Applying Cryptography to Pin-Based 429
INTRODUCTION 429
SECTION ONE: BASIC PIN CONCEPTS 430
Why PINS? 430
PI N Secrecy 431
PIN Length... 432
Allowable PIN Entry Attempts............................... 433
PIN Issuance 434
Bank Selected PIN 434
Cardholder Selected PIN 435
Consider each of these techniques: 436
Comparison of Bank Selected PIN and 438
The Forgotten PIN 439
On-Line PIN Validation 440
Off-Line PIN Validation 441
PIN Validation for Interchange Transactions 441
Conclusions 443
SECTION TWO: EFT FRAUD THREATS 444
E FT Fraud Categories 445
Passive Fraud Threats 446
The Delivery System 447
The Cardholder.. 447
The EFT System.. 448
Relative Risks 448
Active Fraud Threats 449
Communications Lines 449
E DP Systems 450
Fraud and Liability 451
Conclusions 453
SECTION THREE: PRINCIPLES OF 454
Cryptography, The Tool for Fraud Prevention 454
Preventing Passive Fraud Threats 455
PIN Encryption... 455
Protection of Cryptographic Keys 456
Physical Protection of PINS and Cryptographic 456
Preventing Active Fraud Threats 457
Data Modification 457
Replay of Debit Authorization 458
Fraudulent Credits 459
Encrypted PIN Substitution.................................... 459

Fraud Prevention in Interchange 461
Countering the Fake Equipment Threat 462
Conclusions 463
SECTION FOUR: IMPLEMENTATION OF...... 464
Suggested Characteristics of Hardware 464
Suggested Capabilities6 465
Bank Selected Random PIN 466
Customer-Selected PIN 467
PIN Validation 467
Key Management 468
MAC Generation 469
Utilization 469
Figure 10-l. Issuer s PIN Validation - Local....... 471
Figure 10-2. PIN in Interchange........................ 472
Conclusions 473
REFERENCES 473

CHAPTER TEN

Applying Cryptography to Pin-Based Electronic Funds Transfer Systems[1]

Today there are many cryptographic authentication techniques being used and evaluated by major financial institutions for electronic funds transfer systems. Therefore, due to the state-of-the-art, there are divergent opinions as to the order in which problems should be addressed and what methodologies should be used to achieve optimum solutions.

To provide a balanced discussion between the authors' point of view (expressed in Chapter 11) and that of others, permission has been obtained to reprint relevant sections from the *PIN Manual: A Guide to the Use of Personal Identification Numbers for Interchange* [1], which was prepared by the staff of MasterCard International, Inc. (formally Interbank Card Association) in cooperation with MasterCard International's Standing Committees. The material in this chapter, except for two indicated passages, was comprised from the first four sections of the *PIN Manual.* The views expressed and responsibility for the accuracy of the material lies with the originators of that manual.

Helpful footnotes, annotations, and additional material was provided by the authors. (Material added by the authors appears in brackets.) In order to maintain consistency, the original notations for encipherment and decipherment have been changed to conform with the notations used throughout the book.

Pin Manual
A Guide to the Use of
Personal Identification Numbers
in Interchange

INTRODUCTION

In the early 1970's, Interbank Card Association began to investigate the implications of the transition from an off-line paper based funds transfer

[1] By permission of MasterCard International, Inc. (formerly Interbank Card Association). Reprinted in part from *PIN Manual: A Guide to the Use of Personal Identification Numbers in Interchange*, September 1980 [1].

system, exemplified by MasterCard, to an on-line, Electronic Funds Transfer (EFT) system. The investigation soon determined that this transition would present many problems relating to customer acceptance, economic justification, and regulatory policy. However, the only unsolved technological problem was how to insure the system's security.

Interbank soon realized that using secret Personal Identification Numbers, PINs, was the best technique for authenticating customers in EFT. A PIN serves the same role in an electronic system that a written signature serves in a conventional paper based system. While this did not solve the security problem, it did define one major aspect, the need to ensure PIN secrecy everywhere within the EFT environment. Although the assurance of PIN secrecy was the first and foremost EFT security problem, it was not the only one. Insuring the authenticity and integrity of the transaction were also problems.

Since it was apparent that EFT could not progress until these security problems were resolved, Interbank began, in the 1970's, what is believed to be the most extensive study of EFT security ever undertaken. The study, which lasted more than three years, uncovered and assembled a wealth of information regarding virtually every aspect of securing an EFT system. It considered, in detail, the possible fraud threats that could be perpetrated against such a system and developed countermeasures to prevent them. The implementation of each countermeasure was studied in detail to insure that its effectiveness would not detrimentally affect the cost or performance of the EFT system as a whole. The study considered many approaches to the issuance, management, validation, and interchange of PINs, and where choices were available to the financial institution, attempted to determine the pros and cons of the available alternatives. Since the study concluded that most of the required security techniques were cryptographic, considerable thought was given to the practical implementation of cryptography in a retail funds transfer environment. Given special study was the management of the secret keys that are a fundamental ingredient in any secure cryptographic system.

SECTION ONE: BASIC PIN CONCEPTS

Why PINs?

The term PIN refers to personal identification number. It is a secret number assigned to, or selected by, the holder of a debit card or credit card used in an EFT (electronic funds transfer) system and serves to authenticate the cardholder to the EFT system. The PIN is basically the cardholder's electronic signature, and serves the same role in an EFT transaction as a written signature serves in a conventional financial transaction. The PIN is memorized by the cardholder and is not to be recorded by him in a manner that could be ascertained by another person. At the time that the cardholder initiates an EFT transaction, he enters his PIN into the EFT terminal using a keyboard provided for this purpose. Unless the PIN, as entered, is recognized by the EFT system as being correct for this particular account number (read by the EFT terminal from the card's magnetic stripe), the EFT system refuses to accept the transaction. The purpose of all this is so that, should the card be lost or stolen, the finder or thief would be unable to use the card, not knowing the associated PIN. Similarly, it is to prevent someone who would be able to do so from making a usable counterfeit copy of the card. Even if he could make such a counterfeit card he could not use it, not knowing the PIN.

PIN Secrecy

In order for the PIN to serve its required function, it must be known to the cardholder, but to no one else. PIN secrecy is of the utmost importance. If the financial institution wishes the cardholder to be responsible for any compromise of his PIN, and, if a PIN is to be an effective signature substitute, then the institution's own handling of the PIN must be above reproach. It must display to its cardholders extreme care in its PIN management procedures. For example, if a cardholder is given the opportunity of selecting his own PIN and is asked to write the PIN of his choice on the application form containing information identifying him, he will quite likely realize that certain bank employees could ascertain his PIN from this form. This cannot help but influence his own attitude toward the importance of PIN secrecy. On the other hand, if he sees that the institution exercises extreme care to insure that no bank employee can possibly learn his PIN, he will be impressed with the importance of PIN secrecy on his own part.

Some financial institutions tend to view PIN secrecy on a cost-effective basis. That is, they attempt to compare the cost of a certain degree of PIN security with the cost of the fraud losses that might otherwise occur. This is not really a valid comparison, because the impact of fraud due to the compromise of PIN secrecy greatly transcends the actual dollars lost. The most catastrophic type of fraud that can occur because PINs are compromised is the production and use of counterfeit cards, causing the accounts of unsuspecting cardholders to be fraudulently debited. This is not known until the cardholders find their accounts overdrawn or incorrect debits on their monthly statements. Assuming that the fraud losses are not due to negligence on the cardholders' part, the institution must pay not only for the fraud but also for the clerical costs involved in processing cardholder complaints and making restitution. Undoubtedly such fraud would become publicized, and cardholders who had not actually experienced fraud but who could not recall making certain transactions appearing on their statements would suspect that they had been defrauded, and file complaints with the institution. The institution would have no obvious way of distinguishing valid complaints of fraud from invalid ones. As a result, some dishonest cardholders would undoubtedly deny making certain of their transactions, knowing the institution could not prove them wrong. This secondary fraud could be of even greater consequence than the primary fraud. However, the greatest impact of fraud resulting from PIN compromise would probably be on customer relations. A number of honest cardholders would hesitate to trust their funds to such an institution any longer, and would move their accounts elsewhere.

Thus, the net loss to an institution could be many times the loss directly due to PIN compromise.

As electronic banking and other forms of EFT grow as a percentage of total financial transactions, the importance of the PIN, and hence of PIN secrecy, is expected to grow likewise. Only by stringent (though not necessarily costly) security measures can a high degree of PIN secrecy be maintained.

The PIN in its clear (comprehensible) form should never be transmitted over communications lines, because these lines could be tapped. The clear PIN should never reside, even momentarily, in any main frame or any data base, because a clever programmer or computer operator might devise some technique for ascertaining it. It should never be known to, or accessible by, any employee of the institution, not even during the PIN issuing process. (PIN mailers, if used, should be under strict dual control at all times to prevent compromise.)

As stringent as these security measures may be, they can be implemented at modest cost and without noticeable impact upon banking operations. Subsequent sections describe, in detail, security techniques and their implementation.

PIN Length

In order to achieve its intended purpose, the PIN must contain enough digits so that a card finder, thief or counterfeiter would have little probability of hitting the correct PIN by chance, if he simply guessed at values. On the other hand it should not contain very many digits, or it will slow down the EFT transaction time. Therefore it is recommended that the PIN be four, five or six decimal digits in length. A four digit PIN allows ten thousand unique PINs. The criminal has no way of knowing which of these is the correct PIN value for any given stolen or counterfeit card in his possession. Assuming that the number of consecutive incorrect PIN entry attempts per card is limited to a small number (e.g., ten or less), assuming that only one PIN value is usable with any given card, and assuming a best case situation from a card counterfeiter's point of view, namely, an unlimited supply of counterfeit cards (thousands), the unobserved exclusive use of an ATM for hours on end, and no other special system checks to ascertain trial and error PIN determination, he would still require more than forty continuous hours of trial and error (assuming four tries per minute), and nearly one thousand counterfeit cards, before he could determine the PIN for a single card. This is believed to be an unfeasible fraud technique, so a four digit PIN appears adequate. Of course this trial and error procedure would be ten or a hundred times longer for a five or a six digit PIN.

It is assumed that in a properly designed EFT security system, it is impossible for the card counterfeiter to construct an off-line system and use it for trial and error PIN determination. That is, it is assumed that he can attempt this trial and error method only on a terminal connected to the actual EFT network. This assumption is not valid for certain EFT security techniques that have been proposed. Were one of these techniques to be used, a PIN length of six or fewer digits would be extremely non-secure.

Though there is no security disadvantage to having long PINs, there is a practical disadvantage. The longer the PIN, the longer the time the cardholder will require to enter it, and the greater the probability of an entry requiring a repeat. The latter is of special concern in an interchange environment where the PIN must be sent to the card issuer for validation. Several seconds or more could elapse before the cardholder began reentering his PIN. During this time the EFT terminal would be unavailable for other use, and in POS environment, a clerk would also be kept waiting. In addition, there is the delay and inconvenience to the cardholder. Thus long PINs, by increasing the transaction time, are a detriment to the merchant, the cardholder and the financial institution.[2]

Allowable PIN Entry Attempts

It is customary to place a limit on the number of consecutive incorrect PIN entries a cardholder is allowed. This is done to further hinder fraudulent PIN determination by trial and error. Though desirable, this is not as important as it is perhaps believed to be, and would appear unnecessary for all but four digit PINs. Determining a five digit PIN by trial and error would require an average of fifty thousand attempts without such a limit, and this appears unfeasible. If a limit is imposed, it can be either an absolute limit, or a daily limit. An absolute limit gives the cardholder a specified number of attempts to enter his PIN correctly, regardless of the time span. After the allowable attempts have been exhausted, the card is considered invalid. A daily limit restricts the cardholder to a specified number of consecutive incorrect attempts in any one day, but the cardholder starts with a "clean slate" the following day. Only when the number of consecutive incorrect PIN entries in any one day exceeds the limit is the card considered invalid. Of these two approaches, the absolute limit appears preferable, since it more definitively limits criminal attempts at trial and error PIN determination. The benefits of this approach, for a four digit PIN, can be expressed quantitatively. If we let N represent the absolute number of consecutive incorrect PIN entries allowed, where N is small (e.g., ten) relative to ten thousand, then the criminal would have to make an average of about ten thousand tries for each PIN he successfully determined. During this time he would have used up ten thousand divided by N cards. That is, for every card's PIN he successfully determined, he would fail on ten thousand/N cards. Without any type of limit, he would require only a single card, and an average of five thousand tries.

When the PIN is validated using the technique of the American Banking Association PIN Verification Standard, the statistics are somewhat different because this technique uses a "non-reversibly encrypted" PIN, which means that more than one PIN can generally be used with a given card. With this technique and an absolute limit, the criminal requires the average of five thousand trials, and for every counterfeit card on which he succeeds he fails

[2] It is only fair to point out that, at the time of this writing, there are differing opinions as to what constitutes a reasonable and practicable PIN length. Current technology will easily accommodate PINs of up to 16 digits.

on 5,000/N. Without any type of limit he requires a single card and the average of 3,679 tries.

