Electronic Document Distribution

N. F. Maxemchuk
AT&T Bell Laboratories
Murray hill, New Jersey 07974

ABSTRACT

Computers, printers and high rate data transmission facilities are becoming less expensive and more generally available. It is now possible to electronically distribute customized newspapers or magazines, instead of using a network of trucks and vendors to distribute a common paper version. A major obstacle to the use of electronic distribution is the ease of copying and redistributing electronic documents. This capability threatens a publisher's subscription revenues.

We are making it more difficult to redistribute electronic documents by

1. marking documents so that they can be traced to the original recipient,
2. using cryptographic techniques so that the form of the document available to the recipient costs more to redistribute than the form that the publisher distributed, and
3. forcing a person who redistributes a document to give away personal information with the document.

These techniques are being applied in experiments to electronically distribute an issue of a technical journal, and to mark and register paper copies of confidential executive memoranda.

1. Introduction

The electronic distribution of newspapers, magazines, and other printed material can become a major, new application of communications networks. The price of home computers and printers is decreasing, and the quality of the material that can be printed locally is improving. In addition, the transmission rates available to the home are increasing, so that the time required to receive electronic documents is now reasonable. At the same time, distributing paper copies of newspapers and magazines with a network of trucks and vendors is becoming more expensive. Eventually, electronic delivery and local printing will become preferable to printing the document in a central location and delivering paper copies. A number of electronic distribution applications already exist, including an on-line journal of Current Clinical Trials that is published by the American Association for the Advancement of Science, and a version of the New York Times, called Times Fax.

Electronic document distribution can provide services beyond those that are currently available. Paper documents that are mass produced in a central location must meet the needs of all of the intended recipients. Electronic documents, that are printed locally, can be customized for the individual recipients. When a document is read in electronic form, as is now done in NCSA's Mosaic, the information can be organized so that the reader can start with a high level description of the information, and obtain more details as needed. In addition, when a reader is connected to a network while reading one document, pointers to other documents can be used to quickly retrieve additional information on a topic.

Some of the advantages of electronic document distribution are also the major disadvantages. In particular, the ease with which electronic documents are copied and distributed is both an advantage and disadvantage for the publisher. The recipient of an electronic document can make an identical copy and redistribute it by electronic mail or post it on an electronic bulletin board. Redistributing electronic documents threatens an electronic publisher's subscription revenues.

In the near future, coping and redistributing electronic documents will also threaten the publication of paper documents. Scanners that convert paper into electronic form, such as those used in fax machines, are improving in quality and are becoming widely deployed. As scanning technology improves, electronic

publishing will provide an advantage over paper publishing since there are more ways to discourage the redistribution of individually generated electronic copies than there are to discourage the redistribution of mass produced paper copies.

Electronic publishing is different from typical security applications because it is not a publisher's intent to keep information secret. A publisher makes a profit by making information public. The techniques presented here discourage, rather than prevent, copying, and give the publisher advantages over the redistributor:

— Documents are marked so that illicit copies that are recovered can be traced back to the original recipient.

— Cryptographic techniques are used so that documents that are tampered with, to disguise the original recipient, are recognizable.

— A different form of the document is distributed by the publisher than is available to the redistributor, in order to decrease the cost of marking to the publisher and to increase the transmission cost to the redistributor.

— The redistributor may be forced to give away personal information with the document.

In section 2, three techniques are described to mark documents. Each copy of a document that is distributed by the publisher is unique in an imperceptible way, and is registered with the original recipient. If illegitimate copies are found, they can be traced back to the original recipient.

It is possible to remove or hide the identifying marks by processing the electronic version of the document. Our objective is to make the processing needed to remove the markings as difficult as possible, to make the quality of a document become degraded when the marks are removed, and to detect documents that have been tampered with to remove the marking. In many instances, the ability to identify a bootleg copy of a document is a sufficient reason for a document recipient to obtain a legitimate copy from a publisher.

In section 3, two systems for electronic distribution are described. The systems use cryptographic techniques to protect a document from being intercepted during transmission, to make different representations of the document available to the publisher and recipient, and to reduce the cost to the publisher of generating unique documents. The publisher distributes the same encrypted copy of the document to everyone. With each copy, the publisher also distributes a unique identifier and instructions for marking the document. In the first system, tamper resistant decryption hardware is used in the printers so that the electronic version of a document is not available to the recipient. The recipient only has access to a marked paper copy. In the second system, the document from the publisher is decrypted in software, but the recipient is only given access to a marked copy of the electronic document. In addition, when the document is decrypted the representation is changed to increase the number of bits used to represent the document. This gives the publisher an advantage in transmission cost and time over an illegal redistributor.

