ARTICLE:

BANK CARD FRAUD IN SPAIN[1]

By **Ricardo M. Mata y Martín** and **Antonio Mª. Javato Martín**

Technological progress over the last two decades, in combination with the opening up of international borders across the internet that has further developed human and commercial relations, have led to the appearance of new payment systems in the form of bank cards. These instruments may be used at commercial shopping centres, at the network of cash dispensers, and now, with the development of telecommunications networks, in the context of the internet. The mass use of cards as a means of payment inevitably gives rise to a significant amount of fraud. The legal treatment of fraud under criminal law involving bank cards requires penalties to be set for the ways in which these categories of offenses may be committed, and which may be applied under these circumstances: conventional fraud, computer fraud and burglary or housebreaking.

Introduction

At present, electronic telecommunication networks may also be applied to payment systems. These technological advances mean that the most extensively used are the multiple versions of the bank card. The dynamics of economic globalization will, in turn, expand modern payment systems even further.

Naturally, as an effective instrument in commercial relations, the new payments systems are subject to a more or less complete set of legal regulations, which in certain situations requires the implementation of criminal legislation. For such legislation to be effective, it is essential to study and to define the elements that constitute the offences that may be applied to these still relatively novel acts. This article considers the issues relating to the fraudulent use of bank cards, because of the scale of their use – it is significant and the most widely used means of payment, even greater than cash payments – and their specific regulation.

As this article will demonstrate (and the work that this article is taken from), there is no specific provision in Spanish criminal law relating to the use of bank cards as a means of payment, except in the case of misrepresentation. As a means of payment in a criminal context, it is necessary to distinguish between the use of the bank card at commercial establishments, for payment over telecommunications networks, and card abuse at automatic cash dispensers (ATM). Finally, the fraudulent use of banks cards can imply the application of some types of criminal offence related to misrepresentation.

Payment in person at commercial premises

In this part, cases are considered in which a card held in the name of an individual is used without that person's consent as a means of payment at a commercial establishment from which a product or service is acquired, in such a way that the salesperson accepts the payment under the belief that the real card holder is in fact present. Doctrine[2] and jurisprudence[3] has implicitly equated this method of impersonation with the conventional offence of fraud, as regulated under article 248.1 of the Codigo Penal (Penal Code), which states that:

> Cometen estafa los que, con ánimo de lucro, utilizaren engaño bastante para producir error en otro, induciéndolo a realizar un acto de disposición en perjuicio propio o ajeno.

Fraud is committed by whoever, for personal benefit,

[1] *This paper forms part of the research into Electronic means of payment - Proyectos de Investigación sobre Medios electrónicos de pago VA111/04 (Programa General de Apoyo a Proyectos de Investigación de la Junta de Castilla y león) and SEJ2004-03704 (Planes Nacionales I+D/I+D+I, del Ministerio de Educación y Ciencia). Abbreviations (where used): CP: Código Penal/Penal Code; LOPJ: Ley Orgánica del Poder Judicial/Organic Law on Judicial Power; TS: Tribunal Supremo/Supreme Court; STS: Sentencia del Tribunal Supremo/Judgment of the Supreme Court; ATS: Auto del Tribunal Supremo/Order of the Supreme Court; SAP: Sentencia de la Audiencia Provincial/Judgment of the Provincial Court; RJ: Aranzadi (Repertorio de Jurisprudencia del TS)/Collection of Supreme Court Jurisprudence; ROJ: Repertorio Oficial de Jurisprudencia/Official Collection of Jurisprudence; CENDOJ: Centro de Documentación Judicial. Consejo General del Poder Judicial/Centre for Judicial Documentation General Council of Judicial Power; RGDP: Revista General del Derecho Penal/General Journal of Criminal Law.*

[2] *Jesús Fernández Entralgo, 'Falsificación y utilización fraudulenta de tarjetas electrónicas' in Tarjetas bancarias y Derecho penal. Cuadernos de Derecho Judicial, VI-2002, 58.*

[3] *See, amongst others, STS of 30-10-2003 -La Ley Juris 2004,10845-; STS 21-1-2003 La Ley Juris 2003, 1269-.*

From the perspective of the need for 'engaño bastante (sufficient deceit)' in the description of the offence, and in accordance with the general prerequisites of the modern principle of objective accusation, it is necessary that certain self-protection procedures be complied with in carrying out the payment correctly.

practices sufficient deceit to the extent that they mislead another, inducing the latter person to make an act of disposal to his own detriment or to that of a third party.[4]

Criminal deception under Spanish law requires, in the first place, deceitful conduct on the part of the offender (in this case the presentation of a card thereby affirming both an apparent ability to pay and sufficient solvency). The deceit practiced by the active subject must be sufficient to lead another party to be misled (the seller is misled into thinking that they are dealing with the person to whom the card was issued, and trusts in the solvency of the legitimate card holder, but is not in fact dealing with the person to whom the card was issued). The erroneous situation in which the other party is placed leads to an act of disposal (the transfer of goods or the provision of services by whoever receives the payment), which causes a loss for that person or for a third party (the seller, the card issuer or the card holder, according to whoever is liable to cover the costs of the amount that is defrauded). From the subjective point of view, the offender must act with an economic interest in mind and with the sole aim of personal enrichment.

In recent years, it has also been made clear that there is an obligation on the receiver of the payment to comply with certain procedures when accepting payment. From the perspective of the need for 'engaño bastante (sufficient deceit)' in the description of the offence, and in accordance with the general prerequisites of the modern principle of objective accusation, it is necessary that certain self-protection procedures be complied with in carrying out the payment correctly.[5] In cases where the card is presented as a form of payment at a commercial establishment, the basic procedure requires that the seller satisfy themselves that the person in possession of the card is the person to whom the card was issued, and should also check the expiry date of the card.

This tendency has been accepted in modern jurisprudence, which has consistently failed to apply the legal definition of fraud and has punished the offence, where applicable, solely as misrepresentation, under circumstances in which the victim of the deceit failed to act with due diligence that is expected in commercial practice when verifying the identity of the subject. A good example of the approach taken by the judiciary is the STS of 3 June, 2003,[6] which declared as abnormal the act of paying with a stolen bank card belonging to a person of the opposite sex, because the sales person made no effort to verify the identity of the card holder, not even to establish whether the person that presented the card was a man or a women, such that the deceit could not be qualified as sufficient to be held as a causal factor that helped to cause the economic transfer.[7]

As well as conventional fraud, the offence of misrepresentation of a commercial document (article 392 of the CP) may be considered, where the manuscript signature of the actual card holder is forged by another person on the sales receipt issued by the bank card reader. Similarities will occur between both categories of criminal offence. As much is established in the Agreement of the 2nd Chamber of the Supreme Court (Sala 2ª del Tribunal Supremo) dated 18 July 2007, subsequently applied in the Judgment of 19 July 2007 (nº 451/2007), in which the accused, a Romanian national, entered a jeweller's shop in the locality of

[4] *The penalty established for the crime of conventional fraud ranges from a six-month to a three-year prison term (article 249). This same penalty also applies to computer fraud.*

[5] *Jesús María Silva Sánchez in Pablo Salvador Coderch and Jesús María Silva Sánchez Simulación y deberes de veracidad, (Civitas, Madrid, 1999), 98 and following, 387; Francisco Muñoz Conde, 'De la llamada estafa de crédito' in RGDP 9, 2008, Iustel, 2 and following; Mercedes Pérez Manzano, 'Acerca de la imputación objetiva de la estafa', in Hacia un Derecho penal económico europeo. Jornadas en honor del Profesor Klaus Tiedemann, (BOE, Madrid 1995), 285 and following.*

[6] *Nº 807/2003 Actualidad Penal. Nº43. 17 - 23 November 2003, 2310 and following. Along the same lines, supported by the same judgment, the SAP of Barcelona of 25 January 2007, which deals with practically identical circumstances (a card bearing the name of a woman fraudulently used by a man).*

[7] *The failure to notice the gender of a person reflects on the accuracy of the observations about the accuracy of a manuscript signature, as noted in Stephen Mason, Electronic Signatures in Law, (2nd edition, Tottel, 2007), 1.2, footnote 1.*

Tavernes Blanques (Valencia) where she made purchases to a value of 1,399 euro and 860 euro, paying for these purchases with the credit card of another person. To do so, she presented the Swiss National Identity Card of the legitimate holder of the credit card, but which bore the photograph of the accused. Subsequently, the accused signed the sales receipts imitating the signature of the legitimate cardholder. This decision followed and endorsed the comments made in earlier judgments made by the same judicial organ. However, if the card reader finally fails to authorize the attempted payment once the card had been swiped and in such a way that the perpetrator was finally unable to sign the sales ticket, it would amount to an attempt to falsify a commercial document (STS 25-6-98 nº 882/98). In the 1998 judgment, (STS 882/1998 25 June), the following ruling was made where the accused entered a jeweller's shop in Barcelona, and expressed an interest in buying a watch. To pay for it, the accused handed over a Master Card to the sales assistant in the name of a United States citizen, together with the legitimate passport of the US citizen. The accused had replaced the photograph of the passport holder with his own photograph. When the sales assistant requested authorization from the bank, it was refused via the POS (Point-of-Sale) terminal. The accused did not, therefore, place a false signature on the sales ticket. In view of the above, the accused handed over a second card (Visa), which had been cloned, which enabled him to pay for the watch. With respect to the first card, the Supreme Court considered it as an attempted misrepresentation of a commercial document.

As regards aiding and abetting misrepresentation, the TS has made it clear that where more than one person practices deceit in a commercial outlet by purchasing goods or services with another person's card, it does not matter which of those accused actually signs the sales slips; they are all guilty of misrepresentation, because the offence does not solely consist of having signed the sales receipt. Thus, the guilty parties are all those who benefit from the proceeds of the crime where there is a joint decision to commit the crime (STS de 26-5-2002 nº 661/02).

The bank card and remote payments

As the technical possibility of making remote payments with cards became more widely used without the need for the physical presence of the card holder, certain problems have arisen that have affected the law. Electronic procedures, especially over the internet, have facilitated remote commercial transactions that are normally settled with payment made by means of the electronic transfer of the data stored on a card.

The prevailing jurisprudence provides that the deceit at the heart of criminal deception is necessarily of a personal nature. It may only arise as the result of a direct relation between two people. Likewise, the error must also be a consequence of the deceitful act being of a psychological nature, which is only possible where there is close personal proximity.[8] Due to these assumptions, classic or conventional estafa (fraud) is, in such circumstances, impossible. Thus, when the new Penal Code was approved in 1995, the legislator included a different set of circumstances for computer-aided criminal deception (article 248.2). Given the personal nature of deceit and error under Spanish law, no references were made to them in article 248.2. In their place, it was provided that there must be a prerequisite of manipulating computer data. The subject must achieve the unauthorized disposal of an asset through the manipulation of computer data. Property assets are thus construed as objects, the manipulation of which will affect their value in such a way as eventually to cause loss to the property of a third party. The transfer implies that accountable assets pass initially to the property of the offender, and that the effect is to cause actual loss.

The offence of electronic fraud describes the circumstances relating to fraudulent payments made over the internet, in which the offender uses a cloned card or the information obtained from a legitimate card to obtain goods or services using the card details of another person, thereby causing an innocent person to be charged for the payment. The broad concept of computer manipulation basically corresponds to that proposed by Romeo,[9] in the sense of a wrongful modification of the result of an automated process at any of the stages of computer processing or programming with the aim of personal benefit and causing loss to a third party.

An alternative to this broad concept of computer manipulation has arisen with the Judgement of Malaga Criminal Court nº 3 of 19th December, 2005.[10] The court excluded the input of inappropriate data into the

[8] *A detailed presentation may be found in M. L. Gutierrez Francés, Fraude informático y estafa (Ministry of Justice 1991), 336 and following, and Ricardo M. Mata y Martín Los delitos de estafa convencional, esstafa informática y robo en el ámbito de los medios electrónicos de pago, 57. Jurisprudentially, STS nº 533/2007 of 12 June Id Cendoj: 28079120012007100455: and STS 369/2007 of 9 of May Id Cendoj: 28079120012007100374.*

[9] *C.M. Romeo Casabona, Poder informático y seguridad jurídica, (Madrid 1987), 47.*

[10] *ARP 2006/43.*

information system as an element of electronic fraud. The case refers to acts in which the defendants:

> ... puestos previamente de común acuerdo en fecha 28 de noviembre del 2000 a través de la página www.tododvd.com de la empresa Red Fénix Sistemas, SL realizaron el pedido de un reproductor de DVD marca Pionner modelo 530/535 con precio de venta 438 ? a nombre de Luis Pedro, ... y realizando el pago con la tarjeta VISA núm. NUM006, de la que era titular un tercero ajeno a los hechos, quien no había autorizado a los acusados a utilizarla.
>
> ... having previously come to a common accord on 28th November 2000, they placed an order in the name of Luis Pedro... on the Red Fénix website www.tododvd.com for a Pioneer brand DVD player model 530/535 at a sale price of €438 and made payment for it with a VISA card number NUM006, which belonged to a third party unconnected with the facts, who had not authorised its use by the accused.

The court only considered the subject-matter of this form of fraud in terms of the actions affecting the existing data (alteration, modification, deletion) in the system, and not the fact that the data provided, although correct, was not provided with the authority or agreement of the actual person whose data was used:

> Por ello no cabe incluir la conducta de los acusados en el párrafo segundo del art. 248 del Código Penal pues los mismos no manipularon sistema o programa informático alguno sino cuando se les solicita el número de una tarjeta bancaria para cargar en la cuenta asociada a la misma el importe de la compra efectuada designan el número de una tarjeta de la que no es titular ninguno de los acusados y es en la creencia de que todos los datos introducidos en la página web al hacer el pedido del reproductor de DVD son correctos por lo que la empresa Red Fénix SL, procede a hacer la entrega de dicho aparato en el domicilio indicado al hacer el pedido.
>
> It is for this reason that the conduct of the defendants is not to be included in the second paragraph of art. 248 of the Penal Code, as the latter did not manipulate the system or the computer programme in any way, but when they were asked for the number of a bank card against which to charge the said amount to the associated bank account, they inputted the number of a card that was not held in any of their names. It was in the belief that all the correct data was inputted into the web page when the order for the DVD player was placed that led the firm Red Fénix SL to proceed with the dispatch of the said device to the address they specified when the order was placed.

Were such a distinction to be upheld, all such conduct involving the introduction of misappropriated data to make purchases over the internet would be excluded from the category of computer fraud, offences which even today are being punished under that criminal category, as applied by the Supreme Court. Thus, STS of 20.11.2001 points out that computer manipulation:

> bien puede consistir en la alteración de los elementos físicos, de aquellos que permiten su programación o por la introducción de datos falsos
>
> may either consist in the alteration of physical elements, or of those that allow it to be programmed or by inputting false data.

To date, the jurisprudence has only dealt with circumstances referring to the use of credit cards and bank passwords, although in the case of on-line banking, there are no decisions, or at least none that the authors have found, that refer to payment by mobile telephone and what is known as electronic money. Thus, for example, the Judgment of the Provincial Court of the Balearic Islands num. 30/2005 (Section 2º) of 14 of April 2005, convicted a person for computer or electronic fraud that used personal passwords without the authorization of the account holder to make multiple transfers using the Línea Oberta de la Caixa website to accounts held by the banks of Banesto and La Caixa.

A peculiarity arises in this field, with regard to electronic fraud in association with a commercial outlet. It is a question of the circumstances under which the offender in various ways manages to persuade the owner or employee of an outlet to facilitate an irregular payment. Normally, the offer involves a half share of the benefits obtained from the sales in exchange for collaboration. This circumstance is dealt with, for example, in STS num. 2175/ 2001 of 20 November 2001. Specifically, it refers to an employee of a firm who was responsible for sending out credit cards to their owners and who appropriated a card and proceeded to a sales outlet, where, according to the testimony of the sales

assistants, he used it to make purchases, which were subsequently charged to the card holder's account. The TS upholds the similar nature of the offence described in article 248-2 as:

> quien aparenta ser titular de una tarjeta de crédito (...) y actúa en connivencia con quien introduce los datos en una máquina posibilitando que ésta actúe mecánicamente está empleando un artificio para aparecer como su titular ante el terminal bancario a quien suministra los datos requeridos para la obtención de fondos de forma no consentida por el perjudicado.
>
> whosoever appears to be the holder of a card (...) and acts in collusion with whoever inputs the data into the machine, thereby making it possible for it to work automatically is using an artifice so as to register as the owner of the bank card at the bank terminal by inputting the owner's data to obtain funds without the consent of the party incurring the loss.[11]

At other times, that the offender creates a fictitious commercial entity by requesting a Point-of-Sale (POS) terminal with which to commit the fraud. Thus, in the trial leading to the Judgment of the Provincial Court of Valencia of 2-11-1999[12] (num. 4/1999), various individuals by mutual accord considered installing a POS terminal for a fictitious business, and by making use of the terminal and credit cards stolen from their owners (which they possessed in great number), made fictitious transactions, thus obtaining the money from the transactions. The Provincial court appreciated the existence of a continuing offence of electronic fraud and the continuing offence of the falsification of commercial documents.

The use of credit cards in ATMs by thieves

There is yet another area in which the fraudulent use of bank cards takes place: at ATMs owned by banks that enable people to us the facilities offered at any hour of the day. These systems have also prompted the illicit use of bank cards, usually to obtain quantities of cash from cash dispensers. The emergence of such new attacks[13] lacked specific provisions in relation to the offence set out in the Penal Code. However, the response from the judges was to analyse the offence in relation to the physical layout of the ATMs. To begin with, the cash dispensers were placed in an enclosed space that required the same bank card to gain entry. This led the courts to define the offence as burglary using false keys.[14]

With the approval of the Penal Code of 1995, the legislator understood that the solution proposed by the courts made it possible to consider a card as a false key, and amended the definition of false keys with the inclusion of a final paragraph which sought to establish a comparison between a magnetic stripe and false keys:

> A los efectos del presente artículo, se consideran llaves las tarjetas, magnéticas o perforadas y los mandos o instrumentos de apertura a distancia.
>
> For the purposes of the present article, both cards, whether magnetic or perforated, and remote control opening devices or instruments are considered keys (article 239 in fine).

This treatment constitutes consolidated case-law;[15] although the reservations expressed in legal doctrine are not, it appears, altogether dismissed by the amendment to the legislation. In these cases, the application of the specific provision in the final paragraph of article 239 has normally led to charges of 'robo con fuerza (burglary)' in criminal proceedings, without entering into some of the more debateable points that might complicate an appraisal of the actual offence of burglary or housebreaking.[16]

On this point, it is worth pointing out that the provision clearly states that it is to be considered 'for

[11] *See also STS of 26 June 2006 (RJ 2006, 4925). At the Provincial Court level, similar criminal behaviour to those set out here may be appreciated, for example, in the SAP of Granada of 10/11/2006 (ROJ: SAP GR 2008/2006), and in the SAP of Alicante of 27 November 2007 (ROJ: A 2931/2007).*

[12] *ARP 1999/4239, consulted on the Aranzadi Westlaw Database.*

[13] *This class of criminal offence appeared in the second half of the 1980s as a consequence of the proliferation of automatic cash dispensers by banking entities. Enrique Bacigalupo Zapater, 'Utilización Abusiva de Cajeros automáticos por terceros no autorizados' in Poder Judicial, Número Especial IX: Nuevas formas de delincuencia, 85 and following; A.M. Javato Martín, 'Análisis de la Jurisprudencia Penal en Materia de Medios Electrónicos de Pago', in Los medios electrónicos de pago. Problemas jurídicos (Ricardo Manuel Mata y Martín and Antonio María Javato Martín), Comares, Granada, 2007, 375.*

[14] *The crime of burglary (articles 237 and following of the CP) consists in the misappropriation of goods using methods assessed as housebreaking or breaking and entering, which includes the use of false keys.*

[15] *On all these, STS of 22 January 2004, nº35/2004 (Supreme Court Sentence 22nd of January, 2004), EDJ 2004/8295 that rectifies the criteria of the Provincial Court of Madrid that in judgements that led to convictions for theft and not burglary due to it not having taken into account that the cash dispenser from which the money was withdrawn was situated in a booth which would have been opened, or that it would have been necessary to open a door or gate with the magnetic stripe.*

[16] *For further detail on this problem, see Ricardo M. Mata y Martín, Los delitos de estafa convencional, estafa informática y robo en el ámbito de los medios electrónicos de pago. El uso fraudulento de tarjetas y otros instrumentos de pago. (Aranzadi 2007), 142 and following.*

the purposes of the present article', that is, in the case of burglary using false keys, which means it is necessary to provide all of their general features, specifically that force must be used 'para acceder al lugar donde éstas se encuentran (to gain entry to the place where these are found)' (article 237). The provision in question refers to devices or instruments for remote 'opening', which brings us back to the specific context of burglary, which could also be applied to this category of offence. Furthermore, the keys in the Spanish Penal Code are considered false when they are used to open a lock in the normal way in order to allow entry into an enclosed space. However, the cards used in cash dispensers do not have to have previously facilitated access to an enclosed space, and in addition, they involve other aspects that go beyond the definition in the Code of a false key as being merely an opening device.

In reality, the very nature of this type of offence relates to a fraudulent act, to which the conventional offence of fraud as defined under Spanish law does not apply, because of the absence of personal deceit that is required by legal doctrine and judicial precedent. In addition, beyond any possible use as an opening device, the purpose of the card is, naturally, to be used in the exercise of a right to credit or to withdraw funds through the financial entity – based on the pre-existing legal contract between the card holder and the issuing entity – which by using the PIN, the issuer obtains sufficient evidence to assure itself that the legitimate holder of the card wishes to initiate a transaction in the ATM.

Judgment of the TS of 9 May 2007

Showing some sensitivity to the reasoning set out above, the recent judgment of the TS of 9 May 2007 moves away from what is accepted as consolidated jurisprudence and considers it possible to include the improper use of bank cards at automatic cash dispensers within the definition of computer fraud as defined in article 248.2.[17]

The factual circumstances to which the judgment refers concerned a group that was dedicated to copying credit cards and to making fraudulent use of them for the purposes of personal profit. To do so, they used a procedure known as 'skimming' consisting of the substitution of a magnetic band on an original or new credit or debit card for data on an existing one which they surreptitiously obtained by means of card readers. Having created forged cards, they used these in commercial establishments – presenting forged documents to identify themselves as the owners of the cards, and to withdraw money from the bank account of the customers whose data they had stolen.

They also used the procedure known as 'la siembra (sowing)', which consists in obtaining the victim's PIN number and credit card by placing somebody at a suitable distance from a card holder at a cash dispenser to observe the PIN, and to distract him in such a way that he loses sight of the card when it is returned by the machine, by which time it is removed and replaced by another card. The victim does not notice the switched bank card until he uses it to carry out further operations. The bank cards obtained in this way are used to withdraw money from cash dispensers; when these became invalid because they have reached the preset maximum withdrawal limit, they are used as a resource to manufacture other cards with which to carry out further operations.[18]

A number of criticisms have been made by legal commentators on the inclusion of these circumstances within the offence of burglary (absence of access into an enclosed space, lack of consent to hand over the goods), yet the judgment upholds the definition of the facts as elements of computer fraud. The Spanish High Court have established that the offender, by inputting the PIN or secret number of the stolen card into an ATM, is dishonestly identifying himself to the bank as the rightful owner of the card, thereby prompting the bank to transfer an amount of money voluntarily. Such an identification

> ... ha de ser considerada bajo la conducta de manipulación informática a que se refiere el tipo de la estafa del art. 248.2 CP
>
> ... has to be considered as behaviour that amounts to manipulation of computer data to which the category of fraud defined under art. 248.2 Penal Code refers.