The situation for a daily rather than an absolute limit is essentially the same as in the no limit case, except the criminal is restricted daily to one less than the maximum number of attempts allowed. In this way the card is never declared invalid, and the criminal can make additional attempts the next day. The intent is that long before he has made the five thousand (or 3,679 average) tries required, the legitimate cardholder will have noticed that the card is missing, and report the loss. However, if the criminal is using a counterfeit copy of the card, there is no loss to report.

If some type of limit is imposed on incorrect PIN entries, then the question is whether or not the card should be retained when this limit is reached. In an off-line system, card retention may be necessary to prevent unauthorized card usage. However, as a general rule, it appears better not to retain the card if this can possibly be avoided. In an on-line system, the account can be flagged as invalid in the data base, so there is no practical need to retain the card. If the imposed limit is a daily, card related limit, retaining the card does not significantly affect the criminal who is attempting trial and error PIN determination. He is well aware of the limit, and is careful to stay below it. The only effect of card retention at the limit is to reduce by one the number of tries per card per day that can be made. On the other hand, the cardholder is unaware of the card retention threat, and may keep trying to remember his PIN until he has reached the retention threshold. Thus, retaining the card when this type of limit is used does not appear desirable, unless required by off-line usage of the card.

In summary, some limit on the number of consecutive incorrect PIN entries seems desirable when four digit PINs are used, though unnecessary when the PIN length is longer. This should not be an argument to use a longer PIN, because four digits is, in many ways, the optimum PIN length. With a four digit PIN, the first choice is to use a per card absolute limit, without regard to time. A value in the range of three to ten would seem reasonable. The second choice is to use a per card daily limit, in the range of three to four. The third choice is no limit. This choice introduces some risk, however, this risk is not significant, since it would require the average of fifteen to twenty hours of trial and error at an ATM (assuming the ATM allows four tries per minute) for each four digit PIN thus determined. Finally, retaining the card (should the limit, of whatever type, be reached) appears undesirable, unless there is no other method available for restricting future use of the card.

PIN Issuance

There are two basic PIN issuance techniques. In the first, the financial institution determines what the PIN will be, and conveys it to the cardholder. In the second, the cardholder determines what the PIN will be and conveys it to the institution.

Bank Selected PIN

Again, the card issuing institution has two choices. The PIN can be cryptographically derived from the account number, or it can be a random value.

PIN Cryptographically Derived from the Account Number. In this case, the account number is processed using a cryptographic algorithm so as to produce a decimal value of the appropriate number of digits. With proper generation techniques there is no discernible correlation between the derived PIN and the account number, and the PIN is completely unpredictable to anyone who knows the account number but does not know the secret key (defined in Section Three) used in the cryptographic process.

The advantage of this technique is that it eliminates the necessity for maintaining any record of the PIN. When the PIN of reference is needed to validate the PIN as entered by the (alleged) cardholder, it may be regenerated by simply processing the same account number through the same cryptographic process (utilizing the same secret cryptographic key). The main disadvantage of this technique is that the PIN cannot be changed unless the account number is changed. If a cardholder fears that his PIN may have been compromised and requests a new PIN, the only way to give him a new PIN is to give him a new account number. Another disadvantage is that the cryptographic key cannot be changed without changing every cardholder's PIN.

Random PIN. The use of a random number for the PIN overcomes the disadvantage of having a PIN that is inherently linked to an account number and to a specific cryptographic key. With the random number technique, the card issuing institution generates, in a highly secure manner, a random decimal number that serves as the cardholder's PIN. The disadvantage of this technique is that the institution must maintain a record of the random PIN it issued to serve as the PIN of reference for subsequent validation of the PIN as entered from EFT terminals. As indicated previously, it is unacceptable to store the PIN in its clear form. It must be encrypted, as described in subsequent sections. The encrypted PIN may be (1) stored in the issuer's data base, (2) encoded on the magnetic stripe of the card, or (3) both.

Regardless of which technique is chosen to generate a bank selected PIN, it must be conveyed to the cardholder. This is normally accomplished by means of a PIN mailer, a printed document containing the clear PIN. This document must be printed under conditions of very high security, and dual control throughout must be utilized to insure that no bank employee opens, reads, or even closely examines such a mailer. The most secure PIN mailer is a multi-part sealed form with the PIN printing visible only inside the form. (The form can then, if desired, be placed in a windowed envelope to hide any impression which may have been made upon the top surface of the form.)

The PIN and the associated cards should never be mailed together. Preferably, the PIN is mailed after the cardholder has signed a confirmation receipt for the card.

It is recommended that bank selected PIN should be four digits (not five or six) to simplify its memorization and lessen the probability that it will be carried, in written form, with the card.

Cardholder Selected PIN

Many financial institutions, for reasons to be discussed later, prefer to have the cardholder select his own PIN. In this case, a technique is required where the cardholder can convey his PIN to the institution. There are three such techniques:

1. The PIN may be solicited by mail, with the selected PIN mailed back to the institution.
2. The PIN may be entered by the cardholder via a secure terminal located at one of the institution's offices.
3. The PIN may be selected when the potential cardholder visits the issuer's facility.

Consider each of these techniques:

PIN Solicited by Mail. With this technique the institution prepares a two part form to be mailed to the cardholder. The first part contains the cardholder's name and address, and serves only for mailing purposes. The second part contains a reference number and a place for the cardholder to write the desired PIN. The cardholder is instructed to write only the PIN on the second portion and mail it back to the institution in an opaque envelope provided for the purpose. The first part is to be discarded. The reference number bears no discernible relationship to the cardholder's account number, nor to any other information that would identify the cardholder. Thus, a clerk at the institution may open the returned envelope and manually enter the PIN and the reference number into a special security system. This system can, through a cryptographic process, ascertain the account number from the reference number. Then it passes the PIN, encrypted, and the clear account number to the institution's EDP system. At no point in this process is the clear PIN ever associated with the clear account number, nor with any other information which would serve to identify the cardholder.

PIN Entered via a Secure Terminal. Perhaps the simplest and best way for a cardholder to convey his selected PIN to the financial institution is by entering it via the PIN pad of a secure terminal. A secure terminal is one that encrypts the PIN as soon as it is entered. Such a terminal operating in conjunction with special cryptographic equipment at the institution's EDP facility, provides an environment where the clear PIN is never thereafter available (except within the security system).

With this method of cardholder PIN selection, safeguards must be implemented to protect against someone who steals a card from the mail, impersonates the legitimate cardholder in the PIN selection process, and then draws against the legitimate cardholder's funds. To guard against this, one or more officials at the office where the terminal is located should be responsible for validating the cardholder's identity. Such an official has a special PIN, which he enters into the terminal just before the cardholder begins the PIN selection process, provided the official is satisfied with the cardholder's identity. Only when the cardholder's PIN is preceded by a legitimate official's PIN, is the cardholder's PIN accepted.

Another approach to cardholder PIN selection at a secure terminal is to assign the cardholder an initial PIN with a secure PIN mailer, then allow the cardholder to replace it with a PIN of his own choice. In this case, the cardholder must first correctly enter the assigned PIN, then enter the PIN of his choice. This procedure precludes the necessity of a bank official authenticating the cardholder. The fact that the cardholder knows the assigned PIN is probably sufficient proof of identity.

Regardless of which cardholder authentication method is used, it is suggested that the cardholder be required to enter the PIN identically two consecutive times. This is to prevent accidental errors in entering the selected PIN.

PIN Selected at the Issuer's Facility. Today, a large number of financial institutions utilizing the cardholder selected PIN technique have the cardholder select the PIN at the time he applies for the card. This is logistically simpler than mailing PIN solicitation forms. Furthermore, it allows the card to be prepared immediately thereafter, even if the magnetic stripe is to contain an encrypted version of the PIN. A widely used technique is to have the cardholder write the PIN on the application. This is not a recommended procedure, as certain bank employees would be able to relate the PIN to the name, and then to the account number.

A secure technique for cardholder PIN selection at the issuer's facility uses a prepared form, produced by a special security system. This is a sealed, multi-part form, similar to that suggested previously for PIN mailers. On the top layer the security system prints a clear reference number, and inside the form (where it cannot be seen) it prints this number encrypted. There is no discernible relationship between the clear and the encrypted versions of the reference number, as they are related only via a cryptographic process utilizing a secret key known to no one.

The customer applying for a card is given this sealed form. Privately, he removes the inner portion of the form containing the encrypted reference number, and on this portion writes the PIN of his choice. He writes nothing else, places this portion of the form in an opaque envelope and seals and deposits it in a locked container provided for the purpose. The outer portion of the form containing the clear reference number is submitted along with the application. This clear reference number is entered into the data base of the institution's EDP system as a part of the application and becomes the account number, or is associated with the account number as soon as this number is assigned. At some subsequent time, a bank employee enters the encrypted reference number and the PIN into a special security system which encrypts the PIN, decrypts the reference number, passing the result to the institution's EDP system. This system uses the clear reference number to relate the encrypted PIN to the account number. At no point in this process is the clear PIN associated with any data that identifies the cardholder.

If a secure terminal is available at the office where the customer makes application for the card, another PIN selection procedure is possible. Under this alternate technique, the application is assigned a number. This number is written on the application, and entered into the secure terminal. The customer then enters the PIN of his choice into this terminal, via its PIN pad. The PIN is encrypted at the secure terminal, and remains encrypted thereafter during transmission through, or storage in, communications or EDP equipment. The encrypted PIN will be related to the account number, when the latter is assigned, via the application number.

Another issue concerning a cardholder selected PIN relates to the nature of the PIN itself. Should it be a number or a word? It is often held that a word is preferable, being easier to remember than a number. Since PIN keyboards have (or will have) letters (in the manner of a telephone) associated with each digit, the entry of letters is as convenient as that of numbers. However, the use of a single word for the PIN is not recommended. There are simply not enough different words cardholders would be likely to choose to adequately preclude trial and error PIN determination. For example, one could probably make a list of two hundred words, and have a reasonable probability that any given cardholder would select a PIN from this list.

The recommended technique, where an alphabetic PIN is desired, is to instruct the cardholder to select two unrelated words, using the first two letters of each word to form his four character PIN. If a six character PIN is desired, the first three letters of each word should be used.

When a numeric PIN is used, it is also advisable for cardholders to be given instructions on how to select a PIN. For example, they should be instructed not to select a telephone number that might be readily associated with them (e.g., their own or that of a close relative or business associate). Similarly, they should not select a date that might be readily associated with them (e.g., birthday, anniversary of themselves or a close relative), nor any other number closely associated with them (license number, social security number). These admonitions are designed to guard against the interception of a card in the mail (or the theft or counterfeiting of a card), and determination of the associated PIN by trial and error using readily available information about the cardholder. (While this fraud threat is certainly possible, it is not considered to be a major one.)

Comparison of Bank Selected PIN and Cardholder Selected PIN

There is considerable difference of opinion among financial institutions as to which method of PIN selection is preferable. Operational simplicity favors the institution specifying the PIN. PIN mailers can be printed in an automatic fashion at a very rapid rate. When the PIN is selected by the cardholder, however, each PIN must be individually entered into the system. This is a manual and time consuming procedure with some cost consequences for the financial institution, unless the cardholder can perform the entry without manual assistance.

The advantage of having the cardholders select their PINs is that they will more easily remember such PINs, and be less inclined to write them someplace where they might be associated with their cards. The disadvantage, in addition to the above indicated cost consideration, is that the cardholders may tend to select values which might be surmised, despite admonitions to the contrary. Another factor favoring the cardholder selected PIN is customer relations. It is, presumably, less onerous for customers to memorzie a value they have selected, then one that has been imposed on them.

The recommended approach for PIN selection is one whereby the financial institution issues a PIN to the cardholder with the option of selecting an alternate value. This alternate value is selected, desirably, via a secure terminal with the bank issued value authenticating the cardholder for the PIN change procedure. The bank selected initial value, which most cardholders will probably elect to use as their permanent PIN, should be four digits long, whereas the cardholder might be allowed to select a four, five, or six digit value, depending upon his personal preference.

In many situations it is not feasible to give the cardholder a PIN with the option of changing it. This can be true if the PIN, in an encrypted form, is to be encoded on the card. In such a case the recommended procedure would be for the bank to issue the cardholder a four digit PIN, or else use the previously discussed technique of PIN selection at the time of application. Though the latter may be viewed as somewhat preferable, the former appears acceptable, and is logistically much simpler.

Regardless of the PIN selection technique chosen, the cardholder should be advised of the importance of the PIN and PIN secrecy. The cardholder should be warned against recording the PIN value where it might be located by a finder or thief of the card.

The Forgotten PIN

Regardless of how the PINs are conveyed to the cardholders, it is possible they will forget the PINs. When this happens, there are three possible courses of action for a financial institution.

1. Send the cardholder a PIN mailer advising him of the forgotten PIN.
2. Send the cardholder a completely new PIN.
3. Allow the cardholder to select a new PIN.

From a human factors point of view, the most desirable procedure is probably the first and the least desirable alternative is the second. It is a psychological fact that it is easier to rememorize something than to memorize something completely new. Thus, sending a PIN mailer reminding the cardholder of the original PIN is the preferred technique, even if the cardholder had originally selected the PIN. (Once the cardholder sees the PIN, he will most likely recall why it was selected, and this will reinforce it in the cardholder's mind all the more.) Sending a completely new bank selected PIN is not recommended. If the cardholder could not remember the initial PIN, there is little reason to believe he will remember a different one. Finally, allowing the cardholder to select another PIN is probably acceptable, but less desirable than reminding him of the PIN already selected. Of course, when the PIN in encrypted form has been encoded on the magnetic stripe of the card, the financial institution has no choice but to advise the cardholder of the original PIN, unless the card is to be reissued when the new PIN is chosen.

