence of an active (and in some cases malicious) warden, provided that information such as secret and public keys can be traded beforehand. In section 3, we will see how a protocol, described by Ross Anderson in [1], allows steganography in the presence of a passive warden with only one prisoner knowing the other's public key. A modification of this protocol will be shown to admit *pure* steganography in the presence of a passive warden. Finally, section 4 will describe what are here called *supraliminal* channels, which allow pure steganography in the presence of an active (but not malicious) warden. The paper will close with a discussion of the feasibility of supraliminal channels.

2 Private-key steganography

Let us assume that Alice and Bob are allowed to share a secret key prior to imprisonment, or even to trade public keys. This gives them the opportunity not only to communicate covertly, but to defeat an active warden. In the former case, steganography consists merely of encrypting a message in such a way that the ciphertext appears statistically random, and embedding the bits of the text in a known subliminal channel. The embedded information, of course, must be made to have the same distribution as the channel noise in order to foil statistical tests.

In the presence of an active warden, it is not enough to embed a message in a known place. If Alice can subtly alter the bits in an image, it follows that Wendy could scramble those same bits with as little impact, erasing whatever was being sent via the subliminal channel. In this case it is possible to use what is referred to in [1] as a "selection channel." Essentially, the secret information shared between Alice and Bob is used to determine *where* the message is hidden. A cryptographically secure pseudo-random generator, seeded by a secret key, can be used to pick a subset of pixels in an image, for instance, to be used to conceal the data. If Wendy attempts to make subtle changes to the image, she may only be able to scramble a small percentage of the actual channel bits, since she doesn't know exactly where they are. This scrambling can then be fixed using an error-correcting code.

The sharing of keys before imprisonment, however, is a requirement that we would ultimately like to see removed. It allows a great deal of freedom on the part of Alice and Bob, but it is not reassuring to think that if two people ever need to communicate covertly, they must know so far in advance that they can trade secret keys before a real-world "warden" starts listening in.

3 Public-key steganography

3.1 Boiling out the impurities

In public-key cryptography, it is not necessary for two people to share a secret key to establish a secure channel. One only needs to know the the other's public key. This suggests a possible approach to steganography in which a secret key does not have to be agreed upon by Alice and Bob prior to imprisonment. Some information must still be known *a priori*— one prisoner must know the other's public key — but from a practical perspective this is a much more reasonable requirement.

A protocol which allows public-key steganography in the presence of a passive warden is described by Ross Anderson in [1]. It relies on the fact that an encrypted message can be random enough to hide "in plain sight:"

1. Alice, knowing Bob's public key, encrypts her message with it to obtain an

apparently meaningless ciphertext C.

2. Alice embeds C in a channel known to Bob (and, hence, to Wendy). The resulting stego-object is sent to Bob.

3. Bob has no idea if any message is hidden in the channel or not, but we can assume that the technique is standard enough that if he suspects a message, he will look for it. He retrieves a random-seeming string, attempts to decrypt it with his private key, and out pops Alice's message.

One problem with this approach is that Bob has no idea if anything is being sent: he may not even know Alice, and certainly does not know if she intends to use a steganographic channel. If the two traded a private key before imprisonment, at least Bob would know that some secret transmission was pending. In this case, Bob will just have to suspect that a hidden message might be present in any cover object he receives.

This is not too serious a problem, however: it is already assumed (as it usually is in cryptography) that the information-hiding technique used by Alice is known to all, and standard enough that Wendy would suspect its use. Certainly Bob can, too. As long as hidden content is suspected and can be easily extracted by a known method if it does exist, it is not unfair to assume that it will be discovered.

A more practical, related problem is that in a large group of possible recipients every single recipient must suspect hidden content in each object. Only the intended recipient will find it, of course; but in an environment such as a USENET newsgroup, where a large number of people may send stego-objects, each of which may be targeted toward an unspecified recipient, all the parties involved may spend the better part of each day looking inside every object. This is related to the previous problem, since it is the result of the message sender not being able to give the intended recipient any warning that a message is being sent.