This interpretation is supported, in the words of the Spanish High Court, by the Council Framework Decision

[17] *Note that is the sole judgment of the TS pronounced in this direction. Formerly, some decisions by the High Court (STS 185/2006) pronounced in favour of the solution of computer fraud although in a hypothetical or merely dialectical manner, as the category of crime in article 248.2 of the Penal Code has not been the subject of the accusation.*

[18] *For more descriptions of similar attacks, together with additional case law from across the world, see Stephen Mason, editor, Electronic Evidence: Disclosure, Discovery & Admissibility (LexisNexis Butterworths, 2007), 4.04-4.15.*

The courts have begun to move away from the criteria they initially upheld, and begun to punish such conduct as counterfeiting of legal tender.

of 28 May 2001 combating fraud and counterfeiting of non-cash means of payment,[19] because article 3, relating to offences and computers, covers the following:

> Each Member State shall take the necessary measures to ensure that the following conduct is a criminal offence when committed intentionally:
>
> performing or causing a transfer of money or monetary value and thereby causing an unauthorised loss of property for another person, with the intention of procuring an unauthorised economic benefit for the person committing the offence or for a third party, by:
>
> - without right introducing, altering, deleting or suppressing computer data, in particular identification data, or
>
> - without right interfering with the functioning of a computer programme or system.

The defining characteristics of data manipulation as provided for in article 3 includes identification by means of a secret number or PIN.

Counterfeiting and the alteration of cards

The treatment of counterfeiting and the alteration of cards in case law has varied over time. In the Penal Code of 1973, and in the absence of specific regulation, bank cards were accorded the status of a mercantile document. As a consequence, the creation of a cloned card by forgery and the manipulation of legitimate cards were subsumed under the articles dedicated to this type of counterfeiting, as determined by the Supreme Court in its judgment of 3 December 1991.[20]

Greater difficulties were involved in the assessment, alteration or manipulation of the magnetic stripe on the card, as the element that it incorporates is difficult to equate with the concept of a document. The problem was corrected in the Penal Code of 1995, which provides an extensive and broad concept of a document under article 26 that now covers magnetic stripes on cards. However, the specific consideration of credit and debit cards as money in article 387 of the New Penal Code will raise questions over such an approach to the problem in case law. The courts have begun to move away from the criteria they initially upheld, and begun to punish such conduct as counterfeiting of legal tender. An especially controversial point is the alteration of the magnetic stripe, as the Penal Code of 1995 decriminalised the conduct previously defined as alteration of legal tender, such that the card that has been manipulated can only be compared in a rather laboured way to the category of offence in article 386-1, which is the manufacture of money, understood as the creation of new money by counterfeiting legal tender.

The Supreme Court has put an end to debate on the question through the Acuerdo del Pleno no Jurisdiccional de la Sala Segunda (Agreement of the Non-Jurisdictional Full Court Session of the Second Chamber) issued on 28-6-2002. It opted to subsume alterations to the magnetic stripes of an authentic card under the offence of counterfeiting, putting forward the following argument:

> ... las tarjetas de crédito o débito son medios de pago que tienen la consideración de 'dinero de plastic', que el artículo 387 del Código penal equipara a la moneda, por lo que la incorporación a la 'banda magnética' de uno de estos instrumentos de pago, de unos datos obtenidos fraudulentamente, constituye un proceso de fabricación o elaboración que debe ser incardinado en el art. 386 del Código penal.

[19] *2001/413/JHA: Council Framework Decision of 28 May 2001 combating fraud and counterfeiting of non-cash means of payment OJ L149, 2.6.2001, p. 1–4.*

[20] *Emilio Manuel Fernández García and Juana López Moreno, 'La utilización...', in Cuadernos de Derecho Judicial VI-2002, 81.*

... credit and debit cards are means of payment that are considered 'plastic money', which article 387 of the Penal Code equates with money, such that the incorporation of data obtained in a fraudulent manner on the 'magnetic stripe' of one of these instruments of payment constitutes a process of production or preparation that should included under art 386 of the Penal Code.

This view was later be confirmed by a Judgement of the Supreme Court num. 948/2002 (Criminal Chamber) of 8th July,[21] in which the Supreme Court proceeded to differentiate between the behaviour in question and computer fraud. The reform of Organic Law 15/2003 has endorsed the criteria of the Spanish Supreme Court, by reintroducing the alteration of legal tender as a form of criminal offence in article 386 of the Penal Code. Likewise, the judicial interpretation of the concept of money was also extended to 'las demás tarjetas que puedan utilizarse como medios de pago (the other cards that may be used as means of payment)' (cash card and such like).

The classification of falsifying electronic bank cards as a crime of counterfeiting legal tender is open to criticism from two points of view. First, from the point of view of punishment, a very harsh sentence in the case of counterfeiting legal tender (a prison term of between 8 to 12 years, article 386 Penal Code) would be applied to circumstances that are much less serious, such as forgery of an isolated card or the mere possession of a forged card to use it as an instrument of payment. Second, from the point of view of authorization by virtue of article 65. b. LOPJ, the competency to judge these facts falls on the Audiencia Nacional (National Court) with a specialized jurisdiction covering terrorism and organized crime, which appears to be questionable.[22]

Hence, the Supreme Court has subsequently modified its general doctrine in the AATS of 18 February 2004[23] and 21 April 2004,[24] insofar as it identifies two types of circumstances:

a. The forgery of the card (alteration or manipulation of its magnetic stripe) and possessing forged credit cards for making purchases or distribution constitute counterfeiting of money and the competent court is therefore the Audiencia Nacional (High Court).

b. Mere possession of one or various forged cards for their use as an instrument of payment are subsumed under the offence of falsification of mercantile documents and thus do not amount to the counterfeiting of money that comes under the jurisdiction of the Audiencia Nacional.

The European Community perspective

There is an interest in a more effective assurance of security for the means of payment that may clearly be seen in the international context, especially respecting electronic payments. In this respect there are two areas of action of great importance for criminal regulation, the Cybercrime Convention of 2001 and various actions of the European Union.

The Convention on Cybercrime, drawn up in Budapest in 2001, deals with the complex problem of computer crime in the international context. Among its proposed measures, it includes the harmonization of punishable acts linked to computing that should be the subject of criminal offences in the signatory countries. The Convention establishes various groups of infractions that should be incorporated into national legislation and which it classifies into four broad categories of illicit offences. Among these, in a second group of behaviours, the Convention refers to computer crimes, which include computer-related forgery and computer-related fraud. Computer fraud (article 8) refers to the input, alteration, deletion, or suppression of computer data or any interference with the functioning of a computer system, with a view to procuring an economic benefit for oneself or for another person.

Furthermore, especially from European Union institutions, the importance of payment systems has been highlighted, which have a bearing on the criminal legislation of the Member States. In reality, the perspectives of the European Union are not strictly penal, but aim to guarantee and to stimulate economic activity, consumer protection and, to some extent, to prevent and deal with organized crime. However, through certain community measures that have an effect on domestic criminal legislation, it proposes the criminalization of certain conduct and other measures with the aim of protecting this means of payment. Thus,

[21] *Nº 948/2002 in Actualidad Penal. Nº45. 2 al 8 de diciembre de 2002, p. 3141 and following.*

[22] *Carolina Villacampa Estiarte, ' La falsificación de medios de pago distintos del efectivo en el Proyecto de Ley Orgánica de Reforma del CP de 2007: ¿respetamos las demandas armonizadoras de la Unión Europea?' , in Diario La Ley, nº 6994, 22 of July, 2008, 3 and following.*

[23] *Id Cendoj: 28079120012004200356.*

[24] *Id Cendoj: 28079120012004200586. See also on this point ATS 10-3-2004, Id Cendoj: 28079120012004200420; of 1-4-2004, Id Cendoj: 28079120012004200545, of 7-12-2004 Id Cendoj: 28079120012004202326.*

the European Union has not ceased to show concern and to adopt measures to prevent illicit acts with what it refers to as 'non-cash means of payment'.[25]

In view of the importance that is given to electronic commerce for the future economic development of the zone, various initiatives have been taken, each having a greater degree of definition and penetration. Thus, the Commission, on 16th April 1997, in a Communication to the European Parliament, the Council, the Economic and Social Committee and the Committee of the Regions 'A European Initiative in Electronic Commerce' COM(97)157, proposed that certain actions be defined and set in motion aiming to maximise the advantages of the new technology involved in electronic commerce. The Council also invited Member States to set up awareness-raising campaigns and training on practical improvements, and to create transparent consultation mechanisms with the aim of drawing up the legal framework and the specific actions for the promotion of this type of commerce. Finally, it called on European regulatory bodies to draw up more efficient working methods with a view to ensuring interoperability and to respond to consumer needs.

Subsequently the Communication from the Commission to the European Parliament, to the Council, to the Central European Bank and to the Economic and Social Committee on 'A Framework for action on combating fraud and counterfeiting of non-cash means of payment' was approved. The Communication approved by the Commission, on 1st July (COM (1998)395) is in response to the proposal from the Council of Europe on June 1997, in which the Commission examined the question of fraud and counterfeiting in relation to all non-cash means of payment, including electronic payments, which means it will encompass facts relating to conventional criminal activity and facts relating to the use of the new technologies.[26] The Communication proposes a two-pronged plan in the strategy to prevent and deal with fraud.

The first point is a Joint Plan of Action directed, on the one hand, at ensuring that frauds referring to all non-cash means of payment are categorised as criminal offences and made punishable through effective, proportionate and dissuasive sentences in all Member States and, on the other hand, at setting up appropriate mechanisms for cooperation that will enable the effective prosecution of the crimes. To that effect, classes of behaviour are described which are considered advisable to classify as criminal offences, whatever the means of payment might be. The following in particular are included among the offences listed: theft or the forgery of a means of payment, the possession of altered or counterfeited means of payment, the use or acceptance of a payment in full knowledge of the facts with the aid of a forged or stolen means of payment. The second point of the action plan against fraud presents various preventive measures to be studied by all interested parties (payment card schemes, issuers, card users, and competent authorities). Thus, from the standpoint of prevention, it is thought that one of the Communication's objectives is to urge operators to adopt more effective protection measures for the payment instruments that they manufacture.

The concern of Community institutions for the success of the information society, as a prerequisite for growth, competitiveness and employment opportunities, is expressed in the Communication from the Commission on Creating a Safer Information Society by Improving the Security of Information Infrastructures and Combating Computer-related Crime-COM(2000) 890 FINAL. The Communication considers different initiatives with respect to a wide range of objectives that comprise part of the Information Society, which aim to improve information infrastructures as a way of preventing and dealing with computer crime. In reference to the Lisbon summit of March 2000, it underlines the importance of a transition to a competitive, dynamic, knowledge-based economy as well as to the centrality of information infrastructures in present-day economic life, on which society increasingly depends, while noting, at the same time, that these technologies may be used to commit and to facilitate criminal activities. Security measures must focus on adapting to these new forms of criminality.

This makes the ever-greater proximity of computer crime to the new categories of organized crime very clear. Associated criminality increasingly involves a greater number of offences among which computer fraud is increasingly apparent. The community institutions stated as much at the Tampere Summit, in October 1999, at which high-Tech crime was included in a list of areas in which a special effort had to be made to agree on definitions, types of offences and common sanctions. All these points are contained in recommendation 7 of the strategy of the European

[25] *On this matter, see Lafuente Sánchez, R. Los servicios financieros bancarios electrónicos, (Tirant lo Blanch 2005), 337, and Francesco Buffa, "Moneta digitale e tutele". Commercio elettronico e tutela del consumatore a cura di Giuseppe Cassano (Giuffré 2003), 178 and following.*

[26] *Lafuente Sánchez, R., Los servicios financieros bancarios electrónicos (Tirant lo Blanch 2005), 337.*

Union for the new millennium on prevention and control of organized crime, adopted by the JHA Council in March 2000.

An important impetus was given with the approval by the Commission of a Framework Decision in matters concerning non-cash means of payments. Indeed, the Framework Decision, together with other instruments of the European Union, will be greatly heeded in the reform of the Penal Code proposed in the Draft Law of 2006. In accordance with the provisions of article 34 of the Treaty on European Union (the former article K.6), the proposal was to replace the joint action proposed by the Commission in its Communication of 1 July 1998, on combating fraud and counterfeiting of non-cash means of payment with a Framework Decision. Equally, the Proposal for a Framework Decision also had as its aim the inclusion of legislative changes that have been enacted since the approval of the Communication. Thus the Council Framework Decision of 28 May 2001, relating to the fight against fraud and the counterfeiting of non-cash means of payment, is intended to complete a series of measures already adopted by the Council with the same aim. For the purposes of the Framework Decision, means of payment are considered to be all corporeal instruments except for legal tender, the specific nature of which is to allow, by itself or with another instrument, the holder or user to transfer money or a monetary value, and which is protected against counterfeiting or fraudulent use. This description precludes not only money in cash (banknotes and legal tender) but also electronic money in its strictest sense that has no material presence.

The objective of the Framework Decision continues to be that of ensuring, on the one hand, that all fraud with non-cash means of payment becomes an offence subject to effective penalties in all Member States and, on the other hand, that mechanisms are created for cooperation between Member States and between services and public or private bodies with the objective of successfully prosecuting such offences. In the Framework Decision, any fraud involving a non-cash means of payment is considered a criminal offence punishable by effective, proportionate and dissuasive sentences throughout the Member States of the Union.

With respect to the criminal conduct, the approach of the Framework Decision is to avoid resorting to categorical definitions already strictly defined in the criminal law of the Member States, because it varies by country. Thus, the Framework Decision limits itself to drawing up a list of different intentional behaviours that should be considered criminal offences throughout the Union. Different behaviours are defined according to whether they are primarily concerned with the actual instrument of payment or the counterfeiting of instruments of payment, and whether it is a question of one or more payments or of the clearing system used to execute, collect, process, or settle payment transactions. Thus, it includes:[27]

a. theft or misappropriation of an instrument of payment,

b. the alteration or counterfeiting of an instrument of payment with a view to its fraudulent use,

c. receiving, obtaining or transporting, sale or transfer to another person or possession of instruments of payment that have been misappropriated or altered or counterfeited for fraudulent use, and

d. fraudulent use of a means of payment that has been stolen, misappropriated, altered or counterfeited.

Offences that will also be subject to prosecution are those using computers to make or cause a transfer of money or monetary values that lead to unauthorized loss of property, with the intention of procuring economic benefit through the unauthorized inputting, alteration, suppression or deletion of computer data – especially personal data, or unauthorized interference in the operation of a computer system or programme. Other criminal offences include the manufacture, receipt or transfer of computer programs and other devices prepared for the commission of the former offences.

With respect to the nature of the penalties to be adopted in this field, it is envisaged that the list of conducts be categorised as criminal offences throughout the Member States. As a consequence, Member States should establish criminal penalties for these offences, according to whether they are committed by natural or by legal persons. The expression that is so well liked in EU documents reiterates that the penalties must be effective, proportionate and dissuasive. They will not necessarily imply prison terms, except for the most serious cases for which extradition can be justified. The Member States enjoy a certain leeway when defining the seriousness of an offence and the nature and severity of

[27] *Francesco Buffa, 'Moneta digitale e tutele' in Commercio elettronico e tutela del consumatore (Editor Giuseppe Cassano) (Giuffré 2003), 179-80.*

the applicable penalties.[28]

Finally, as the work of the Commission on these means of payment has continued, a further Communication was issued from the Commission.[29] The Commission considers that cooperation between all the agencies involved is a fundamental principle in order to prevent and deal with fraud in an effective manner. In fact, greater cooperation is desirable between public authorities and the private sector in the Member States. With the aim of ensuring an effective exchange of information at a European level, the Commission stated that clarification of community and national legislation in the field of data protection is needed in the area of fraud prevention.

The draft reform of the Penal Code

The economic significance of payment systems and the high volume of fraud drew the attention of the legislator to this field, and particularly the attention of the criminal legislator. Hence, the Draft Law to reform the Penal Code of 15 December, 2006,[30] is intended to amend criminal regulation of these matters. It sought to add a specific element to the field of frauds under article 248: the use of bank cards or related data. The draft law fell into abeyance as the legislative term came to an end, but parts of it may be found in the programme of criminal measures for the present legislature, and in any case, it points to the way in which possible criminal reforms may be introduced in this field.

In a general way, the Explanatory Memorandum of the draft law goes a long way to justifying its proposals on the basis of the commitments and obligations that European integration implies for the criminal justice system. Among the areas subject to community harmonisation is that of the means of payment, which lies behind the new regulation. The Explanatory Memorandum of the project points out that:

> La causa central que explica su acotado alcance ha de ser buscada fundamentalmente en los compromisos y obligaciones que la integración europea suponen para la justicia penal en toda su dimensión penal, procesal, judicial y policial. La importante vertiente del derecho penal ha venido recogiendo al paso de su aparición cuantas orientaciones comunes, plasmadas en los diferentes instrumentos jurídicos de la Unión Europea, determinaban modificaciones u adiciones al Código penal, y eso explica buena parte de las alteraciones del Código. Pero además, en los últimos años, especialmente a partir del Tratado de Ámsterdam en 1997, el llamado Tercer Pilar fortaleció la importancia de hacer efectiva la cooperación policial y judicial en materia penal, lo cual exigía necesariamente la armonización o aproximación de las leyes estatales en materia penal, y por esa razón se han ido produciendo Decisiones marco sobre un amplio catálogo de problemas penales, Decisiones que empujan a una necesaria similitud de las formulaciones de delitos y responsabilidades en los derechos internos.

The central reason that explains its highly defined scope is to be found in the commitments and obligations entailed by European integration for criminal justice in all of its dimensions, be they criminal, procedural, judicial or police related. This important vector of criminal law has brought together many common perspectives since its emergence, expressed in the different legal instruments of the European Union, which determined modifications and additions to the Penal Code, and these explain a good part of the amendments to the Code. But in addition, in recent years, especially since the Treaty of Amsterdam in 1997, the so-called Third Pillar strengthened the importance of ensuring effective police and judicial cooperation in criminal matters, which necessarily required the harmonisation or approximation of State laws on criminal matters, and for that reason Framework Decisions have been drafted in response to a wide range of criminal problems, Decisions that work towards a much-needed similarity in the formulation of offences and liabilities in domestic rights.

Committing fraud through the use of misappropriated

[28] *For an incomplete treatment of how some Member States have implemented the various EU Directives (that nevertheless runs to over 450 pages), see a paper by Stephen Mason, ' The implementation of Community regulations in national legislation: IT offences in the strict sense of the word and offences committed using IT' , prepared for a judicial seminar entitled: Investigation, Prosecution and Judgment of Information Technology Crime: Legal framework and criminal policy in the European Union, held for judges and public prosecutors specializing in dealing with cybercrime, organized within the framework of the European Judicial Training Network, (Tuesday 25 November 2008 to Friday 28 November 2008 at the Hôtel Jean de Bohême, Durbuy, Belgium), and available as a free download from http://www.stephenmason.eu/training-for-lawyers/judicial-training/.*

[29] *Communication from the Commission to the Council, the European Parliament, the European Economic and Social Committee, the European Central Bank and Europol - A new EU Action Plan 2004-2007 to prevent fraud on non-cash means of payment {SEC(2004) 1264} (Text with EEA relevance)/* COM/2004/0679 final */.*

[30] *http://www.congreso.es.*

cards or their corresponding data has been added to the reform of criminal legislation relating to the group of criminal offences that constitute fraud under article 248. Thus, the first paragraph of article 248 maintains the conventional definition of estafa (fraud by false representation), but in the second paragraph, a list of comparable circumstances is included, through the expression 'También se consideran reos de estafa' (also considered crimes of fraud), followed by four letters under which four other cases of fraud are specified. Electronic fraud is consigned to letter a), the second paragraph of the article in question, which previously constituted the sole circumstance envisaged under this specific legal reference. The regulation of electronic fraud does not vary, but its applicability does as a consequence of the newly incorporated criminal categories. Under letter c), the misuse of credit and debit cards or the data recorded thereon to carry out transactions of any kind causing loss to another person are specifically defined as criminal offences. One consequence of this proposed regulation, incorporated as a further individual category of fraud, that of making a fraudulent payment through the use of card – differing from the offence of computer fraud, is that the criminal regulation of fraudulent uses of means of payment are widened to an even greater extent.

The use of the identifying data on a card is expressly included under the punitive category, in a different way than the use of the card itself. This implies, on the one hand, the reinforcement of the thesis – now upheld by the courts – that fraudulent use of the data on a card that falsely attributes the payment of a transaction to the card holder is punishable as an offence of fraud. In addition, it supposes the absence of the actual possession of the card by the criminal, which therefore implies an on-line or remote payment, which in principle should be treated as computer fraud. However, the specific prevision of this new circumstance supersedes this category – on the basis of the speciality principle – for which reason this particular provision would be applied on the grounds of the fraudulent use of another person's card. By doing so, the offence of electronic fraud currently described under article 248.2 is, to a great extent, devoid of any practical content.

With reference to the act of forging instruments of payment, the project chose to separate them from the framework of counterfeiting legal tender, bringing them under the offence of forgery of documents. To that end, a new section was created, section 4º, in Chapter II of Title XVIII of Book II that has as its title 'The falsification of credit and debit cards and travellers cheques'. A new article, 3999 bis, is included in it, which contains three types of criminal behaviour:

a. forgery, either by copying or by reproduction, of credit or debit cards or travellers cheques;

b. possessing such forged items in an amount that suggests they are destined for distribution or trafficking; and

c. the fraudulent use of these forged instruments of payment on the part of whoever has not intervened in their forgery.

The considerable reduction of penalties should be highlighted among the positive aspects of the draft law[31] as well as the exclusion under all circumstances involving the forgery of bank cards of the procedural competency of the Audiencia Nacional or National Court.

Ricardo M. Mata y Martín is Vice-Deacon and Titular Professor of Criminal Law at Valladolid University (Spain), coordinator of the Research Group on Computer Criminality, Director of 6 Research Projects, 6 monographs, editor of 2 collective works, with over 30 publications to his name in journals and book chapters.

rimata@der.uva.es

Antonio Javato holds a law degree awarded by the University of Valladolid and gained a doctoral degree from the same university, where he works as a Temporary Professor of Criminal Law. He has published various articles, many of which are on computer crime, and has received FPI scholarships from the Spanish Ministry of Education and the D.A.AD.

javato@der.uva.es

[31] *The punishment is a prison term of between four to eight years for the first offence – as against eight to twelve years and a fine as contained at present in article 386, with the possibility of selecting the upper half of the range when the forged items affect a great many people or when the acts were committed as part of a criminal organization dedicated to these activities. The same punishment is imposed in the case of a repeat offence, whereas the use of forged credit or debit cards or travellers cheque is punishable by a prison term of between two to five years.*

ARTICLE:

BREAD AND DONKEY FOR BREAKFAST

HOW IT LAW FALSE FRIENDS CAN CONFOUND LAWMAKERS: AN ITALIAN TALE ABOUT DIGITAL SIGNATURES

By **Ugo Bechini**

False friends are a well known hazard. The same word can sometimes even be pronounced the same way in two different languages, but the meaning can be utterly different. For instance, Italian and Spanish are very similar languages, and the word 'burro' is pronounced the same way in both, but actually means 'butter' in Italian and 'donkey' in Spanish. An Italian tourist who is having breakfast in a Spanish hotel, the popular story goes, should not be surprised to be presented with some bread and a donkey, if he asks for bread and butter in his mother language.