A PIN mailer, advising the cardholder of a forgotten PIN, must be printed under rigid physical security to prevent bank employees from opening the mailer.

PIN Validation for Local Transactions

Local transaction refers to a transaction in which the institution that issued the card also controls the terminal. By contrast, an interchange transaction is one in which the terminal is controlled by an institution other than the card issuing one. For a local transaction, there are two possible techniques for PIN validation, on-line and off-line.

On-Line PIN Validation

On-line validation refers to PIN validation at the institution's EDP facility, whereas off-line PIN validation refers to PIN validation in the terminal itself. On-line PIN validation is possible only when the terminal in question is on-line to the institution's EDP system (i.e., communicating with an operational EDP system).

In any PIN validation procedure, the PIN as entered by the cardholder is compared against the PIN of reference as recorded by the financial institution. There are three possible techniques for obtaining the PIN of reference. First, it may be stored in encrypted form in the data base of the financial institution. In this case either the encrypted PIN of reference is decrypted and compared with the clear PIN as entered, or else the PIN as entered is encrypted using the same procedure and key as was the PIN of reference, and the two encrypted values are then compared. Second, it may be recorded in encrypted form on the magnetic stripe. Again, either the encrypted PIN of reference is decrypted and compared to the clear cardholder-entered PIN, or else the cardholder-entered PIN is encrypted and compared against the encrypted PIN of reference. Finally, it may be a cryptographic function of the account number, obtained by employing the account number with a cryptographic process.

Encrypted PIN from Data Base. In this approach, the PIN must be encrypted at the terminal, then transmitted to the institution's EDP system along with the other elements of the transaction. Using the account number, the EDP system locates, in its data base, the encrypted PIN for this account. Special security equipment is then (desirably) used to decrypt both the PIN of reference from the data base, and the PIN entered by the cardholder from the terminal. These two decrypted versions of the PIN are compared, and an indication is sent to the EDP system as to whether or not they agree. If the two versions agree, the cardholder entered the correct PIN, and is presumed to be a legitimate user of the card.

Alternately, the PIN as entered by the cardholder is decrypted, then immediately reencrypted using the same key and in the same manner as is the PIN or reference. These two encrypted versions of the PIN are then compared.

Encrypted PIN from the Card. In this approach, the cardholder's PIN is encrypted at the terminal and transmitted, along with the other elements of the transaction, to the EDP system. Included in this transaction data are the contents of the card's magnetic stripe. One of the fields in the stripe contains the encrypted PIN of reference. These two encrypted verions of the PIN, the PIN entered by the cardholder and the PIN of reference from the magnetic stripe, are passed (desirably) from the EDP system to special security equipment which performs appropriate cryptographic operations and then compares both versions as described above. Note, the PIN of reference is encrypted utilizing the account number (which must be passed to the security equipment before it can decrypt this version of the PIN) so that if two cardholders have identical clear PINs, their encrypted PIN will be different.

PIN a Cryptographic Function of the Account Number. As before, the PIN entered by the cardholder is encrypted at the terminal and transmitted to the EDP system, along with the other elements of the transaction. Then the EDP system (desirably) transfers to special security equipment both the encrypted PIN from the terminal and the account number. This security equipment decrypts the encrypted PIN and cryptographically processes the account number to generate the PIN of reference. The two versions of the PIN are then compared, as before.

Off-Line PIN Validation

When the PIN is validated within the terminal (or other remote facility), the terminal (or facility) must have the means to compare the PIN of reference with the PIN entered by the cardholder. Thus the PIN, in an encrypted form, must be encoded on the card's magnetic stripe, or the PIN must be a cryptographic function of the account number. In either case, the terminal must have the cryptographic capability, as well as the necessary encryption keys, to perform the required comparison. The terminal allows the transaction to proceed only if the two versions of the PIN agree.

The use of off-line PIN validation at other than a highly secure terminal like an ATM (automated teller machine) is not recommended. Should a terminal with off-line PIN validation ever be compromised, and the secret encryption keys stored within it ascertained by anyone intent on fraud, they would be able to determine the correct PIN for any lost or stolen card issued by that institution. Furthermore they could, with some additional sophistication, produce usable counterfeit copies of any or all of this institution's cards.

Today off-line PIN validation is widely used in ATMs. This enables the ATM to perform most of its functions (excluding balance inquiry) even when the ATM cannot communicate with the EDP system. This is a useful and valid mode of operation, yet care must be taken, as indicated above, that the secret cryptographic keys are always kept highly secure.

PIN Validation for Interchange Transactions

At some future time it is anticipated that the on-line interchange of EFT transactions will become as common as the present off-line interchange of credit card transactions. It is anticipated that a combination of ATMs, POS (point of sale) terminals, and POB (point of banking) terminals would be included in a nationwide EFT network. Whenever cash is dispensed, the use of a PIN will probably be required. Thus, PIN validation in interchange will become, in the future, a matter of considerable importance.

The use of PINs in interchange poses special security problems. This is due to the fact that the PIN is entered into a terminal under the control of one institution, the acquirer, whereas the card was issued by another institution, the issuer. Should the acquirer's negligence allow the issuer's PINs to be ascertained Ly someone intent on fraud, the issuer would bear the fraud loss, and there would be no obvious way of determining the identity of the negligent institution because hundreds or thousands participate in a nationwide interchange network.

To provide the highest possible protection for the PIN in an interchange environment, several basic principles appear evident:

1. An acquirer should not be able to validate the PINs of other issuers. Were every institution in an interchange network able to validate the PINs of every other institution, the compromise of a single institution could compromise every PIN of every other institution in the interchange network. This is an unacceptable risk. Therefore, each issuer must validate its own PINs, though an issuer may delegate this responsibility to someone else.
2. Clear PINs should not be allowed over any communications line nor in the EDP system of any acquirer. If clear (intelligible) PINS were transmitted over communications lines, these lines could be tapped and the PINs ascertained. This would not only compromise the PINs of the institution whose lines were tapped, but those of every other institution whose cardholders used terminals of the institution in question. Similarly, clear PINs should not be allowed in any acquirer's or switch's EDP system, because some clever programmer or computer operator might determine a technique for recording the PINs along with the corresponding magnetic stripe information. Even though an issuer might trust its own EDP system to store and/or process its own PINs in a secure manner, such an issuer would quite likely not similarly trust the EDP systems of perhaps thousands of other institutions to be similarly secure. Thus, the PIN should be encrypted at all times as it traverses communications circuits and EDP systems from the terminal where it was entered to the issuer's facility.

Therefore to provide very high security:

1. Every PIN-using terminal in an EFT network should have an integral encryption capability that is physically secure.
2. Every acquirer, as well as certain other nodes that a transaction may traverse, should have a special, physically secure, cryptographic capability to translate the PIN from one cryptographic key to another, in order not to perform any cryptographic operations that might expose clear PINs in a general purpose EDP system.

A "physically secure" cryptographic capability in the above context has a very specific meaning if very high security is desired. The cryptographic capability in question is enclosed, and if the enclosure is penetrated by any means, the cryptographic keys stored within the enclosure are automatically erased. If someone should penetrate the enclosure in an attempt to commit fraud, the cryptographic capability would be rendered inoperative (because it can no longer decrypt PINs) and all the information that could possibly be used to commit the intended fraud would be destroyed.

In order to assure PIN security at all points in the interchange process, it appears that PIN validation in interchange should operate as follows if very high security is desired:

1. PIN is encrypted at the entry terminal, using a secret cryptographic key. The encrypted PIN is then transmitted to the acquirer's EDP system, along with other transaction elements.
2. The acquirer's EDP system routes the encrypted PIN to special security equipment, a security module. Within this physically secure module the PIN is decrypted using the cryptographic key of the terminal, and is immediately reencrypted with a cryptographic key used for interchange. The PIN, thus encrypted, is returned to the acquirer's EDP system. It is then routed to the issuer's EDP system via normal communications channels.
3. The issuer also has a security module, and the PIN from the interchange transaction is routed to this module where it is decrypted, then validated, using any one of the three techniques previously discussed for on-line, local PIN validation.

Although use of the above suggested hardware module provides very high security for PINs in interchange, this degree of security may exceed that which will actually be required by Interbank rules. Main frame software may be acceptable for the decryption and the reencryption of such PINs, especially in the initial steps of nationwide interchange.

The above discussion of interchange applies primarily to the eventual nationwide interchange between hundreds or even thousands of financial institutions. Regional interchange among approximately a dozen institutions is quite likely at an earlier date, and might not operate in conformity to the relatively rigid PIN security principles indicated above. For example, off-line PIN validation within ATMs might be possible in regional interchange, provided all participants clearly understand that the compromise of any one of these shared ATMs could compromise every PIN of every participant. Basically, in regional interchange of this sort, the various institutions trust one another in a way that would not be realistic when interchanging with hundreds or thousands of institutions across the country.

Conclusions

PINs are an essential ingredient of any EFT system, serving as the cardholder's electronic signature to authenticate his right of access to his account. To serve this purpose, the PIN must be kept secret and must be known by no person other than the cardholder. PINs should be from four to six digits in length, long enough to preclude trial and error PIN determination, but not so long as to impede transaction time. When four digit PINs are used, some limit on the number of PIN entry trials is desirable, though it is preferable not to retain the card if this limit is exceeded.

The PIN may either be determined by the institution or selected by the cardholder. There are advantages and disadvantages to each approach, but techniques exist to implement either approach in a secure manner so that no one can determine cardholder PINs. Secure techniques are also available to advise a cardholder of his PIN should he forget it.

A cardholder's PIN is validated by comparing his PIN as entered via an EFT terminal with his PIN of reference. This comparison may be either on-line at the institution's EDP center, or off-line within the EFT terminal itself. On-line PIN validation can be implemented more securely than off-line validation by the use of special security equipment at the institution's EDP center. However, in many cases off-line PIN validation is a necessity, to permit off-line operation of ATMs, for example.

In the forthcoming nationwide interchange of EFT transactions, the PIN cannot be validated off-line in the terminal, but must be transmitted securely to the facility of the issuer (or some institution serving on behalf of the issuer) for validation. Specific security requirements must be placed on acquirers and switches in an interchange environment to insure that an issuer's PINs are not compromised by negligence on the part of another institution.

SECTION TWO: EFT FRAUD THREATS

The preceding section considered general PIN management concepts, and dealt with most of the issues faced by a financial institution that wishes PINs to be used with its debit or credit cards. Insofar as possible, this discussion has been nontechnical, and has avoided the detailed discussion of how security techniques can be implemented.

The remaining sections consider in detail the technical aspects of PIN management, as well as other aspects of EFT security. These sections are intended for those who wish to pursue the subject in greater detail, especially those who are concerned with the design of systems and equipment to be used in an EFT environment, and those who wish to evaluate different equipment designs. This present section serves as an introduction to this detailed technical discussion by considering the general fraud threats against which an EFT system must be protected.

Major EFT fraud is not expected to become a significant risk until a nationwide system for the interchange of EFT transactions is in operation. The relatively small scale of most of today's EFT systems, and the diversity between such systems, tends to discourage fraud. Considerable study and development effort would be required to compromise any one EFT system. Even if such a system were compromised, it would most likely be shut down and/or its security techniques upgraded, long before the criminal had recouped his investment in fraud technology. Though the shutdown of an EFT system would be a mild catastrophe for the institutions involved, it would be preferable to sustaining a substantial fraud loss since EFT is currently more of a convenience (e.g., ATMs as a source of after hours cash) than a necessity. Thus, the potential payoff in compromising an EFT system today would not appear to justify the investment in fraud technology which it would require.

When the nationwide interchange of EFT transactions becomes well established, however, EFT will become a tempting target for a concerted fraud effort. Such a nationwide network will, of necessity, use standardized security techniques throughout. Thus, if organized crime could develop the fraud technology to defeat these techniques, it could be applied against financial institutions all across the country. Furthermore, by this time EFT will be well entrenched, with many thousands of supposedly secure terminals, so the retrofitting of these terminals to counter the exploited vulnerability would be almost prohibitively expensive, and impose nearly insurmountable transition problems. By this time EFT could well have become a major payment system, comparable to checks and credit cards, so shutting down such a system would be virtually unthinkable. Thus, the situation faced by the banking industry would be similar to, though far more serious than, the one faced by the telephone company when Blue Boxes (electronic devices used to make unbillable long distance calls) were first developed. If the banking industry were essentially defenseless against certain fraud threats, these threats would become very attractive to potential perpetrators.

It should also be noted that much of the fraud loss in today's payment system (e.g., bad checks) is borne by merchants. In an EFT system it would be borne primarily by financial institutions.

EFT Fraud Categories

There are three main types of fraud threat to which an EFT system might be susceptible:

1. Fraudulent use of lost or stolen cards.
2. Production and use of counterfeit cards.
3. Manipulation of data.

The use of lost or stolen cards, assuming the PINs for them can be ascertained, is probably the most obvious fraud threat. It requires no technological sophistication (except whatever might be required to determine the associated PIN) and could enable funds to be withdrawn from the corresponding accounts at ATMs or other EFT terminals. However, this exposure to fraud is limited, in an on-line EFT system, to the time between the loss of a card and the reporting of this loss by the cardholder.