3.2 Pure steganography using the Anderson protocol

The assumption made above is not to be taken lightly, for it suggests that the protocol described in [1] is actually more powerful than may initially be suspected. If we can assume that Bob will attempt to decode what might or might not be a hidden message, then we can also assume (at least in the two-prisoner case) that Bob may attempt to participate in what might or might not be an ongoing steganography protocol. For example, suppose that Alice and Bob are astronomers residing on separate planets, each searching for signs of intelligent life in the universe. Bob receives from Alice's planet what might be a message, or what might be noise. He may decide to treat the signal as if it were a message, send one back in the same "language," and wait for a second signal to confirm his suspicion.

In the context of steganography, the situation is complicated by the fact that the signals being traded must be meaningless to everyone but Alice and Bob. We can imagine that if the search for extra-terrestrial Alice ends up beaming a meaningful signal to the nearby warlike Wendy, Bob may give away that his planet is inhabitable and ripe for colonization. On the other hand, the situation is made less complicated by the cryptologic assumption that whatever data-hiding or encryption algorithm is used is known to all.

In any case, the possibility that Bob may participate in a protocol on suspicion of a steganographic message suggests a method by which Alice may send Bob her public key while they are both imprisoned, *entirely eliminating the need for prior communication.* A passive warden is still required for this to happen, but this too can be relaxed, as will be demonstrated in the next section.

Protocol 1 *A modification of the protocol described in [1]*

This is essentially Anderson's protocol with an extra step at the beginning. It is assumed that a method of information-hiding is known to all, as well as a public-key encryption algorithm. Wendy is assumed to be passive.

1. Alice generates a random public key/private key pair.

   If she is using RSA, for instance, she could generate the large prime numbers used for the key by picking cryptographically pseudo-random strings and testing for primality until one is found. She then selects another pseudo-random string for the public key. This is usually the way RSA keys are created.

2. Alice embeds the key in the subliminal channel, viewable by both Bob and Wendy, and sends the resultant stego-object to Bob.

   Neither Bob nor Wendy can determine if the channel contains noise or a key. If Wendy can only snoop, then Bob can attempt to communicate:

3. Bob suspects that the data in question is a public key, uses it to encrypt a short note of acknowledgement with a randomly chosen secret key for future covert transmissions, and sends this back to Alice embedded in a stego-object.

4. Alice suspects that the returned data is an encrypted key, decrypts it with her private key, and out pops a message from Bob. Now the two can communicate by an appropriate secret-key protocol.

At no point in the protocol does Wendy have any idea that the randomness means anything: Alice alone can deduce content from Bob's reply. As long as she is not allowed to insert her own information (and thus either foil the transmission or maliciously insert her own key to catch the two in the act), she can not conclude that communication has taken place. □

If Wendy is capable of writing to the channel, then there is no way communication can take place: if Wendy does not utterly destroy the "in plain sight" bits, she can attempt a man-in-the-middle attack by overwriting Alice's key with her own. A more malicious warden could entirely spoof either Alice's or Bob's response. In that case, however, the original protocol would not work either, since in neither case would Bob have any way of identifying the author of the original message.

4 Public-key steganography in the presence of an active warden, via perceptually significant channels

Cover-plots and Supraliminal Channels

If we assume that the Warden can only make minor modifications to the possible stego-objects sent between Alice and Bob, then we can assume that there is some amount of *perceptually significant* information that the warden cannot change whatsoever. For instance, if Alice sends Bob a picture of a cow and Wendy can only modify 1 bit in every 100, we can assume that Wendy will not be able to turn the cow into a pig. In a novel, there could be explicit states of affairs or descriptions of characters so relevant to the plot that no information about those states can be changed without a significant rewrite of a number of portions of the book. If we develop some formal encoding of object and state-of-affairs descriptions, we have the makings for a channel through which Alice can send a small amount of information to Bob out in the open, but with high integrity. What we are describing here is a *supraliminal* channel rather than a *subliminal* one: information is hidden in plain sight, so obviously, in fact, that it is impossible to modify without gross modifications to the transmitted object.