False friends can also be a danger in the IT law world. The same words often have different meaning in IT law and in the general practice.

Consider the meaning of 'copy', for instance. A copy in the physical world is an object that can, generally, be recognised as such, something in itself different from the original. This is not at all true in the digital world. A file, whatever its content, is just a string of zeros and ones, or of letters and numbers, if you want. If you ask Alice for Bob's mobile telephone number, you will not expect Alice to answer that she cannot give it to you, because Bob kept his number for himself, and all she has is a accurate copy of that number. A copy of a number is the same number again. The same is true for digital files: they are numbers. The verb 'to copy' can be employed in order to describe the process that allows a computer user to replicate a file from the hard disk to her USB device, but the output of such a process is not a copy in any sense of the word: it is, in fact, a perfect duplicate. There is no way to tell for certain which is the 'original'.

This has significant legal implications. In some countries, such as Italy, there is no formal provision that prevents a person from executing digital cheques. The cheque is basically a text: any technician will tell you that it can be digitally signed very much like anything else. The lawyer will not find anything against it in the law, either. But still the answer is no: a cheque cannot be digitally signed. A digitally signed document is just a file, as any other file, and can be duplicated endless times. One cheque could be duplicated one hundred times and cashed in one hundred different banks, and nobody would be able to identify the original one. A digital signature is, therefore, an unsuitable tool whenever the legal properties of a document stem from its uniqueness.[1]

This is a field where neither the law nor IT can walk alone. A digital signature affixed to a cheque is technically feasible, and the law (at least in some countries) does not forbid it. What happens here is that *a legal feature of the cheque is incompatible with a technical feature of a digital signature*. The question is whether the proposition in italics belongs to IT or the law. The point is, the lawyer must understand the technology, because the of the interaction of technology and law, as Albert de Lapradelle, a professor of International Law, is understood to have written on the changes in the law of naval warfare for the Conference

[1] *The problem can be solved creating infrastructures that hold an authoritative copy of the document.*

on The Hague in 1907:

> Ce ne sont pas les philosophes avec leurs théories, ni les juristes avec leurs formules, mais les ingénieurs avec leurs inventions qui font le droit et le progrès du droit.
>
> It is not the philosophers and their theories, and lawyers with their formulas, but the engineers with their inventions which are the right and the progress of law.

The 'signature' is another dangerous false friend. Unless biometric technologies are in place (and the quality of the biometric technology may be the subject of a challenge), anybody who gains control of the token and the PIN can create signatures that are, in themselves, genuine digital signatures. A mansucript signature links a document to a person, while a digital signature does not: it links a document to a device. The missing link is provided by the law; it is the law (in some countries, and if some conditions are fulfilled) that determine whether the document is binding to a particular person. It is a virtual legal technique that holds somebody responsible for a statement even if it does not come, in any meaningful sense, from the same person. There is nothing inherently wrong in this. In most jurisdictions, for instance, companies are liable for the actions of their executives, even if they act against the resolutions of the board. This is a reasonable burden for business organisations, in the interest of providing for the speed of a transaction.

The burden would be deemed quite acceptable in the case of digital signatures, if adequate use policies were in place and duly followed in everyday life. This means that each user would have to retain both the signature token, and secure the PIN without recording it. In this perspective, such a practice could somehow fill the existing gap between IT (that cannot guarantee that the signature comes from a given person) and the law (that assumes so). This is not about theoretical legal concepts, but about their acceptability in the context of a well-functioning and consistent legal environment.

The Italian case is rather special. Millions of smart cards have been issued, and basically every owner of a business (including small rural shops) has one. They are used for tedious bureaucratic chores that can only be performed with digital signatures. It is not surprising that the owners of the signature tokens are not thrilled by the burden. Most of the smart cards, that are usually blue in colour, are retained in piles in accountants' offices, each of them with a small yellow Post-It note with the PIN written on it (perhaps look-conscious Italians would go for more subtle and fashionable nuances, if they considered the smart cards really important). If on-line rumours[2] are to be taken at face value, most of the people do not even know that a smart card exists in their name.

Nevertheless, lawmakers go on assuming that documents signed with such smart cards are tantamount to documents signed with a manuscript signature. This is what the law provides, and a new significant implementation was introduced in 2008 that pushed things even further with Decreto-legge 25 giugno 2008, n. 112 Disposizioni urgenti per lo sviluppo economico, la semplificazione, la competitività, la stabilizzazione della finanza pubblica e la perequazione Tributaria (Decree June 25 2008, n. 112),[3] approved with Legge 6 agosto 2008, n. 133, Conversione in legge, con modificazioni, del decreto-legge 25 giugno 2008, n. 112, recante disposizioni urgenti per lo sviluppo economico, la semplificazione, la competitività, la stabilizzazione della finanza pubblica e la perequazione tributaria[4] (Law August 6 2008, n. 133, article 36, paragraph 1bis). The text of the law is full of technicalities that require a deep knowledge of some obscure details of the Italian legal system. The relevant part of article 36, paragraph 1bis reads as follows:

> 1-bis. L'atto di trasferimento di cui al secondo comma dell'articolo 2470 del codice civile può essere sottoscritto con firma digitale, nel rispetto della normativa anche regolamentare concernente la sottoscrizione dei documenti informatici, ed e' depositato, entro trenta giorni, presso l'ufficio del registro delle imprese nella cui circoscrizione e' stabilita la sede sociale, a cura di un intermediario abilitato ai sensi dell'articolo 31, comma 2-quater, della legge 24 novembre 2000, n. 340.
>
> 1-bis. The transfer deed mentioned by Article 2470 of the Civil Code can be signed with a digital signature, in accordance with the rules about the signature of electronic documents, and filed within thirty days, at the office Registration Court in whose area is established the head office of the company, through an authorized agent according to the provision of Article 31, paragraph 2-c, of the Law of 24 November

[2] *http://punto-informatico.it/423980/PI/Lettere/chi-smart-card-ai-commercialisti.aspx; http://www.interlex.it/docdigit/faq/faq42.htm - http://www.interlex.it/docdigit/nonlosa.htm.*

[3] *Pubblicato nella Gazzetta Ufficiale n. 147 del 25 giugno 2008 - Suppl. Ordinario n.152/L.*

[4] *Pubblicata nella Gazzetta Ufficiale n. 195 del 21 agosto 2008 - Suppl. Ordinario n. 196.*

The possibilities are almost endless: the employee you just dismissed signs; the employee that your accountant just dismissed signs too, and dead people might also sign.

2000, n. 340.

Briefly put: since 1993,[5] every sale of a share in an Italian limited liability company (srl, società a responsabilità limitata) must be notarized. This requirement can appear to be far too formal, but it was part of an attempt to prevent the mafia and other criminal organisations buying into legitimate businesses. It is difficult to deny that such a strategic goal justifies much more than a few annoying bureaucratic steps. Moreover, the problem, as will be demonstrated later in this article, lies not the security level in itself, but in the equivalence (or, better, lack thereof) between two different procedures, both of them requiring the use of digital signatures.

In the traditional procedure, still in use, people must sign a deed before a Civil Law Notary,[6] usually drafted by the CLN himself; it is the notary's duty to prepare a digitally signed copy of the deed and send it to the Registro delle Imprese.[7] The data are introduced automatically into the register, as they are already presented in XML format and do not require any kind of manual editing. In the new procedure, the deed is digitally signed by the parties themselves, and sent to an accountant, who forwards it to the Registro delle Imprese. The data processing is the same, but there is a significant difference: in the new procedure, nobody can be certain who really signs the deeds. The possibilities are almost endless: the employee you just dismissed signs; the employee that your accountant just dismissed signs too, and dead people might also sign.

In a country where Civil Law Notaries operate, there is an additional set of differences between a notarized document and a document that has not been notarized. The Civil Law Notary is a publicly appointed official who usually drafts the document, and is responsible to ascertain that the parties fully understood the document. Without a CLN, people may sign files they never read. People might sign files they did not understand. People may sign poorly drafted files. There is a lack of proper and impartial legal information. This is exactly what Mr Giuseppe Limitone in the Vicenza Court considered in Ordinanza del Giudice del Registro, April 21st 2009, n. 6/09,[8] in which he refused the registration of a transfer that had been executed in accordance with the new procedure. The details of the case are not available in the decision. However, it is certain that an application was made to delete the registration of the share transfer because it was not notarized. It appears the action was initiated by the seller. This seems to be the case, because the judge is on record as responding to an argument presented by

[5] *Legge 12 agosto 1993, n. 310: Norme per la trasparenza nella cessione di partecipazioni e nella composizione della base sociale delle società di capitali, nonchè nella cessione di esercizi commerciali e nei trasferimenti di proprietà dei suoli. (Pubblicata nella G.U. n. 195 del 20 agosto 1993) (Law 12 August 1993 number 310).*

[6] *Civil Law Notaries (CLN) are to be found in countries that adopt the Latin Notarial system: about 90 counties that sum up about 55 per cent of the world's population. The Civil Law Notary is a lawyer that has already (albeit not always) been admitted to the bar; he or she is, at the same time, an officer of the state and a professional. The foremost task of the CNL is not the mere identification of the parties. He is also responsible, and liable, for an array of different issues related to the contract. For instance, in real estate transactions, if the seller was not the legitimate owner of the estate, the CLN will be required to refund the buyer. The same will occur if he fails to properly take care of the mortgages. The CLN must ensure that the results of the agreement are in accordance with the provisions of every applicable law, and explain to the parties the value, legal effects and consequences of the agreement. In most countries, he is also required to collect taxes, and the CLN is personally responsible for paying the taxes if the job is not properly done. In some jurisdictions, a CLN is even liable upon failing to inform the parties about an available tax deduction. If a house does not match the building and zoning regulations, liability can sometimes arise. If a sum of money comes from a source that cannot be clearly identified, state agencies in charge of money laundering investigations are informed. These tasks are performed not only in the interest of the parties, but in the general public interest, as it keeps litigation at comparatively incredibly low levels in all the areas covered by the work of the CLN.*

[7] *The Italian Companies' House; it records a wide array of events during the life of a company, including share transfers, that are not legally effective until registered.*

[8] *The full text appeared in Le Società (Milan), 6/2009, p. 738, with an assenting comment by Vincenzo Salafia, former President of the Corte d' Appello of Milan, the most authoritative Italian court in company law. Every Italian Court (Tribunale) has a judge called the Giudice del Registro, who is in charge of the Registro delle Imprese. If any dispute arises about a registration, the Giudice del Registro decides; the decision can be overturned by a full Tribunale.*

the resistant (only the company is allowed to make an application to have a registration deleted) by stating that anybody who as an interest in the matter can take action, and this would be enough. Nevertheless, he goes on to make it absolutely clear that the 'preteso cedente' (the purported seller) can take action.

The court began by pointing out that, if the legislature intended to make share transfers that are only digitally signed by the parties fit for registration, they fell short of their target. It was the view of the court that the new law, seen in the context of the Italian legal system, was a failure. The traditional procedure provides a check of the lawfulness of the contract and verification of the actual (not virtual) identity of the real signer, and this is vital in order to preserve the reliability of the register. The new procedure does not prescribe any of the safety features that have been in place for some time, but at the same time it does not explicitly state that they are not required: therefore the court held that the general rules apply, which means that no data can be entered in the register without the controlling mechanisms. In other words, as the old and the new procedure live side by side, it cannot be imagined that the law may want to leave people free to choose to be scrutinized or not.

The Vicenza court resolved the matter in a straightforward manner. The new law does not mention notarization, but this is a general requirement for any document presented for registration. This means that the new procedure requires the document to be notarized. The court determined that the only possible applicaiton of the law would be the following:

1. the parties digitally sign the deed;
2. the digital signatures are executed before a notary;
3. the notary certifies the digital signature;
4. the document is sent to the accountant's office;
5. the notary relays it to the Registro delle Imprese.

The first, fourth and fifth steps are provided by the new law and are retained; however, the court added steps two and three as requirements to enable a share transfer to be registered – the digital signature must be executed before a Civil Law Notary and officially certified.[9]

It is not known at the time of writing if this interpretation will be widely accepted by Italian courts, or whether Parliament will modify the legislation in the light of the decision by the Vicenza Court. The framework in which this case arose may be unique to the Italian legal system, but the underlying message is not. A digital signature cannot always be considered as equal to a manuscript signature, especially a notarized one. Whether the passing of Decreto-legge 25 giugno 2008, n. 112 indicated a deliberate change in the legal philosophy of the Italian state, or whether this was a mistake, it was big enough to make a judge sitting in the Vicenza Court of a small (albeit historical) Italian city stand up, and present the overwhelming majority of his country's Parliament with a breath of reality: that if a digital signature is to have any legal effect, it is necessary to demonstrate as false the proposition asserted by technicians that the private key of a digital signature, when used, proves it has been used the person whose key it is. This presumption can only carry any weight in law if a notary attests to the fact that the private key was used by the person whose key it was. If Parliament decides to change the law and overturn this decision, it will, in effect, be overturning the laws that were enacted to prevent criminal organizations from buying into legitimate business.

Ugo Bechini is a Civil Law Notary in Genoa, Italy and Chairman of the New Technologies Working Group, Conference of the Notariats of the European Union, Brussels.

ugo@bechini.net

[9] *Digital signatures have been notarized in Italy since 1997: Decreto del Presidente della Repubblica 10 novembre 1997, n. 513 Regolamento contenente i criteri e le modalità per la formazione, l' archiviazione e la trasmissione di documenti con strumenti informatici e telematici a norma dell' articolo 15, comma 2, della legge 15 marzo 1997, n. 59 (G. U. 13 marzo 1998, serie generale, n. 60) (Presidential Decree 10 November 1997 number 513, article 16).*

ARTICLE:

THE ESSENTIAL ELEMENTS OF AN EFFECTIVE ELECTRONIC SIGNATURE PROCESS

By **Greg Casamento** and **Patrick Hatfield**[1]

Companies conducting business throughout the United States wanting to implement an electronic signature process (for customers, employees or suppliers) are provided little guidance from the electronic signature statutory schemes across the country. Those electronic signature laws, essentially two bodies of statutory law, provide that electronic signatures and electronic records may not be denied legal effect *solely* because they are in electronic form.[2] These laws do not give greater status to signatures or records in electronic form. Further, and more significantly, these laws do not describe any processes which, if followed, would result in enforceable contracts.

This article seeks to help those wanting to design and implement an effective electronic signature process by describing six perspectives from which a proposed electronic signature process should be evaluated. This six-point framework takes into account the legal and practical aspects beyond just the electronic signature laws, such as the rules of evidence. Examining the risks of an electronic signature process from these six perspectives allows one to match the mitigation measures for each risk with the level of risk acceptable for a given electronic signature process. For example, most will agree that an electronic signature process to buy a low priced book need not be as secure as an electronic signature process to buy an expensive item or for authorizing the disclosure of very sensitive information. This six-point framework helps to distinguish each risk to focus more clearly on the optimal way to mitigate each distinct risk.

The framework will help to answer the three fundamental questions that should be addressed for any proposed electronic signature process. This article includes an in-depth discussion of the essential elements that should be included in an electronic contracting process that will result in admissible evidence to enforce terms and conditions in records signed electronically in the United States of America, whether the electronic signature process is governed by the Federal electronic signature law or a particular state's enactment of the model electronic signature law.[3]

Critical risks and questions

Framework for evaluating the risks

Companies often approach their legal advisors for guidance on what steps an effective electronic signature process[4] should include. In seeking this guidance, companies have described a range of risks they associate with an electronic signature process. There are essentially five distinct risks for an electronic signature process, each of which should be examined relative to those same risks in dealing with paper and

[1] *The authors would also like to thank the editor for his valuable input in editing this article as well as Vita Zeltser, an associate of Locke Lord Bissell & Liddell LLP, for her assistance.*

[2] *The two bodies of laws are the federal act, the Electronic Signatures In Global and National Commerce Act, 15 U.S.C §70001 and following (referred to as ESIGN) and the various state enactments of the version of the Uniform Electronic Transactions Act, as published by the National Conference of Commissioners on Uniform State Laws (referred to as UETA). Forty seven states and the District of Columbia have enacted some version of UETA.*

[3] *There are significant differences between ESIGN the federal electronic signature law and the version of UETA adopted by forty-seven states and the District of Columbia. Except as expressly identified in this article, the differences are not significant for the topics described in this article. For example, UETA addresses when an electronic record is deemed to be sent by the sender and received by the addressee. The federal ESIGN law is silent on the topic.*

[4] *Throughout this article references to an 'electronic signature process' should be read to include the required disclosures of required terms, the delivery of the executed documents to the other party as well as the archival process for these records.*

manuscript signatures. These five risks and the benefit of examining each risk in context in this fashion comprise the Six-Point Framework identified below and discussed in more detail further below:

1. Authentication Risk – This is the risk that the signer[5] signing a record, accepting delivery of a record or providing a record is an imposter using a false identity; the records then being unenforceable by the user[6] against the person the user thought it was dealing with via electronic means.

2. Repudiation Risk – This is the risk that the signer claims that the electronic records that were signed were altered after they were signed, such that the person against whom enforcement is sought attempts to repudiate the actual terms and conditions in the signed electronic record.

3. Admissibility Risk – This is the risk that the other party to a transaction successfully challenges the admissibility of the necessary records, such as the signed contract, acknowledgment of receipt of certain disclosures, on the grounds of reliability.

4. Compliance Risk – This is the risk that the records signed or presented do not comply with other substantive laws, such as laws mandating certain content in documents to be presented or signed or do not comply with the basic requirements of ESIGN and UETA for delivery for such records.

5. Adoption Risk – This is the risk that in managing the risks above, an electronic signature process is so burdensome that the intended users are not satisfied with the process or find ways to avoid certain steps in the process, thereby undermining the process.

6. Relative Risk - In examining the risks above, users should evaluate the risk with a proposed electronic signature process *relative* to the corresponding risk in the process using paper and a manuscript signature, in the belief that an electronic signature process may not be risk free, but should not, on the whole, be any riskier than the paper and manuscript signature process, if feasible.

For the reasons explained below, it is possible to design an electronic signature process which is no riskier than, and in some areas, significantly less risky than, using paper and a manuscript signature. By examining the risks from these perspectives, it is easier to assess the particular risk and then determine the optimal means to mitigate the risk.

Critical questions

In reviewing a proposed electronic signature process, the following three fundamental questions should be considered:

a. Will the transactions executed using the proposed electronic signature process *be in compliance* with the applicable laws governing the use of electronic signatures and delivery of related electronic records, including the required consumer disclosures and consents, if any?

b. Will the records presented, signed, secured, archived and retrieved using the proposed electronic signature process be *admissible* in court (or arbitration) to enforce the terms and conditions in such records?

c. Will the terms and conditions in electronic documents signed using the proposed electronic signature process *be enforceable* against each signing party?

Subject to a subtle but important caveat, if each of the three questions above cannot be answered affirmatively, the electronic signature process should be re-examined, and appropriate changes made to the process. The transactions conducted through electronic means should be as compliant, generate records as admissible and result in terms as enforceable, as would be the case if those same records were completed on paper with manuscript signatures. In other words, aside from all the other applicable contract principles, such as capacity, fraud, duress, mistake, unconscionability, the records signed using the electronic signature process should be as enforceable as would be the case for those same records signed using a manuscript signature on paper.

[5] *The term 'signer' refers to the person, often a consumer, signing the electronic record, whether the record is a contract, application, consent, authorization or acknowledgement of receipt of terms.*

[6] *The term 'user' refers to the person, often a company, that has established the electronic signature process for enforceable and compliant transactions.*

Excluded areas

The focus of this article is on transactions between private parties, which include consumers. Excluded from the scope of this article are the following areas:

a. transactions dealing with the specific areas of the law expressly excluded from the federal ESIGN law and the state enactments of UETA, as described immediately below;

b. transactions dealing with any governmental agency where that agency is acting as a market participant;

c. execution of documents required by any governmental agencies which are not related to transactions between private parties, even if those documents are permitted to be filed with such an agency exclusively through electronic means, such as documents required to be filed with or maintained for inspection by the SEC or FDA;[7] or

d. records subject to any other law which specifies a particular method of verification or acknowledgment of receipt, such as requiring delivery by registered mail, return receipt required.

ESIGN and UETA do not apply to contracts and records that are governed by laws and regulations in only a few select areas.[8] Given the preemption provisions in ESIGN, the federal law, the states have limited authority to expand the scope of the areas excluded from ESIGN or the state enactment of UETA.[9]

Notwithstanding the foregoing exceptions, ESIGN (and UETA), when applied with other laws such as Revised article 9 of the Uniform Commercial Code, provide a mechanism for the use of electronic signatures and records in many of the most common business and consumer transactions, including contracts and records involving: (i) sales and leases of goods; (ii) insurance applications; (iii) mortgage loan documentation; and (iv) banking and investment transactions. ESIGN, the federal law, specifically applies to the business of insurance.[10] Given the similarity between ESIGN and UETA, insurance companies and other firms regulated under the state insurance codes, may adopt a uniform, national electronic signature process.

Legal analysis

ESIGN and UETA compared

For all purposes relevant to the analysis in this article, except as noted otherwise, the analysis under ESIGN (the federal statute), the relevant enacted version of UETA in 47 states and even under the non-UETA states, is essentially the same.[11] For those states that have adopted electronic signature laws governing interstate commerce inconsistent with ESIGN in areas relevant to the issues discussed in this article, ESIGN's broad preemption provisions will preempt such state laws.[12] For those states that have not adopted any electronic signature laws, ESIGN will govern as a result of its broad preemption provisions.[13]

The legal effect of electronic signatures

ESIGN recognizes that an electronic signature may be as legally effective as a signature applied on paper with a manuscript signature. ESIGN does not give electronic signatures a special status in the law. Rather, ESIGN states that a signature may not be denied legal effect *solely* because it is in electronic form. The foundational provision of ESIGN acknowledging electronic signatures provides, at § 101(a), the following:

(a) In General.--Notwithstanding any statute, regulation, or other rule of law (other than this title and title II), with respect to any transaction in or affecting interstate or foreign commerce-

(1) a signature, contract, or other record relating

[7] *This is not to say that the concepts described in this article do not apply to transactions with governmental agencies. Rather, this caveat is simply to alert the reader that certain governmental agencies may take the position that documents not related to transactions between private parties may not be within the scope of ESIGN and UETA.*

[8] *Excluded areas are: wills, codicils, and testamentary trusts; a state statute, regulation, or other rule of law governing adoption, divorce, or other matters of family law; the Uniform Commercial Code, as in effect in any state, other than sections 1-107 and 1-206 and Articles 2 and 2A; court orders or notices, or official court documents required to be executed in connection with court proceedings; notices for cancellation or termination of utility services (including water, heat, and power); notices of default, acceleration, repossession, foreclosure, or eviction, or the right to cure, under a credit agreement secured by, or a rental agreement for, a primary residence of an individual; notices of cancellation or termination of health insurance or benefits or life insurance benefits; recall notices of a product, or material failure of a product that risks endangering health or safety; and any document required to accompany any transportation or handling of hazardous materials, pesticides, or other toxic and dangerous materials. ESIGN § 7003 (a), (b), UETA § 3.*

[9] *ESIGN § 102(a).*

[10] *ESIGN § 101(i).*

[11] *The states that have not adopted UETA are Illinois (adopted Electronic Commerce Security Act), New York (adopted Electronic Signatures and Records Act), and Washington (adopted Electronic Authentication Act).*

[12] *ESIGN § 102(a).*

[13] *ESIGN § 102(a).*

Another form of electronic signature is to say or select 'yes' over the telephone to accept terms and conditions contained in a writing acknowledged by the person so signing.

to such transaction may not be denied legal effect, validity, or enforceability solely because it is in electronic form; and

(2) a contract relating to such transaction may not be denied legal effect, validity, or enforceability solely because an electronic signature or electronic record was used in its formation.