The production and use of counterfeit cards is potentially a far more serious fraud threat. In a nationwide EFT system, the flooding of the country with many thousands of counterfeit cards could have a potentially disastrous effect. Unlike the losing or stealing of a card, which is likely to be promptly reported, the use of a counterfeit version of a legitimate card would not be detected until the legitimate cardholder examined his next statement or received notification that his account was overdrawn. Thus, thousands of unsuspecting cardholders would find funds missing from their accounts, an obviously catastrophic occurrence. As indicated in Section One, not only would the banking industry be faced with the resulting direct fraud loss, but, once the fraud had become publicized, inadvertent claims of fraud, or outright fraudulent claims of fraud (cardholders who claim fraud because of transactions they have honestly forgotten, and cardholders who claim fraud knowing that they themselves withdrew the disputed funds) would without doubt occur. This secondary fraud could result in even greater losses than the original direct fraud. Perhaps most serious of all would be the loss of customer good will, and customer confidence in the EFT system. Thus, mass fraud of this type is clearly in a different class from the type which could occur through the fraudulent use of actual issued cards.

The most difficult task in the production and use of counterfeit cards would probably be determination of PINs and corresponding account numbers. The actual encoding of counterfeit cards would not appear to be especially difficult, employing either stolen or handmade card encoders, and using either stolen card stock, plain bank cards, or reencoding expired cards. It should be noted that the same physical card could be reencoded and reused for many different accounts.

The final threat, the manipulation of data, is perhaps the most sophisticated fraud threat of all. In this threat, the EFT system is penetrated and data are inserted or modified in real time. For example, funds may be withdrawn without any account being debited, or with the wrong account being debited. Similarly, credits may be routed into the incorrect account, or credits spuriously originated when no actual deposit or return of merchandise took place. This subject will be considered in greater detail under "Active Fraud Threats."

Fraud techniques can be categorized as either passive or active. Passive techniques are those which simply ascertain information, presumably to enable the subsequent use of lost or stolen cards, or the production and use of counterfeit cards. Active techniques, in contrast, modify or insert data, as indicated above.

Passive Fraud Threats

In order to use lost or stolen cards or to produce and use counterfeit cards at PIN-using EFT terminals, it is necessary to learn PINs and associated account numbers. In the absence of appropriate security safeguards, an identifiable cardholder's PIN might be subject to determination for possible fraudulent use by:

1. The card issuing institution.
2. The PIN delivery system.
3. The cardholder himself.
4. The EFT system.

The following discussion considers each of these, then briefly evaluates the relative risks involved.

The Card Issuing Institution

Improper PIN management and PIN issuing techniques on the part of the card issuing institution can expose the PIN to possible determination. There is a risk of this type of exposure whenever even one member of the institution's staff:

1. Has the opportunity to see or access any cardholder's PIN.
2. Has the ability to change cardholder PINs (enabling him to change PINs to values known by him).
3. Is in a position to authorize others to access or change cardholder PINs.
4. Has the capability to ascertain the cryptographic keys used to protect or derive cardholder PINs, or to enable others to ascertain these keys.

Another area of possible risk within the card issuing institution is the method used to generate cardholder PINs. An improper PIN generation technique might enable information about some or all PINs to be determined from obtainable data.

The Delivery System

The technique used to convey the PIN to the cardholder when the institution selects the PIN, or to the institution when the cardholder makes the selection, is a possible area of PIN compromise. For example, if PINs are sent to cardholders via PIN mailers, there is the possibility that someone might ascertain PINs from these mailers. Risk exists whenever the PIN, together with cardholder identifying information, is conveyed via non-secure channels.

The Cardholder

Careless or improper actions on the part of the cardholder could cause his PIN to be compromised. For example he might:

1. Fail to destroy or secure the PIN mailer. The cardholder might allow his PIN mailer to fall into unauthorized hands.
2. Record the PIN. The cardholder might write the PIN on the card, or in some other place where it could be found and associated with the card.
3. Divulge the PIN. He might tell his PIN to others, who in turn, could allow it to be compromised. Similarly, he might be tricked into giving his PIN to someone purporting to have the authority to receive it.
4. Allow the PIN to be observed. He might allow his PIN to be observed as he enters it into the EFT terminal.

5. Make a poor PIN selection. If the cardholder himself selects his PIN, he might select a value which could be surmised.

The EFT System

The EFT system itself could be vulnerable to PIN compromise if it does not include adequate security techniques.

1. Wire tapping. If PINs are unencrypted or improperly encrypted, they could be ascertained from taps on appropriate communications lines.
2. Computer tapping. If PINs in their unencrypted form exist within a general purpose EDP system even momentarily (perhaps while being decrypted and then reencrypted), the computer software could be surreptitiously modified to cause PINs and corresponding cardholder identifying information to be recorded or otherwise divulged. Similarly, if the cryptographic keys used to encrypt PINs are available within such a system, these keys could be subject to compromise, which would then allow the PINs encrypted under them to be determined. These problems are especially critical in an interchange environment where one institution's PINs pass through EDP systems of switches and other institutions over which it has no control.
3. Terminal tapping. If PIN-using terminals without adequate safeguards are used in a non-secure environment (e.g., point of sale), it would be possible to tap the PIN pad and the magnetic stripe reader, and thus ascertain PINs and corresponding cardholder identifying information.
4. Fake equipment. In a non-secure environment without specialized safeguards, it might be possible to place fake equipment to record PINs. For example, a PIN pad could be installed with a non-PIN using terminal. This PIN pad would go only to a recorder, which would operate in conjunction with a tap in the terminal itself or on the communications line for recording magnetic stripe information.
5. Trial and error PIN determination. If the number of consecutively invalid PIN entry attempts is not appropriately limited, trial and error PIN determination might be possible using the actual EFT system, especially if the finder or counterfeiter of the card had some insight into which PIN values were most likely. Furthermore, PIN determination by exhaustion could be quite feasible if the card finder or counterfeiter could employ some off-line simulation of the EFT system to make and evaluate PIN trials automatically.
6. Compromise of cryptographic keys. If it is possible to ascertain any of the cryptographic keys used to encrypt PINs as they traverse the EFT system; such keys could then be used to decrypt, and thus reveal, the PINs encrypted under them.

Relative Risks

It appears impossible to develop any type of payment system which completely eliminates all fraud risks. In an EFT system, special attention must be paid to those risks which could result in mass fraud, especially the flooding of the country with many thousands of counterfeit cards. Less concern need be focused on those risks which expose only an occasional, specific account.

Though the most difficult risks to counter are those which involve cardholder negligence in the handling of his own PIN, these risks are not of extreme concern. Such risks do not appear to be a source for mass fraud, and furthermore it appears that cardholders can be instructed to treat their PINs with appropriate care.

The greatest fraud risks appear to be in the issuing institution, and in the EFT system itself. Security weaknesses in these areas could result in the compromise of thousands of PINs. For example, security weaknesses which would allow PINs to be determined from an institution's PIN management techniques, or by wiretapping, terminal tapping, or computer tapping the EFT system itself, would appear to pose very great risks. A security weakness cannot be discounted simply because a considerable investment would be required to exploit it. In a nationwide EFT system, the fraudulent payoff from such an investment could be tremendous.

Finally, it should be noted that, once a large-scale nationwide EFT system is in operation, upgrading its security features could be an almost impossibly difficult task because of the standardization and cost inherent in such a system. Thus the system cannot necessarily take advantage of advancing technology. The criminal, however, operates under no such handicap, and can employ the latest technology in his efforts to defraud the system. Therefore an EFT security system must be carefully designed at its inception, with this realization in mind.

Active Fraud Threats

Active fraud threats, the manipulation or insertion of data in real time, require a high degree of technical sophistication. Nevertheless, the continuing advance of computer and electronic technology will make the required technology increasingly available. The two areas in an EFT system where such technology might be applied are the communications lines and the EDP systems.

Communications Lines

There are many ways in which transmitted data could be manipulated to commit fraud in the absence of appropriate security safeguards. The amount field, the transaction type, or the account identifier could be modified in various ways to commit different types of fraud. However, the most likely active fraud threat is probably the simulation of the response message from a host to a terminal, causing every transaction to be approved, but no account to be debited. An ATM would be a likely candidate for this type of fraud. The communications line from the ATM to the host would be found, and a microprocessor system would be placed in series with this line. To the host, this system would look like an idle ATM. To the ATM, this system would look like the host. The criminal would then initiate a transaction from the ATM. The transaction would be intercepted by his microprocessor and would never reach the host. Instead, the microprocessor system would respond to the ATM with "transaction approved" indication, causing the ATM to dispense cash. This procedure would be repeated time after time until the ATM had been depleted of cash.

There are a number of other ways in which such a microprocessor system inserted in a communications line could be used to commit fraud. It could be programmed to pass all transactions unaltered except those for certain specified account numbers. In the case of these accounts, it would modify the message from the terminal to the host then modify the response from host to terminal in the inverse manner. For example, an ATM dispense cash request for $150 might be reduced to $10 by the microprocessor, so the account in question would be debited by only this latter amount. The amount field in the approved response would then be changed from $10 back to $150. Similarly, the amount of a credit transaction could be increased or a debit transaction turned into a credit. Also it might be possible to cause transactions to be misdirected. The account number of a debit transaction might be changed so that the wrong account was debited. Similarly, credit transactions might be misdirected into accounts controlled by the criminal. Finally, it might be possible to introduce spurious credit transactions, or to fraudulently replay previously valid credit transactions.

Another type of active wiretapping would be the substitution of one encrypted PIN for another in an EFT system which encrypted the PIN but did not preclude such substitution. For example, a transaction including an encrypted PIN could be recorded via a passive tap. A counterfeit version of the associated card would be produced and then used at the same terminal with a fictitious PIN. By means of active wiretapping, the encrypted fictitious PIN would then be replaced by the previously recorded encrypted true PIN, causing the transaction to be accepted as valid by the issuer.

EDP Systems

The same types of active fraud which could be perpetrated via communications lines could also be perpetrated within the EDP systems, acquirer's, switch's, and issuer's, which the transaction traverses. This could be accomplished through surreptitious modifications of the CPU software, causing data to be inserted or modified as suggested above. This type of computer fraud would be especially attractive were transactions cryptographically protected only over communications lines and not within EDP systems.

Other opportunities for active fraud exist within the EDP system of the issuer. For example, it might be possible to cause debits against accounts controlled by the criminal to be applied against other accounts instead, with the corresponding journal entries similarly adjusted. In a totally electronic system without backup paper documents, this type of fraud could be quite effective.

Another fraud scenario which might be perpetrated within an issuer's facility is encrypted-PIN substitution. This scenario is possible, for example, when an ATM encrypts the PIN under the PIN KEY (a key shared by all of the institutions' ATMs), then transmits this encrypted PIN to the EDP system for comparison against an identically encrypted PIN of reference in the data base. The criminal would make a counterfeit version of a valid card, then use this card at an ATM with a PIN of his own choosing. The transaction would, of course, be rejected (invalid PIN). However, the criminal would have an accomplice, a skilled programmer at the bank's EDP facility, who would record the criminal's PIN in its encrypted form. At some later time the programmer accomplice would replace the encrypted PIN of reference in the data base with the criminal's encrypted PIN. Thereafter, the criminal could withdraw funds at will from this account, because his PIN would have become the PIN of reference for this account in the data base.

Fraud and Liability

In an interchange environment, the possibility of fraud brings with it the obvious question of which institution is liable in the event that fraud does occur. The concept of liability is that a negligent party must pay any losses which other parties incur because of this negligence. Although liability may be viewed primarily as a legal issue rather than a technological one, liability can be established only if the party responsible for the fraud can be determined. Since technology is generally required to make this determination, the security techniques used in an EFT interchange environment serve not only to counter anticipated fraud threats, but also to provide a basis for establishing liability in the event of fraud.

Considering first the use of lost or stolen cards and the production and use of counterfeit cards, the liability would appear to rest with the card issuer for all transactions in which PINs are used. The main prerequisite for this type of fraud is the ability to ascertain PINs for known accounts. This is most likely to occur in the issuer's PIN management system, PIN distribution system, or as a result of negligence on the part of the cardholders. Since it is impossible for an interchange system to oversee all these aspects of an issuer's internal operations, the interchange system has little choice but to assume that, if PINs are compromised, the issuer is responsible. The result of the above assumption is that the PINs must be extremely well protected in the interchange environment. Were PINs to be compromised in interchange, it would be virtually impossible to establish responsibility. Thus, all aspects of interchange PIN handling must be made especially secure if an ambiguous liability situation is to be avoided.

For example, it may eventually become a fundamental principle of interchange that one institution's clear text (unencrypted) PINs should never exist, even momentarily within the general purpose EDP facility of any other institution. Were this principle to be violated, a devious programmer at an EDP system would be able to ascertain PINs in such large quantities that he could pick and choose among them for fraudulent use so as to give little evidence as to the location of compromise. In such a case many issuers would experience fraud losses, perhaps substantial ones, because of a security compromise over which they had absolutely no control, and for which the guilty party could not be determined.