Formally, we can define a *cover-plot* function $f : \mathbb{S} \rightarrow \{0, 1\}^N$, where $\mathbb{S}$ is a set of possible encodings of state-of-affairs descriptions, or "cover-plots" — so called because they can be considered the plot, or content, of a cover-object. A cover-plot S is simply a formal description of content appropriate to the medium of the cover-object. Such a description could be used by an author or content-creator to compose a covertext C containing an *embedding* of S, and hence an embedding of the bit-string $f(S)$. The cover-plot function would have the following properties:

- The function would be known to both Bob and Alice (and hence we must assume that it is also known to Wendy).
- A member of the inverse image under f of any particular bitstring is feasible to compute, as is f itself.
- f is onto. It is not necessary for f to be one-to-one—indeed, if f is many-to-one Alice's job will be significantly easier, since she will have more freedom in choosing a cover-object which embeds a particular bit-string.
- A cover-object C can be created to contain a state-of-affairs description S such that an active warden cannot significantly change the encoding of S (that is, she cannot change S to S' such that $f(S) \neq f(S')$) without making unreasonable, wide-sweeping changes to the entire cover-object.
- A human being (or, ideally, a computer program) can deduce the exact state-of-affairs description embedded within a cover-object. The composer of the covertext can facilitate this deduction by making the embedding very clear, but this should not be so necessary that an object with an easy-to-interpret embedding is conspicuous.
- Clarifying the last property, an innocent cover-object without an explicitly embedded state-of-affairs description S can yet be interpreted as having one. That is, innocent cover-objects should have clearly deducible cover-plot descriptions (which map under f to random garbage) so that the presence of a clearly deducible cover-plot description is not indicative of a hidden message, unless the bit string $f(S)$ possesses some meaning.

This can be summarized as a channel that is **robust**, **blatant** and **inconspicuous.** Here, a **robust** channel is one whose content cannot be altered without making unreasonably drastic changes to the stego-object (i.e., requiring a malicious, instead of an active, warden). A **blatant** channel is one whose content is publicly accessible, a necessity if Alice is to send Bob information when no key is shared between them prior to transmission. For a channel to be **inconspicuous** as well as blatant may sound like a contradiction, but it is not: an inconspicuous channel is one whose existence alone is not indicative of a hidden message. A perfectly innocent image, containing no covert information, will seem to the warden to contain random cover-plot bits. If Alice sends Bob a cover-object with an obvious message in this supraliminal channel ("Bob: here is the secret plan to break of out of jail. Love, Alice") then Wendy can throw the two into solitary confinement. However, if Alice sends Bob apparently random cover-plot bits, perhaps a random session key to be used later, Wendy cannot conclude that the message is anything other than innocent, since innocent messages also contain random cover-plot bits. Only when that key is later used can Wendy know that covert communication is taking place.

One cannot help but notice the connection to digital watermarking, a form of information hiding in which the hidden data is required to be robust to significant modification of an image. In particular, a recent approach to watermarking, advocated by Ingemar Cox, *et al* in [2], has resulted in robust watermarking schemes using what largely resembles a subliminal channel with supraliminal aspirations. Cox, *et al* emphasize the importance of hiding a watermark in *perceptually significant* components of an image, such as high-magnitude DCT matrix coefficients, so that an ownership label can survive a significant amount of abuse by a forger. Further, the watermark can still be made invisible, as the technique presented in [2] demonstrates. One may wonder if a similar scheme would allow the embedding of a supraliminal channel invisibly in an existing image, so that Alice and Bob need not compose a new image for each message for the explicit purpose of embedding a particular string of cover-plot bits.

Unfortunately, conceptual differences between robust watermarking and embedding bits in a supraliminal channel prevent the application of one to the other. In the former case, some sort of secret key is often used to embed a watermark, so that only those who know the key can detect or remove the watermark. A cover-plot, on the other hand, must be readable by all but removable by none. Also, the purpose of a supraliminal channel (and, indeed, steganography in general) is to hide a specific message of some meaning to the recipient; In the case of invisible watermarking, the embedded labels need not have any semantic content at all, or may be a function of the image. The scheme described in [2], for instance, embeds a vector of pseudo-random numbers picked from a normal distribution. In [3], it is discovered that in order for this scheme to become secure it may be necessary for the vector to be a function of a one-way hash of the image itself. In short, we have a difference of priorities: in watermarking, an ownership label can be meaningless, or picked to fit the cover-object, while the cover-object's content is given beforehand and is important. In steganography, the image can be meaningless, or picked to fit the embedded message, while the hidden message's content is given beforehand and is important.

Finally, if it was possible to embed a supraliminal channel in an existing image (for example) without significantly altering it, it would be equally possible for a warden to scramble the channel without significantly altering the image, since everybody knows where the information is and how to bury cover-plot bits inside. Hence the content of the cover-object should itself be a function of the information embedded within it. It may be hasty to declare it impossible for a supraliminal channel to be subliminal, but the two concepts seem to have irreconcilable differences.