In section 4, three experiments in document marking and distribution are described. The first experiment is an application of one of the marking techniques to an actual document, the second an application to marking and registering private memos, and the third a trial distribution of a professional society journal.

2. Marking

Identifying different versions of a document by marking them is not a new idea. A cartographer who makes a street map of an area adds or deletes a small detail to identify his map. The ability to process and print individual copies, rather than just making identical copies, makes it possible to extend this idea from identifying versions to identifying individual copies.

The text analogy to cartography is to change the content slightly, possibly by substituting synonyms. This type of technique is not considered. Our techniques do not change the content or perceptibly alter the appearance of the document. Three marking procedures are considered

1. modifying the space between adjacent lines of text,

2. modifying the space between words, and

3. modifying the edges of characters or boundaries in figures.

Every copy of a document has a different set of markings to identify it. The marking techniques send encoded information through a document, as if the document were a communication channel. However, document marking is different from the classical communications problem because the decoder can be given access to information at the source as well as the received signal. For instance, the decoder is given the location of the encoded information, while an opponent who wants to remove the marks may not be able to determine their position.

The marking techniques are most effective when the information can be decoded after the markings have been distorted by printing the document on paper, copying and faxing the paper, then scanning the paper back into the computer. These processes introduce a variety of nonlinear types of distortion into the document. For instance, after copying, the shading may be darker on some sections of the page than on others. The types of distortion that have been observed vary gradually over a page. This makes it possible to reduce the effect of the distortion by predicting its value from adjacent regions. It has also been noted that when documents are copied or faxed, the replica is almost never the same size as the original. It is usually a few percent larger or smaller. Differential encoding mechanisms, that will be described, are effective against the types of distortion encountered in documents.

The number of bits of information that can be included in a document is proportional to the number of lines, the number of words, and the number of characters in the document. On a single page with only 40 lines of text, and 12 words or 75 characters per line, on the order of 10^{12}, 10^{144} and 10^{300} distinct markings can be produced, respectively. Clearly, in a typical document, there are many more bits available than are needed to assign a unique number to each document. The extra bits are used to:

1. design encoding mechanisms that are easier to decode and are more robust against distortion,

2. avoid inserting information in locations where it may be noticeable or easily lost because of distortion,

3. insert redundancy for error or erasure correction,

4. hide the few bits that carry real information in a large number of bits that carry random signals, and

5. make it difficult for anyone except a publisher to create a valid marking.

The first two points are specific to the individual marking techniques, and are addressed in the respective sections. The last three points are common to all of the marking techniques, and are addressed here.

The function of both error correction and hiding marks is to force an opponent who is trying to eliminate an identifier by adding distortion, to add more distortion, and thereby further degrade the quality of the document. Codes can correct more erasures than errors, and the procedures for filling in erasures are simpler than those for correcting unknown errors. Therefore, when possible, the decoder should detect the location where an opponent has added distortion. Hiding information among a large number of random bits also forces an opponent to add more noise than he would have to if the location of the information were known. Hiding information among random bits does not provide as much protection as correction techniques, but is much simpler to implement.

One attack on the marking system is for an opponent to change the mark before redistributing a document so that cannot be traced back to the original recipient. This attack can be prevented by using only a small subset of the possible marks, so that it is unlikely that an opponent who change the marking will select a valid mark. For instance, assume that there are 32 bits available for information, but only 16 are required to number a document. The publisher can pick the 2^{16} numbers at random from the 2^{32} available numbers. If an opponent removes the publisher's number, and inserts one of his own, in order make it appear to be someone else's copy, only one in about 8000 numbers that the opponent picks is a valid number. When a document doesn't have a valid number, it is detected as being illegitimate.

Rather than picking the valid numbers at random and keeping a list of the numbers that are used, valid numbers can be created using techniques that are similar to digital signatures. Specifically, encode a message that includes both the document number and an encrypted function of both the message and other

information. In a digital signature system, the encrypted information is only a function of the message, so that the receiver can verify that the message was sent by a transmitter that could perform the encryption. In the publishing environment, the encrypted information includes information that is specific to the document to prevent an opponent from taking a number off a copy of one document and inserting it on a copy of a different document. Including information that is not transmitted through the communication channel takes advantage of the ability of the decoder to obtain additional information from the encoder in the document marking environment.