Another fundamental principle of interchange would appear to be that an institution's PINs are validated only by the issuing institution itself, or by an institution explicitly authorized to do so by the issuing institution. Inherent in the capability to validate PINs is the capability to determine PINs. (There are exceptions to this, but they require the use of very long PINs, eight digits or more.) Thus, in an interchange environment in which every acquirer has the capability to validate the PINs of every issuer, a one-time compromise on the part of a single acquirer could compromise every PIN of every issuer, and leave absolutely no residual evidence as to which acquirer was at fault. Again, many institutions would suffer losses, perhaps very substantial, because of negligence on the part of an institution whose identity could not be determined. Thus the interchange system must be designed to prevent the negligence of one unidentifiable institution from resulting in loss to other institutions.

Fortunately, it is essentially only the passive threats in an EFT interchange environment which can result in losses to some institutions as a result of untraceable negligence on the part of others. In the case of most active threats it is possible to pinpoint the guilty institution and make it financially responsible. Therefore the countering of such active threats in an interchange environment can be left up to the discretion of each participating institution.

In the case of two active threats it is impossible to distinguish negligence on the acquirer's part from negligence on the issuer's part. The first such threat is the substitution of a previously recorded PIN replayed as part of a counterfeit transaction. There is no apparent way to distinguish this fraud threat, due to the acquirer's negligence, from the use of a counterfeit card with a PIN ascertained through negligence on the issuer's part. Probably the simplest resolution to this ambiguity is for the acquirer, who chooses not to use terminals which preclude the active version of this threat, to assume liability unless it can be shown that it was the passive version of the threat which occurred.

The other active threat is the fraudulent replay of a previously valid credit transaction when this replay takes place between an issuer and an acquirer who communicate directly. It would not always be possible to determine whether the replayed message originated from the acquirer's EDP system, the issuer's EDP system, or on the communications link between them. It appears that, in this case, the issuer must assume responsibility for the fraud, since he is better equipped to detect it, as will be shown subsequently.

In other situations it is not possible to distinguish fictitious claims of fraud on the part of dishonest cardholders, the issuer's problem, from certain active fraud threats for which the acquirer should be responsible. For example, the fraudulent misdirecting of credits could, unless precluded by appropriate security measures, occur because of an active wiretap on an acquirer's terminal. On the other hand, a cardholder could claim a credit which he did not receive, and produce a fictitious receipt to prove his claim. Since credits in interchange should be relatively rare (probably limited to the return of merchandise), it is suggested that all terminals with a credit capability protect against this threat. In the absence of such protection, the acquirer should assume liability unless it can be proven otherwise.

Another similarly ambiguous situation might be as follows: A cardholder institutes a debit transaction for what he claims is $20, and has a receipt to prove it. However, his account is debited for $200. The possibilities are

(1) that the cardholder himself falsified the receipt and actually received the $200, or (2) that active wiretapping of the acquirer's terminal caused the amount to be increased for transmission to the host, then correspondingly decreased in the response message back to the terminal, with a teller or some other acquirer's employee pocketing the $180 difference.

Fortunately the ambiguous active fraud threats are judged to be relatively improbable. The most probable active threats are unambiguous. For example, the previously suggested "draining" of an ATM (or other EFT terminal) by cutting this terminal off from its host and giving an "approved" response to every request, is clearly the acquirer's liability. The use of active wiretapping to decrease the amount of a debit as reported to the issuer is also clearly the acquirer's liability. Even when a ambiguous situation does exist, it is between two, or at most three potentially responsible parties. In the case of passive fraud threats, it might be impossible to ascertain which one of a thousand or more institutions was responsible for fraud.

Finally, the flooding of the country with tens of thousands of counterfeit cards, the ultimate EFT catastrophe, is possible through passive, not active threats. Thus, it appears acceptable to allow each participating member of an interchange system to decide for itself the degree of protection against active wiretapping which it considers cost effective between its terminals and EDP systems. Whenever an instance of ambiguous fraud occurs (which should be seldom), the associated liability can be stipulated in the Interbank rules.

Another aspect of fraud and liability concerns a cardholder's liability if his PIN is compromised. Consumer protection legislation will make it increasingly difficult for financial institutions to hold the cardholder financially responsible for protection of his own PIN. However, an interchange system which is extremely secure against passive fraud threats, coupled with a highly secure PIN management system on the part of the issuing institution (in which no employee of the institution knows any cardholder's PIN, and all PIN related operations are under strict dual control) can significantly reduce the risk that lost, stolen, or counterfeit cards will be used other than as the result of cardholder negligence. Furthermore, a well publicized, highly secure PIN system should discourage cardholders, who would otherwise fraudulently deny transactions they had in fact actually made (today's main fraud threat), and also convince cardholders to treat their PINs with considerable care. If the precautions taken by the financial institution to prevent any of the institution's employees from learning any cardholder's PINs are highly visible to the cardholder, it cannot help but influence the cardholder's own attitude toward the importance of PIN secrecy, and may, in time, also affect the judicial attitude toward the liability for PIN compromise.

Conclusions

Fraud is not expected to become a major problem for EFT until a nationwide EFT interchange system is in operation. At this point the payoff for fraud technology could be very great, and the result could be mass fraud of perhaps catastrophic proportions. Potentially, the most serious fraud threat is the production and use of counterfeit cards, and the widespread dissemination of large numbers of such cards could be truly disastrous. The use of lost and stolen cards is a potential problem but the use of any sort of PIN system could control this fraud threat. The remaining threat is the modification or insertion of data in real time. Though a serious threat, it would not have nearly as devastating an effect as the massive use of counterfeit cards.

The two main types of fraud threats are passive, the determination of PINs and corresponding account identities to enable the use of lost, stolen, or counterfeit cards, and active, the fraudulent modification or insertion of data in real time. Of these two, the passive is of greater concern in an interchange environment, not only because the passive threats could lead to mass fraud in the form of counterfeit cards, but also because it would be virtually impossible to ascertain the negligent party and thus establish liability for such fraud. Active threats, though potentially serious, are not judged to be catastrophic, and quite often can be traced back to the offending institution.

As a result, it is concluded that an EFT system should protect against passive threats in all aspects of interchange. If very high security is desired, it appears that an issuer's clear PINs should not be allowed, even momentarily, in any general purpose EDP equipment of an acquirer or switch but rather than such clear PINs should be restricted to physically secure cryptographic hardware where they, and the cryptographic keys used to encrypt them, can be physically protected. By imposing stringent security requirements upon all acquirers and switches, it seems more reasonable to then place liability for any use of lost, stolen, or counterfeit cards on the issuer, any compromise of his cardholders' PINs being assumed to result from negligence on the part of the issuer or its cardholders.

A further conclusion is that protection against active fraud threats can be left up to an acquirer's discretion, provided he is willing to assume responsibility for any such fraud against which he does not provide protection. However, it does appear desirable to protect transactions in interchange between institutions against these active threats.

Perhaps the most important conclusion of all is that fraud in a nationwide interchange environment could eventually become a serious problem, if not a major catastrophe. Thus, in planning for interchange, and during the evolution toward it, careful provisions must be made to ensure that adequate security is indeed realized. Security must be an inherent characteristic of EFT from the start. It cannot be added on after the fact.

SECTION THREE: PRINCIPLES OF FRAUD PREVENTION

Cryptography, The Tool for Fraud Prevention

Most of the techniques used to prevent fraud in an EFT system are based upon the use of cryptography. Cryptography involves encryption and decryption. Encryption is the transformation of clear (comprehensible) data by means of an algorithm (i.e., a defined procedure) and a secret number called a key into a form called cipher, which bears no resemblance to the original, clear data. Only someone else possessing the identical secret key is able to decrypt the transformed data and recreate its original clear form.

The Interbank recommended encryption algorithm for use in EFT is the Data Encryption Standard (DES), which is the cryptographic algorithm sponsored by the National Bureau of Standards for data security. It is the only publicly available algorithm which has been certified as highly secure by the United States Government. DES is a block encryption technique. That is, given an input data block of 64 binary bits, a 56 bit secret binary key, and the encrypt command, the algorithm produces 64 cipher bits. These bits bear no obvious relation to the input. In fact, a minor (i.e., one bit) change in the input produces a drastic change in the output. Given these 64 cipher bits, the same 56 bit key, and the decrypt command, the algorithm produces the original clear data.

Though inherently a binary, block encryption technique, DES can be applied in various ways to implement virtually any desired type of encryption.

Preventing Passive Fraud Threats

As described in the preceding section, passive fraud threats are those which enable the PIN for known accounts to be ascertained. As a result, the use of lost, stolen, or counterfeit cards is possible. Given today's technology, the fraud threat can be virtually prevented by:

1. Insuring that the PIN is encrypted at all times except when within a physically secure environment.
2. Insuring that the keys used to encrypt PINs are never available in "clear" form except within physically secure environments.
3. Insuring that physically secure environments, in which clear PINs and clear PIN encrypting keys are found, are in fact secure against physical compromise.

PIN Encryption

To be protected against compromise the PIN should be encrypted using DES and a secret key. Furthermore, the PIN should be encrypted as a function of some quantity which varies from transaction to transaction, or at least from account to account. Were this not done, identical clear PINs encrypted under the same key would produce identical ciphers. This fact could be exploited by a criminal, who would, for example, open ten accounts under fictitious names. He would then know ten PINs and institute transactions from a specific EFT terminal against each of his ten accounts. He would also have tapped the line from his terminal, and would record the encrypted PIN from each of his ten transactions. Next he would use the tap to record transactions from other cardholders. Whenever the encrypted PIN from one of these transactions exactly matched the encrypted PIN from one of his accounts, he would know that the clear PINs were identical. In an environment in which most or all PINs are four digits, he would be able to ascertain the PIN for approximately one account in a thousand which used this particular terminal. For each of these cases, from other information in the recorded transaction, he could determine the contents of the associated card's magnetic stripe, and thus make a counterfeit copy of this card and use it to draw against the cardholder's funds.

One method of encrypting the PIN as a function of a varying quantity (and thus preclude the above indicated fraud threat) is to concatenate the clear PIN with a value six decimal digits or longer, which is:

1. A random or pseudorandom number,
2. A counter, which increments on each transaction, or
3. The least significant digits of the account number.

The result, which must be no more than 64 binary bits in length, is block encrypted using DES and a secret key. The entire 64 bit resulting cipher thus serves as the encrypted PIN for this transaction.

Other equally secure PIN encryption methods exist in which the encrypted PIN is a decimal value of the same number of digits as the clear PIN.

Protection of Cryptographic Keys

PINs should always be encrypted under secret 56 bit DES keys. Maintaining the secrecy of these keys is of the utmost importance, because if any such key becomes compromised the PINs encrypted under it can be similarly compromised. There are two problems associated with key secrecy. The first is the generation and distribution of keys in such a way as to preclude compromise. This is called key management, and will be discussed subsequently. The other, protection of the key while it is within the cryptographic device, will be considered now.

Physical Protection of PINs and Cryptographic Keys

PINs and keys must be physically protected whenever they are in clear (unencrypted) form within cryptographic devices. (For high security, clear PINs and keys should exist nowhere else.) The most effective solution to this problem appears to be the interlocking of the device's enclosure. All of the cryptographic logic is placed within a physically secure enclosure and tamper detection circuitry, built into the enclosure, detects any attempt to gain access to the internal circuitry of the device. The secret keys are interlocked by means of these tamper circuits so that any act of tampering causes the keys to be erased.

The immediate erasure of the keys obviously protects them from compromise. Tapping of the device to ascertain future PINs is prevented also, because opening the device to install the tap erases the secret keys and this renders the device inoperative (i.e., unable to decrypt incoming PINs). As a result, the tap will not successfully capture PINs.

In order to protect the PIN at its point of entry into the system, the PIN keyboard must be a part of this protected enclosure. If it is, and if the enclosure is properly protected via the above suggested interlocks, the terminal tapping threat discussed in the preceding section is precluded.

Preventing Active Fraud Threats

Though passive fraud threats may be countered by simply protecting the PIN from disclosure, countering active threats is more difficult, with different countermeasures needed for different threats.

Data Modification

Many active fraud threats involve the modification of data in real time. These threats can be countered by either of two techniques, message encryption or message authentication. Message encryption, as the name implies, is simply the encryption of all, or at least most, of the message which conveys the transaction. This technique has the advantage of providing privacy as well as security. The disadvantage, however, is that the encrypted message cannot be comprehended or processed by the EDP systems (acquirer's, switch's, issuer's) through which the transaction passes. Thus, the message must be decrypted prior to being processed, but once decrypted, it loses its cryptographic protection, and therefore is susceptible to fraudulent modification within the EDP system.

Message authentication is a technique which produces cryptographic check digits which are appended to the message. These digits are analogous to a parity check or cyclic redundancy check, except that in this case the check digits are cryptographically generated, using DES and a secret key. These digits, called the message authentication code or MAC, are generated by the originator, appended to the transmitted message, and then checked by the recipient, who also holds the same secret key used in the generation process. Should anyone attempt to modify the message between the time the MAC is generated and the time it is checked, he would be detected. Not knowing the secret key, he would be unable to generate the correct MAC for his modified message. Similarly, no one can successfully introduce a spurious message because he could not generate the proper MAC for this message.