Thus, assuming ESIGN or UETA[14] applies to the transaction, each gives equal recognition to electronic signatures as given to manuscript signatures on paper.

Electronic signature defined

When a signature is created using a 'sound, symbol or process' that is 'attached to or logically associated with' a contract or other record by a signer with intent, such signature will be legally effective. For clarity, phrases such as 'legally effective' are used, rather than the statutory language of ESIGN, which states, 'not be denied legal effect solely because such signature is an electronic signature.' ESIGN § 106(5) defines an 'electronic signature' as:

> Electronic signature.--The term "electronic signature" means an electronic sound, symbol, or process, attached to or logically associated with a contract or other record and executed or adopted by a person with the intent to sign the record.

ESIGN § 106(4) defines 'electronic record' as:

> Electronic record.--The term "electronic record" means a contract or other record created, generated, sent, communicated, received, or stored by electronic means.

ESIGN § 106(9) defines 'record' as:

> Record.--The term "record" means information that is inscribed on a tangible medium or that is stored in an electronic or other medium and is retrievable in perceivable form.

Thus, an electronic signature may consist of an electronic sound or symbol, such as an individual saying 'I agree,' or typing 'I agree' or the person's name or following some other process, such as clicking 'I agree,' which is attached to or logically associated with information inscribed: (i) on a tangible medium, such as the tangible, hard copy of an authorization; or (ii) stored in an electronic medium retrievable in a perceivable form, such as the electronic record containing the identical information as contained in the tangible hard copy delivered to the consumer. Another form of electronic signature is to say or select 'yes' over the telephone to accept terms and conditions contained in a writing acknowledged by the person so signing.[15]

Users may select from a variety of ways to generate the signer's actual signature. The electronic signature process should clearly inform the signer that using such an electronic sound, symbol, or process is how the signer expresses his or her consent to sign such documents thereby evidencing his or her intent to be bound to such terms and conditions.

Evidence of the signer's intent to sign the record (which is required if the signer signs on paper with a manuscript signature) may be inferred (as it is with a manuscript signature on paper) from words close to the place of the signature where such words indicate in

[14] *ESIGN § 101(a); UETA § 7(a).*

[15] *See for example, Shroyer v. New Cingular Wireless Serv., Inc., 498 F.3d 976 (9th Cir. 2007) where an electronic signature process was recognized whereby terms and conditions contained in a printed booklet in a box in the consumer's possession for a consumer product are accepted by the consumer selecting 'yes' over the telephone. While the court recognized the electronic signature process, the terms of the contract were not enforced for reasons having nothing to do with the electronic signature process.*

clear and conspicuous terms the signer's intent to sign and be bound. For example, the text in the ESIGN Consent[16] could include the following text to explain the legal significance of the signer using the electronic signature process to create his or her electronic signature:

> By [describe method used to consent, e.g., selecting 'I AGREE'], you confirm that you have the computer hardware and software to obtain access to electronic records in the form that important disclosures will be provided to you in connection with [describe transactions], and you consent to receiving consumer disclosures related to [describe transaction] exclusively through electronic means.

Accordingly, pursuant to ESIGN and UETA, an electronic signature process where the significance of the process is clear may not be denied legal effect solely because it is in electronic form. Similarly, a document relating to such a transaction may not be denied legal effect, validity or enforceability solely because an electronic signature was used to sign such a document and subsequently stored as an electronic record, rather than in hard copy.

Consumer disclosures

On this topic of providing consumer disclosures exclusively by electronic means, there is a significant difference between the Federal ESIGN Act and UETA as enacted by many of the states. Both bodies of law (the federal ESIGN and the state enactments of UETA) permit consumer disclosures which are required by some other law to be given exclusively through electronic means, but the Federal ESIGN Act, and some, but not all, states which have enacted UETA, specify the process and the content for obtaining the consumer's consent to receive certain consumer disclosures exclusively through electronic means.[17]

ESIGN provides that, upon consent by the consumer, certain information relating to a transaction or transactions in or affecting interstate or foreign commerce, which is required by a statute, regulation, or rule of law (other than ESIGN) to be provided or made available to a consumer in writing (referred to as a Special Consumer Disclosure) may be delivered exclusively via electronic means, provided that the recipient of the Special Consumer Disclosure is first provided, and agrees to, the ESIGN Consent.[18] Whether a particular transaction requires a Special Consumer Disclosure, and how the ESIGN Consent is provided in connection with the required Special Consumer Disclosure, must be determined on a transaction-by-transaction basis. The user should identify which documents are Special Consumer Disclosures that require the need for the ESIGN Consent in the context of each type of transaction to be completed using the electronic signature process. The ESIGN provisions describing a Special Consumer Disclosure and the contents of the ESIGN Consent are set out in § 7001(c):

> If a statute, regulation, or other rule of law requires that information relating to a transaction or transactions in or affecting interstate or foreign commerce be provided or made available to a consumer *in writing*, the use of an electronic record to provide or make available (whichever is required) such information satisfies the requirement that such information be in writing if--
>
> (A) the consumer has affirmatively consented to such use and has not withdrawn such consent;
>
> (B) the consumer, prior to consenting, is provided with a clear and conspicuous statement--
>
> (i) informing the consumer of (I) any right or option of the consumer to have the record provided or made available on paper or in nonelectronic form, and (II) the right of the consumer to withdraw the consent to have the record provided or made available in an electronic form and of any conditions, consequences (which may include termination of the parties' relationship), or fees in the event of such withdrawal;
>
> (ii) informing the consumer of whether the consent applies (I) only to the particular

[16] *'ESIGN Consent' refers to the disclosure required by ESIGN to be provided to a signer who is a 'consumer' as defined by ESIGN, to which that signer must consent as a condition to the user providing one or more consumer disclosures required by law (referred to as a Special Consumer Disclosure) to that signer exclusively via electronic means, where such consent is given in a way that demonstrates the signer's ability to reasonably obtain access to information in electronic form the Special Consumer Disclosures will be provided.*

[17] *UETA §8(a) and ESIGN §7001(c). Providing Special Consumer Disclosures exclusively through electronic means is slightly complicated by the fact that a few states have included in their enactment of UETA provisions similar to those in the Federal ESIGN Act, as well as the fact that a federal law may require many Special Consumer Disclosures. For this reason, users are well advised to comply with the Federal ESIGN Act ESIGN Consent provisions discussed further below. See for example, Ala. Code 1975, § 8-1A-8(e).*

[18] *ESIGN §7001(c).*

transaction which gave rise to the obligation to provide the record, or (II) to identified categories of records that may be provided or made available during the course of the parties' relationship;

(iii) describing the procedures the consumer must use to withdraw consent as provided in clause (i) and to update information needed to contact the consumer electronically; and

(iv) informing the consumer (I) how, after the consent, the consumer may, upon request, obtain a paper copy of an electronic record, and (II) whether any fee will be charged for such copy;

(C) the consumer--

(i) prior to consenting, is provided with a statement of the hardware and software requirements for access to and retention of the electronic records; and

(ii) consents electronically, or confirms his or her consent electronically, in a manner that reasonably demonstrates that the consumer can access information in the electronic form that will be used to provide the information that is the subject of the consent; and

(D) after the consent of a consumer in accordance with subparagraph (A), if a change in the hardware or software requirements needed to access or retain electronic records creates a material risk that the consumer will not be able to access or retain a subsequent electronic record that was the subject of the consent, the person providing the electronic record--

(i) provides the consumer with a statement of (I) the revised hardware and software requirements for access to and retention of the electronic records, and (II) the right to withdraw consent without the imposition of any fees for such withdrawal and without the imposition of any condition or consequence that was not disclosed under subparagraph (B)(i); and

(ii) again complies with subparagraph (C).

The signer's affirmative consent to the ESIGN Consent must exhibit the signer's ability to obtain access to information in the manner that the Special Consumer Disclosures will be provided. For example, if the required disclosure (a truth in lending disclosure for example) will be posted at a secure web site accessible only after the signer is given a unique access code, the signer should be given that unique access code during the ESIGN Consent process to confirm that the unique access code in fact allowed the signer to obtain access to the secure site where the Special Consumer Disclosures, such as the truth in lending statement, will be posted.

If the signer consents to receive such disclosures electronically but does not reasonably demonstrate his or her ability to obtain access to the information in the manner the Special Consumer Disclosures are provided, then the Special Consumer Disclosures are likely to be ineffective and therefore the basis for providing the required disclosures exclusively by electronic means could fail. Failure to comply with the ESIGN consumer disclosure requirements does not, however, render void or voidable the underlying transaction. ESIGN § 101(c)(3) provides:

> Effect of failure to obtain electronic consent or confirmation of consent.--The legal effectiveness, validity, or enforceability of any contract executed by a consumer shall not be denied solely because of the failure to obtain electronic consent or confirmation of consent by that consumer in accordance with paragraph (1)(C)(ii).

Failure to comply with the ESIGN consumer disclosure requirements could, however, subject the user to regulatory sanctions for failing to provide the required disclosures (such as the truth in lending notice in the example above) in accordance with applicable law. There may also be civil remedies available to signers if the disclosures are deemed to have not been given effectively. Not all notices or documents that users are required to provide to signers are Special Consumer Disclosures subject to the ESIGN disclosure requirements above. For such notices and documents which are *not* such Special Consumer Disclosures, the signer only needs to agree to receive such notices and documents exclusively via electronic means.

The user must first determine whether, for a given transaction, there are any Special Consumer Disclosures and, where there are, the electronic signature process:

(1) must present the appropriate ESIGN Consent to the signer, (2) should record that the signer consented to receive Special Consumer Disclosures exclusively through electronic means in a way that reasonably demonstrates the ability of the signer to obtain access to information in the electronic format the actual Special Consumer Disclosures will be provided or made available to the signer, and (3) for the Special Consumer Disclosures, provide or make available to the signer such disclosures in that same format. Taking these actions would allow the user to provide Special Consumer Disclosures in accordance with the requirements of ESIGN.

Use of an electronic process to complete transactions requiring Special Consumer Disclosures, or other documents containing mandated terms such as pre-approved forms, can actually reduce the user's compliance risk, compared to the conventional approach of paper and manuscript signatures. An automated electronic signature process allows the user to specify each document which must be presented and signed, as an acknowledgment of receipt or otherwise, as a condition to completing the transaction. Further, for an automated electronic signature process, the user can specify each particular which field in a record, such as an application for insurance, which must be completed as a condition to completing the entire transaction (as well as the nature of the information completed in such field, such as state of residence in a state where the user's products are not available). Thus, the user may configure the electronic signature process to prevent incomplete or non-compliant transactions from being submitted to the user for review. This can significantly improve the user's ability to comply with the requirements for such regulated transactions, and reduce risk while at the same time improve the rate of successfully completed transactions.

Verifications and acknowledgements

Verifications and acknowledgments required by law are expressly permitted to be delivered in electronic form under ESIGN in certain circumstances. ESIGN § 101(c)(2)(B) provides:

> Verification or acknowledgment.--If a law that was enacted prior to this Act expressly requires a record to be provided or made available by a specified method that requires verification or acknowledgment of receipt, the record may be provided or made available electronically only if the method used provides verification or acknowledgment of receipt (whichever is required).

Thus, if a law requires a disclosure to be provided by a certain method, which requires acknowledgment of receipt, such as delivery by first class mail, with proof of delivery required, such verification or acknowledgment may be given electronically if, and only if, the electronic method for providing such verification or acknowledgment also provides verification or acknowledgment of receipt. For example, the electronic signature process should be configured so that the consumer, before reviewing the verification or acknowledgement, must confirm receipt.

Record retention – sufficiency of electronic records

There are two record retention issues addressed by ESIGN. The first relates to the requirement that, where a statute requires a contract or other document to be in writing, the electronic record may be denied legal effect if all the parties or persons cannot reproduce it for reference entitled to the contract. The relevant section of ESIGN § 101(e) provides:

> Accuracy and Ability To Retain Contracts and Other Records.-- Notwithstanding subsection (a), if a statute, regulation, or other rule of law requires that a contract or other record relating to a transaction in or affecting interstate or foreign commerce be in writing, the legal effect, validity, or enforceability of an electronic record of such contract or other record may be denied if such electronic record is not in a form that is capable of being retained and accurately reproduced for later reference by all parties or persons who are entitled to retain the contract or other record.

Thus, if a user is going to rely exclusively on the archived electronic record to satisfy the statutory requirement that a contract or other document be in writing, failure to maintain the record in a form capable of being retrieved by all parties for later reference, could jeopardize the enforceability of the transaction to which such record relates. Users may satisfy this requirement by making the electronic record available to the signer for the required period of time, or the user may send a copy of the document or documents, in hard copy or electronically, so the user is not relying on the signer's

ability to obtain access to the electronic record maintained by the user.

In contrast, the second record retention issue relates to the user satisfying statutory record retention obligations. The user may electronically store the record (whether that record was initially in tangible form and later converted to an electronic form or initially in electronic form) of a transaction and thereby satisfy the statutory record retention requirement, provided certain conditions are met. ESIGN, § 101(d) provides:

> Retention of Contracts and Records.--
>
> (1) Accuracy and accessibility.--If a statute, regulation, or other rule of law requires that a contract or other record relating to a transaction in or affecting interstate or foreign commerce be retained, that requirement is met by retaining an electronic record of the information in the contract or other record that--
>
> (A) accurately reflects the information set forth in the contract or other record; and
>
> (B) remains accessible to all persons who are entitled to access by statute, regulation, or rule of law, for the period required by such statute, regulation, or rule of law, in a form that is capable of being accurately reproduced for later reference, whether by transmission, printing, or otherwise.
>
> (2) Exception.--A requirement to retain a contract or other record in accordance with paragraph (1) does not apply to any information whose sole purpose is to enable the contract or other record to be sent, communicated, or received.
>
> (3) Originals.--If a statute, regulation, or other rule of law requires a contract or other record relating to a transaction in or affecting interstate or foreign commerce to be provided, available, or retained in its original form, or provides consequences if the contract or other record is not provided, available, or retained in its original form, that statute, regulation, or rule of law is satisfied by an electronic record that complies with paragraph (1).

ESIGN permits a user to satisfy its record retention obligations relating to transactions by retaining documents exclusively through electronic means. These ESIGN record retention requirements do not affect the user's record retention practices, except for those records relating to transactions to be retained exclusively through electronic means. Thus, if the user is satisfying the record retention obligations imposed on it by other laws by storing hard copies, ESIGN will not impose additional obligations.

As noted above, ESIGN does permit the user to satisfy its record retention obligations under applicable laws by retaining only the electronic records if the requirements of Section 101(e) of ESIGN are met. Thus, if documents in the audit trail,[19] which are required by law to be retained, are retained exclusively in electronic media, and are available to the regulators having jurisdiction over the user and such electronic records are available as described in Section 101(e) of ESIGN, the user may not be required to print and retain hard copies of these documents.

Notarizations

Signatures to be notarized may be notarized using an electronic notary process, providing that all other requirements of the notary laws are satisfied. ESIGN § 101(g) provides:

> Notarization and Acknowledgment.--If a statute, regulation, or other rule of law requires a signature or record relating to a transaction in or affecting interstate or foreign commerce to be notarized, acknowledged, verified, or made under oath, that requirement is satisfied if the electronic signature of the person authorized to perform those acts, together with all other information required to be included by other applicable statute, regulation, or rule of law, is attached to or logically associated with the signature or record.

As stated further above, with limited exceptions, signatures will not be denied legal effect solely because they are electronic. Thus, if a law requires a signature to be notarized, either or both the signature to be notarized and the signature of the notary may be electronic signatures. All the other requirements for notarizing signatures (such as the notary must witness

[1] *'Audit trail' is a collective reference to the records containing the processes and details involved in each significant step of a given transaction involving a user including, the process of each signer accessing, completing, executing and transmitting each document to be signed in connection with the transaction, the user's process for authenticating each signer of each document for that transaction and all documents executed or resulting from the process, all as cryptographically sealed.*

Having signatures notarized is another form of authentication of the identity of the signer.

the person sign the document) must be met.

The official commentary to relevant provision in UETA (which is consistent with the notary provision in ESIGN) explains more about satisfying the notary requirement:

> This section permits a notary public and other authorized officers to act electronically, effectively removing the stamp/seal requirements. However, the section does not eliminate any of the other requirements of notarial laws, and consistent with the entire thrust of this Act, simply allows the signing and information to be accomplished in an electronic medium.
>
> For example, Buyer wishes to send a notarized Real Estate Purchase Agreement to Seller via e-mail. The notary must appear in the room with the Buyer, satisfy him/herself as to the identity of the Buyer, and swear to that identification. All that activity must be reflected as part of the electronic Purchase Agreement and the notary's electronic signature must appear as a part of the electronic real estate purchase contract.[20]

While ESIGN and UETA permit the notary requirements to be satisfied exclusively through electronic means, this does not require notaries to use electronic signatures or obligate private third parties requiring notarized signatures to accept the electronic signature of the notary.

Risk analysis framework and mitigation

Different categories of transactions present different risk profiles. For example, a transaction where a consumer authorizes the release of highly sensitive health or financial information to the person signing the release, presents a much greater risk of forgery than does a transaction for the purchase of a low-priced book. Likewise, a transaction for a consumer to sign an authorization to release sensitive health or financial information to an insurance company for underwriting purposes presents, as a practical matter, a lower forgery risk than if the sensitive information were to be released to the person signing, because the forger has less opportunity to benefit from the disclosure to the third party than from the disclosure directly to the forger, and therefore there is less incentive for a forger in the first instance. Because of these differences, when designing an electronic signature process, one should critically review the risks from various perspectives. The framework below identifies the six perspectives.

Authentication risk

This is the risk that a signer is in fact not the person he or she claims to be. A user may authenticate the identity of each signer in various ways. The identity of each person to sign should be verified. Such verification steps may include confirmation of the identity of such person from a trusted source, such as a single sign-on process deployed by, or otherwise determined to be reliable by the user. Alternatively, the results from an identity verification process conducted by an independent third party can be used for this purpose, such as a consumer reporting agency or other trusted third party offering such services. A further method can be used, such as the answer to a shared secret question that the user determines adequately verifies the identity of the signer. Having signatures notarized is another form of authentication of the identity of the signer. If there are documents required to be notarized, the electronic signature process should allow the notary verifying another signer's signature to enter the notary's signature and other credentials, in accordance with applicable state notary laws.

The method and results used to authenticate each

[20] *UETA § 11, Official Commentary.*

signer should be included in the archived signing session, or audit trail, which should then securely archived and capable of being retrieved securely. Where the user opts not to include the authentication process in the audit trail, the user may need to have access to other reliable evidence to establish the actual identity of the person completing the transaction.

As a practical matter, users should also critically evaluate the likelihood of forgers, or even signers who seek to disavow a given transaction claiming that a forger signed the documents. Consider, for example, the authentication risk in the context of applications for automobile insurance. The question that needs addressing is the likelihood of a consumer seeking to recover a payment for a covered claim contesting that he or she did not sign the application documents (which would include certain elections and waivers of coverage). To claim that a forger signed the documents would result in there being no cover, albeit for a different reason. Furthermore, it might also be useful to assess what motive a person have to forge the signature of another person for insurance cover on the car of the person whose signature is forged.[21]

At least one court has addressed this risk.[22] In Kerr, the employer sought to enforce a mandatory arbitration provision against an employee. The question was whether the employee did in fact sign the electronic record agreeing to be bound to the mandatory arbitration provisions. The court held that in light of the employee's credible claims that she did not sign the record containing the mandatory arbitration provisions combined with the employer's opportunity to sign such record using the employee's credentials, the mandatory arbitration provisions would not be enforced against the employee. Had the employee's supervisor not had such ready access to the employee's user name and password to obtain access to the secure site where the record in question was presented for signature, the court may have reached a different conclusion.

Repudiation risk

This is the risk of a signer acknowledging that he or she signed a document, but claims that the electronic signature is attached to or logically associated with a document containing terms and conditions different than those in the document the signer signed. The risk is that the signer repudiates the terms and conditions in the document attached to or logically associated with his or her signature and thereby reduces the chance that the document will be admissible and, if admitted into evidence, that the tier of fact will be persuaded that the signer did not agree to be bound by all such terms and conditions.

The electronic signature process should deploy readily available technology that can reduce the repudiation risk far below the repudiation risk associated with paper documents and manuscript signatures. The electronic signature process should cryptographically seal each document upon execution of that document by each signer, thereby rendering such document unalterable without detection. Documents electronically sealed in this fashion are likely to pass the admissibility threshold (for which, see the discussion below) and once such documents are admitted into evidence, users are likely to have meaningful, persuasive evidence as to why such document could not have been alerted without detection.

Each encrypted document should be securely stored in such a way that it cannot be viewed without overcoming at least industry standard security safeguards applicable to the document in question. For each transaction, whether the transaction involves two or more parties, the electronic signature process should record the date and time of each significant step and the identity of the person taking each such step and each particular step taken by that party, where such record is part of the audit trail. The audit trail for each transaction should include each document presented and signed during a given transaction where each such document signed having been encrypted as described above. Relevant parts of the audit trail should also be encrypted using industry standard encryption technology to render those portions of the audit trail unalterable without detection.[23]

Admissibility risk

This is the risk that a court refuses to admit into evidence copies of electronic documents generated, presented, signed, secured, archived and retrieved by

[21] *Admittedly, there is fraud in the automobile insurance sector, some of which involves forgery. Distinguishing the types of fraud and when fraud occurs in this area is essential to determine the mitigation measures with the actual risk presented in a given scenario.*

[22] *Kerr v. Dillard, 2009 U.S. Dist. LEXIS 11792 (D. Kansas 2009).*

[23] *The reader should be aware of the long-term viability of digital signatures when archiving digital documents protected by a digital signature, for which, see Stefanie Fischer-Dieskau and Daniel Wilke 'Electronically signed documents: legal requirements and measures for their long-term conservation', Digital Evidence and Electronic Signature Law Review, 3 (2006) 40 – 44.*

the electronic signature process. As a preliminarily point, it is important to recognize that all of the rules of evidence and evidentiary foundations that apply to paper documents and manuscript signatures also apply to electronic documents signed electronically, stored electronically and retrieved electronically. The Federal Rules of Evidence, or their state equivalents, govern the admissibility of evidence and thus would govern the admissibility of a copy of a document presented, signed, secured, archived and retrieved by the electronic signature process.[24] The electronic signature process should be able to satisfy the admissibility standards in the Federal Rules of Evidence to prove the authenticity of a document retrieved if the electronic signature process creates a reliable record of the entire signature process, including:

(a) the terms and conditions presented to the signer with which the electronic signature will be logically associated;

(b) the specific act of the signer expressing his or her intent to be bound to those terms and conditions, as called for in those same terms and conditions; and

(c) the circumstances under which signatures were obtained.