2.1 Line Space Encoding

In line spacing encoding, alternate lines of text are held stationary and the position of every other line in the document is moved up or down by a small amount, as shown in figure 1. One bit is transmitted in each line that is moved. For instance, if the line is moved up, a one is transmitted, and if it is moved down, a zero is transmitted. In order to make decoding easier, the first and last complete lines in paragraphs are stationary, and partial lines at the end of paragraphs are not used at all. More information can be transmitted by line spacing if every line is moved, and if more than two positions are allowed for each line. However, our approach provides sufficient information to mark most documents and is easier to decode.

To decode the information from a paper copy of the document, the copy is scanned into a computer and conventional document processing programs are used to reorient the horizontal lines in the document. The centroid of each line is calculated as the center of mass of the line about a horizontal axis. The centroids of lines in an unmarked document are not uniformly spaced because their position is dependent upon the number of characters that extend below the line or above the middle of the line. Encoding is performed with movements of a centroid that are less than the difference from uniform spacings that may occur naturally. In order to decode with these small movements, the decoder is given the position of the centroids in the unencoded document as well as the received document.

Decoding is performed by comparing the differences in centroids in the received document and the original. In the received document, let $\Delta_{R,+}$ be the distance between the centroid of a line that has been moved and the centroid of stationary line above it, and $\Delta_{R,-}$ be the distance between the line that is moved and the stationary line below it. $\Delta_{X,+}$ and $\Delta_{X,-}$ are distances between the centroids of the same lines, in the unmarked document. If $\frac{\Delta_{R,+}-\Delta_{R,-}}{\Delta_{R,+}+\Delta_{R,-}} > \frac{\Delta_{X,+}-\Delta_{X,-}}{\Delta_{X,+}+\Delta_{X,-}}$ then the distance above the line was increased, and the line was most likely moved down. Similarly, if $\frac{\Delta_{R,+}-\Delta_{R,-}}{\Delta_{R,+}+\Delta_{R,-}} < \frac{\Delta_{X,+}-\Delta_{X,-}}{\Delta_{X,+}+\Delta_{X,-}}$ then the distance above the line was decreased, and the line was most likely moved up.

Note that if the $\Delta_{R,.}$'s are multipled by the same constant, α, due to a change in size during reproduction, then the α's cancel out during detection. Changes in print density in the vertical direction effect all centroids approximately the same, and changes in the print density in the horizontal direction that vary slowly with respect to the character height do not change the location of the centroids. Technique is resistant to errors caused by the expected distortions.

In an experiment to test this line space encoding, line movements of $(1/300)^{th}$ of an inch have been used. $(1/300)^{th}$ of an inch was the smallest movement that could be obtained with the printer that was used in the experiment. Virtually 100% correct decoding decisions were still made after ten generations of copying. We stopped trying to decode after 10 generations of copying because the document was no longer considered usable. The details of the experiment are described in reference 1.

The original recipient of a copy can remove the markings, and avoid being recognized, by detecting the locations of lines and changing their positions. Baselines, the line that the bottom of characters is written on, are uniformly spaced in an unmarked document, but not a marked document. The encoding can be removed by making the distance between all of the baselines the same. We have found that baselines cannot be accurately located after a document is distorted by printing and copying, and are only useful in documents that have remained in electronic form.

Line encoding can be hidden, without accurately locating baselines, by randomly moving all of the lines that may contain information up or down by an amount that is greater than the movement used to encode the information. This approach introduces more distortion in the document, and makes it less desirable

than the publisher's copy.

Detecting lines in a plain text is straightforward and hiding the encoding can be automated. However, when documents contain a variety of line spacings for headings, text in figures, tables, figure captions, etc., automatic detection of all of the lines becomes more difficult, and hiding the encoding may require human intervention.

2.2 Word Space Encoding

Word space encoding is implemented by altering the horizontal space between words, see figure 2. Word spacing is currently altered in order to right justify columns, and this results in nonuniform spaces between some words. The alterations for encoding can be made undetectable, by making the nonuniform spaces that result from encoding less than those which may result from right justification.

