Electronic Document Distribution

Nicholas F. Maxemchuk

Computers, printers, and high-rate data-transmission facilities are becoming less expensive and more generally available. It is now possible to distribute customized newspapers and magazines electronically, instead of using a fleet of trucks and a network of vendors to disseminate a common paper version. A major obstacle to the use of electronic distribution is the ease of copying and redistributing electronic documents. This capability can affect a publisher's subscription revenues. The AT&T Distributed Systems Research Department is making it more difficult to redistribute electronic documents by:

- Distinctive marking, so documents can be traced back to the original recipient;
- Using cryptographic techniques, so the form of the document, available to a recipient, costs more to redistribute than that disseminated by the publisher; and
- Requiring someone who redistributes a document to divulge personal, identifying information with the document.

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

Introduction

The electronic distribution of newspapers, magazines, and other printed material is a new and economically significant application of communication networks. Home computers and printers are decreasing in cost, and the quality of material that can be printed locally is improving. In addition, the rate of data transmission—available to home computer users—has increased steadily over the last few years, accelerating the delivery of electronic documents. Moreover, distributing paper copies of such periodicals as newspapers and magazines, by using a fleet of trucks and a network of vendors, is becoming considerably more expensive.

Eventually, electronic delivery and local production will become preferable to printing documents at a central location and delivering paper copies. A number of electronic-distribution applications already exists, including an on-line journal of *Current Clinical Trials*, published by the American Association for the Advancement of Science, and a version of the *New York Times*, known as *Times Fax*.

Electronic document distribution can provide services beyond those currently available. Mass-produced, paper documents, originating at a central location must meet the needs of all intended recipients. Electronic documents—printed locally—can be customized for individual readers. When a document is read in electronic form—as is now done in the Mosaic Interface to the Worldwide Web—the information can be organized so that a 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

Panel 1. Abbreviations, Acronyms, and Terms

IEEE—Institute of Electrical and Electronics Engineers
latex—a UNIX-based text-processing language
Mosaic—an interface to information available on the Worldwide Web
S_D—representation of a secret key used to encrypt a document and its identification information
S_{Pr}—the public portion of a printer's public key, used to encrypt S_D
troff—a UNIX-based formatter used for typesetting text

while reading one document, pointers to other documents can be used to retrieve additional information on a topic quickly.

Paradoxically, 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 in this manner can substantially reduce an electronic publisher's subscription revenues.

In the near future, copying and redistributing electronic documents will also affect the publication of printed documents. Scanners that convert print material into electronic form, such as those used in fax machines, are improving in quality and are becoming widely deployed. As scanning technology improves, electronic distribution will provide an advantage over paper publishing because there are more ways to discourage redistribution of individually generated electronic copies than of mass-produced paper copies.

The safeguarding of electronically published material differs from typical security processes because it is not a publisher's intent to keep information secret. Publishers profit by making information public. The following techniques discourage rather than prevent the copying of electronic documents, giving a publisher certain advantages over a redistributor:

- Documents are marked so that illicit copies, recovered later, can be traced back to the original recipient.
- Cryptographic techniques are used, so documents that have been tampered with to disguise the original recipient are recognizable.
- The form of a publisher's document differs from that of a redistributor, decreasing the publisher's marking cost and increasing the redistributor's transmission cost.
- A redistributor may be forced to provide personal, easily traced information together with a document.

In the "Marking" section, three techniques are discussed whereby each copy of a document—distributed by a publisher—is registered to the original recipient and marked in an imperceptible way. If unauthorized copies are found, they can be traced back to the original recipient.

It is possible to conceal the identifying marks by specially processing the electronic version of the document. Such processing has three objectives:

- To make the removal of security markings as difficult as possible;
- To detect documents that have had their security markings removed; and
- To degrade a document's quality when the marks are removed.

In many instances, the ability to identify a bootleg copy of a document is a sufficient reason for a recipient to obtain a legitimate copy from a publisher.

In the "System Overview" section, two electronic-distribution systems are described. Using cryptographic techniques, these systems:

- Protect a document from interception during transmission;
- Provide different representations of a document to a publisher and recipient; and
- Reduce the publisher's cost of generating unique documents.