The suggested technique for MAC generation is as follows: The first 64 bits of that portion of the transaction to be protected are block encrypted using DES and the secret key. Then the next 64 transaction bits to be thus protected are Exclusive-ORed (modulo 2 added) with the just produced cipher. The result is then block encrypted using the same key, producing a new 64 bits of cipher. This procedure is continued until all critical transaction fields have been included. (The final data block will likely be less than 64 bits, so it is padded with zeros to make a full 64 bits prior to being Exclusive-ORed with the just produced cipher.) Some subset of the final cipher, at least six decimal digits or five hexadecimal digits, serves as the MAC.

As a minimum, the following fields should be included in the MAC generation for a transaction request message:

1. Transaction type (debit, credit, etc.).
2. Cardholder's account number.
3. Amount.

4. Transaction identification information (that information which uniquely identifies a specific transaction from a particular terminal).

In the case of a transaction response message, the equivalent fields plus the response code (approved, disapproved) should be protected. Alternatively, the MAC can be generated only if the transaction is approved, because no known fraud threat could exploit an unprotected disapproved response message.

Though the MAC approach does not provide privacy, its advantage over message encryption is that the protected message is also intelligible, and thus can be processed by EDP systems. With message encryption, the protection is lost when the message is decrypted, so the message is protected against fraudulent modifications only over communications lines but not within EDP systems. Thus message authentication, by protecting the transaction against fraudulent modifications both over communications lines and within EDP systems, is the recommended approach.

As indicated in the preceding section, message authentication can be optional within an acquirer's own network, provided the acquirer is willing to assume any fraud loss which might result from the failure to protect against any "active" fraud threats. However, it is recommended that all transactions in interchange be protected against active fraud by means of message authentication. If subsequent privacy legislation requires encryption, this can be accomplished, for the interchange network, by using link encryption devices, or by encrypting the six or so least significant digits of the account number to conceal the cardholder's identity.

Replay of Debit Authorization

Perhaps the most likely active fraud threat is the isolation of a terminal (especially an ATM) from its host, giving it an "approved" response to every transaction request. Message authentication alone cannot necessarily solve this problem, because it might be possible to record the "approved" response to a valid transaction prior to isolating the unit, then institute identical fraudulent transactions, and replay the previously recorded "approved" response. This fraud threat can be countered only if there is something unique about each transaction (even two transactions for the same amount against the same account), and the transaction-approved response includes this uniqueness, which is checked by the terminal. Furthermore, this unique characteristic must not be something which the criminal can duplicate. For example, the terminal itself can insert a sequence number into the transaction request message. This same number is included in the transaction authorization message back to the terminal protected under the MAC. The terminal checks for agreement between the two values and authorizes the completion of the transaction (e.g., the dispensing of cash) only if the sent and received sequence numbers are identical. Thus, it is impossible to replay the "approved" response to a previous transaction because the sequence number is invalid for any other transaction. Furthermore, the replay version cannot be modified to include the current sequence number, because the MAC would not check.

For this approach to provide the required degree of protection, the sequence number should not repeat within the life of the cryptographic key used in the MAC generation. Alternately, it could be a random, rather than a sequential value, provided it is truly unpredictable, and is at least six decimal digits in length.

Other approaches can provide the same effect. For example, if the terminal key is changed after every transaction, a MAC on the transaction authorization message is unique to a given transaction.

Fraudulent Credits

The four fraud threats associated with credit transactions are:

1. Modified credits (amount field increased, or debit made into a credit).
2. Misdirected credits (account number modified).
3. Spurious credits (totally fictitious credit transaction).
4. Fraudulent replay of previously valid credit transaction.

The first three fraud threats may be countered by the use of message authentication. The use of MAC prevents undetected modifications in any critical transaction field, and also prevents the origination of totally spurious transactions. The fourth fraud threat, however, is somewhat more difficult to counter. Though message sequence numbers are commonly used to detect duplicated messages, these numbers are not considered a part of the security system and thus cannot be relied upon for fraud prevention purposes. That is, they can only be checked in the presumably non-secure main frame, and cannot be checked in special security equipment.

It appears that the responsibility for detecting replayed credits must rest with the issuer. Under this approach, the issuer would be expected to check the previous real time EFT credit transaction for the account in question and verify that it had occurred earlier than the current credit transaction. (This assumes that real time EFT credit transactions are stored in chronological order and that the date/time field in every such credit transaction is protected by the transaction's MAC.) The issuer could then immediately detect a replayed transaction.

It should be noted that real time credits in a retail EFT system should be relatively infrequent. Normal bank deposits are not real time because they are subject to verification and to check clearances. The only expected real time credits would result from the return of merchandise, a relatively uncommon occurrence, so placing the responsibility for detecting the replay of such credits upon the issuer should not be an undue burden.

Encrypted PIN Substitution

Unless appropriately precluded, it would be possible for a criminal to record an encrypted PIN as it leaves a terminal in a valid transaction, then, by active wiretapping at a later time, replay this recorded encrypted PIN as part of a fraudulent transaction. This fraud threat is prevented by techniques which insure that the same PIN when encrypted by the same terminal on two or more different occasions always produces a unique cipher each time. This can be achieved by encrypting the PIN as a function of a variable quantity, such as a terminal generated transaction sequence number (in addition to the secret key). To be fully effective, this terminal should utilize message authentication for the transaction authorization message. This message should include the same variable quantity (e.g., the transaction sequence number) which is used in the encryption of the PIN in the transaction request message. Only if the variable quantity in the authorization message matches the one the terminal used for PIN encryption in the request message, and only if the authorization message is successfully authenticated by the terminal, does the terminal complete the transaction. At the host end, PIN validation and authentication of the authorization message should be performed as a single operation in a physically secure environment, the message authentication code for the authorization message being generated only if the cardholder entered PIN is successfully validated.

If the above indicated procedures are followed, the criminal is prevented from fraudulently replaying a previously recorded encrypted PIN. This PIN would have been encrypted as a function of a previously used variable quantity which could not be successfully reused.

Another technique which can be used to prevent the substitution of the encrypted cardholder entered PIN is key transformation. With this technique, the terminal's key is changed after every transaction. This is accomplished by generating the new key from the old key via a cryptographic procedure. Since the key used to encrypt the PIN changes on each transaction, the criminal is unable to successfully replay the encrypted PIN from a previous transaction. To insure this completely, however, the authorization message to the terminal must be authenticated using the current key.

Other fraud threats are possible if the criminal is able to substitute the encrypted version of a PIN, which the criminal himself knows, for the cardholder's encrypted PIN of reference. Two techniques are required to preclude this threat. First, the cardholder's PIN must be encrypted as a function of his account number. This prevents the criminal from substituting his encrypted PIN of reference for that of the targeted cardholder. Second, the PINs of reference must be encrypted under a cryptographic key never used to encrypt cardholder entered PINs. This prevents the criminal from using an invented PIN with a counterfeit card for the targeted account, then replacing the PIN of reference for this account with the encrypted version of the invented PIN.

It should be noted that the prevention of this fraud threat requires that the PIN of reference and the entered PIN both be decrypted (or that the latter be decrypted, then reencrypted like the former) before a comparison is made since both are encrypted under different keys. This does not permit a commonly used PIN validation technique in which the encrypted cardholder entered PIN is compared in a non-secure environment against a similarly encrypted PIN of reference.

Fraud Prevention in Interchange

The fruad threats associated with interchange are basically no different than those already considered, namely, passive threats to ascertain PINs (for use with lost, stolen or counterfeit cards), and active threats to modify or insert data in real time. Thus, the fraud prevention techniques considered above apply to interchange just as much as to local operations. However, interchange poses a practical problem concerning the implementation of these fraud prevention techniques, the necessity to translate from one cryptographic key to another.

A typical interchange transaction begins when the cardholder enters his PIN at some EFT terminal which also reads his card. At this point the PIN must be encrypted in a key unique to this particular terminal.

Similarly, if message authentication is used, it must be based on a key, the same or different, but still unique to this particular terminal. In an EFT system with thousands or tens of thousands of encrypting EFT terminals, it is not feasible, or even desirable, for every key of every acquirer's terminal to be known at every issuer's facility. Thus, each acquirer must have the capability to translate from the terminal's key to an interchange key known either by the issuer, or by the switch which serves the acquirer (in which case the switch will make a second translation into a key known to the issuer). Furthermore, in the case of the PIN, this translation must take place under conditions of very high security, desirably using special, physically protected, cryptographic hardware. As indicated previously, it may eventually be considered unacceptable for one institution's clear PINs, or the clear keys used to encrypt such PINs, to reside even momentarily in the general purpose EDP equipment of any other institution. Thus, for high security, the acquirer should not perform this translation function of the PIN using CPU software, but rather should use the above indicated physically secure hardware.

In the case of message authentication, however, this does not necessarily apply. Since message authentication is optional on the acquirer's part, he may use CPU software for MAC translation. However, in this case a completely different key must be used for MAC generation than is used for PIN encryption or the clear PIN key would exist in the acquirer's CPU. Since the acquirer may utilize special cryptographic hardware for PIN translation, it is suggested that this same hardware be used for MAC translation as well.

Should an acquirer or a switch find an invalid incoming MAC while performing MAC translation, it is considered acceptable for this institution simply to generate an invalid outgoing MAC. In this way, only the issuer, who performs the final MAC check, need implement the error paths which handle the invalid MAC situation.

Another acceptable technique is to superimpose the MAC on the encrypted PIN and avoid an additional field to the message. This may be accomplished by Exclusive-ORing the MAC and the encrypted PIN. Should the MAC not check, the issuer finds a garbled PIN, and the PIN check fails. While this technique cannot distinguish between an invalid PIN and a modified message, this is of little consequence since the net result in either case is to disallow the transaction.

It is assumed that the EFT system as a whole provides an adequate degree of error control, so that the MAC is relied upon for fraud detection rather than error detection. Since fraud attempts are expected to be virtually nonexistent because of the use of message authentication, there is virtually no inefficiency in relaying a message with an invalid MAC all the way to the issuer.

Returning again to a typical interchange transaction, the encrypted PIN is transmitted from the EFT terminal, encrypted in this terminal's secret key. Optionally a MAC, to protect the critical message fields, is also included. The transaction message, including the encrypted PIN, and optionally the MAC, reaches the acquirer's facility. Here the above indicated cryptographic transaction takes place. The PIN is decrypted using the terminal key, and reencrypted using an interchange key. If there is an incoming MAC, this too is translated into the interchange key. If there is no incoming MAC, one is generated.

The interchange key indicated above is either a bilateral key shared by the issuer and the acquirer, or a key shared by the acquirer and his EFT switch. In this latter case, the transaction goes to the switch where a second cryptographic translation occurs, translating the PIN and MAC from the key used between switch and acquirer to that used between switch and issuer. The message with this new encrypted PIN and MAC is then transmitted from switch to issuer, where the MAC is checked and the PIN decrypted and validated. If the transaction is approved by the issuer, a transaction authorization message is sent from issuer to switch. This message is protected by means of a MAC using the key shared by switch and issuer. The switch, upon receiving the transaction authorization message, translates it into the key used between switch and acquirer. The MAC in this key reaches the acquirer. If the MAC is valid, the appropriate authorization response is transmitted to the terminal where the transaction originated. If this terminal expects a MAC with the authorization message, such a MAC is generated, in the terminal key.

Countering the Fake Equipment Threat

Perhaps the most difficult fraud threat to counter is the fake equipment threat, in which a dishonest merchant induces unsuspecting cardholders to use EFT terminals with PIN pads which are fake. Either the entire terminal is fake, or a fake PIN pad is added to a non-PIN-using terminal. In either case the PIN pad output goes to a recorder, as does the output of the magnetic stripe reader. From the information thus collected, the criminal is able to produce usable counterfeit cards. Fortunately, this fraud threat is considered too blatant to be especially probable, but must be considered nevertheless.

This threat can be countered only with the help of the cardholders themselves, some of whom can be induced into cooperating by the offer of substantial rewards (but only after the fraud threat has actually materialized). To enable the cardholders to detect that something is suspicious, a code printed on the EFT receipt must indicate whether or not a PIN was used with the transaction in question. For example, a Transaction Proof Code with a non-zero leading digit printed on the receipt, can indicate that a PIN was used. If the leading digit is zero, a PIN was not used. Thus, an observant cardholder who uses such a terminal with a PIN pad and finds the leading digit of his Transaction Proof Code is zero, would know that he could receive a reward for informing the financial institution of this fact without alerting the merchant.

In a similar manner, an account-related sequence number, maintained by the issuing institution and printed on the EFT receipt, would allow the alert cardholder to immediately detect a totally fake terminal (which did not communicate with his institution) because the sequence number printed on the receipt would not be the number he was expecting.

It is possible that the dishonest merchant could have a non-PIN terminal (with a fake PIN pad) modified so as to change, in real time, the leading Transaction Proof Code digit from zero to some other value. However, the Transaction Proof Code as recorded by the financial institution also prints on the cardholder's monthly statement. Thus, the alert cardholder would also know that he could receive a reward for reporting discrepancies between his EFT receipts and his statement (provided these discrepancies could be confirmed by the institution). The totally fake terminal (which resulted in no statement entry) and the terminal which actively modified the Transaction Proof Code, would be discovered through such cardholder reports. In an interchange environment with on-line EFT, it would be virtually impossible for the dishonest merchant to predict just how soon after using his fake equipment an alert cardholder would receive a statement and report his suspicions. The dishonest merchant would be exposed to an unacceptably high level of detection and therefore would not likely attempt this type of fraud in the first place.