This information all goes to establish the authenticity of the document (containing the terms and conditions) retrieved by the electronic signature process. The electronic signature process should enable users to securely archive and retrieve the documents in a way to show that the documents containing the signatures could not have been altered without detection. The electronic signature process should also enable the appropriate witness on behalf of the user to provide an affidavit or live testimony as to items (a) – (c) above. For the reasons described below, such copies of documents generated by the electronic signature process based on documents presented, signed, secured, archived and retrieved by the electronic signature process should be as admissible under the Federal Rules of Evidence as such documents containing the same terms and conditions generated, presented, signed in hard copy and manuscript signature, where such paper copy is secured, archived and retrieved using conventional archival and retrieval methods.[25]

Federal Rules of Evidence

The standard for the authentication of evidence under the Federal Rules of Evidence is contained in Rule 901, Requirement of Authentication or Identification, which provides that 'the requirement of authentication or identification as a condition precedent to admissibility is satisfied by evidence sufficient to support a finding that the matter in question is what its proponent claims.'[26] As stated throughout the case law regarding the admissibility of computer generated information, '"reliability must be the watchword" in determining the admissibility of computer generated evidence.'[27] The 'factors [must] effectively address a witness' familiarity with the type of evidence and the method used to create it, and appropriately require that the witness be acquainted with the technology involved in the computer program used to generate the evidence.'[28]

Certain subparts of Sections 901 and 902 of the Federal Rules of Evidence are particularly suited to address the admission of electronic signatures and records: Sections 901(b)(1), (3), (4) and (9), and 902(7) and (11). Rules 901(b)(1), (3), (4) and (9) require witness testimony to authenticate proffered evidence, while 902(7) and (11) allow for self-authentication.[29]

F.R.E. 901

A witness with direct knowledge, pursuant to F.R.E.

[24] *Many states have adopted rules of evidence that track the Federal Rules of Evidence (FRE). For purposes of this discussion, all cases cited are based on the FRE or state law that follows the FRE.*

[25] *This would require the user to identify who, by name and title, is qualified to testify (in person or via an affidavit) as to how each document was presented, signed, secured after signature to render it unalterable without detection, archived, retrieved and printed. This person will also testify as to the integrity and security of each system involved in creating, securing, archiving, retrieving and printing the document.*

[26] *Lorraine v. Markel American Insurance Company, 241 F.R.D. 534, 541-42 (D.Md 2007).*

[27] *State v. Swinton, 268 Conn. 781, 812 (CT. 2004) (applying the federal standard to a state case).*

[28] *State v. Swinton at 813, 814.*

[29] *Magistrate Judge Paul W. Grimm's opinion in Lorraine v. Markel American Insurance Company provides ones of the best analysis to date of the admissibility of electronic evidence, which broadly could include electronic signatures, 241 F.R.D. at 542; Brian W. Esler, ' Lorraine v Markel: unnecessarily raising the standard for admissibility of electronic evidence, Digital Evidence and Electronic Signature Law Review' , 4 (2007) 80 - 82. See also In Re Vee Vinhnee, 336 B.R. 437 (proponent failed properly to authenticate exhibits of electronically stored business records); United States v. Jackson, 208 F.3d 633, 638 (7th Cir. 2000) (proponent failed to authenticate exhibits taken from an organization's website); St. Luke's Cataract and Laser Institute PA v. Sanderson, 2006 WL 1320242, at *3-4 (M.D. Fla. May 12, 2006) (excluding exhibits because affidavits used to authenticate exhibits showing content of web pages were factually inaccurate and affiants lacked personal knowledge of facts); Rambus v. Infineon Tech. A.G., 348 F. Supp. 2d 698 (E.D. Va. 2004) (proponent failed to authenticate computer generated business records); Wady v. Provident Life and Accident Ins. Co. of Am., 216 F. Supp. 2d 1060 (C.D. Cal. 2002) (sustaining an objection to affidavit of witness offered to authenticate exhibit that contained documents taken from defendant's website because affiant lacked personal knowledge); Indianapolis Minority Contractors Assoc. Inc. v. Wiley, 1998 WL 1988826, at *7 (S.D. Ind. May 13, 1998) (proponent of computer records failed to show that they were from a system capable of producing reliable and accurate results, and therefore, failed to authenticate them).'*

901(b)(1), or an expert witness with learned knowledge, pursuant to F.R.E. 901(b)(3), are certainly two fairly straightforward methods a user could use to admit hard copies of documents signed using the electronic signature process. F.R.E. 901(b)(4), which permits exhibits to be authenticated by appearance, contents, substance, internal patterns, or other distinctive characteristics 'is one of the most frequently used [rules] to authenticate [electronic signatures] and other electronic records.'[30] F.R.E. 901(b)(9), which authorizes authentication by '[e]vidence describing a process or system used to produce a result and showing that the process or system produces an accurate result', is 'one method of authentication that is particularly useful in authenticating electronic evidence stored in or generated by computers' and is frequently used as a litmus test for admissibility of computer-related information.[31] '[It] dictates that the inquiry into the basic foundational admissibility requires sufficient evidence to authenticate both the accuracy of the image and the reliability of the machine producing the image.'[32]

The electronic signature process should secure each document after it is signed, as discussed above relating to the risk of repudiation. This would also allow the user to meet the admissibility standards under the subsections in F.R.E. 901. The testimony of a witness with knowledge of the specific transaction will satisfy F.R.E. 901(b)(1), and a learned expert witness should suffice under F.R.E. 901(b)(3). A witness knowledgeable about the contents, substance and distinctive characteristics of the electronic signature process of creating, presenting, signing, securing, archiving and retrieving the documents in question should satisfy F.R.E. 901(b)(4), while testimony describing how the electronic signature process accomplishes the foregoing accurately should suffice under F.R.E. 901(b)(9).

In addition to the express language of F.R.E. 901(b)(9), Imwinkelried's Evidentiary *Foundations* provides an eleven-step process under the Rule for the admission of computer generated records.[33] Most of the testimony proffered under these eleven steps is a simple recitation of facts. More challenging is step four, which requires proof that the 'procedure has built-in safeguards to ensure accuracy and identify errors … regarding computer policy and system control procedures, including control of access to the database, control of access to the program, recording and logging changes, backup practices, and audit procedures to assure the continuing integrity of the records.'[34] In satisfying this requirement or making arguments for admissibility under 901(b)(4), the user would need to provide expert technical testimony as to the functionality and safeguards in the electronic signature process.

Witness testimony seeking the admission of signatures and documents from the electronic signature process pursuant to F.R.E. 901(b)(9) would, in all likelihood, need to include:

a. The manner in which the user's server or servers, as appropriate, are used to generate electronic signatures and documents;

b. The reliability of these servers;

c. Procedures for manual data entry and system controls; and

d. Safeguards to ensure accuracy and identify errors (that is, safeguards, access rules and other controls on the environment that govern the flow of information through its system), tamper resistant software, use of cryptographic technology, and that all of these meet or exceed industry standards.

Presumably, after a number of court decisions recognizing the safeguards of a particular electronic signature process, such as by selecting "yes" in a recorded interactive voice recognition process as in the *Shroyer* case or a clear and conspicuous online process as in the *Bell* case, parties to transactions will be more inclined to stipulate, and not disagree about the authenticity of electronic signatures created using a given electronic signature process. If this were to occur, the need for witness testimony to authenticate

[30] *Lorraine at 544.*
[31] *Lorraine at 549.*
[32] *Swinton, 268 Conn. at 811.*
[33] *Edward J. Imwinkelried, Evidentiary Foundations, (LexisNexis 6th ed. 2005) 58-59, and see Stephen Mason, Electronic Evidence: Disclosure, Discovery & Admissibility (LexisNexis Butterworths, 2007), 4.23 for further comments on Professor Imwinkelried's list: 1. The business uses a computer; 2. The computer is reliable; 3. The business has developed a procedure for inserting data into the computer; 4. The procedure has built-in safeguards to ensure accuracy and identify errors; 5. The business keeps the computer in a good state of repair; 6. The witness had the computer readout certain data; 7. The witness used the proper procedures to obtain the readout; 8. The computer was in working order at the time the witness obtained the readout; 9. The witness recognizes the exhibit as the readout; 10. The witness explains how he or she recognizes the readout; 11. If the readout contains strange symbols or terms, the witness explains the meaning of the symbols or terms for the trier of fact.*
[34] *In re Vee Vinhnee at 447. Opposing parties often allege that computer records have been tampered with and thus lack authenticity. Such claims have been viewed as 'almost wild-eyed speculation…without some evidence to support such a scenario….' United States v. Whitaker, 127 F.3d 595, 602 (7th Cir. 1997).*

documents may not be required in those later cases.[35]

F.R.E. 902

Although in a major dispute, testimony may be necessary regarding the electronic signature process and the authenticity of its process as noted above, documents presented, signed, secured, archived and retrieved using the electronic signature process may also be admitted as self-authenticating documents under F.R.E. 902(7). Judge Grimm in his opinion in *Lorraine v. Markel*, stated, at 549, that: '[e]xtrinsic evidence of authenticity as a condition precedent to admissibility is not required with respect to the following:...(7) Trade inscriptions and the like. Inscriptions, signs, tags, or labels purporting to have been affixed in the course of business and indicating ownership, control, or origin.' 'Under Rule 902(7), labels or tags affixed in the course of business require no authentication. The electronic signature process should collect and record information showing the entire signature ceremony. The identification markers alone stored in the secure container may be sufficient to authenticate an *electronic record* and *electronic signature* under Rule 902(7).'[36]

F.R.E. 902(11) of the Federal Rules of Evidence is the other subsection that might be considered for authentication of documents presented, signed, secured, archived and retrieved using the electronic signature process' electronic signatures. As Judge Grimm noted at 552: 'Rule 902(11) also is extremely useful because it affords a means of authenticating business records under Rule 803(6), one of the most used hearsay exceptions, without the need for a witness to testify in person at trial.' The primary reason for seeking to authenticate electronically stored information using this rule is that it permits a written declaration by a custodian rather than oral testimony, which under most circumstances makes it preferable to F.R.E. 901(b)(4) or (b)(9). F.R.E. 902(11) addresses:

> Certified domestic records of regularly conducted activity. The original or a duplicate of a domestic record of regularly conducted activity that would be admissible under Rule 803(6) if accompanied by a written declaration of its custodian or other qualified person, in a manner complying with any Act of Congress or rule prescribed by the Supreme Court pursuant to statutory authority, certifying that the record-
>
> (A) was made at or near the time of the occurrence of the matters set forth by, or from information transmitted by, a person with knowledge of those matters;
>
> (B) was kept in the course of the regularly conducted activity; and
>
> (C) was made by the regularly conducted activity as a regular practice.

Rule 902(11) was designed to work in tandem with an amendment to Rule 803(6) to allow proponents of business records to qualify them for admittance with an affidavit or similar written statement rather than the live testimony of a qualified witness. In addition to the affidavit requirements, there is a notice requirement to afford opposing parties an opportunity to review the document and affidavit to challenge its authenticity.[37] Thus, assuming no challenge, F.R.E. 902(11) is one of the best ways to secure the admission into evidence of signatures and documents executed using an electronic signature process.

As explained above, critical in the analysis of admissibility and the overall enforceability of documents executed using a given electronic signature process, is the requirement of a secure method to archive and retrieve the documents so they cannot be altered after signature. In addition to the method or process, there must be a credible person called by the user who is suitably qualified to explain the process:

a. the documents submitted to enforce the transaction are true, accurate and complete hard copies of each document signed by each signer that accurately reflect what the signer was presented with in connection with each signer using the electronic signature process;

b. the electronic signature process generates a true, accurate and complete hard copy of the audit trail for each transaction; and

[35] *For example see, Shroyer v. New Cingular Wireless Serv., Inc., 498 F.3d 976 (9th Cir. 2007) and Bell v. Hollywood Entm' t Corp., 2006 Ohio App. LEXIS 3950 (2006).*

[36] *Lorraine at 549, quoting Weinstein's Federal Evidence § 900.07[3].*

[37] *Federal Rules of Evidence 902 (11) at 773 at footnote 4.*

The audit trail should record each step required to meet the regulatory requirements, such as the sequence and timing of presenting certain forms and the actual contents of records presented.

c. the documents submitted to enforce the transaction were generated from electronic records that were cryptographically sealed in such a way that each record, as accurately represented by such hard copies, could not have been altered without detection, in the absence of a person using supercomputing power to break the encryption method used, currently thought to require several years of such supercomputing power.

Users should consider who would be qualified, willing and able to testify on the above items in designing the electronic signature process.

Compliance risk

The electronic signature process should assure that:

a. Each document presented or signed by a signer complies with the legal requirements for the content, presentation, sequence and information to be obtained for each such document;

b. For Special Consumer Disclosures, the signer is provided the appropriate information to enable them to make the informed consent in a way that complies with the consumer disclosure requirements of ESIGN, where such Special Consumer Disclosure Requirements will be provided exclusively via electronic means;

c. Each document required to be presented and signed is in fact presented and signed as required by law governing the particular transaction, and

d. The significance of each step in the signature process (whether on an acknowledgement of receipt, unilateral consent, application for goods or services, or contract) is abundantly clear to each signer.

The audit trail should record each step required to meet the regulatory requirements, such as the sequence and timing of presenting certain forms and the actual contents of records presented. The electronic signature process with the audit trail containing reliable, admissible evidence that each step was taken using the required content, a user may reduce the compliance risk considerably lower that the risk in transactions using paper and manuscript signatures.

The courts have been presented with a variety of disputes where a person alleged to have electronically sign a record disputes having signed the record. Where the significance of the steps involved in signing a particular record was made adequately clear to the person challenging the enforceability, the courts have enforced the electronic signature process. Where the significance was not sufficiently clear to the challenger, the courts have not enforced the terms against the challenger.[38]

Adoption risk

The adoption risk refers to the risk that the electronic signature process, in an attempt to reduce the authentication, repudiation, compliance and admissibility risks, is overly burdensome, such that the intended signers do not use the process or find alternatives that undermine the overall effectiveness of the proposed electronic signature process. This risk can, and should be, managed by conducting a series of pilot tests before introducing the electronic signature process

[38] *For example, see Bell v. Hollywood Entm' t Corp., 2006 Ohio App. LEXIS 3950 (2006) where the court enforced a mandatory arbitration provision against an executive of the defendant employer. The court found that it was sufficiently clear to the executive what the consequences were of selecting 'yes' in the electronic signature process. See also Brueggemann v. NCOA Select, Inc., et al., No.08-80606, 2009 WL 1873651(S. D. Fla. June 29, 2009), where the court enforced an electronic signature comprised of the process of continuing to use the website where the significance of proceeding was made sufficiently clear to a consumer purchasing consumer goods. In contrast, see Campbell v. Gen. Dynamics Gov' t Sys. Corp., 407 F.3d 546 (1st Cir. 2005), where the court concluded that the significance of not objecting to the terms was not sufficiently clear. The court refused to enforce the mandatory arbitration terms against the employee.*

to potential signers for the user. By conducting tests, the user can obtain feedback from the signers and make the appropriate adjustments.

Relative risk

As noted throughout this article, the risks of a given electronic signature process should be considered relative to the risks associated with a paper and manuscript signature. This allows the user to better assess the risks inherent in the particular electronic process. It is often easy to configure the electronic signature process to reduce the risks considerably below the corresponding risks of using paper and a manuscript signature. For example, the electronic signature process can be configured to prevent a record from being signed by the signer if there are any blanks in the record and prevent any document relating to a transaction from being submitted to the user unless all the required steps, including execution of or acknowledgement of receipt of all Special Consumer Disclosures, are fulfilled and then once signed and securing documents from being altered without detection. This can significantly reduce the compliance risk below that for paper and manuscript signature.

Conclusion

The overall effectiveness of a given electronic signature process depends on how well the user determined the means to mitigate the risks for particular documents and records to be presented, signed and archived. The user who carefully considers the risks associated with the types of transactions to be processed can design and implement an electronic signature process that is no riskier than, and in most cases, less risky than the same transaction using paper and a manuscript signature. Doing so provides greater confidence that the electronic signature, when affixed within US, will be admitted into evidence in a US court.

From the court decisions to date, there appears to be a premium placed on making it very clear to the person against whom enforcement is sought, the significance of the act comprising the electronic signature. The clearer the significance to the person signing, the more likely enforcement of the electronic signature process. Enforcement of the electronic signature process will not, however, overcome terms and conditions otherwise unenforceable for reasons having nothing to do with the electronic signature process, such as unconscionable terms in mandatory arbitration agreements.

It is to be expected that as the significance of actions comprising the electronic signature are made clearer, persons aiming to avoid obligations in signed agreements will look for other ways to avoid liability, such as challenging the admissibility of the electronic records for various reasons. The framework described in this article should help companies critically evaluate those risks with the aim of determining what measures to implement that are appropriate within the risk assessment profile discussed in this article.

Greg Casamento and Pat Hatfield are both partners in Locke Lord Bissell & Liddell LLP, a national law firm with offices across the United States. Greg practices in the area of electronic commerce and related litigation matters, including e-discovery and e-admissibility. Pat practices in the electronic commerce, intellectual property and technology areas. The views expressed in this article are those of the authors and do not constitute legal advice regarding any particular set of facts, products or services.

PHatfield@lockelord.com
GCasamento@lockelord.com

http://www.lockelord.com/

ARTICLE:

RELIABILITY OF CHIP & PIN EVIDENCE IN BANKING DISPUTES

By **Steven J Murdoch**

Smart cards are being increasingly used for payment, having been issued across most of Europe, and they are in the process of being implemented elsewhere. These systems are almost exclusively based on a global standard – EMV (named after its designers: Europay, Mastercard, Visa)[1] – and commonly known as Chip & PIN in the United Kingdom. Consequently, the reliability of the Chip & PIN system, and the evidence it generates, has been an increasingly important aspect of disputes between banks and their customers. A common simplification made by banks when deciding whether to refund a disputed transaction, is the assertion that cloned smart cards will be detected, and that the correct PIN must be entered for a transaction to succeed. The reality is more complex, so it can be difficult to distinguish the difference between customer fraud,[2] a third party criminal attack, and customer negligence. This article will discuss the situations which may cause disputed transactions to arise, what may be inferred from the evidence, and the effect of this on banking disputes.

The replacement of magnetic stripe cards with smart cards for credit and debit card payments has changed the nature of disputes between banks and their customers over unauthorized transactions. Previously the operation and weakness of cards was well understood, and there was ample evidence of criminal practice. Now, with the implementation of Chip & PIN, the situation has become uncertain: the system is much more complex, the level of security is less clear, and little is known about the capabilities of criminals in terms of committing fraud. This complicates the task of a bank in identifying whether a customer is entitled to be refunded.

Chip & PIN offers greater resistance to fraud when compared with the previous magnetic stripe system, and unlike earlier domestic smart card payment standards, it works across national boundaries. However, the implementation is not infallible, and its complexity increases the likelihood of flaws. In several respects there has also been a trade-off between cost and security, leading to the creation of weaknesses, some of which have been exploited by criminals, some have been demonstrated by researchers, and the remainder are currently assumed to be merely theoretical.

Customers who notify their bank of unauthorized transactions are often recompensed, but sometimes the disputed transactions are not reversed. One possible reason is that the bank believes that the customer authorized the transaction, and is attempting to defraud the bank by making a spurious complaint. Statistics on this type of fraud are not publicly reported by the banking industry, but a fraud investigator working for a major bank, speaking under the Chatham House rule,[3] did perceive that levels are high. For example, a group of people have been accused of committing, with the assistance of bank insiders, fraud in the region of US$422,000, where they opened banks accounts and then claimed their ATM cards had been lost or stolen,

[1] *EMV Specifications for Payment Systems, available at http://www.emvco.com/specifications.aspx.*

[2] *The term ' first-party fraud' is used within the banking industry to describe fraud by a customer.*

[3] *The Chatham House Rule reads as follows: ' When a meeting, or part thereof, is held under the Chatham House Rule, participants are free to use the information received, but neither the identity nor the affiliation of the speaker(s), nor that of any other participant, may be revealed. The Chatham House Rule may be invoked at meetings to encourage openness and the sharing of information' : http://www.chathamhouse.org.uk/about/chathamhouserule/*

and that certain ATM withdrawals were not authorized by them.[4] The bank may alternatively believe that the customer has acted negligently, in violation of the account terms and conditions, by inadequately protecting their card or PIN, or both their card and PIN. If challenged over such a decision, arguably the bank ought to be required to show that their position is defensible, and that the transaction was not in fact performed by a third-party criminal exploiting a security vulnerability.

The bank's decision will be based on the evidence they have regarding the disputed transaction, the value of the customer's relationship with the bank, and the perceived security of the Chip & PIN system. Much of this evidence will be in digital form, and requires processing and interpretation before it can be understood. While this was also the case with magnetic stripe payment cards, Chip & PIN increases the amount of evidence that could be made available and its level of complexity.

Almost all of this evidence will be held by the bank, as is the information necessary to interpret it. Thus during a dispute, if the bank is required or volunteers to give this evidence to the customer, there will be questions as how to verify the accuracy of the information, and what conclusions can be safely drawn from a forensic analysis. First, this article provides a simplified introduction to Chip & PIN. Then the article sets out the evidence created regarding transactions, and the interpretation of the evidence is discussed to discover whether and how card fraud has been performed.

Introduction

In addition to the visible security mechanisms – such as the hologram, embossing, and fluorescent ink – UK credit and debit cards incorporate a magnetic stripe. This stores the data which is visible on the face of the card (name, expiry date, card number, and such like). It also holds the CVV (Card Verification Value; not to be confused with the CVV2, which is printed on the signature strip of the card). Prior to the use of Chip & PIN, the data from the magnetic stripe would be read by the point-of-sale (PoS) terminal or automated teller machine (ATM) and sent to the bank that issued the card to their customer (the card-holder). This bank (the issuer) would be capable of verifying whether the CVV they received corresponded to the one expected for that particular card number. Thus, based only on information which is visible on the card or a receipt, a criminal should not be able to produce a cloned card which evades detection.

The data read from the magnetic stripe only offers assurance that the card is authentic. It is also necessary to confirm that the genuine card-holder has authorized the transactions. For PoS transactions, the cashier would ask the customer for their signature, which they can then compare to the one on the card. ATM transactions are authorized by PIN. Here, the customer enters their PIN at a keypad, which the ATM encrypts and sends, along with the data from the magnetic stripe, to the issuer, potentially via networks operated by parties such as Visa, Mastercard, or VocaLink. The bank can then compare the PIN entered with the one stored in their records.

Magnetic stripe cards have well-known weaknesses. Using commercially available equipment, it easy to read details from the magnetic stripe of a card, including the CVV, and write a perfect copy of it to a blank card. Such a cloned card would work at an ATM, because only the magnetic stripe is used. Criminals have exploited this weakness in numerous ways, for example adding a 'skimmer' to ATMs, which records the magnetic stripe of the card as it is inserted, and incorporates a camera to record the PIN being entered. Together, this yields enough information to make and use a clone in an ATM. To use a clone in a PoS transaction, the visible security features would also need to be copied, which takes more effort but is well within the capabilities of criminals, and has the advantage that the PIN is not required.

The explanation above has been somewhat simplified for brevity. In fact the authorization systems which verify the CVV and PIN can be quite complex, consisting of many components built and operated by different parties; there will also be significant variation between banks and even more between countries. It can be that the issuer does not authorize the transaction at all, but delegates this responsibility to a third party. Card and PIN details are also likely to pass through several different systems between the PoS terminal or ATM, and the authorization system. However, despite this complexity, the cards themselves use the same technology as video and audio tapes, which means there is good intuitive understanding of their main security vulnerability – that if someone obtains

[4] *'Gang charged in $400,000 ATM scam', Finextra News, 31 July 2009, http://www.finextra.com/fullstory.asp?id=20328. See also Stephen Mason, editor, Electronic Evidence: Disclosure, Discovery & Admissibility (LexisNexis Butterworths, 2007), 4.04–4.15 for a discussion of cases regarding ATM fraud and banking fraud across the world, including insider fraud. The cases in this text predate the introduction of Chip & PIN.*

During card authentication at the PoS, the card submits a cryptographic certificate to the terminal, incorporating the card's account number and a digital signature.

possession of the card, even briefly, they can create a perfect copy.