Potentially, the space between every word can be modified to encode information. The only constraint on the modifications is that the sum of the movements on a line equal zero, so that the line remains right justified. For our application it is sufficient to encode less than one bit per line. Unencoded lines are included to detect, and compensate for non-linearities that occur in printing and copying. In the encoded lines, a single bit of information is inserted by moving space from between two words near the beginning of the line to two words near the end of the line. Moving space between the beginning and the end of the line results in a large difference between the lines and improves the distortion immunity.

Word spacing can be removed or hidden in the same way that line spacing is removed. To remove word spacing, the beginning and end of words are determined with a greater degree of accuracy than the changes in spacing, and all spaces are made equal. To hide word spacing, a courser determination of word boundaries is made, and the space between words is randomly increased or decreased by an amount that is greater than the encoding. Removing word spacing is preferable to hiding the spacing, since it does not distort the document. However, it may not be possible to accurately determine word boundaries in a document that has been distorted by printing, copying and scanning.

Before determining word spacing, the location of lines must be determined. Therefore, removing word spacing requires more processing than removing line spacing. There are more locations where word spacing may be located. Therefore, to hide word spacing requires that more changes in the document than hiding line spacing.

2.3 Noise Placement Encoding

Information can be inserted in a document by adding a noise signal that is barely visible. Different noise patterns can signify zeros or ones. Noise is least noticeable when it occurs at a natural boundary in an image like the edge of a letter or a boundary in a figure. One way to insert noise at the edge of characters is to design two fonts, in which the characters look alike but differ in a small number of bit positions. In the unencoded document the fonts are randomly selected for each character. In the encoded document, the font that has been selected for the unencoded document is reversed or not, to transmit a bit of information.

The small amounts of noise that are added do not survive printing or copying. However, of the three approaches, this approach is the most difficult to remove from a document that has remained in electronic form. To remove information that is encoded into the characters, the characters must be recognized and the document reset. When information is also transmitted by placing noise near edges in figures, additional noise must be added to the figures to hide the signal.

2.4 Comparison

Of the three techniques, line spacing is the most immune to the distortion that occurs during printing and copying documents. However, it is the easiest of the three techniques for an opponent to remove. Line spacing is useful to discourage some recipients from redistributing a document, but it is unlikely to discourage anyone who seriously wants to subvert the system. Since this technique is immune to distortion, we can reliably determined when the document has been tampered with to remove or hide the information. Therefore, it is extremely useful for identifying modified, illicit copies.

Word spacing is the least noticeable of the techniques, since the information is carried in patterns that may occur in an unencoded document. It is more difficult to remove than line spacing, but is not as robust

against distortion.

Noise placement does not survive copying and printing, but is the most difficult of the three techniques to remove from a bit map version of the document. In an environment where bit maps may be captured, it is useful to include this type of marking.

There is nothing in any of the three of the marking techniques that precludes the possibility of using the others. The three techniques have different characteristics, and can survive different types of attacks. It is most difficult to subvert the system when all three techniques are used.

3. System Overview

In order for document marking to be useful, it must be part of a complete document distribution system. The system cannot place an unreasonable burden on a publisher in terms of either transmission or processing. At the same time, the system should make the processing and transmission costs of illegitimate redistributors as high as possible. Cryptographic techniques accomplish these contradictory goals by making different representations of a document available to the publisher and recipient. The publisher encrypts and distributes a convenient representation of the document. When the document is decrypted, the representation is changed before making it available to the recipient.

Two document distribution systems are considered, one uses tamper resistant hardware in the recipient's printer and the other is limited to software in the recipient's computer. In both systems the publisher distributes a high level description of the document that is the same for every recipient and an information packet that is different for each recipient. The high level description of the document can be in a structural language, such as LATEX or TROFF, in which the document is primarily text and commands. The publisher encrypts both the document and information packet with a secret key. The recipient cannot use the document before it is decrypted, and when it is decrypted, the representation is changed so that the recipient only has access to a marked bit map or marked paper copy of the document.

In a large distribution system, the publisher encrypts a document and distributes it in a tree structured network to end offices. The publisher only generates a single document. The end office sends the encrypted document and an encrypted information packet to each recipient. The unique copies are generated by the recipient. This mechanism limits the long distance transmission and processing costs for the publisher.

Changing the representation of the document from a high level language to a marked bit map increases the transmission cost to redistribute the document. For instance, if a page has 50 lines with 100 characters per line, the high level description requires $4x10^4$ bits. If the bit map is 300 dots per inch, and a page is 10 inches by 10 inches, the bit map requires $9x10^6$ bits. For some pages, particularly those with graphics, the high level description requires more bits. In addition, standard compression techniques can reduce the number of bits needed to represent a bit map. However, bit maps require more bits than high level descriptions of a document, so that the conversion from one representation to the other provides a transmission advantage for the publisher.