Example 1 *A rudimentary supraliminal channel for audio/video clips*

In this example, the cover-object may be an audio-video clip in which people are speaking. The noise within the clip could be used to embed a subliminal channel. The set $\mathbb{S}$ of cover-plots is the set of all texts which can be spoken by Alice and unambiguously understood by Bob. The cover-plot function $f : \mathbb{S} \rightarrow \mathbb{N}$ is computed by hashing the text of each sufficiently long word, assigning each a numeric value.

For instance, suppose that each letter of the alphabet is assigned some agreed-upon numeric value. When an audio clip is received by Bob, he types into his computer all words of (say) at least six letters. For each word the computer calculates the product of the letters' numeric values, modulo some small prime p. These values can then be arranged left-to-right and treated as the base-p representation of an integer[2].

Alice must be able to compose a convincing body of text which hashes to a desired array of values. She has some flexibility in that smaller words can be used with impunity to generate a context for the larger words. She can then use a dictionary (or, more likely, a thesaurus) which can be searched by cover-plot values. Notice that the larger the number of values to which a word can hash (the larger the value of p in the above example), the harder Alice's job will be, and the more conspicuous a cover-plot might be given a small amount of time to create one. On the other hand, if the number of possible hash values is made small, the ease in composition on Alice's part is offset by a lower bandwidth.

[2]It would be necessary to keep in mind the frequency of letters and letter groups in the language used when deciding upon their numeric values (not to mention making sure that no value is a multiple of p), so as to prevent patterns from showing up in the resulting array of hash values.

Once this composition step is complete, Alice turns on a video camera and records herself (or other people) reciting the body of text she composed.

This channel is *blatant,* as long as the technique is made publicly available and the letters' numeric values are a known standard. The channel is *inconspicuous,* since every text document, innocent or not, has a cover-plot. Finally, the channel is *robust* unless Wendy is capable of seamlessly altering both the sound of Alice's voice and her lip movements so as to change the text of her speech. While this is possible with today's technology, there exist real-world domains where Wendy does not possess the time to do this (say, where she has to monitor hundreds of such transmissions per day or per hour), or where the cost in time and computation is too high to justify the altering of video clips on the mere suspicion that their text may have some future steganographic significance. Alternatively, she could completely re-record the message herself, hoping that Bob does not know what Alice looks or sounds like, but this would likely classify her as a malicious warden. □

A formalized method for describing the plot, characters, etc. of a letter or book, or a describing the overall content of an image may also serve as good cover-plot functions; one could use details so central to a cover-object's content that to change them would require significant changes throughout the object. Perceptually significant components of certain transforms of the cover-object may also do the trick, provided one can feasibly compose a cover-object when given particular transform components that it must contain. This could make Alice's job much more difficult, but Bob's would be so simple that it could be completely automated. If a robust supraliminal channel can be entirely automated, requiring no (or negligible) human intervention in composing or interpreting a work, it could be quite a significant contribution to the field of steganography.

Public-key steganography with a supraliminal channel

Of course, a supraliminal channel is not appropriate for *subliminal* communication between Alice and Bob, for a number of reasons. First, the bandwidth will likely be very small, since the channel is engineered to be highly robust. The number of deducible states-of-affairs may be larger for a novel than for an image or video clip, but in either case it is unlikely that a single object could hold a reasonably sized message. A robust channel of this kind is more appropriate for transmission of a session key, as it will be used below.

Secondly, it is expected that Wendy also knows how to compute f, so any information passed through the channel is available to her. Unless the cover-plot bits have no apparent meaning, Alice and Bob are in serious trouble (or at least more serious than what previously got them thrown in prison).

With public-key cryptography, however, such a channel could be used to defeat an active warden. Usually it is assumed that Alice and Bob share a secret key. With a supraliminal channel, however, *a secret key can be exchanged covertly with no prior information,* right in front of an active warden. The protocol is described below and in figures 2 and 3.