Chip & PIN

Chip & PIN was designed to mitigate vulnerabilities in magnetic stripe cards, albeit with increased costs, as well as requiring much infrastructure to be upgraded. The cards include a magnetic stripe and the same visible security features as before, but incorporate an additional computer chip underneath the cards' surface. A terminal can interact with the chip through electrical contacts on the face of the card. This chip is a computer with processing power comparable to desktop computers of the 1980s, but with additional security functionality.

The chip has a program loaded into it, which is designed to follow the communication conventions (a protocol) specified by the EMV documentation, and so be able to communicate with terminals that comply with the EMV standard. This specification is complex, consisting of several thousand pages, but there also will be many thousands of additional pages which describe the design of the chip and its software. National industry bodies and industry members may also extend the specification with additional material.

The chip performs three main operations: card authentication (establishing that the card is authentic), card-holder verification (establishing that the person presenting the card is the authorized account holder), and transaction authorization (establishing that there are enough funds to complete the transaction and the card is not cancelled).

Card authentication

The aim of card authentication is to allow the operator of a PoS terminal (the merchant) to establish whether a card presented is legitimate, without contacting the issuer. This is important because in a small proportion of UK PoS transactions, the terminal is 'offline' and does not communicate with the issuer until after the customer has left with the goods. However, since ATM transactions should always be carried out online, card authentication is not performed here. During card authentication at the PoS, the card submits a cryptographic certificate to the terminal, incorporating the card's account number and a digital signature. The terminal can then check whether this certificate was issued by a bank recognized by a payment system (e.g. Visa or Mastercard) supported by the terminal, and validate the digital signature.

Card-holder verification

Once the merchant is satisfied that the card is authentic, both the card and merchant must be assured that the person presenting the card is the legitimate account holder. This is the role of card-holder verification, which is normally achieved by using a PIN. The customer first enters their PIN on a PIN entry device attached to the PoS terminal or ATM. For PoS transactions, the PIN is sent to the card and the card compares the PIN against the one it stores, and returns the result of the comparison to the terminal. If the PIN entered is incorrect, the card will allow the PIN entry to be re-attempted, but only up to a maximum number of tries – normally three. For ATM transactions, the PIN is not sent to the card, but encrypted and sent back to the issuer, as with magnetic stripe transactions.

Transaction authorization

The final step is transaction authorization, where the issuer, card, and merchant are assured that the card is authentic, card-holder verification succeeded, the card has not been cancelled, and there are adequate funds in the customer's account. Here, the terminal or ATM sends the card a summary of the transaction (amount, date, and such like). The card appends its own data, such as

the result of card-holder verification, and also its application transaction counter (ATC), which is a value maintained by the card, counting how many transactions have been initiated. The card then responds with a cryptographic authentication code. For offline transactions, the authentication code (the transaction certificate – TC) is stored by the terminal for later transmission to the issuer, and the transaction is complete. However, for online transactions, the card sends a different type of authentication code, an authorization request cryptogram – ARQC. The ARQC is sent to the issuer, and it responds with a message stating whether the ARQC is valid, incorporating an authorization response cryptogram – ARPC. Finally, the ARPC is sent to the card for verification, and it responds with a TC indicating that the transaction has succeeded. Alternatively the card can at any time send an application authentication cryptogram (AAC) which means the transaction has been declined.

The issuer and card share cryptographic keys which allow them to generate and verify the cryptographic authentication codes (ARQC, ARPC, TC, and AAC). These keys are loaded during its 'personalization' process. However, the merchant does not have these keys, so must rely on the issuer or card to perform the verification. This is because the digital signature keys used in transaction authorization are symmetric, meaning that the same key is used for both generation and verification, and so merchants could not be trusted with the keys. In contrast, the cryptographic keys used for card authentication are asymmetric, meaning that one key is used for signature generation and another key for signature verification, and it is infeasible to convert the latter key into the former. Thus the merchants are all given a verification key (the public half), but the generation key (the private half) is kept by the bank.[5]

Security failures in Chip & PIN

As noted above, the process of a Chip & PIN transaction is much more complex than magnetic stripe transactions. In fact the description above is a simplified version, which shows how transactions should normally happen in the UK; for a variety of reasons the process may diverge from the steps above, and other countries may have different procedures. This complexity is largely for good reasons: the additional verification catches more types of fraud, and so allows transactions to proceed in situations where magnetic stripe cards could not be safely used. However, the complexity also increases the number of ways in which security failures could occur, and makes it more difficult to establish what has happened when they do. This section will summarize some of the potential security vulnerabilities in Chip & PIN, how they may come about, and what their effect might be.

Card vulnerabilities

While magnetic stripe cards merely act as storage, the security of Chip & PIN depends on the cards implementing a set of security constraints, such as not releasing cryptographic keys or the PIN, and performing card-holder verification correctly. If, due to a bug in the software running on the chip, it is possible to violate these security constraints, criminals could exploit the weakness to commit fraud. Even if the software is correct, before a card can be used, it must be configured during the personalization process. If there is a mistake or oversight in this process, the card may be left in an unlocked state in which some security constraints are not enforced.

Criminals must discover the vulnerabilities in order to exploit them. This may be achieved with the help of an insider, who learns about the vulnerability after the cards with the security vulnerability are already issued. The insider may even create the security vulnerability themselves, by interfering with the software or configuration process, or by disclosing the cryptographic keys needed to unlock a Chip & PIN card. Alternatively criminals could discover vulnerabilities on their own. One technique for doing so is 'fuzzing', where an automated process is used to discover security vulnerabilities. This does not need any knowledge of the software being tested. Fuzzing has been widely used in other contexts by both security researchers and criminals, and is a very effective technique.

A further approach to compromising card security is to attack the chip itself, rather than the software. One set of techniques are known as invasive and semi-invasive attacks, where the chip is removed from the card and manipulated using laboratory equipment.[6] These techniques can discover confidential information or create carefully chosen failures in the enforcement of security constraints. Non-invasive attacks are also possible, which do not require the chip to be removed from the card. For example, by measuring minute

[5] *The details of the cryptography used are not important for the purposes of this article, but for further information on the design and use of digital signatures and authentication codes, refer to Ross J Anderson, Security Engineering, (2nd edition, Wiley, 2008).*

[6] *Sergei P. Skorobogatov, Semi-invasive attacks – A new approach to hardware security analysis, University of Cambridge Technical Report UCAM-CL-TR-630, April 2005: http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-630.html.*

variations in the power consumption of smart cards, it is possible to extract cryptographic keys.[7] While smart cards do commonly incorporate defences against attacks, they are not always effective, and criminals regularly use these techniques to clone the smart cards used for subscription television.[8]

Regardless of how the criminal has discovered the security vulnerability, if they can extract the card's cryptographic keys used for transaction authorization, they can create a clone of the card which will be undetectable to the bank systems. A criminal does not need to know the correct PIN to use the cloned card for PoS transactions, because the PIN is verified by the card, and it can be programmed to accept any PIN. Another way a criminal could use a card would be if the correct PIN could be extracted from a card, or if the PIN stored on the card could be changed without the authorization of the issuer.

Other attacks against Chip & PIN do not require the exploitation of security vulnerabilities at all, but rely on inherent limitations of the cards. One such approach is the 'relay attack', which makes use of the fact that smart cards do not have a display to inform the card-holder which transaction they are authorizing.[9] The attack works as follows: the card-holder inserts their authentic card into a compromised Chip & PIN terminal, and at approximately the same time, the criminal inserts a special relay card into a real Chip & PIN terminal or ATM. As the relay card is interrogated, it passes on messages to and from the authentic card via the compromised terminal. Thus the real terminal or ATM will believe the relay card is authentic. The customer will think they are authorizing one transaction, but actually the criminal is carrying out a far larger one, potentially on the other side of the world.

Personalization failures

It may not be necessary to compromise the card in order to clone a card, because all the information needed is available at the personalization bureau (where blank cards have keys and customer data loaded), and at the authorization centre (where transaction authorization messages are sent). Personalization and authorization are both performed on behalf of the issuer, but they are commonly sub-contracted (in whole or in part) to specialist service providers. If a criminal is able to interfere with or extract information from either of these processes, they could create a cloned card without having seen the real one.

When considering disputed transactions, banks commonly make the assumption that exactly one copy of each card has been produced. Therefore, if bank records show that the transaction authorization succeeded, then they infer that the particular card issued to the customer was used. The bank may then consider the customer negligent for allowing their card to be used without authorization, and therefore liable for the transaction. However, the assumption that cloned cards cannot exist is not valid, even if it is assumed that the security vulnerabilities above, which allow card cloning, cannot or have not been exploited.

This is because the personalization bureau must have the ability to produce cloned cards, because the process of personalization occasionally fails due to mechanical problems. For instance, the personalization of the chip may have failed, or the printing on the card may be imperfect. An operator should notice the failure, and if they do, they will request that a second card with the same data be produced. Procedural controls should ensure that the damaged card is destroyed, and all cards are accounted for. If these procedures are followed correctly, each customer should receive exactly one card, which complies with quality assurance standards.

However, these procedures occasionally fail. For example, two bank customers have contacted the author to inform him that they each received two identical cards in the post. This is, presumably, due to a technical or procedural failure at the personalization bureau. The author has read the data from the chip on these cards. In one case, both chips appear to contain identical information, including cryptographic keys, and therefore are perfect clones of each other. The customer had used one of the cards successfully, but had not used the second one. In the other case, one chip was functional, but the other was not active. This customer reported that he used both cards for successful Chip & PIN transactions, so it could be that the bank eventually

[7] *Stefan Mangard, Elisabeth Oswald and Thomas Popp, Power Analysis Attacks: Revealing the Secrets of Smart Cards, (2007, Springer-Verlag New York, Inc.).*

[8] *Kevin Poulsen, 'DirecTV hacker sentenced to seven years', SecurityFocus, 10 December 2004, http://www.securityfocus.com/news/10103; Kim Zetter, 'From the Eye of a Legal Storm, Murdoch's Satellite-TV Hacker Tells All', Wired News, 30 May 2008 (Condé Nast Digital), http://www.wired.com/politics/security/news/2008/05/tarnovsky.*

[9] *Saar Drimer and Steven J. Murdoch, Keep Your Enemies Close: Distance Bounding Against Smartcard Relay Attacks, Proceedings of 16th USENIX Security Symposium on USENIX Security Symposium (2007, USENIX Association Berkeley, CA, USA), http://www.usenix.org/events/sec07/tech/drimer.html. A demonstration of the attack on a Chip & PIN terminal was also filmed by BBC Watchdog, and aired on 6 February 2007, 19:00, BBC One.*

noticed the cloned card and remotely de-activated it. In both cases the cards were visibly identical, had the same information recorded on the magnetic stripe, and had the same details printed on the card, including the CVV2.

In these cases no harm was done, because the legitimate card-holder was sent both clones of the card. However, these instances raise the possibility that a malicious insider could trigger the issue of a cloned card, and retain the cloned card in order to commit fraud. Procedural controls are supposed to stop such activity, but clearly they are not infallible, otherwise cloned cards would not be seen. It is unclear what caused the cloned cards to be sent to these customers, because both had no visible problems. It was confirmed that both chips in one pair worked correctly; for the other pair, the chips presumably worked correctly, otherwise they should not be able to complete transactions. It could be that a software bug or human error triggered the creation of clones, but another possibility was that a malicious insider caused it, but failed to intercept the clone before it was dispatched.

Attacks on hardware security modules

In both the personalization and authorization centres, cryptographic keys and PINs are processed within hardware security modules (HSM). These are computers running specialized software, which will disable themselves should unauthorized interference be detected by their enclosure. Their storage capabilities are limited, and it would be infeasible for them to store separate cryptographic keys for every card that was issued. Therefore a single master key is stored, and a cryptographic procedure called 'key derivation' is used to generate a different key for each card. The key derivation procedure takes as input the master key and identifying information of the card (account number, sequence number), and produces a unique derived key (UDK). The procedure is designed so that if a person did find out the UDK for one or more cards, such knowledge would not provide any help in discovering the UDK for any other cards.

The security of HSMs is therefore of critical importance to the integrity of Chip & PIN. In addition to their tamper resistance, the software running in HSMs must enforce security constraints, such as permitting cryptograms to be verified, but not allowing cryptographic keys to be extracted. However, the complexity of the software has led to the discovery of numerous security vulnerabilities. Initially these were found only by academic researchers,[10] but more recently criminals have been exploiting security vulnerabilities in order to commit fraud. In one case reported by Verizon, criminals had extracted cryptographic keys from an HSM and used these to decrypt customer PINs as they were being processed, presumably at an authorization centre.[11] If it was possible for a criminal to extract the cryptographic keys used during the personalization or authorization process of Chip & PIN cards, they could create undetectable cloned cards or discover the correct PIN for a card, or both. For example, in February 2008, Citibank reported to the FBI that one of their ATM authorization systems was compromised by criminals, account details collected, and US$750,000 of fraudulent ATM transactions carried out.[12]

There have been persistent rumours of a system sometimes termed 'Bergamot' which is claimed to allow criminals to obtain the PIN for a stolen card. Neither the operation nor the existence of such a device has been verified, although reports of its use exist.[13] A journalist for ARD Germany also investigated 18 cases of unauthorized ATM withdrawals committed in La Palma in January 2005. In all cases the cards were stolen, but the customers claimed that their PIN was not written down. Nevertheless, the bank records show the correct PIN was used, and the customers were considered liable. One of the criminals responsible (they were convicted in January 2009) was interviewed by a journalist, but refused to say how they used the card without a PIN. Files from the Guardia Civil assumed that the criminals used a 'dispositivo' (device) to obtain the PIN, but did not give further details.[14]

Design constraints of EMV

The discussion above has assumed that the authorization system will always detect cloned cards and an incorrect PIN. However, this is not always the case; this section will describe some scenarios in which the authorization system will fail to detect even imperfect cloned cards. Some of the vulnerabilities discussed below may be a consequence of errors made

[10] *Mike Bond and Ross Anderson, API-Level Attacks on Embedded Systems, Computer, Volume 34, Issue 10, (October 2001), 67–75, available at http://www.cl.cam.ac.uk/~rja14/Papers/API-Attacks.pdf.*

[11] *Kim Zetter, 'PIN Crackers Nab Holy Grail of Bank Card Security', Wired News, 14 April 2009 (Condé Nast Digital), http://www.wired.com/threatlevel/2009/04/pins/.*

[12] *Kevin Poulsen, 'Citibank Hack Blamed for Alleged ATM Crime Spree', Wired News, 18 June 2008, (Condé Nast Digital), http://www.wired.com/threatlevel/2008/06/citibank-atm-se/.*

[13] *Steve Gold, 'A PIN to go with that stolen card sir', IT Pro Portal, 16 August 2006, http://www.itproportal.com/security/news/article/2006/8/16/a-pin-to-go-with-that-stolen-card-sir/.*

[14] *Sabina Wolf, 'Sicherheitsrisiko EC-Karten: Wie Banken mit geschädigten Kunden umgehen', Report MÜNCHEN, 15 June 2009, http://www.br-online.de/das-erste/report-muenchen/report-sicherheit-eckarten-ID1244812929699.xml.*

during design and implementation, which remained undiscovered until the system was in use. However, others may have been identified earlier, but permitted to remain because the risk of the vulnerability was perceived to be smaller than the cost to resolve it, after other checks and balances were put in place. In designing a system, the bank will try to find an appropriate compromise by applying only those security measures that will at least reduce fraud by their cost. For example, cards issued in the UK are vulnerable to attack in offline transactions, but the UK banks agreed that the cost of the more secure cards would be higher than the fraud they would resolve. This is because it is cheaper to put high-value transactions online, and put fraud detection algorithms in place. The problem from the perspective of the customer is that these trade-offs may not be known by the fraud investigation team. For instance, while the weakness of UK cards in offline transactions is well known, other vulnerabilities might exist as a consequence of one department making a cost saving, without informing other departments.

Cost-benefit trade-offs may have been considered during the design of EMV if the vulnerability was identified then, but if the vulnerability was discovered during the implementation phase, a decision would have been made not to fix it, because it was not cost effective for the banks. Other vulnerabilities are not inherent to all systems that implement EMV, but are a property of a particular implementation; again, these may be because of an oversight or due to a deliberate design decision.

Static PIN

One inherent design decision in EMV is to use a single PIN for the card, which must be entered in its entirety. This means that if someone can see the PIN being entered by the customer, and subsequently steals the card, the criminal can easily commit fraud. Other countries, for example Brazil, have adopted a different approach. In addition to the PIN, an ATM prompts the customer for a number of letters from a password. This makes it unlikely that if a person is able to look over a customer's shoulder, that they will obtain enough information to commit fraud.

It may also be possible for a criminal to guess the right PIN for a card. From a single guess, if it is assumed that every combination of PIN is equally likely, a thief who steals a card has a 1 in 10,000 chance of guessing the PIN (and perhaps even more because some PINs are much more popular and the issuer may not permit certain easy-to-guess combinations of numbers). But the thief actually has six guesses because the card permits three tries before the card will lock, and an ATM will permit a further three, making the chance of success 1 in 1,666. If the customer has multiple cards with the same PIN (a practice recommended by banks to prevent customers having to write down their PIN), the odds for criminals can be even better. With four cards in a stolen wallet, the thief could have five attempts on each card without locking them, and thus have a 1 in 500 chance of finding the PIN, and if successful, they will then be able to use all the cards.

These estimates are assuming that each card only has one PIN which will be accepted. This is certainly the case (assuming the card functions correctly) for PoS transactions where the card verifies the PIN. However, this is probably not the case for ATM transactions if the PVV (PIN verification value) technique of verifying PINs is used. This approach is used to reduce the risk that a compromise of the authorization system will lead to PINs being discovered. Rather than storing the PIN, the customer's PIN is encrypted and then truncated to 4 digits of ciphertext – the PVV. The PINs entered at the keypad are also encrypted, truncated, and compared to the PVV. If the correct PIN is entered, the two will match, but if an incorrect PIN is entered, there is still a chance of a match. Most cards will have two or more PINs which will trigger a PVV match, and some will have as many as ten.[15]

Yes cards

During the process of card authentication, the card presents a cryptographic certificate to prove that it is a legitimate card. This certificate can be verified with information which is available publicly, and therefore can be carried out even by PoS terminals which are offline. However, in order to allow anyone to verify the certificate, the card must also permit anyone to read its certificate and all the other information it presents. Therefore anyone who can read the certificate can, for most UK cards, produce a cloned smart card that will present identical information, and so pass card authentication. Criminals could produce such a clone by reading data from a Chip & PIN card and writing it to a

[15] *Mike Bond and Jolyon Clulow, Encrypted? Randomised? Compromised? (When Cryptographically Secured Data is Not Secure), Workshop on Cryptographic Algorithms and their Uses, July 2004, (Queensland University of Technology), http://www.cl.cam.ac.uk/~mkb23/research/Enc-Rand-Comp.pdf.*

generic smart card. Equipment and software to achieve this, along with programmable smart cards, are commercially available, and cloned smart cards created in this way have already been found in Europe.[16]

For PoS transactions, verification of the card-holder is performed by the card. The terminal sends the PIN entered to the card, and the card responds whether it is correct. Therefore a criminal does not need to know the correct PIN when using a cloned card, because clones can be made which simply respond that any PIN is correct – known as 'Yes-Cards'. Clones such as this would not contain the correct keys for generating the ARQC or TC, and so could be detected by the issuer. However, the TC is not sent to the bank until long after the transaction for offline transactions, so by that stage the thief will have left with the goods, although the fraud can be detected afterwards. For online transactions, the incorrect ARQC should be detected and the transaction declined.

Copying the certificate to circumvent card authentication, as described above, is possible in cards which support static data authentication (SDA). As of 2009, most UK cards are of this type, but some banks are distributing out a more secure alternative – dynamic data authentication (DDA). This provides some resistance against card cloning, but was not issued in the UK, in part due to concerns about the increased costs of the cards and longer transaction times. DDA works by adding an additional step to card authentication, where the card proves that it is the legitimate owner of the certificate it presents. This feature requires giving the card an asymmetric key (both the public and private half), and the ability to produce its own digital signatures, which is more expensive, because asymmetric cryptography is much more complex than the symmetric cryptography used in transaction authorization.

The wedge attack

However, DDA does not prevent yes-cards completely, because card authentication can occur before card-holder verification, and so may not include the result of the PIN verification. A simple yes-card cannot be used, because it would fail card authentication, but an alternative technique might still be effective against offline transactions. Here, a stolen card is plugged into a device (a 'wedge') that can modify the data as it flows between the terminal and the card. The terminal is permitted to communicate directly with the legitimate card during card authentication, which will therefore be successful. But during card-holder verification, the wedge suppresses the messages as they are sent to the card and, regardless of the PIN entered by the thief, the wedge tells the terminal that the PIN was correct. The wedge can either pass through the TC from the real card, or create a fake one of its own. In this way, a criminal who has stolen a Chip & PIN card (SDA or DDA) can use it in offline transactions without knowing the correct PIN.[17]

The wedge attack also works against online transactions, due to an oversight in the design of the transaction authorization stage. In the EMV specification, the ARQC and TC message includes the result of card-holder verification. However, the result only indicates whether the verification was attempted but failed; it does not distinguish between whether the verification succeeded or whether it was not attempted. Therefore a wedge could suppress card-holder verification, and then relay the ARQC and TC between the legitimate card and terminal. The issuer would receive these cryptograms, and since they were from the legitimate card, the authorization would succeed and the bank would accept the transaction.

This flaw was eventually identified, and banks produced a proprietary extension to EMV which included an additional result in the ARQC and TC, stating whether PIN verification was attempted; a similar extension was later included in the revised EMV specification, but has yet to be widely implemented. However, these only allow the issuer to establish that PIN verification was not attempted; in the wedge attack, the merchant's PoS terminal will still believe PIN verification succeeded, even though the wrong PIN was entered. A further extension – combined DDA and application cryptogram generation (CDA) – can prevent the wedge attack even in offline transactions by combining card authentication and transaction authorization, but this further extension has yet to be adopted, at least in the UK.

Stand-in authorization

As noted in the discussion above, transaction authorization is of critical importance: it is the only way to reliably detect cloned cards and whether the correct

[16] *Dave Birch, 'I didn't want to write about fraud yet again, but...', Digital Money Forum, 15 October 2008, http://digitaldebateblogs.typepad.com/digital_money/2008/10/i-didnt-want-to.html.*

[17] *Chris Mitchell, 'Payment and e-commerce applications (Part B2)', Lecture notes for IY5601, 2005 (Royal Holloway, University of London), http://www.isg.rhul.ac.uk/cjm/IY5601/IY5601_B_060205_83-156.pdf.*

PIN has been entered. Most transactions in the UK (estimates of 80–90 per cent have been given) are processed online. Despite this, the issuer may not process the authorization message, because of the possibility of 'stand-in authorization'. Here, if the issuer cannot be contacted in sufficient time, an intermediate party such as the payment system or an outsourced processing centre may authorize the transaction on behalf of the issuer. The party that provides the authorization is sometimes contractually obliged to accept liability for the transaction if it is fraudulent. However, if there is an equipment failure, it still may be more cost-effective to authorize the transaction and accept the risk without performing all the checks. Issuers may not be aware of their own policy (or that of any outsourced provider) on how authorizations are handled when equipment fails, or the times at which such failures may have occurred. They may even fail to disclose this information to customers who are disputing a transaction.