A distribution system with tamper resistant hardware in the printer is shown in figure 3. The publisher sends documents that can only be printed by a specific printer. Each printer has a public key that is registered with a publications board. The publisher sends the printer the document and the identification information encrypted with the secret key, S_D, and S_D encrypted with the public portion of the printers public key, P_{Pr} . At the printer, the secret portion of the public key is used to determine S_D, and the document is decrypted. P_{Pr} is used to encrypt S_D rather than the document because public key decryption requires more processing than secret key decryption, and should not be used on large documents. In addition, if documents were encrypted with the printer's public key, the publisher would have to encrypt a document for each recipient.

In the hardware system, the recipient only has access to the marked paper copy of the document. Many printers have sufficient processing power to perform the decryption and marking tasks, however, tamper resistant hardware is needed to prevent the recipient from obtaining printer's secret key and decrypting the document outside the printer.

In a system without special purpose hardware, figure 4, the functions performed by the special purpose hardware in the printer are performed in a program in the recipient's computer. The output of the program is a marked bit map. Public keys are not used in this system. Instead, a decryption and marking program is redistributed regularly and has a secret key imbedded in it.

Normally, encryption is used to transmit information between two trusted end points, across a hostile domain. However, in this application, the information is being decrypted in the middle of a hostile domain. In this environment, the recipient must be discouraged from

1. giving away the program with the document, and
2. reverse engineering the program to determine the secret key.

In order to discourage the recipient from giving away the program, the program requires personal information to operate. For instance, the secret key may be exclusive-or'ed with the recipient's credit card number so that the decryption can only be performed by someone who knows that number.

It is unlikely that a recipient would give his credit card number to anyone who is willing to receive illegal documents. In order to discourage anyone from reverse engineering the program, determining the secret key and decrypting unmarked documents, new versions of the program are distributed frequently that use different secret keys, and store the keys in different locations. Reverse engineering a program involves considerable effort. An opponent is more likely to expend the effort if he gains access to all documents, rather than a single document. A difference between general software distribution and the publishing environment, is that useful life of a program can be controlled in publishing.

4. Experiments

Marking documents with line spacing can be implemented by a straightforward modification of a postscript file. When the first internal memorandum describing the technique was distributed, it was also marked using the technique and registered to the recipients. This experiment showed that the technique is reasonable to implement, did not distort the document in any noticeable way, and when documents, or copies of documents, were returned we were able to reliably identify the original recipient.

It is clear from the first experiment, that document marking is useful outside of electronic distribution. It can also be used to control the recopying and redistribution of sensitive paper documents whenever the number of copies is small enough to print them individually. In the first experiment, the lines in the document that were shifted were selected manually, and a secretary kept a list of the recipients of each document. In a second experiment we are automating this process. The automated system will be used to mark executive memoranda.

An agreement has been reached with the IEEE to try electronic distribution for a single issue of the Journal On Selected Areas in Communications in the second quarter of 1995. The document marking and distribution techniques will be used in that trial. In addition, a Mosaic interface will be used to compare the use of paper and hypertext documents. There may also be a trial of an experimental billing system that is based upon an anonymous credit mechanism. The anonymous credit card provides a means to charge for services in a network environment without disclosing your credit card number to the vendors.

5. Conclusion

The electronic distribution of newspapers and magazines can become a major new application of telecommunications if a publisher's revenues can be protected from illicit copying and redistribution. The work described here is the first attempt to protect a publisher's property by making it preferable to obtain the information from legitimate, rather than illegitimate suppliers.

Document marking techniques are used to trace copies of documents to their original recipients, and to detect documents that have been tampered with to modify the marking. Three marking procedures are described that can withstand different attacks. All three techniques should be used simultaneously.

Two distribution systems are described for marked documents. Both limit the processing per document required by the publisher, and make it more expensive for a second party to redistribute documents than it

is for the publisher to distribute it. One system has tamper resistant hardware in each printer, and the other only uses software. The hardware system provides better security for publishers. However, it is unlikely that, in initial implementations of document distribution, recipients will purchase special purpose devices for this application. Therefore, a software system is required. Both systems distribute the same representation of a document and identifier, so that they can coexist, and a hardware system can evolve from a software system without any changes for the publisher.