Protocol 2 *Key exchange in the presence of an active warden*

In this protocol, Alice and Bob have no information prior to imprisonment. It is assumed that a public-domain cover-plot algorithm (as well as this protocol) is known to all, including Wendy. Alice and Bob perform the following steps:

1. Alice generates a random public key E and a private key D.

2. Alice then finds a cover-plot $S \in f^{-1}[E]$ within the inverse image of E.

3. Alice composes a cover-object C_1 containing an embedding of the cover-plot S, and sends this to Bob.

   At this point, Wendy can inspect the channel, and will find a random-looking string of cover-plot bits. There is no reason to expect anything unless the data is later used in some way, say as a session key. The cover-plot bits pass unmolested through the channel to Bob.

4. Bob extracts the cover-plot bits, suspecting that they represent a public key. He creates a pseudo-random secret key r, which he will later use to embed information in a subliminal channel, and encrypts it with the purported public key E to get a pseudo-random string $s = E(r, K_{\text{Alice}})$[3]. He then finds a cover-plot $T \in f^{-1}[s]$, and creates a cover-object C_2 containing an embedding of T.

5. Bob now uses his secret key r to securely embed a message in the *subliminal* channel of the image, using an existing secret-key-based technique which the active warden cannot defeat.

[3]He may also wish to append to r a note of acknowledgement before encrypting, so that Alice will know that the protocol is working.