A publisher distributes the same encrypted copy of a document—along with a unique identifier and instructions for marking the document—to all recipients. In the first system, tamper-resistant decryption hardware is used in the printers. As a result, the electronic version of a document is unavailable to recipients, who only have access to a marked paper copy. In the second system, published material is decrypted in software, but a recipient can access only a marked copy of the electronic document. Moreover, when a document is decrypted, it is changed to increase the number of bits used to represent it. The publisher gains advantages, therefore, in transmission cost and time over illegal redistributors.

In the "Experiments" section, three demonstrations in document marking and distribution are

described:
- Marking techniques as applied to an actual document;
- Marking and registering private memos; and
- Trial distribution of a professional society journal.

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 the map. The ability to process and print individual copies, rather than just producing 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. The techniques discussed in this paper do not change the content or perceptibly alter the appearance of a document. Three marking procedures, however, are considered:
- Modifying the space between adjacent lines of text;
- Modifying the space between words; and
- 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 was a communication channel. Document marking is different from the classic communications problem, however, because the decoder can be given access to the received signal, as well as information at the source. For instance, the decoder is given the location of the encoded information, while an opponent who wants to remove the marks might not be able to determine their position.

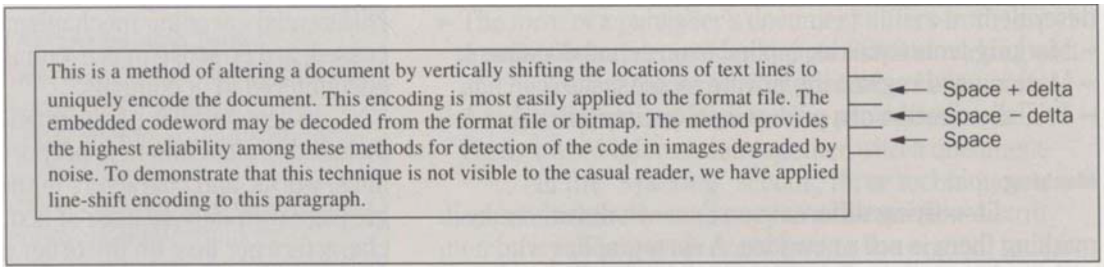
Marking techniques are most effective when the information can be decoded after the markings have been distorted by printing the document on paper, copying and faxing the paper, then scanning the paper back into the computer. These processes introduce a variety of nonlinear types of distortion into the 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, which will be discussed, are effective in preventing the types of distortion encountered in documents.

The number of information bits that can be included in a document is proportional to the number of lines, words, and 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:
- Design encoding mechanisms that are easier to decode and are more robust against distortion;
- Avoid inserting information in locations where it may be noticeable or easily lost because of distortion;
- Insert redundancy for error or erasure correction;
- Hide the few bits that carry real information in a large number of bits that carry random signals; and
- Make it difficult for anyone except a publisher to create a valid marking.

The first two points are specific to individual marking techniques and are addressed in their respective subsections. The last three points are common to all of the marking techniques, and are addressed in this section.

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, thereby further degrading document quality. Codes can correct more erasures than errors, and the procedures for filling in erasures are simpler than those for correcting unknown errors. Whenever possible, therefore, 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 would otherwise be necessary if the location of the information was known. Hiding information among random bits—although simple to implement—does not provide as much protection as correction techniques.

One attack on the marking system is for an opponent to change the mark before redistribution, so that a document cannot be traced back to its original recipient. This attack can be prevented by using only a small subset of all possible marks, making it unlikely that an opponent, who changes the marking, will select a valid mark. For instance, assume that there are 32 bits



This is a method of altering a document by vertically shifting the locations of text lines to uniquely encode the document. This encoding is most easily applied to the format file. The embedded codeword may be decoded from the format file or bitmap. The method provides the highest reliability among these methods for detection of the code in images degraded by noise. To demonstrate that this technique is not visible to the casual reader, we have applied line-shift encoding to this paragraph.

Figure 1. This drawing is an example of line-shift coding. The second line is shifted down by approximately 1/150th of an 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.

available for information, but only 16 bits are required to number a document. A publisher can pick the 2^{16} numbers at random from the 2^{32} available numbers. In order to make a document appear as someone else's copy, an opponent might remove the publisher's number and insert another one. Only one in about 8,000 numbers the opponent chooses, however, is a valid number. Against such odds, the chance of guessing the correct marking number is remote. A document is detected as being illegitimate, therefore, when it lacks the valid number.