Where the transaction value is low, and the costs of communications are high, it may be cost-effective to not attempt to contact the issuer at all. This is especially likely to happen in international transactions, but the prevalence is decreasing because of improved reliability and the lower cost of data communications. Each type of intermediate party is able to check different aspects of the transaction. For instance, some have the keys to verify the ARQC and TC, some can verify the PIN (for ATM transactions), some can check if the card is reported stolen, and some may not be able to check any of these. Issuers will have different policies on which types of intermediate party is able to perform stand-in authorization. In addition to establishing liability, contracts will also impose service level agreements that will set out the speed by which an authorization message must be processed, and in such a manner control the circumstances in which stand-in processing is appropriate.

Fallback

Another set of vulnerabilities exist because the magnetic stripe system is still operational, even with Chip & PIN cards. UK cards continue to have magnetic stripes, to enable them to work in terminals and ATMs without chip readers (e.g. outside the UK), or when the chip or chip reader has failed. UK PoS terminals also have magnetic stripe readers for use with foreign cards or as a backup when the chip cannot be read. This means a criminal who cannot clone a chip can simply copy the magnetic stripe from a smart card, and produce a magnetic stripe clone. From the perspective of an ATM or PoS terminal, this clone will appear to be a legitimate card, but the chip on the card might be damaged, or the chip reader in the terminal might have failed. Since chips regularly break and chip readers frequently get dirty and fail, this is not very suspicious, and the transaction may be permitted to proceed regardless. This is known as a 'fallback transaction'.

The criminal does not have to read the magnetic stripe to clone the card, because the chip contains a copy of the data on the magnetic stripe. This data is also commonly sent to the issuer during a transaction. Therefore a criminal who can read the chip or intercept the communication between a terminal and the issuer, can also copy the magnetic stripe. A criminal who can intercept the communication between the PoS terminal and chip can copy the same data, and also can obtain the PIN entered by the customer, as it is sent to the chip during card-holder verification. PoS terminals have tamper resistance measures to prevent this, but due to design errors, it is quite simple to circumvent the protection in place and connect a 'tap' built with off-the-shelf electronic components.[18] This device reads all the information necessary to produce a cloned magnetic stripe card and use it in an ATM. Chip & PIN terminals have even been discovered with taps having been added during or soon after manufacture.[19] For these reasons, more recent cards do not store the full CVV on the chip, instead replacing it with an alternative termed the 'iCVV'.

This general approach has been widely exploited for committing both fraudulent ATM and PoS transactions. For instance, Maxwell Parsons was convicted in November 2006 of having collected card details by connecting a MP3 player to the back of ATMs. With this information he was able to produce cloned cards, and use them to perform unauthorized transactions.[20] In

[18] *Saar Drimer, Steven J. Murdoch, and Ross Anderson, Thinking Inside the Box: System-Level Failures of Tamper Proofing, Proceedings of the 2008 IEEE Symposium on Security and Privacy, (2008, IEEE Computer Society Washington, DC, USA), 281–295, available at http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-711.pdf. A demonstration of this attack on a Chip & PIN terminal was filmed by BBC Newsnight, and aired on 26 February 2008, 22:30, BBC Two.*

[19] *Henry Samuel, 'Chip and pin scam' has netted millions from British shoppers'', The Daily Telegraph, 10 October 2008, http://www.telegraph.co.uk/news/newstopics/politics/lawandorder/3173346/Chip-and-pin-scam-has-netted-millions-from-British-shoppers.html.*

[20] *'Cash machine bug scam expert jailed', Manchester Evening News, 15 November 2006, http://www.manchestereveningnews.co.uk/news/s/228/228286_cash_machine_bug_scam_expert_jailed.html; Stephen Mason, editor, Electronic Evidence: Disclosure, Discovery & Admissibility, 4.10.*

It is likely that there may be weaknesses in the card processing infrastructure because of the complexity of the system, and because it is continually being upgraded to accommodate new equipment and additional operational requirements.

October 2008, Anup Patel was convicted of committing fraud to the value of £2 million. This was achieved by defeating the physical protection put in place to protect Chip & PIN terminals, and to record both PINs and card details. This attack was so successful that it enabled cloned magnetic stripe cards to be produced.[21]

Back-end failures

The description of the vulnerabilities in the section above assumed that the back-end systems controlled by the card issuer, together with other processing infrastructure, operate correctly. However, if the bank systems are not perfectly designed and correctly operated, these assumptions will not be true. It is likely that there may be weaknesses in the card processing infrastructure because of the complexity of the system, and because it is continually being upgraded to accommodate new equipment and additional operational requirements. A recent illustration of a failure was demonstrated where a couple in Essex, UK, discovered that they could withdraw cash from a particular ATM without the transaction being recorded against their account. Even though these transactions should have been declined because the couple's account was overdrawn, a failure at some point in the processing allowed them to be accepted. This failure eventually became public when the couple were convicted, having withdrawn over £61,000 in this way.[22]

For online transactions, if the ARQC or TC is wrong, the issuer should decline a transaction, and for offline transactions an incorrect TC should be detected when the terminal goes online at a later time, the fraud discovered, and the customer refunded. In this way, customers should not lose money from the use of yes-card attacks, although for offline fraud, the merchant will probably have to pay once the fraudulent transaction is reversed. But if the issuer fails to detect an ARQC or TC which was generated with the wrong key, or contains the wrong information, the use of a yes-card or wedge attacks will not be detected for online or offline transactions. This failure could occur simply because of a programming error, but it could also be an intentional decision; for example if the HSMs which validate the ARQC and TC are overloaded, the issuer may decide to accept transactions without checking. Also, in some circumstances, the ARQC or TC will be corrupted before being sent to the issuer; in these situations the issuer may decide to accept the transaction rather than risk insulting the customer by declining it, and accept the risk of fraud. Such corruption can occur due to random errors, or because of a format translation error such as the one believed to be the reason that Visa debited customers' accounts by US$23 quadrillion.[23]

Because almost all UK cards, ATMs and PoS terminals have been upgraded to support Chip & PIN, it is common for the issuer to automatically decline fallback transactions on cards which have a chip when used in the UK. However, a failure in processing may also allow a fallback transaction to succeed when it should be declined. This could be because the issuer is informed that a fallback transaction has occurred, but a software bug causes it to be accepted. Alternatively, a bug in the ATM or PoS terminal may cause it to identify a fallback transaction as a chip transaction, and the authorization system does not decline the transaction on the basis of it having a missing or incorrect ARQC or TC. Alternatively, a criminal could modify the 'service code' on the magnetic stripe to indicate the card does not have a chip, and hence a fallback transaction should be permitted. The issuer should detect the tampered service code, but in 2005 someone working for the London Programme made a magnetic stripe clone or a smart card, altered the service code, and successfully

[21] *Steve Bird, ''Catch me if you can,' said student behind biggest chip and PIN fraud', The Times, 19 October 2008, http://www.timesonline.co.uk/tol/news/uk/crime/article5034185.ece.*

[22] *Stephen Bates, 'Couple who took £61,000 from faulty ATM sentenced', The Guardian, 21 April 2009, http://www.guardian.co.uk/uk/2009/apr/21/cash-machine-theft-essex.*

[23] *Dan Goodin, 'Reg readers crack case of the $23 quadrillion overcharge', The Register, 16 July 2009, http://www.theregister.co.uk/2009/07/16/visa_programming_error_cracked/.*

used it for an ATM cash withdrawal.[24]

Logging failures

Most banking systems produce extensive log files that record their actions, which makes it easier to identify malfunctions and understand the cause. When transactions are disputed by the customer, these logs may be examined. However, the systems which produce the log files are complex, and their output often requires further processing before it can be easily understood. It is therefore common to have reporting systems, which take the raw log input (potentially from multiple sources), interpret them, and produce a new file which is intended to be easier to understand. A failure in either logging or reporting systems could cause the result of a transaction to be misinterpreted; for example the operator may believe it to be a Chip & PIN transaction when in fact it was fallback. If a malicious person has gained access to logging or reporting systems, they could also tamper with the result in order to cover their tracks, because these systems are commonly less well protected than authorization systems.

Even after a reporting system has processed a log file, it can still be difficult to interpret the output. Output is generally presented using terse codes, and their meaning must be found within documentation. Sometimes this documentation is not available, or it may be out-of-date following changes to the system concerned. Therefore, the operator may interpret them by comparing the log output with similar output observed in the past, and then they may draw conclusions. For example, in the case *Job v Halifax plc*,[25] the witness for the bank examined the format of the log entry of the disputed transaction, and pointed out that it was similar to other legitimate Chip & PIN transactions, and different from other legitimate fallback transactions. From this, the witness inferred the disputed transaction must have been a legitimate Chip & PIN transaction. The bank did not refer to any documentation on the meaning of the data, or discuss what the log entry would look like should one or more security checks have failed.

PIN verification

As discussed with respect to the yes-card attack, for PoS transactions, the card is responsible for verifying the PIN, and if a cloned card is used, the criminal need not know the correct PIN. In contrast for ATM transactions, the PIN is sent back to the issuer or a stand-in processor for verification. Even so, the criminal would not need to know the correct PIN if a stand-in processor that cannot verify the PIN authorizes the transaction. If there is a malfunction which allows transactions to be authorized if the PIN verification is not attempted or fails, a card could be used without the correct PIN. An insider may also try to trigger such failures, for example by gaining access to the authorization system.

Terminal failures

The discussion above has been about failures of the processing and authorization systems. These are very important, because the correct functioning of these systems is of critical importance to the integrity of the Chip & PIN system. PoS terminals and ATMs are relied upon to a lesser extent because they are under the control of potentially untrustworthy merchants, and their correct functioning cannot be guaranteed. It is for this reason they are tamper resistant, to prevent malicious people from extracting confidential information (although these measures can easily be overcome as noted above, and criminals have been caught doing so). Nevertheless it is still possible to commit fraud because the terminal fails to operate properly.

During transaction authorization, the PoS terminal or ATM generates an unpredictable number. The number is sent to the card, and incorporated into the cryptographic process which generates the TC and ARQC. If this number is predictable, a criminal could clone a card by asking the legitimate card for a number of TC and ARQC cryptograms, then writing these values to a generic smart card. This clone could then be used for a transaction, and provided the thief guessed a correct value for the unpredictable number, the clone can produce a TC and ARQC which will pass the check by the issuer. In this way, the criminal can put through online Chip & PIN transactions at both PoS and ATMs, given only temporary access to the legitimate card. The criminal may, however, need to know the correct PIN if cryptograms are required that indicate that card-holder verification succeeded.

While there are well established techniques for

[24] *Chip and PIN security flaw uncovered, London Programme, ITV1 London, 15 March 2005, 19:30–20:00.*

[25] *Job v Halifax plc, Nottingham County Court (case number 7BQ00307), 30 April 2009, the judgment is published on page 235.*

securely generating unpredictable numbers, it is notoriously difficult to verify whether a generator is working correctly, so failures do regularly occur. For example, one version of Linux had a feature which was supposed to generate unpredictable random numbers, but was in fact relatively easy to predict. This flaw was introduced in 2006, but remained undetected until 2008.[26] Linux is used in both ATMs and PoS terminals, but it is not clear whether such devices ran an affected version. A criminal can also tamper with an ATM or PoS terminal to reduce the unpredictability of the random number generator. Research has shown that it might not even be necessary to open the device to do so; manipulating the power supply or transmitting a radio signal may be sufficient.[27]

Whistleblowing and insiders

Examples of security failures in banking systems are hard to find for a variety of reasons, such as the restrictions imposed by non-disclosure agreements; the banks are reluctant to admit vulnerabilities and the complexity of systems, thus making it challenging to discover vulnerabilities in the first place. The examples discussed above became public either because a customer noticed, or because of legal proceedings. In other fields, most notably safety critical systems such as aerospace and medical equipment, there are legal requirements on companies to report failures, and to investigate serious failures. The banks are not subject to such requirements; only whistleblowers and researchers acting outside the banking system can notify the public of problems.

However, there are substantial obstacles to this. For example, a French engineer, Serge Humpich, discovered a way to make forged banking smart cards, and reported this to the banks involved. Having demonstrated his technique worked by purchasing ten Paris Metro tickets at the request of the banks, he was arrested and convicted for counterfeiting.[28] Also, a journalist in the UK whom the author assisted with reporting on vulnerabilities in Chip & PIN was threatened by his own bank that they might cancel his mortgage (though the bank in question eventually withdrew the threat).[29] Cases such as this create a chilling effect, preventing people from testing the existence of vulnerabilities or reporting those they become aware of.

Banking insiders who are aware of security weaknesses may decide to exploit the vulnerability for fraudulent purposes. They may discover the problem directly, or be notified of it in a security testing report. Discovering a vulnerability does not, however, require insider information; while one of the criminals in the Essex case referred to above worked for a bank, it is believed they discovered the vulnerability by accident. The ATM in question was old, which offers a possible explanation as to how the fraud happened. In order to integrate legacy infrastructure with new systems, an established technique developed by Brodie and Stonebreaker is to build a 'gateway',[30] which translates between the old and new conventions, but potentially loses information in the process. This component is then modified to fix bugs until it passes the necessary tests.

Integrating legacy systems is notoriously difficult, and tests cannot be guaranteed to find every problem. Although it can only be hypothesized that this class of flaw allowed the fraud, another example of where integrating a legacy component causes failure is with the Ariane 5 satellite launch vehicle. Here, despite extremely rigorous testing, an integration problem between software originally designed for Ariane 4 and the Ariane 5 navigation system remained undiscovered, until it caused a failure soon after launch, destroying the rocket.[31] The writers of the software made assumptions that were true at the time the component was designed, but were invalidated when the surrounding system was upgraded, leading to its catastrophic failure.

Evidence in Chip & PIN cases

When a transaction is disputed, there may be disagreement between the customer and the bank as to who should be liable for it. A common example is where the bank believes that a customer's real card and PIN have been used, and hence argues that either the customer performed the transaction (and is attempting to defraud the bank), or has been negligent in protecting their card or PIN or both card and PIN. The

[26] *Robert Jaques, Debian flaw exposes communications breakdown, V3, 28 May 2008, (Incisive Media Ltd), http://www.v3.co.uk/vnunet/news/2217710/linux-security-flaw-should-wake.*

[27] *A. Theodore Markettos and Simon W. Moore, The Frequency Injection Attack on Ring-Oscillator-Based True Random Number Generators, Workshop on Cryptographic Hardware and Embedded Systems, LNCS 5747, September 2009 (Springer).*

[28] *Cedric Ingrand, ' French credit card hacker convicted' , The Register, 26 February 2000. http://www.theregister.co.uk/2000/02/26/french_credit_card_hacker_convicted/.*

[29] *Personal communication.*

[30] *Michael Stonebraker and Michael L. Brodie, Migrating Legacy Systems: Gateways, Interfaces & the Incremental Approach, (Morgan Kaufmann Publishers, 1995).*

[31] *Professor Jacques-Louis Lions and others, Ariane 501 Inquiry Board report, ESA, 19 July 1996: http://esamultimedia.esa.int/docs/esa-x-1819eng.pdf.*

One very effective technique is penetration testing, where a skilled team are given access to the system and given the task of finding security vulnerabilities in any way they see fit.

customer may believe that they did not carry out the transaction, and they were not negligent with their card or PIN, and argue that the transaction is due to an error having been made by the bank, or a security vulnerability having been exploited by criminals. Evidence may be requested to corroborate each party's position, but drawing conclusions from it must be performed with care. The evidence itself could be insufficient, and mistakes in interpretation might be made. Also, bank employees might try to cover up embarrassing security failures, and criminals may attempt to tamper with evidence.

Evidence can be collected from all the systems that have been discussed in this article, commonly in the form of log files. These include the manufacture and personalization facilities, where the chips are produced, placed on cards and loaded with data. The PoS terminal or ATM will also contain a log, generally on paper and informally called the 'till-roll', which records transactions and other important events. The card itself also contains useful information, such as the ATC, but depending on the bank, there may also be summary information available about transactions. The most important logs are kept at the authorization centre, recording the type of transaction and result of authorization. However, because there is potential for errors in all of these items due to mistakes or tampering, it is prudent to collect as much evidence as possible in order to show that records are consistent.

Audit and compliance

Another way to establish the reliability of evidence is to examine whether the system producing the logs was operating correctly. This is commonly achieved through an audit, where experts (possibly internal to the bank or external) will examine documentation or the system itself (or both the documentation and the system), and check it against requirements. The result of the audit is a report, which can be very informative as to the dependability of the system. If any potential problems are identified, it will highlight these, and where an external audit is undertaken (by a payment system for instance), the auditor may require the bank to undertake changes. However, if a vulnerability discovered by a payment system is considered to only affect the bank itself and not other members of the scheme, it may be possible for the bank not to deal with the problem and accept the risk.

Not only is the report itself important, but also the changes which were carried out in response to the report. For each change, there should be information on how and when the modification was applied, and how it was established that the modification properly fixed the issue that was identified. The methodology for performing the audit is also significant, because approaches vary in how effective they are at identifying problems. One very effective technique is penetration testing, where a skilled team are given access to the system and given the task of finding security vulnerabilities in any way they see fit. At another extreme, some audits only examine documentation and not the system itself, and so will miss implementation errors. However, regardless of the methodology, audits do miss critical vulnerabilities; for example the OpenSSL cryptographic library was subjected to an extensive audit under the FIPS 140-2 scheme, and passed, even though a serious security vulnerability existed in the random number generator.[32]

Logs of changes to bank systems are important, even if modifications were not the result of an audit report. This is especially true if the modification was known to have an effect on security, but even apparently innocuous changes can cause security vulnerabilities which are hard to identify. Obtaining documentary evidence is important because there may be a dispute, even within a bank, as to when a particular change is

[32] *OpenSSL FIPS Object Module Vulnerabilities, 29 November 2007, http://www.openssl.org/news/secadv_20071129.txt.*

made. For example, in the case of *Job v Halifax plc*, a number of dates were given as to when magnetic stripe transactions were disabled. The witness for Halifax, Ian Brown, stated in paragraph 5.4 of his statement dated 6 February 2008, that 'If the transaction is presented in 'fallback' mode, then the transaction will be declined.' The inference of this comment was that Halifax had disabled fallback before February 2006 (the date of the disputed transactions). The expert witness for Halifax plc, David Baker (Head of the APACS Cards Technical Unit), stated in paragraph 5.1 of his second expert's report dated 14 February 2008 that 'To our knowledge all UK issuers will routinely decline transactions flagged as magnetic stripe read and have been doing so since 2005.' When Mr Brown gave evidence, his barrister questioned him about the statement he made in paragraph 5.1, and he amended his evidence to the effect that the comments he made were correct in September 2006. The author has confirmed that some banks permitted fallback as late as May 2007.[33]

One explanation for this discrepancy is that there can be a delay between when a change is mandated, and when it is actually applied; this has also been seen with iCVV, which APACS announced to be fully in place by 2008:[34]

> 'All UK issued cards issued after 1 January 2008 include an updated iCVV (Integrated Circuit Card Verification Value) which means that if one of these cards were compromised in the method described, the data would be useless to the fraudster (i.e. a fake magnetic stripe card created via a compromise of this type would not work in a cash machine, even overseas in a non-chip and PIN country).'

The author has confirmed that several banks, including Halifax Bank of Scotland (with a card issued in March 2008) and Barclays (with a card issued in February 2008), have not implemented iCVV, and they were still issuing cards that do not comply with the iCCV standard past this date.

Before a card transaction can take place, the chip must be manufactured and the software loaded. Logs from this process will be useful to establish which version of the chip hardware and EMV software was used. Internal audit reports may indicate if any versions were known to be vulnerable to attack. Even if these audit reports do not exist, it would be informative to examine the 'change log' documentation accompanying the chip software, as this should indicate the differences between versions. If this documentation indicates that a new version of software was released in order to fix a bug in a prior version, the possibility that such a bug would allow a fraudulent transaction should be investigated.

The personalization process is another area in which logs would be valuable, for example whether a cloned card was produced because the first one might apparently fail quality assurance. Procedures should also be examined, to ensure that cryptographic keys are being safely handled. Also, audit reports and change logs for the software which configures cards should be examined, to ensure it properly locks the card to protect confidential data.

Logs of the transaction and authorization process will be available from a number of places, and the information they contain will be somewhat different. The ATM or PoS terminal will be able to indicate whether the transaction that is in dispute actually took place, and how it was authorized. However, interpretation of these can be difficult. For example, in the Essex fraud case, the bank that operated the ATM originally believed that members of staff were stealing the money. This suspicion must have arisen because logs are maintained of how much money was loaded into the ATM and how much was withdrawn, and the totals were inconsistent. In fact, the CCTV surveillance put in place to catch the thief, actually showed the couple who were later convicted.

During the authorization process, messages are sent via a number of intermediate parties: the acquiring bank who is contacted by the merchant; the issuer (or stand-in processor) which generates the authorization; and the payment system which allows acquiring banks and issuers to communicate and transfer funds. All of these parties probably keep logs, especially the payment system. This is because they are responsible for providing the communication infrastructure, and they also offer dispute resolution services between their members. In these cases, where there is an inconsistency in the records of two members, the payment system can examine their logs to establish the facts. The logs of payment systems are particularly valuable in this case, because they are from a neutral party. Similarly, in the case of customer disputes, collecting logs from a third party increases the chance

[33] *Demonstration on ITV Manhunt, aired on ITV1, 29 May 2007.*

[34] *Statement from APACS, in response to BBC Newsnight, 26 February 2008, http://news.bbc.co.uk/1/hi/programmes/newsnight/7265888.stm.*

of detecting insider attacks.

Transaction authentication is the most important step of the EMV transaction, and for this stage the logs are kept with the issuer. These should include a description of the transaction, the result of authorization, and the cryptograms involved (ARQC, TC, and ARPC). The description should include all the usual information, such as the amount and date, but also will have the result of card-holder verification, and importantly whether PIN verification was attempted and whether it succeeded. Two versions of this are given: one by the terminal in its description of the transaction, and one by the card as part of the ARQC and TC, in its issuer application data section (IAD). These should be compared to establish consistency. Because of the cryptographic processing, these logs can be subjected to enhanced verification, but they do not replace the other logs discussed above, because they do not contain all the necessary information and still can be tampered with.

Principles for design of secure systems

In the field of computer security, the Trusted Computing Base (TCB) is the part of a system which must be relied upon in order for the overall system to function securely. A widely accepted principle of security engineering is to minimize the size of the TCB, in order to improve the robustness of the system.[35] Following this principle means that during the design and implementation of a system, the available testing resources can be focused more intensely on the TCB, increasing the chances of identifying bugs. Additionally, this principle aids forensic analysis, because if a component can be shown to be outside the TCB, there is no need to waste effort in establishing whether it is functioning correctly. In EMV, there is no clearly defined TCB, but analyzing the system from this perspective is a helpful way of deciding what system components should be examined and what they can be relied upon for.

In disputed transaction cases, the issuer will typically have almost all the evidence that is presented to the court or adjudicator. Sometimes audit reports are made public, as occurs for the banking smart cards issued in Germany and evaluated under the Common Criteria scheme.[36] However, in banking, it is more common to keep audit reports and system documentation secret than in other areas of security engineering.

In discussing what evidence should be presented, there may be a question as to whether a bank, by giving an opposing expert witness access to an item of information, would harm the security of the system. An accepted best practice in security engineering is that the security of a robust system should not depend on secrecy of its design. This is because it is difficult to keep design documents secret, and if the detailed functionality of a system cannot be described, questions may be raised as to whether it is in fact secure. It is for this reason that it is common to openly publish details of security systems, often including the source code from which they are built. Even if source code is not published (e.g. Microsoft Windows), lists of known security flaws are publically available.