6. REFERENCES

J. T. Brassil, S. Low, N. F. Maxemchuk, L. O'Gorman, "Electronic Marking and Identification Techniques to Discourage Document Copying," To appear in the Proceedings of Infocom '94, June 13-17, 1994, Toronto, Canada.

7. Acknowledgements

The work reported in this article comprises the efforts of a group of individuals. The document marking techniques are being investigated by Jack Brassil, Steven Low, Larry O'Gorman and the author. The document distribution techniques are being investigated by Abijhit Choudhury, Sanjoy Paul, Henning Schulzrinne, and the author. Marking executive memoranda is being implemented by Aleta Lapone. Anonymous billing in a network environment is being implemented on by Dave Kristol, Steven Low, and the author. All of those involved in this project are also working on the IEEE trial of electronic distribution.

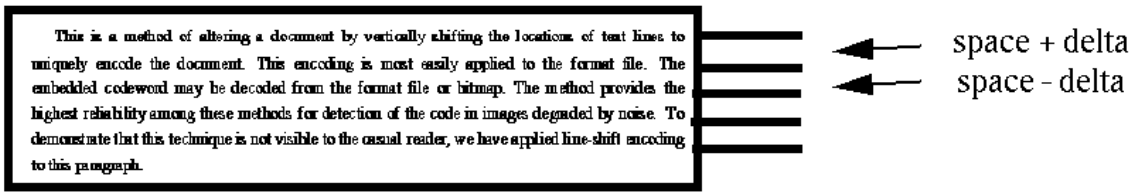

Figure 1 — Example of line-shift coding. The second line is shifted down by approximately 1/150 inch (delta in the diagram). Differential coding causes the spacing after the first line to be larger and the spacing after the second line to be smaller.

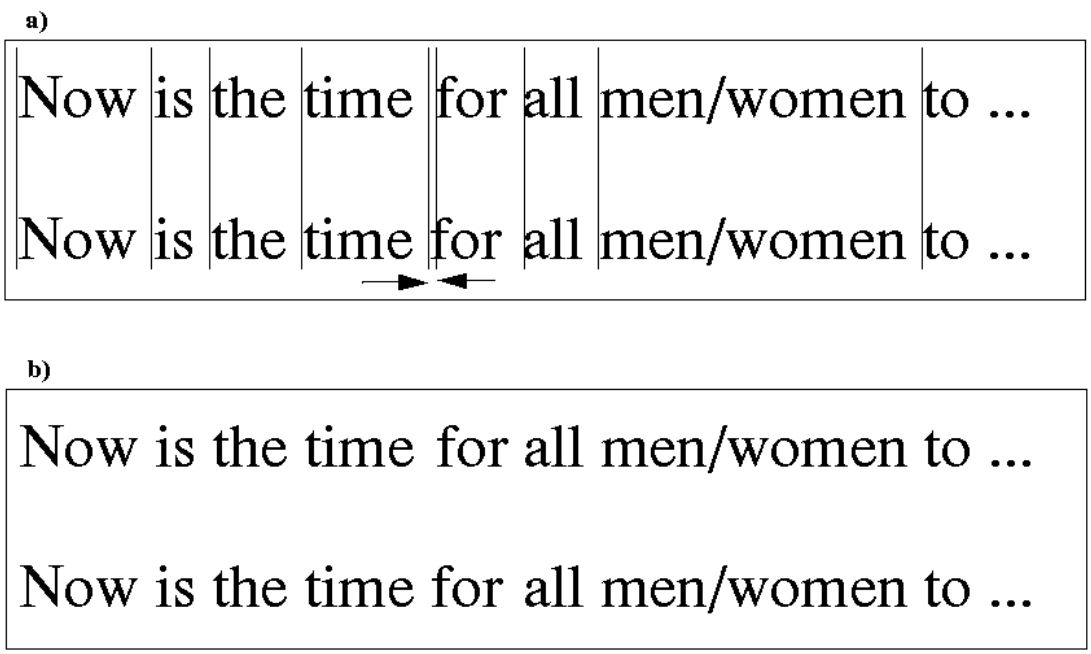

Figure 2 — Example of word-shift coding. In a), the top text line has added spacing before the "for", the bottom text line has the same spacing after the "for". In b), these same text lines are shown again without the vertical lines to demonstrate that either spacing appears natural.