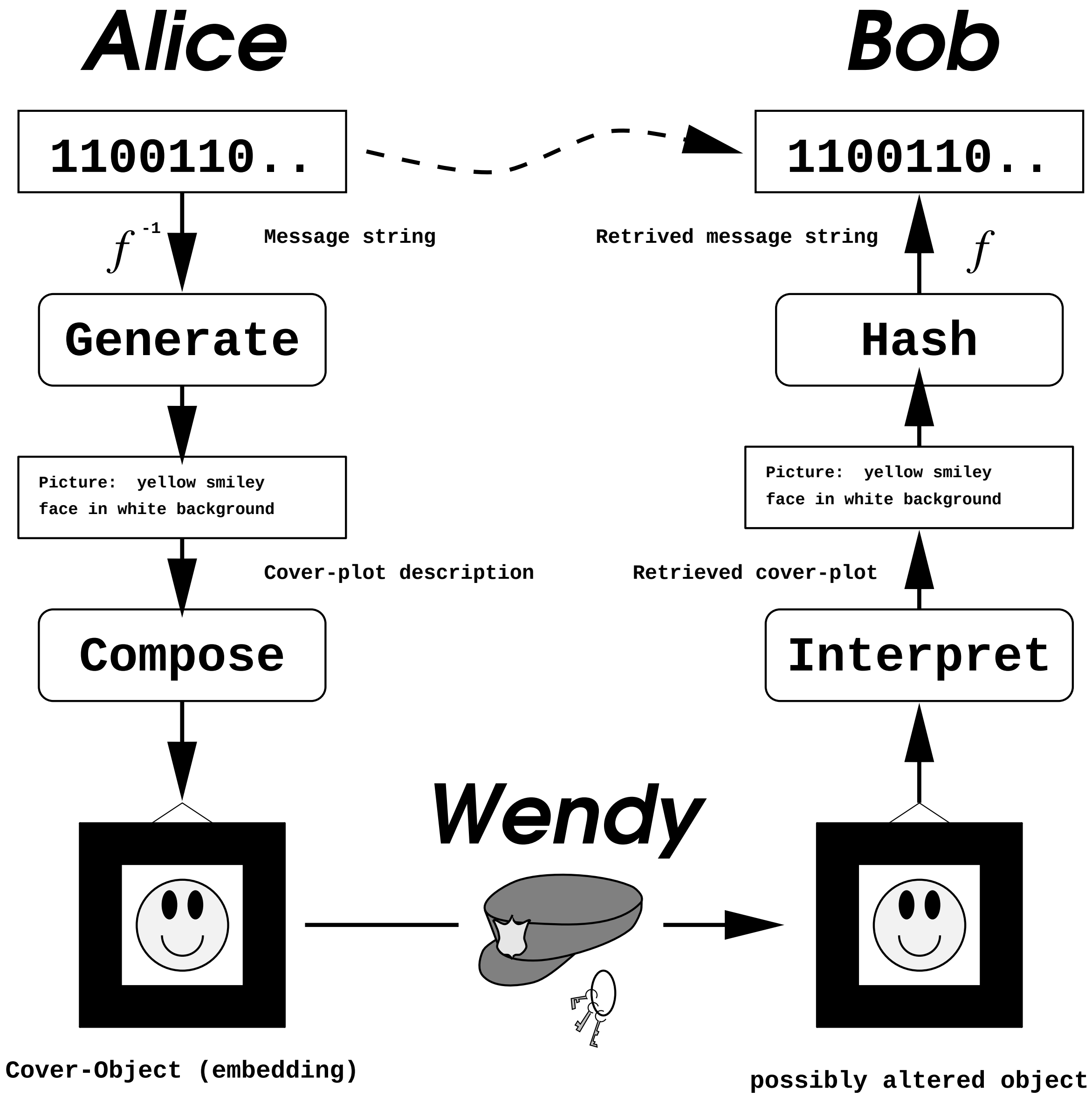

Figure 2: How a cover-plot function is used to establish a supralminal channel.

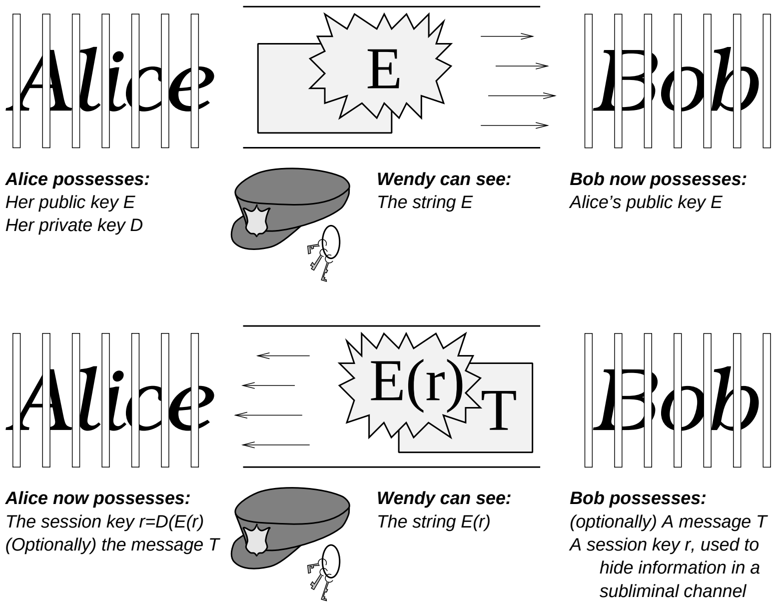

Figure 3: Protocol using supraliminal and subliminal channels. Beneath each character is listed the information he or she knows once the illustrated step has taken place.

If the medium used to implement the supraliminal channel is not rich enough to support a subliminal channel, Alice and Bob can always postpone this step and embed their secret messages in the subliminal channels of subsequent cover-objects.

6. Bob sends C_2 to Alice.

   At this point, Wendy can again snoop the channel, and now has (what she might suspect is) both E and $E(r)$. Even if she does suspect this, there is no indication that the strings are related in this way, and no way she can determine the value of r, and thus no way she can determine the existence of any subliminal message. Further, both strings are random-seeming enough that there is no indication that any covert communication is occurring. She sends the message through the channel to Alice.

7. Alice decrypts what she assumes to be $E(r)$ to get $r = D(E(r))$, and, optionally, snoops the subliminal channel using the secret key r to find a message.

8. Alice and Bob now communicate with impunity using a secure secret-key based scheme, with the random key r that they now both share. Wendy has no reason to suspect that a key exchange has occurred.

Statistical tests on the cover-plot bits may yield some deviation from what would be expected in an innocent object. However, the strings Alice and Bob share are already the product of cryptographically secure pseudo-random generators, and if the cover-plot bits of a certain medium typically fall into a non-uniform distribution, the bits can always be encoded in a manner that matches this distribution. Further, the strings transmitted via this channel are very short, and only two transmissions are needed to share a key. This may make statistical tests difficult. □

5 Are supraliminal channels fair?

There are a number of possible objections that may be made to the protocols presented here. First, just how fair is it to assume that the warden is not malicious? Why *not* impersonate Alice or Bob, or grossly modify their work for the purpose of foiling the possible covert communication? After all, Bob has no idea what cover-object to expect from Alice. He would not necessarily know if Wendy completely changed the transmitted object to something else.