While many cardholders would either not understand, or ignore the above suggested reward offers, a small number of intelligent and alert cardholders should respond. Even these few should be sufficient to pinpoint such dishonest merchants relatively quickly.

Since this fraud threat is considered rather improbable by Interbank, little attention need be paid at this time. Nevertheless, the possibility of the threat, and the techniques for countering it, should be kept in mind.[3]

Conclusions

The two basic techniques used to provide security in an EFT environment are PIN encryption and message authentication. PIN encryption prevents passive fraud threats from ascertaining PINs, and thus precludes the use of lost, stolen, and counterfeit cards. For PIN encryption to be effective, the PIN must be physically protected everywhere that it is not encrypted. Similarly, the cryptographic keys used to encrypt PINs must be protected. Inter-

[3] A fake equipment attack, which cannot be defended against using the above mentioned technique is described in Chapter 11 in the section entitled Threats to the Secrecy of a Key Stored on a Magnetic Stripe Card.

locking these keys with the cryptographic device's enclosure is one obvious technique for providing the required physical protection.

Message authentication insures message integrity, and prevents most active fraud threats. A few active fraud threats, however, require specialized countermeasures, as does the fake equipment.

Interchange does not pose any unique fraud threats, but it does require a special cryptographic capability—translation from one cryptographic key to another under conditions of very high security.

The above discussion has presented only the basic principles of fraud prevention. The following section considers, in some detail, how these principles can be effectively implemented, both in EDP facilities and in EFT terminals. In addition, techniques for secure PIN management and key management will be considered.

SECTION FOUR: IMPLEMENTATION OF FRAUD PREVENTION TECHNIQUES

Section Three considered the general principles which are recommended for preventing fraud in EFT networks. This section will describe in further detail how these principles can be applied, through the use of a "security module." In its preferred implementation, such a module is a physically secure hardware device which serves as a peripheral to an EDP system. Potentially less secure implementations are also possible, in which security module functions are implemented in mainframe software. The implementation used by an issuer for its own PINs is clearly its decision. The implementation used by an acquirer in Interbank interchange may be dictated by future standards, though it is presently premature to indicate what such standards will be.[4]

Suggested Characteristics of Hardware Security Module Implementation

When very high security is desired, hardware implementation of the security module is recommended. The previously mentioned Interbank security study considered in detail how such a hardware device could best be implemented. The hardware implementation as suggested by this study is now considered.

The suggested security module is a self-contained, physically secure, microprocessor controlled cryptographic device programmed to perform the cryptographic functions which the EDP center of a financial institution requires for its EFT operations. A security module interfaces with the institution's EDP system as a peripheral device. Information which requires cryptographic processing is sent from the computer to the module, which almost immediately sends back the results.

Each security module can be programmed to perform virtually any required cryptographic function. These programs are written to insure that the secret ingredients in EFT operations, customer PINs and cryptographic keys, never exist in a clear form outside the physically secure internal circuitry of a security module or of an EFT terminal's cryptographic hardware. In effect,

[4] See Cryptographic PIN Security—Proposed ANSI Method, Appendix E.

the security module takes from a non-secure EDP system all information which must be kept secret to prevent EFT fraud, and concentrates it within a dedicated module where it can be physically protected. Thus, the security module trades computer security, which is now and may always be, elusive, for physical security, which is well understood.

The suggested security module achieves its physical security by means of both locks and interlock circuitry. The module is protected by two different physical locks, requiring two different physical keys. This insures that the module can be legitimately opened only under dual personnel control. Furthermore, whenever the module is opened, whether legitimately using the two keys or by force, interlock circuitry causes all secret data stored within the module (i.e., the cryptographic keys) to be erased. If a criminal breaks into the module, the secret information contained therein will disappear as it is forced open.

The suggested security module system consists of three modules, all electrically and mechanically independent, but sharing a common cabinet. A conventional terminal serves as the keyboard and printer for the system. It is connected to only one of the three modules at any given time, and serves to perform certain subsequently described PIN and key management functions.[5]

Suggested Capabilities[6]

The main functions performed by the security module include PIN management, PIN verification, PIN "translation" in interchange, key management, and message authentication. These functions are based on the National Bureau of Standard's Data Encryption Standard (DES), although proprietary algorithms can be supported as well.

For PIN management, the module can generate a random value for the PIN, then encrypt this value for storage in the mainframe and/or for encoding on the card's magnetic stripe as the "PIN offset" of the "PIN verification field." Alternately, the module can cryptographically derive the PIN from the account number. In either case, the module can print the PIN mailer on a dedicated printer. As a result, the unencrypted PIN is never known to any person and is never present, even momentarily, in the mainframe. If the institution prefers that cardholders select their own PINs, the module implements a system that never allows anyone within or outside the bank, except the actual customer, to associate an unencrypted PIN with the corresponding account number.

For PIN verification, the module decrypts a PIN which has been encrypted by an EFT terminal. It then compares this customer-entered version of the PIN with the "reference" version of the PIN using the technique appropriate to the institution in question, which can be an encrypted PIN from the data

[5] The security module described here was developed and tested in prototype form by Interbank Card Association under the name PINPACK (see also reference 2).

[6] The following six paragraphs are from *Datapro Reports on Banking*, Report No. B61-854-101, "Transaction Security Products Security Module" [3]. Copyright 1980 by Datapro Research Corporation, Delran, New Jersey. All rights reserved. Reprints may be obtained from Datapro Research Corporation.

base, a "PIN offset" or "PIN verification field" from the card's magnetic stripe, or a PIN cryptographically derived from the account number. The module responds with a "valid" or "invalid" indication, but in no case discloses the unencrypted PIN.

For interchange use, the module can "translate" the PIN (and any other data) from the cryptographic key and format used by the EFT terminal to the cryptographic key and format used for interchange. When the module is used for this "translation" function, no unencrypted PINs of any issuer are ever present, even momentarily, within the mainframe of participating institutions, where they might be subject to disclosure.

All functions performed by the DES algorithm are controlled by "keys." A security module is able to generate, control, maintain, and protect all keys associated with the user's network. This includes terminal keys, data storage keys,[7] and interchange keys. Terminal keys are used for encrypting and decrypting PINs and other data transmitted between a terminal and its host. A module can generate terminal keys and can support down-line key loading. Data storage keys protect sensitive data such as PINs stored in a user's data base. Interchange keys are used for transmitting data among various users within a shared system.

The message authentication function performed by a module provides protection against fraudulent modification of messages by cryptographically protecting the text. This is accomplished by processing critical message fields through the DES encryption algorithm, which generates a Message Authentication Code (MAC) appended to the message by the originator and checked by the recipient. Without knowledge of the key used in this process, anyone attempting to modify the message fraudulently would be unable to do so without detection.

PIN management refers to the techniques by which an institution issues, stores, and validates customer PINs. Here the system provides several options.

Bank Selected Random PIN

The security module itself can generate PINs, then print a PIN mailer on its own dedicated terminal printer. This just generated PIN is then encrypted under the PIN Master Key [PMK] and transmitted to the CPU for storage in the EDP system's data base, or, for encoding on the magnetic stripe of the customer's card.

Before the security module will print PIN mailers, it must be put into the "authorized state." To do this, each of two members of the institution's staff must enter a different secret code. Each had previously selected his code, and neither knows the other's code. If both of these codes are entered correctly, the module enters the authorized state, and will print PIN mailers when instructed to do so. If it is not in the authorized state it will respond with an error indication to such an instruction. This would prevent unauthorized personnel from printing PIN mailers at a time when the security module's printer is not properly secured.

[7] Storage keys are equivalent to the master keys discussed below.

PIN Cryptographically Derived from the Account Number

The security module, using DES, can cryptographically derive a PIN from a customer's account number and issue the PIN to this customer on a PIN mailer as discussed above.

Customer-Selected PIN

This may be accomplished by a mailed customer response, by a document given to the customer at the bank's facility, or by having the customer enter his PIN via a secure terminal.

To provide secure management for a customer mailed response, a PIN solicitation document is prepared and mailed to the customer. The portion to be returned, on which the customer writes his PIN, contains no customer identifying information except a reference number. This reference number is really an encrypted account number, intelligible only to the security module. Someone who sees the returned portion of the mailer would be unable to relate the PIN to the account. A bank employee uses this returned portion to enter the selected PIN and the reference number into the security module. The module then decrypts the reference number to determine the account number and encrypts the PIN under the PIN Master Key [PMK]. This resulting information is then transferred to the CPU for storage in the data base. At no point in this process is the clear PIN ever associated with the clear account number.

A similar document can be used by those institutions which have the customer choose his PIN at the time he applies for the account. Another technique which can be used under some conditions is to have the customer convey his PIN selection by entering it into a secure EFT terminal.

PIN Validation

The security module provides a number of different techniques for PIN validation, depending upon the characteristics of the EFT system and how the PIN was issued and stored. In most of today's ATM systems the PIN is validated by the ATM itself. In this case, the only role played by the security module is the preparation, during the PIN issue process, of the encrypted PIN or offset to be encoded on the magnetic stripe of the bank card used to activate the ATM. (Many ATMs use this offset value in the PIN validation process.) In an interchange environment, however, PIN validation must be performed at the EDP system of the card issuing institution, or of some other institution designated by the issuer to perform this function. (As previously indicated, it would be non-secure to have the information needed for the validation of one institution's PINs available to all other institutions.) In this type of environment, the customer's PIN arrives from the EFT terminal or the interchange network, encrypted under a terminal key [TK] or an interchange key [IK] known only to the security module. The module decrypts the customer entered PIN, and then determines the PIN of reference for comparison purposes. In some institutions the security module determines this PIN of reference from the account number, or from data encoded on the bank card (which was previously generated by the module during the PIN issue process). In other institutions there is an encrypted PIN entry in the data base for each account, and this entry is passed to the security module along with the transaction. This entry was generated by the module itself during the PIN issue process, and the security module alone holds the PIN Master Key [PMK] required to decrypt it. After obtaining the PIN of reference and comparing it with the customer entered PIN, the security module informs the EDP system of the comparison result, but does not, under any conditions, output the clear PIN.

Key Management

The preceding discussion has mentioned several types of keys: master keys, terminal keys, and interchange keys. The generation and management of such keys is an important feature of a security module system.

Master keys, of which there are perhaps fifteen in a security module, are common among all the modules which serve a financial institution's EDP facility. However, no two institutions share the same, or even similar, master keys. Master keys are used primarily for encrypting terminal keys [TK1, TK2, ...], interchange keys [IK1, IK2, ...], and PINs. Interbank has developed a special, highly secure Key Management Center for the generation of master keys for use in Interbank interchange. The Interbank personnel who operate this center do not have the capability to ascertain the keys it generates. These keys are conveyed from the center to the institution's security modules via electronic key transfer devices, so no printed record of the keys is ever produced. Though these devices must be transported to the institution under dual controlled conditions of very high security, the use of two independently conveyed devices for each set of keys means that both devices would have to be compromised before any of the conveyed master keys would be revealed.

The keys for an institution's terminals are produced by that institution's security module. The module must first be placed in the authorized state as described previously. Then the module, upon command from the EDP system, generates a random value to use as the terminal key. Ideally, this key is transferred from the security module directly into an electronic key loading device by which it is transported to the terminal in question. Such devices, several versions of which are in use today, prevent the person who loads the key from ascertaining it. As an interim necessity, in the absence of the terminal's ability to interface with such a device, the key may be formed by the addition of two or more sub-keys, each of which is separately printed by the security module's printer, carried to the terminal, manually entered into it, and then the printed record destroyed. Desirably, the key is formed inside the terminal as the sum of two values which are independently conveyed from the security module to the terminal.

After the security module has generated a new terminal key [TK] and transferred it to a key loading device (or to the printer) the module encrypts the just generated key under its Terminal Master Key [TMK], and sends this encrypted value to the EDP system for storage. This eliminates the necessity of internally storing a large number of terminal keys. Every time a transaction comes from a terminal, the EDP system finds the corresponding encrypted key in its data base and passes this to the security module along with the transaction.[8]

Interchange keys are used between the security modules of different institutions to encrypt data in interchange. Such keys for use in Interbank interchange are generated by the Interbank Key Management Center. Usually they are generated on a bilateral basis, two institutions (or an institution and its EFT switch) sharing a common interchange key for transactions between them. The Interbank Key Management Center must first have conveyed to each institution a unique Interchange Master Key [IMK]. The center then generates a random value to serve as this bilateral interchange key [IK], then encrypts this interchange key under the Interchange Master Key of the first institution [IMK1], then under the Interchange Master Key of the second [IMK2]. Thus encrypted [E_{IMK1}(IK) and E_{IMK2}(IK)], the interchange key may be conveyed to each institution via non-secure means.

Note that the security module's PIN management and key management techniques are designed to enforce dual control over all critical manual operations. If the security module is implemented in the suggested physically secure hardware, no cryptographic key can ever be ascertained by anyone, and no PIN can be ascertained except by the customer to whom it is issued, unless there is fraudulent collusion between two explicitly trusted employees of the institution.