Rather than choosing numbers at random and keeping a list of them, valid numbers can be created using techniques that are similar to digital signatures. Specifically, a message can be encoded that includes both the document number and an encrypted function of the message, as well as other information. In a digital-signature system, the encrypted information is only a function of the message. The receiver can verify, therefore, that the message was sent by a transmitter performing the encryption. In the publishing environment, encrypted information includes that which is specific to a document. Such information prevents an opponent from taking a number from one copy of a document and placing it on a copy of a different document. Including information not transmitted through a communications channel takes advantage of the decoder's ability to obtain additional encoder information in the document-marking environment.

Line-Space Encoding. In this process, alternate lines of text are held stationary, and the position of every other line in a 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 a line is moved up, a "one" is transmitted. If a line is moved down, a "zero" is transmitted. In order to make decoding easier, the first and last complete lines in paragraphs are held 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. The approach discussed here, however, provides sufficient information to mark most documents and is easier to decode.

To decode information from a paper copy of a document, the text is scanned into a computer. Conventional document-processing programs are then used to reorient the horizontal document lines. 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 on 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 spacing 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 and original documents. In the received document, $\Delta_{R,+}$ represents the distance between the centroid of a line that has been moved and that of the stationary line above it. $\Delta_{R,-}$ represents the distance between the line that is moved and the stationary line below it. $\Delta_{R,+}$ and $\Delta_{R,-}$ 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

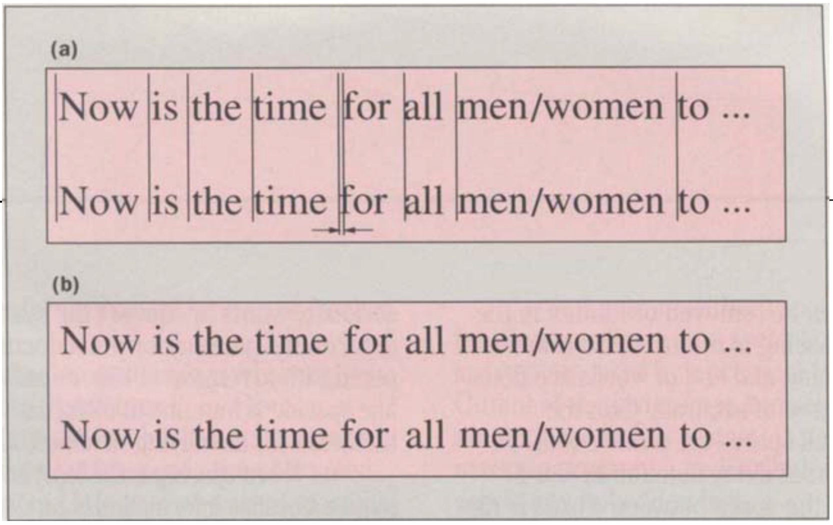

(a)
Now is the time for all men/women to ...
Now is the time for all men/women to ...
(b)
Now is the time for all men/women to ...
Now is the time for all men/women to ...

Figure 2. This drawing is an example of word-shift coding. In illustration a., the top text line has added spacing before the word "for", and the bottom text line has the same spacing after the word "for". In illustration b., the same text lines are shown again, this time without the vertical lines, in order to demonstrate that either method of spacing appears natural.

line most likely was 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 most likely was moved up.

Note that if the Δ_R's are multiplied 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 affect all centroids in approximately the same way. Changes in print density in the horizontal direction, which vary slowly with respect to character height, do not change the centroids' location. These characteristics make line-space encoding resistant to errors caused by the expected distortions.

In an experiment to test this encoding process, line movements of 1/300th of an inch were used. 1/300th of an inch is the smallest movement that could be obtained with the printer used in the experiment. Virtually 100-percent-correct decoding decisions were still made after ten generations of copying. Decoding was stopped after ten generations of copying because the document was no longer considered usable. All details of this experiment are available.[1]

The original recipient of a copy can remove the markings—and avoid being recognized—by detecting the location of lines and changing their position. The bottom of characters is written on baselines, which are uniformly spaced in an unmarked document but not in one that is marked. The encoding can be removed by equalizing the distance between baselines. Research has determined that baselines cannot be accurately located after a document is distorted by printing and copying. Baselines are useful only in documents that have remained in electronic form.