The reason why Wendy can not be malicious in typical real-world situations is quite clear, but is sometimes forgotten when we focus entirely on the situation of two prisoners whose very names indicate that they intend to communicate covertly. A real-world warden, be she a postal censor or a system administrator, may deal with communications between a large number of people. At best she can suspect that trading of secret information is taking place, and may not be sure where or between whom. To significantly alter the meaning of large number of e-mail documents or telegrams is simply unacceptible behavior. Even if the recipients of those documents have no idea what content to expect, all subsequent communication would either give away the ruse or need to be intercepted and modified as well. David Kahn in [4] describes the concept of a "semagram," a secret communication not in written form:

> For example, during World War 2, censors intercepted a shipment of watches. Fearing that the position of the hands of the watches could be spelling out a secret message, they changed the position of the hands.

If no secret message existed in those watch-hands, the recipient would likely not care about the modification. On the other hand, if the recipient ultimately recieved a shipment of radial tires instead of watches, he or she would have legitimate reason to complain, as would everyone else who received a package subjected to a similar censoring methodology. Even if the recipient had no idea what content to expect, the resulting "thank you" note would be confusing indeed. Unless the United States could intercept all future transmissions from the recipient which allude to the shipment's content, the censorship will ultimately be discovered. In any case, such gross modification of potentially innocent content is unreasonable if the channel is to be trusted by innocent people. As Einstein may have put it, "Raffiniert ist der Vorsteher, aber boshaft ist er nicht."[4]

The big question-marks, however, regard the very feasibility of the supraliminal channel: can it be done efficiently? Further, is it fair to allow the cover-object to be a function of the information being sent? It is often assumed that the cover-object is given, and any message could be inserted therein. But the cover-object is assumed to be meaningless to Alice and Bob anyway, a mere front for the *real* message[5]; we might as well use the opportunity to hide actual content.

[4] "Subtle is the warden but not malicious."

[5] One can imagine a sneaky warden successfully catching Bob by watching him pore over a 6 page cover-text for 2 hours and then quizzing him on its content.

Efficiency is the real problem, in this author's opinion. Cover-plot functions mapping plot descriptions of stories or other high-level content would probably require a human being to perform both the composition and interpretation steps, and quite possibly the generation step as well (see figure 2.) One could develop a scheme which makes the interpretation step easy enough for a computer to manage, but in order to automate the entire process it may not help to turn a cryptographic problem into an artificial intelligence problem. If a supraliminal channel could be based on perceptually significant information that a computer is suited to both interpret and convincingly compose, this hurdle would seem much less daunting.

6 Conclusion

The prisoners' problem exists in many variations, often assuming that the prisoners share some information prior to imprisonment. *Pure* steganography, without any *a priori* secret knowledge between the communicating parties, is very difficult, but not necessarily impossible: a public-key steganography protocol has been described in [1] which requires only that one party's public key be known to the other. Furthermore, certain assumptions allow this protocol to be modified so as to allow pure steganography in the presence of a passive warden. Finally, if a *supraliminal* channel as described in this paper can be feasibly implemented, secret key exchange and hence pure steganography can take place in the presence of an active (but not malicious) warden.

References

[1] R.J. Anderson, "Stretching the Limits of Steganography." in *Information Hiding,* Springer Lecture Notes in Computer Science v 1174 (May/June 1996), pp 39–48.

[2] I.J. Cox, J. Kilian, T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," Technical Report 95-10, NEC Research Institute, 1995.

[3] S. Craver, N. Memon, B. L. Yeo, and M. M. Yeung, "Resolving Rightful Ownerships With Invisible Watermarking Techniques:, Limitations, Attacks, and Implications" to appear in *IEEE Journal on Selected Areas of Communications*,

special issue on Copyright and Privacy Protection (IBM Research Report, RC 20755, March 1997.)

[4] D. Kahn, "The History of Steganography," in *Information Hiding,* Springer Lecture Notes in Computer Science v 1174 (May/June 1996), pp 1–6.

[5] B. Pfitzmann, "Information Hiding Terminology." in *Information Hiding,* Springer Lecture Notes in Computer Science v 1174 (May/June 1996), pp 347–350.

[6] B. Schneier, *Applied Cryptograpy: Protocols, Algorithms, and Source Code in C, 2nd ed.* New York: John Wiley and Sons, 1996.

[7] G. J. Simmons, "The Prisoners' Problem and the Subliminal Channel." *Advances in Cryptography: Proceedings of CRYPTO '83,* pp 51-67.