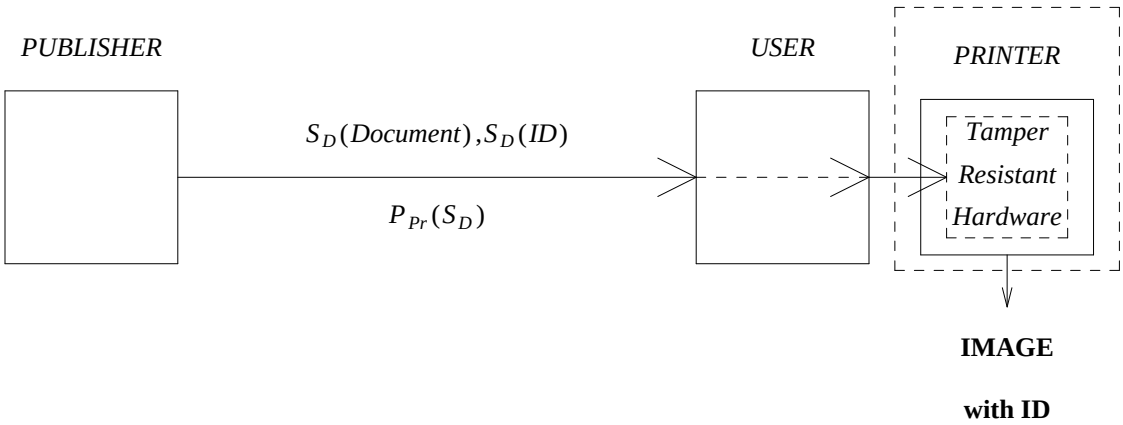

Figure 3 — A transmission system for electronic distribution that includes special purpose hardware in the printer. The publisher transmits a document and identification number encrypted with a secret key, and the inverse of the secret key encrypted with the public portion a secret key that is unique to the printer.

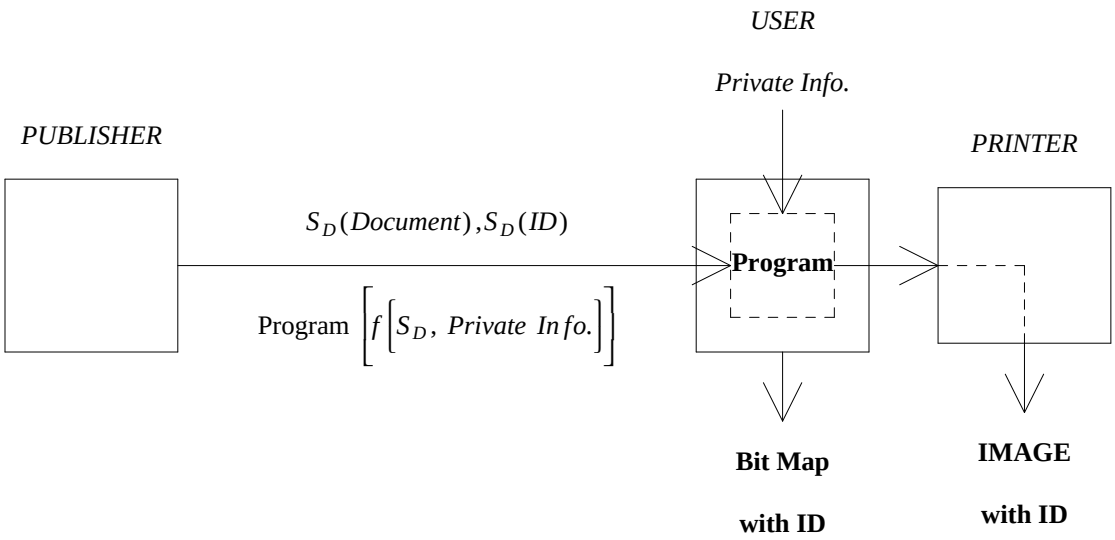

Figure 4 — A transmission system for electronic distribution that uses only software. The publisher transmits a document and identification number encrypted with a secret key, and a program that includes the inverse of the secret key modified by private information about the intended recipient.

CONTENTS

1. Introduction 1

2. Marking 2
   2.1 Line Space Encoding 4
   2.2 Word Space Encoding 5
   2.3 Noise Placement Encoding 5
   2.4 Comparison 5

3. System Overview 6

4. Experiments 7

5. Conclusion 7

6. REFERENCES 8

7. Acknowledgements 8

REFERENCES

CONTENTS