Line encoding can be hidden—without accurately locating baselines—by randomly moving all lines containing information up or down by an amount greater than that used to encode the information. This approach introduces more distortion into a document, making it less desirable than the publisher's original copy.

Detecting lines in plaintext is straightforward, and hiding the encoding can be automated. Automatic detection of all lines becomes more difficult, however, when documents contain a variety of line spacing for headings, text in figures, tables, figure captions, and so forth. Hiding the encoding may also require human intervention.

Word-Space Encoding. This technique is implemented by altering the horizontal space between words, as shown in Figure 2. Word spacing is currently altered in order to right-justify columns. This results in nonuniform spaces between some words. The alterations for encoding can be made undetectable by making nonuniform spaces that result from encoding smaller than those resulting 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 movements on a line must equal zero, so that the line remains right justified. For most applications, it is sufficient to encode fewer than one bit per line. Unencoded lines are included to detect and compensate for nonlinearities 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 end of the line results in a large difference between lines and improves distortion immunity.

Word spacing can be removed or hidden in the same manner 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 coarser determination of word boundaries is made, and the space between words is randomly increased or decreased by an amount greater than that of the encoding. Removing word spacing is preferable to hiding, because it does not distort the document. It may not be possible, however, to accurately determine word boundaries in a document that has been distorted by printing, copying, and scanning.

Before ascertaining word spacing, the location of lines must be determined. Removing word spacing requires more processing than removing line spacing, and there are more locations where word spacing may be located. Hiding word spacing, therefore, requires more changes to a document than hiding line spacing.

Noise-Placement Encoding. Information can be inserted in a document by adding a barely visible "noise" signal. Different noise patterns can signify zeros or ones. Noise is least noticeable when it occurs at a natural boundary in an image, such as 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 few bit positions. In an unencoded document, the fonts are randomly selected for each character. In an encoded document, the font selected for the unencoded document is reversed or not, in order to transmit a bit of information.

The small amounts of noise that are added do not survive printing or copying. Of the three approaches, however, this one 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.

A Comparison of Techniques. Of the three techniques discussed in this section, line spacing is most immune to the distortion that occurs during printing and copying documents. It is the easiest of the three techniques, however, for an opponent to remove. Line spacing is useful for discouraging some recipients from redistributing a document, but it is unlikely to discourage anyone who seriously wants to subvert the system. It can be reliably determined whether or not a document has been tampered with to remove or hide encoding information, because line spacing is immune to distortion. It is extremely useful, therefore, for identifying modified, illicit copies.

Word spacing is the least noticeable of the techniques, because 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 resistant to distortion.

Noise placement does not survive copying and printing, but it 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 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.

System Overview

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

Two document distribution systems are considered here. One uses tamper-resistant hardware in a recipient's printer. The other only uses application software in the recipient's computer. In these systems, a publisher distributes both a high-level document description that is identical in content for each recipient and an information packet that differs. The high-level document description can be composed in a structural language, such as *latex* or *troff*, in which the document contains primarily text and commands. A publisher encrypts both the document and information packet with a secret key. A recipient cannot use the document before it is decrypted. When it *is* decrypted, its representation is changed so that a recipient has access only to a marked bit map or marked paper copy.

In a large distribution system, a publisher encrypts a document and distributes it, in a tree-structured network, to end-offices. A publisher generates only 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 a publisher's long-distance transmission and processing costs.

Changing the document's representation 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"x10", 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. Bit maps require more bits, however, than high-level document descriptions. The conversion from one representation to the other, therefore, provides a transmission advantage for a publisher.

In a distribution system with tamper-resistant hardware in the printer, a publisher sends documents that can be printed only by a specific printer. Each printer has a public key that is registered with a publications board. A publisher sends a printer the document and identification information encrypted with the secret key, S_D, and S_D encrypted with the public portion of the printer's 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, which should not be used on large documents, requires more processing than secret-key decryption. In addition, if documents are encrypted with a printer's public key,.a publisher would have to encrypt documents for each recipient.

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

In a system without special-purpose hardware, functions performed by special-purpose printer hardware are implemented by a program in a recipient's computer. 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. In this application, however, information is being decrypted in the middle of a hostile domain. The recipient, therefore, must be discouraged from:

- Giving away the program with the document; and
- Reverse-engineering the program to determine the secret key.