MAC Generation[9]

Another security technique provided by the security module is called message authentication. Under this technique, the message is cryptographically processed using a secret key. This process produces a residue, which is then appended to the clear message. This residue, called the message authentication code or MAC, is generated by the originator and checked by the recipient. Should anyone attempt to modify such a message while in transit, he would be unable to do so without detection. Not knowing the secret key used in the authentication process, one would be unable to generate the MAC appropriate to the modified message. This technique is used primarily to prevent messages from being fraudulently modified as they traverse nonsecure communications circuits and EDP systems. Since the message is assumed to be in clear (comprehensible) form, it can be comprehended, though not modified, by EDP systems along the way [2].

Utilization

Security module utilization in an operational environment is perhaps best illustrated by means of two examples—its use in a *local* transaction and its

[8] Each ATM which performs PIN validation must internally store the PIN Master Key. To convey this key to an ATM, the security module encrypts it under the terminal key, so that it can be transmitted to the ATM over the nonsecure communications link. Replacement terminal keys can be similarly conveyed from security module to terminal.

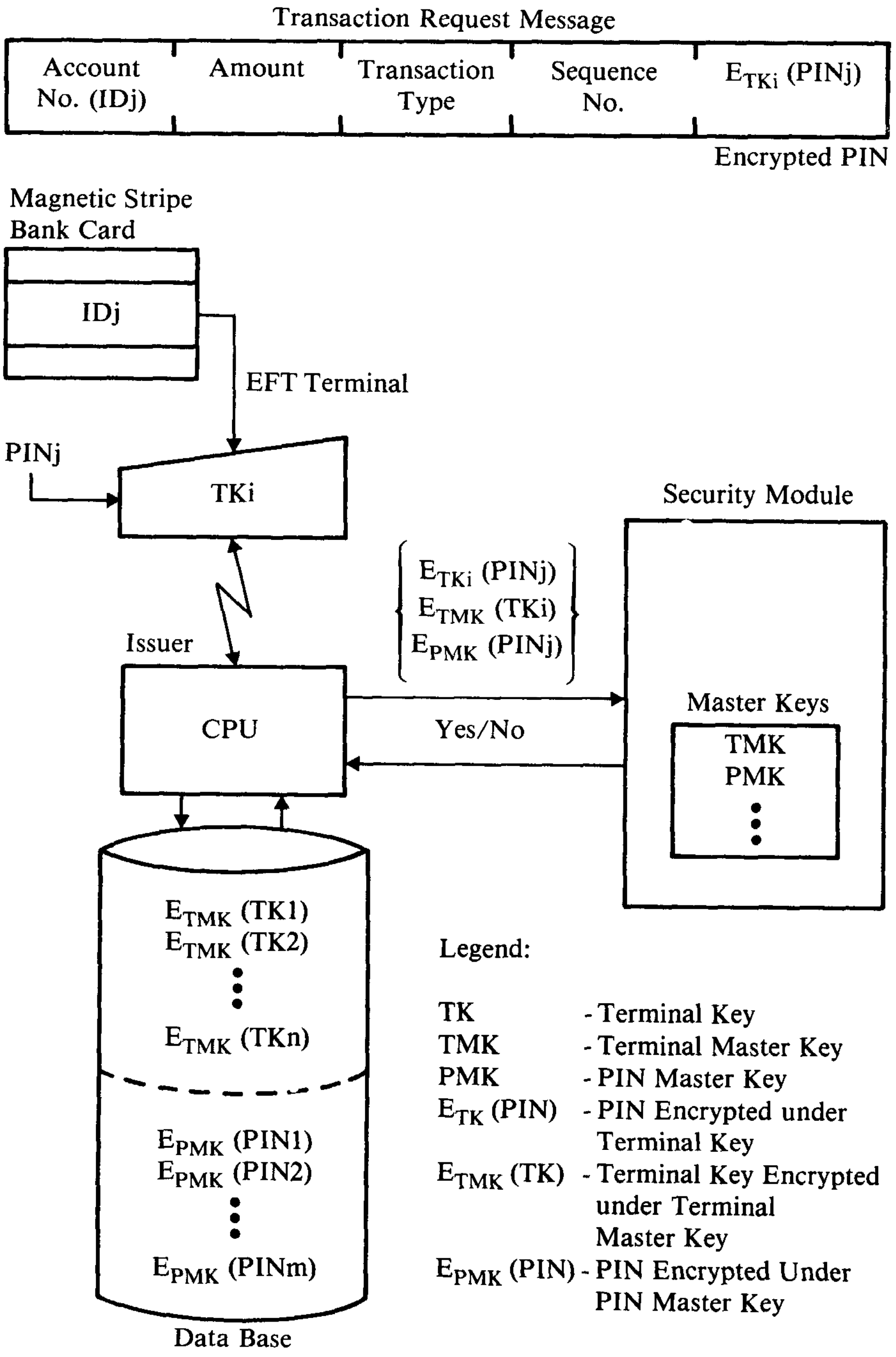
[9] The material in this section is taken from reference 2.

use in an *interchange* transaction. A local transaction is one in which the customer uses a terminal controlled by his own institution. An interchange transaction is one in which the customer uses some other institution's terminal.

Figure 10-1 illustrates a local transaction. The cardholder's PIN is entered into the terminal, where it is immediately encrypted under the terminal key [TK]. This encrypted PIN, together with the other elements of the transaction, is then transmitted to the institution's CPU. The CPU examines the just received transaction and determines the identity of the terminal from which it originated, and the identity of the cardholder who initiated it. (The cardholder is identified by his account number, which is read from the card's magnetic stripe by the terminal.) From this information the CPU finds, in its data base, the terminal's key encrypted under the Terminal Master Key, TMK, and the cardholder's PIN of reference encrypted under the PIN Master Key, PMK. (Alternately the encrypted PIN of reference may be encoded on the card's magnetic stripe.) These two encrypted values [E_{TMK}(TK) and E_{PMK}(PIN)] along with pertinent fields from the transaction are conveyed to the security module. The module contains, in its internal storage, the master keys [TMK and PMK]. Using its Terminal Master Key [TMK] it decrypts the terminal key from the data base. This just decrypted value is used to decrypt the cardholder entered PIN as received from the terminal. Then using the PIN Master Key [PMK] it either decrypts the PIN of reference from the data base or encrypts the just decrypted cardholder entered PIN. Either way, the two versions of the PIN are compared. If they disagree, the module sends the CPU a "no" indication. If they agree it sends a "yes" indication.

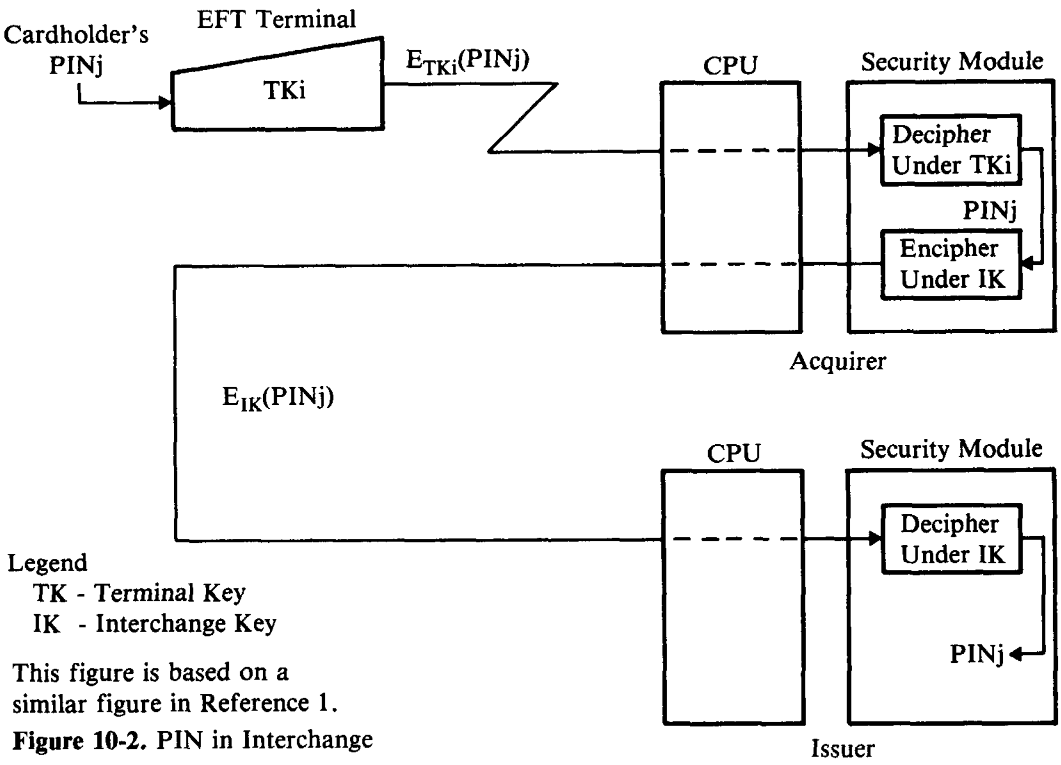
If the terminal in question uses message authentication, the module will output a valid MAC for the response message only if it finds both the incoming MAC and the PIN to be valid. This prevents a clever CPU programmer from substituting an "approved" response to a terminal for the "disapproved" response which always results from an invalid PIN (or modified message).

An interchange transaction, Figure 10-2, begins just as a local one, with the cardholder's PIN being encrypted under the terminal key [TK] and transmitted as part of the transaction to the acquiring institution. This institution's CPU examines the transaction to learn the identity of the terminal and the cardholder. Again, from the terminal's identity, it locates the terminal's key encrypted under the Terminal Key Master Key in its data base [E_{TMK}(TK)]. It notes, however, that this is an interchange transaction and that this particular cardholder's PIN is not on file here. It thus sends the encrypted terminal key and the transaction to its security module with instructions to perform a PIN translation. The security module decrypts the terminal key, TK, then uses this to decrypt the PIN. It immediately reencrypts the PIN under the appropriate interchange key, IK. In some cases, as illustrated in Figure 10-2, this is a key which the acquirer shares, on a bilateral basis, with the issuer. If it does not share a bilateral key with the issuer in question, the interchange key used is the one the acquirer shares with the interchange switch. The switch will then perform a second key translation, decrypting under the interchange key which it shares with the acquirer [say IK1] and reencrypting with the interchange key it shares with the issuer



This figure is based on a similar figure in Reference 1.

Figure 10-1. Issuer's PIN Validation - Local Transaction



This figure is based on a similar figure in Reference 1.

Figure 10-2. PIN in Interchange

[say IK2]. Either way, the issuer receives the PIN encrypted under an interchange key available to its security module. It then validates the PIN in essentially the same manner described for a local transaction.

Though not shown in Figure 10-2, every interchange transaction is protected by means of a Message Authentication Code, or MAC. This code is generated by the acquirer using the same key used for PIN encryption. The code is validated by the issuer. When the acquirer and issuer do not share a bilateral key, the switch's security module performs a MAC translation as well as the above indicated PIN translation.[10]

Message authentication is also used for the response message from issuer to acquirer, and again a MAC translation, if required, is performed by the interchange switch. If the EFT terminal uses message authentication, the acquirer's security module generates the proper message authentication code for the terminal only if the MAC from the issuer is valid.

Note that throughout this interchange procedure, the clear PIN exists only within the physically secure confines of the originating terminal and the two

[10] For example, the switch decrypts the PIN and checks the MAC using the interchange key shared with the acquirer (IK1). If the incoming MAC is found to be valid, it then reencrypts the PIN and generates an outgoing MAC using the interchange key shared with the issuer (IK2). (It is assumed that each interchange key shared with the switch is stored encrypted under the switch's Interchange Master Key, IMK.)

or three indicated security modules, thus providing the highest possible security for the PIN at all points in the interchange environment.

Conclusions

The above suggested "security module" can be implemented by mainframe software, or by a hardware device. Especially when implemented by a hardware device, the module provides very high security to perform those fraud prevention functions which a financial institution requires in order to participate in a secure EFT interchange network. It provides a means to translate encrypted PINs from one cryptographic key to another without allowing either the clear PINs, or the clear keys used to encrypt them, from existing even momentarily outside of the security module, or of the EFT terminal itself.

Not only does the security module protect PINs in interchange, but it also provides fraud protection for the institution's own PINs as well. It enables a system of PIN issuance, PIN management, and PIN validation by which not even one member of the institution's staff has the capability to ascertain cardholders' PINs. In addition, it performs all other cryptographically related functions (e.g., message authentication) which are required of an institution's EDP facility, and also provides for the required key management capabilities.

REFERENCES

1. *PIN Manual: A Guide to the Use of Personal Identification Numbers in Interchange,* Interbank Card Association (September 1980). Distribution restricted. Contact Security Department, MasterCard International Inc. (formerly Interbank Card Association), 888 7th Avenue, New York, NY 10019.
2. Campbell, C. M., Jr., "A Microprocessor-Based Module to Provide Security in Electronic Funds Transfer Systems," *Proceedings COMPCON 79*, 148–153 (1979).
3. "Transaction Security Products Security Module," Report Number B61-854-101, reprinted from *Datapro Reports on Banking*, Datapro Research Corporation, Delran, NJ (1980).