To discourage a recipient from giving away the program, personal information is required for operation. For instance, the secret key may be exclusive-or'ed with a recipient's credit-card number, so that decryption can be performed only by someone who knows that number.

It is unlikely that a recipient would give a personal credit-card number to anyone willing to receive illegal documents. New versions of a program—using different secret keys—are frequently distributed to discourage the reverse-engineering of the program, the determining of the secret key, and the decrypting of unmarked documents. In addition, keys are stored in various locations. Reverse-engineering a program involves considerable effort. An opponent is more likely to expend the effort if access to many documents, rather than only one, is possible. A crucial difference between general software distribution and distribution within the publishing environment is that the useful life of a program can be controlled in publishing.

Experiments

Marking documents with line spacing can be implemented by means of a straightforward postscript-file modification. When the first internal AT&T memorandum describing the technique was distributed, it was also marked using line spacing, thereby registering the recipients. This experiment showed that such a technique is feasible. It did not noticeably distort the text in any way. Furthermore, when the memorandum or copies of it were returned, identification of the original recipient was possible.

This experiment showed that document marking is also useful outside the scope of electronic distribution. It can be used to control the recopying and redistribution of sensitive paper documents whenever the number of copies is low enough to print them individually. In this experiment, the lines of text that were shifted were manually selected, and a secretary kept a list of each document's recipient. In a second experiment, this process will be automated, and the system will be used to mark executive memoranda.

An agreement has been reached with the Institute of Electrical and Electronics Engineers (IEEE) to distribute electronically, during the second quarter of 1995, one issue of the *Journal On Selected Areas in Communications*. 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. Trial of an experimental billing system, based on an anonymous credit mechanism, may also be conducted. Anonymous credit processes provide a means to charge for services in a network environment without disclosing a credit-card number to vendors.

Conclusion

The electronic distribution of newspapers and magazines could become a major new application of telecommunications if a publisher's revenues can be protected from illicit copying and redistribution. The work described in this paper is the first attempt to protect a publisher's property by encouraging the procurement of such material from legitimate suppliers.

Document-marking techniques are used both to trace copies back to their original recipients and to detect documents that have had their markings modified. Three marking procedures that can withstand different attacks were described. All three techniques should be used simultaneously.

Two distribution systems were described for marked documents. Both systems limit the amount of processing, per document, that is required by a publisher. These systems also make it prohibitively expensive for a second, unauthorized party to redistribute documents.

One system relies on tamper-resistant hardware in each printer, and the other uses special software. The system using tamper-resistant hardware provides a superior level of publishing security. It is unlikely, however, that recipients will purchase costly, special-purpose devices for this application during initial implementations of document distribution. A software system, therefore, is required.

Both types of security systems distribute the same representation of a document and identifier, so they can coexist. A hardware-based system can evolve from a software-based system, however, without requiring a publisher to make any distribution-system changes.

Acknowledgments

The work discussed in this paper is the result of the efforts and contributions of many individuals within AT&T Bell Laboratories. Document-marking techniques are being investigated by Jack Brassil, Steven Low, Larry O'Gorman and the author. Abijhit Choudhury, Sanjoy Paul, Henning Schulzrinne, and the author are researching document-distribution techniques. Aleta Lapone is working on the marking of executive memoranda. Research on anonymous billing in a network environment is being completed 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.

Reference

1. J. Brassil et al, "Electronic Marking and Identification Techniques to Discourage Document Copying," Proceedings of INFOCOM, June 13-17 1994, Toronto, Canada.

(Manuscript approved August 1994)

Nicholas F. Maxemchuk *is head of the Distributed Systems Research Department at AT&T Bell Laboratories in Murray Hill, New Jersey. He is responsible for projects related to the technical aspect of heterogeneous networks, electronic document distribution, and techniques for enhancing privacy of communications. In addition, Mr. Maxemchuk is a senior editor of the* Journal of Selected Areas in Communications, *and is a member of the steering committee for the IEEE/ACM Transactions on Networking. He received a B.S. from the City College of New York, and M.S. and Ph.D. degrees from the University of Pennsylvania in Philadelphia, each in electrical engineering. Mr. Maxemchuk, an IEEE fellow, joined AT&T in 1976.*