This practice is historically known as Kerkhoffs' principle,[37] where it was applied to military communication systems. With respect to banking systems, the same principle is described by APACS (the UK banking industry representative body), in their PIN Administration Policy:[38]

> 'The PIN Administration process must not only be secure, but also be demonstrably secure. If PIN Security is publicly challenged, either in the media or in a court of law, it must be possible to respond to such a challenge and for the response to be supported with evidence. Furthermore, the use of that evidence in the public domain must not in itself compromise security.'

Verifying authorization logs

If there is a disputed Chip & PIN transaction, it can safely be assumed that the bank authorization system shows that the correct card and PIN were used (otherwise the customer would be immediately refunded). In which case, the next step in examining the evidence would be to establish whether these logs can be relied upon in concluding that the customer's card and PIN were used. Some generic approaches have been described above, such as corroborating different items of evidence and examining documentation relating to the systems relied upon. However, there is one particularly useful set of techniques which can be applied to authorization system logs, because the EMV

[35] *Butler Lampson, Martín Abadi, Michael Burrows and Edward Wobber, 'Authentication in Distributed Systems: Theory and Practice', ACM Transactions on Computer Systems, Volume 12, Issue 1 (February 1994) (ACM Press), http://research.microsoft.com/en-us/um/people/blampson/45-authenticationtheoryandpractice/acrobat.pdf.*

[36] *Certification Report for ZKA SECCOS Sig v1.5.2, BSI-DSZ-CC-0341-2006, 13 June 2006 (BSI), http://www.commoncriteriaportal.org/files/epfiles/0341a.pdf.*

[37] *Auguste Kerckhoffs, 'La cryptographie militaire', Journal des sciences militaires, 9 January 1883, http://www.petitcolas.net/fabien/kerckhoffs/.*

[38] *APACS PIN Administration Policy, January 2004 (APACS), http://www.apacs.org.uk/resources_publications/documents/PIN_Administration_Policy.pdf.*

cryptograms act as a audit log, allowing their authenticity to be established without having to rely on the authorization system.

Validating the ATC

The simplest item to validate is the ATC, which is sent along with each cryptogram. It will therefore be stored in the authorization system, and may also be recorded by the payment system and at the PoS terminal or ATM. The ATC is a number stored by the card, and incremented by one each time a transaction is initiated. Therefore, logs of transactions should show the ATC increasing by one for each transaction, in chronological order of the transaction time. The ATC may pass over values if a transaction is initiated but aborted before the cryptogram is sent to the issuer, but it should never decrease. Large jumps should be viewed with suspicion.

It is important to examine the ATC sequence for both disputed and non-disputed transactions, because if a clone is being used, and a criminal is not very careful, there will be inconsistencies in the pattern. For example, suppose the criminal creates a cloned card and uses it for a transaction. If the ATC produced for this fraudulent transaction is lower or equal to that of the last legitimate one, logs of ATC values would show up a discrepancy. Even if the criminal is able to guess the correct value of the ATC to use, the logs will still show a discrepancy when the customer next uses the legitimate card, unless it happened to leave out a value due to an aborted transaction.

While the authorization system should detect grossly irregular sequences of ATC values, when investigating disputed transactions, it is advisable to perform more rigorous examination of the information than the authorization system would normally perform. This is because criminals will generally attempt to circumvent the fraud detection measures, but no more (so as not to waste effort). If the process of analyzing logs for disputed transactions is merely to repeat the same checks which it would have had to pass in order for the transaction to succeed, no new type of fraud would ever be detected. Authorization systems might also not enforce tight constraints; for example the author has tested cards which have worked despite large gaps in the ATC sequences.

However, the criminal can still circumvent the process if he has access to the legitimate card. First, the criminal uses the cloned card a few times while the customer is not using the legitimate one. Then the thief obtains the legitimate card, increments the ATC the same number of times that he used the cloned one, adds a few more additional ones, and then returns it to the customer. Finally, the criminal can use the cloned card more times, provided that its ATC remains less than the one he set on the legitimate card. In this way, the thief can interleave two groups of fraudulent transactions, without causing disruption to the pattern of ATC values. With more regular access to the legitimate card, the criminal could effect further fraudulent transactions.

EMV cards can, optionally, contain a record of the ATC value when the card last successfully completed online transaction authentication. This value can be used to help detect whether a criminal has incremented the ATC as described above; in such a case, there would be a significant gap between ATC and the last online ATC. Many UK cards have this feature enabled, and it has proved a useful forensic tool. Another optional feature which would be especially useful for investigating disputed transactions is the transaction log. Here, the card maintains a record of recent transactions, and will return the list when requested. Unfortunately, the author is not aware of any UK bank which has adopted this feature.

Even without the optional additions, the ATC is a useful tool in validating transaction logs, and the interleaving of disputed transactions with non-disputed ones, with a consistent ATC pattern, was used by the First Trust Bank as evidence against their customer in a disputed ATM withdrawal case.[39] While the ATC logs are held by the bank (and potentially other parties), the customer can partially validate this information himself, because ATC values are sometimes printed on receipts. Additionally, if the customer has retained the card which his bank states was used for the disputed transactions, he or someone acting for him can read the current ATC value using specially designed software. The author has attempted to do this in three cases so far, but in two the issuer instructed that the card be destroyed (in one case by the customer, and in the other by the bank which had retained the card in an ATM), and in the third, the bank sent a message to the card instructing it to permanently disable itself before the author could obtain access to the card.

Validating the cryptogram

A further item that can be validated is the cryptogram (ARQC or TC or both). First, having a cryptogram

[39] *The author assisted in this case by attempting to read the ATC from the card; however the bank had electronically disabled the card before returning it to the customer. The customer did not take the case further than the bank's internal dispute resolution process.*

contributes towards evidence that it was a Chip & PIN transaction, not fallback. The transaction data which accompanies the cryptogram includes the type of transaction, date, value, etc. as seen by the card, which should be compared against the version that was sent to the issuer. Most important is the IAD, which is generated by the card and incorporates details on whether the PIN was entered correctly, and if the card has detected any unusual activity. This is the only way to verify whether card-holder verification succeeded; because of the wedge attack, the PoS terminal may have been misled. The detailed meaning of the IAD is specific to the issuer, but it generally follows one of the standards produced by Visa, Mastercard, or the EMV consortium.

However, the records of both the IAD and ATC could be manipulated by the authorization, reporting, or logging systems and networks, so they cannot be trusted unless the reliability of these can be assured. But following from the principle of minimizing the trusted computing base, it is possible to eliminate consideration of these systems by validating the authentication code using independently implemented software, based on the public standards for cryptogram generation (such as one written by the author).[40] Checking a cryptogram requires the UDK of the card, which needs to be kept confidential while the card is active, but after the card is cancelled it can be safely disclosed. This is because knowing the UDK of one card is of no assistance in discovering the UDK of another. This key could, for example, be obtained by requesting the HSM, which generates keys for personalizing newly issued cards, to generate a key for just one card. The key can also be validated by checking an ARQC generated by the card (if the customer still holds it), or receipts which show the ARQC or TC.

Nature of disputes

From the above description, it is clear that the complexity of EMV substantially changes the nature of disputes between customers and banks over unauthorized transactions. While the addition of cryptography offers greater resistance to fraud, this also makes it more likely that customers will be denied refunds by their bank. Not many of these cases make it to court in the UK, because the sums the customers claim for are typically a few hundred to a few thousand pounds, and the claimant risks an order to pay costs that can be significant, should they lose. For example, in *Job v Halifax plc*, the disputed transaction was £2,100, but the bank proved their case to the satisfaction of the judge, and Mr Job was ordered to pay £15,000 in costs. Prosecutions also occur, such as that of Jane Badger, who disputed a transaction and was subsequently charged with making a false statement. She was acquitted, but at the time of writing, the bank (Egg) continues to refuse to refund the disputed transaction.

Despite only a few cases making it to court, the consumer rights organization Which? reports that 20 per cent of customers are not refunded after claiming to be the victim of fraud.[41] There are difficulties with the way customers can seek a resolution in respect of disputed transactions. Initially, they have to defer to the bank's internal dispute resolution process, and then consider adjudication by the Financial Services Ombudsman. However, the customer is in a fairly weak position, because neither the bank nor the Financial Services Ombudsman produces the evidence, unless the customer makes a request under the provisions of the Data Protection Act 1998. The Banking Code is also not very helpful. It states that banks are liable for fraudulent transactions, but this only applies if the bank believes the customer has been either negligent nor is acting fraudulently. A common position taken by banks over disputed transactions is that if a transaction is Chip & PIN, and it does not match the standard patterns of known frauds, then the customer is considered liable. However, the criteria banks use for identifying patterns are not subject to public scrutiny, and may vary between banks and individual fraud investigators.

Another frequent problem during disputes is that evidence is destroyed by the time the case is adjudicated or when legal proceedings are initiated. As mentioned above, in the cases where the author has attempted to read the ATC from cards, the bank has requested that this evidence be destroyed. This appears to be standard procedure, but seems to be unwise now that cards can contain useful forensic evidence. Similarly, in the case of *Job v Halifax plc*, the transaction logs which included the ARQC were destroyed by Halifax after 180 days, even though the transactions were in dispute. Since there was only one log of the transaction presented as evidence, any inconsistency which might have existed would not have been detected. While the

[40] *The author wrote this software in order to be able to verify any cryptograms that might have been produced in Job v Halifax plc. It is not, as yet, publicly available.*

[41] *'Fraud victims struggle to get money back: One in five financial fraud victims not reimbursed', Which?, 25 June 2009, http://www.which.co.uk/news/2009/06/fraud-victims-struggle-to-get-money-back-179150.jsp.*

judgment in *Job v Halifax plc* went in the bank's favour, the judge cautioned that in future cases, the fact that a bank destroys evidence may be considered differently by another judge in different circumstances.

Obtaining evidence held by third parties can also be problematic, such as CCTV footage. A common scenario is that upon reporting a disputed transaction to their bank, a customer is immediately refunded. The customer is then satisfied with the outcome, and does not take the case further. Simultaneously, an internal investigation is initiated by the bank, which could take many weeks. If this investigation decides against the customer, the refund will be reversed. At this point, the customer will be motivated to obtain CCTV evidence and logs from third parties, but by this stage they may have been deleted. Even if they still exist, the CCTV owner may only respond to an application by the police, and since April 2007 the police will only investigate if requested to do so by the bank. From the bank's perspective, a case in which the customer has had their refund denied is resolved, so they are unlikely to take any further action.

These problems have led to many customers contacting the press, and stories on Chip & PIN attract high levels of interest from their audiences. Investigative journalists have worked with researchers in order to discover and demonstrate security vulnerabilities. In some cases they have also contacted bank insiders and reformed criminals to ask for assistance. In this respect, the press performs a valuable role by protecting sources from potential recrimination. The media can also be helpful in obtaining refunds for disputed transactions. For example, Barclays refunded Suzanne Lewis £1,400 following the intervention of BBC Watchdog in February 2007.[42]

The Financial Ombudsman has been criticized for accepting assurances from the banks that Chip & PIN cards cannot be cloned.[43] The banks' opinion is based on their experience that criminals have not been caught either using cloned Chip & PIN cards, or exploiting failures in authorization systems. Care must be taken to ensure that such arguments are not circular: if the definition of a cloned card is one which will evade detection by the bank's anti-fraud measures, then of course they will not have been caught by banks. Similarly, in this article, a number of examples of failures in bank computer systems and procedures have been given, that have become public only because either the customer reported the problem, or there was an associated criminal prosecution. Even though these cases are not complete explanations for how cloned smart cards could be produced, it might be that others exist which have not become public, and which could be exploited by criminals.

Conclusion

A theme throughout this article has been that Chip & PIN greatly increases the complexity of banking systems. This helps deter criminals, but also greatly increases the amount of preparation work necessary when disputed transactions involving Chip & PIN are the subject of litigation. The fact that logs, CCTV footage, and other useful information may be destroyed suggests that requests to preserve evidence should be sent and pursued quickly, even if the disputed transaction is initially reversed. For this reason, and so that opportunities to challenge the evidence are not missed, it is also prudent for customers disputing transactions to obtain legal representation early on in their case.

On the technical side of disputes, the complexity of Chip & PIN offers both advantages and difficulties. The extra evidence available can, potentially, help support a particular interpretation, but the technical nature of the evidence is such that it needs greater precision and effort to interpret and analyze. However, for the evidence to be subject to analysis and interpretation, it must be disclosed. In addition, it is also necessary to adduce sufficient information to establish its reliability, and what conclusions may safely be drawn from it. This presents challenges both to litigators and expert witnesses. It is anticipated that this article provides assistance to both these audiences, should they be involved in such a case.

Dr Steven J. Murdoch is a researcher at the University of Cambridge Computer Laboratory. His areas of interest include cryptography, privacy, and banking security. His publications on these topics are available on his website. Steven has also acted as an expert witness in civil and criminal cases involving Chip & PIN.

Steven.Murdoch@cl.cam.ac.uk
http://www.cl.cam.ac.uk/users/sjm217/

[42] *BBC Watchdog, 6 February 2007, 19:00, BBC One.*

[43] *Submission to the Hunt Review of the Financial Ombudsman Service, Foundation for Information Policy Research, 16 January 2008, http://www.fipr.org/080116huntreview.pdf.*

ARTICLE:

PINs, PASSWORDS AND HUMAN MEMORY[1]

By **Wendy Moncur**
and **Dr Grégory Leplâtre**

Introduction

Automated Teller Machines (ATMs) provide access to cash, confidential information and services for service users of all types, cultures and abilities, across the globe. The standard authentication mechanism by which these users gain access to ATMs consists of use of a token (in the form of a bank card) combined with a password known as a personal identity number (PIN), that can be between 4 and 12 digits (for instance, 4 digits are used in the UK, 5 digits are used in South Africa and 6 in France). Yet this mechanism, which is based on knowledge retained by the person, is unsatisfactory. Passwords are easily forgotten. Users deal with this problem by behaving in a manner that reduces the security: by writing down their PIN, or making them all the same, or disclosing them to friends and family.[2] Whilst security administrators may blame the user for the failure in securing their PIN, in reality the method of authentication chosen by banks as a means of authentication at the ATM disregards users' innate cognitive abilities and limitations. Awareness is emerging of the need to design authentication with these abilities and limitations in mind, with researchers at IBM describing this as 'a critical area for research'.[3] NCR, a leading ATM manufacturer, employs experts to address these issues specifically.

This article explores the literature in respect of current Western authentication systems, together with an overview of some of the main authentication alternatives. In considering current and proposed authentication mechanisms, information is drawn from a range of sources, including journal articles, conference proceedings, company reports and sales literature for security products designed for use on mobile telephones and the internet. These mechanisms are examined under the following headings: authentication; pervasive mechanisms and known user behaviours; usability and universal design at the ATM; how we remember, and a taxonomy of alternative authentication methods.

Throughout this article, reference will be made to the term 'usability'. 'Usability' is a quality that can be measured in relation to the ease of use of a computer application. The elements that make up usability comprise the following:

a. Learnability: this tests the ease with which users can complete basic tasks the first time they encounter the computer system.

b. Efficiency: this measures how quickly users can perform tasks once they are familiar with the system.

c. Memorability: this tests how easily users can re-establish competence after a period of not using the system.

d. The number of times users make mistakes are measured, as is the extent of the errors that are made by users.

e. Measures of satisfaction establish if users find the system enjoyable to use, and utility measures whether the system does what the user needs it to do.

Authentication: pervasive mechanisms and known user behaviours

In this section, how people use authentication

[1] *This article is taken from 'Exploring the usability of multiple graphic passwords', a dissertation written by Wendy Moncur and submitted in partial fulfilment of the requirements of Napier University for the degree of Master of Science in Multimedia and Digital Technology (School of Computing, May 2006), which was awarded a Distinction. Wendy Moncur acknowledges the advice and guidance of Dr. Grégory Leplâtre and Dr.Lynne Coventry in exploring the area of password security and usability. A useful text on this topic is Lorrie Faith Cranor and Simson Garfinkel, Security and Usability: Designing secure systems that people can use (2005, O' Reilly Media, Inc.).*

[2] *Anne Adams and Martina Angela Sasse, 'Users are not the enemy', Communications of the ACM, Volume 42, Issue 12, July 1999, 41-46.*

[3] *Clare-Marie Karat, John Karat, Carolyn Brodie, 'Why HCI research in privacy and security is critical now', International Journal of Human-Computer Studies, Volume 63, Issue 1-2, July 2005, 41-46.*

mechanisms and their known behaviours are reviewed across the broad range of prevailing mechanisms, rather than limiting the review to ATMs. Whilst ATMs have very specific requirements for speed, security and usability in a small space, knowledge-based authentication mechanisms in general share common issues, regardless of which form of device the person is required to interact with. It is pertinent that an account can be viewed through an ATM, the internet, the counter at a bank, and from point of sale machines in retail outlets. A different authentication mechanism might be required for each avenue by which a person can obtain access to their account. This can cause problems for the person, such as *memory interference*, which in turn causes *memory confusion*, which in turn leads to insecure behaviour towards ATM security.[4] By understanding the broader picture, it is possible to more fully understand the behaviour of users when interacting with authentication mechanisms, which in turn helps to formulate future research.

Context

ATMs provide access to cash, confidential information and services to consumers across the globe. Machines may be located indoors or outdoors, in a wide range of climates, but are usually in a public place. The banks authenticate a customer with the use of a token, in the form of a bank card, and a PIN.

Mechanism

Current user authentication systems are based on the knowledge of the user, yet this approach is known to be error-prone.[5] Knowledge-based authentication systems require selection of a strong password to resist attack. Ideally, a password should consist of a set of eight or more randomly allocated characters, incorporating upper and lower case characters, digits and special characters.[6] Yet this is extremely difficult for users to remember. Similarly, a numeric password – or PIN number – made up of a series of digits appears challenging for users to remember, and can be easily changed to a code that is easy to crack. In addition, up to 50 per cent of users write down their PIN number and store it in close proximity to the matching bank card.[7]

An extra layer of complexity is introduced by the variety of mechanisms applied to a single account, depending on how the customer is required to obtain access to the account. For example, one leading UK bank requires customers to use four different mechanisms to obtain access to the same account:

At the bank: presentation of the bank card, plus a manuscript signature, is considered adequate.

ATM: presentation of the bank card and entry of the correct PIN number.

Internet: a combination of customer number, a random selection of digits from the user identification code (this is different to the PIN number or the account number) plus 'secret' personal information.

Making a payment via the internet: in addition to the above, the customer must insert their bank card into a separate card reader supplied by the bank, enter their PIN, then enter the random number displayed on the card reader into their internet banking session.

From the bank's perspective, this may be simple to administer, but from a usability perspective, it is doubtful that it provides user satisfaction. For instance, in a study of Canadian banks, on-line banking customers reported that they found documented security practices confusing and extremely difficult to comply with.[8]

People

People involved in the use of authentication systems can be divided into three groups: administrators, legitimate users and unauthorized users.

Administrators understandably want to protect 'their' systems from attack by unauthorized users. It is common for security departments in organisations not to have any contact with their legitimate users, and to fail to communicate with them. Such detachment can lead to a failure to understand users' needs and

[4] *Anne Adams and Martina Angela Sasse, 'Users are not the enemy', Communications of the ACM.*

[5] *Rachna Dhamija and Adrian Perrig, 'Deja vu: A user study using images for authentication', Proceedings of the 9th conference on USENIX Security Symposium, Volume 9 (Denver, Colorado, 2000).*

[6] *Susan Wiedenbeck, Jim Waters, Jean-Camille Birget, Alex Brodskiy and Nasir Memon, 'Authentication using graphical passwords: effects of tolerance and image choice', Proceedings of the 2005 symposium on usable privacy and security, (Pittsburgh, Pennsylvania) (ACM International Conference Proceeding Series, Volume 93, 2005), 1-12.*

[7] *Anne Adams and Martina Angela Sasse, 'Users are not the enemy', Communications of the ACM, cited by Karen Renaud and Antonella De Angeli, 'My password is here! An investigation into visuo-spatial authentication mechanisms', Interacting with Computers, Volume 16, Issue 6, December 2004, 1017-1041.*

[8] *Mohammad Mannan and P. C. van Oorschot, 'Security and Usability: The Gap in Real-World. Online Banking', Proceedings of the 2007 New Security Paradigms Workshop, September 2007, available from http://www.scs.carleton.ca/~paulv/papers/pubs.html.*

objectives, resulting in the creation of unusable security systems. It may seem reasonable to an administrator to create a system that enforces regular password changes, because users are unlikely to change them otherwise. Yet the same system may be perceived as a hindrance by users, because frequent password changes add to the burden placed on human memory, known as 'memory burden'. Similarly, administrators may assign individual passwords when it may be more appropriate to share a common password across an organisational unit. A parallel may be found in joint bank accounts, where each card holder is given a different PIN or password to the same account.

As a result of security being forced upon them by the organisation rather than tailored to their needs, users do not consider themselves to be accountable for security as much as the bank might wish them to be. When the purpose of the security mechanism is unclear or inappropriate, the motivation to comply is weakened, thus eroding the culture that ought to accompany security. The system may be perceived as an obstacle to 'real' work, which therefore must be circumvented.[9] The failure is compounded when security administrators advise users to write down their passwords: this is a clear indication that the authentication mechanism is not viable.

The error rate, a measure of usability, may be high because legitimate users have difficulty remembering passwords that are less vulnerable to cracking – known as 'strong' passwords. While 94 per cent of users can remember semantic passwords, they can remember syntactic passwords only 35 per cent of the time.[10] The 'Power Law of Forgetting' explains that not only do individuals forget a great deal quickly, but their memory is further eroded over time.[11] For a password to be remembered without resort to an insecure aide-memoir of some form (usually by writing it down), it must be encoded within long term memory. To achieve this, the password must be meaningful or easy to work out, practiced, based on information that is personal to the user that is already familiar, and incorporated into a special memory scheme. It is difficult to apply these criteria to strong passwords.

Given that most people in the twenty-first century are required to remember a number of passwords for different purposes, users may mix up which password applies to which authentication mechanism. This phenomenon is known as 'interference'. At one industry site, where users had 16 different passwords for use within the organisation, it was extremely common to forget passwords, or mix up which password was used for which system, because of intra-password interference.[12] An added layer of confusion is also generated when different rules for the creation of passwords are enforced in different systems.

Standing at an ATM in a public place and observed by others, users may feel pressurised when trying to recall their password. This pressure can itself adversely affect recall. Failure to recall their password can generate embarrassment in the user when they are observed by others, but it may also generate a suspicion of wrong-doing amongst observers.[13] Further pressure may be added if a user is aware of the risks of being observed either directly or by a hidden camera when entering their password: a phenomenon known as 'shoulder-surfing'.

The card holder may give another person authority to use their card and PIN. The authority may be given expressly or by implication, such as how they deal with their card within the family, for example. A person that has authority to use the card and PIN is not expected to use the card beyond the authority given to them by the card holder. When a user shares their password with a colleague, friend or family member, they are in effect sharing the method by which they are authenticated and, by extension, authorisation to use the facilities that may be afforded to the user when using an ATM, for instance. Depending on the terms and conditions of use, the organisation operating the authentication mechanism may consider this practice to be insecure. However, it may be reasonable from the perspective of the user to share their authentication details with others. It is certainly extremely common – 36 per cent of users admit to sharing their PIN with someone,

[9] *Susan Wiedenbeck, Jim Waters, Jean-Camille Birget, Alex Brodskiy and Nasir Memon, 'PassPoints: Design and longitudinal evaluation of a graphical password system', International Journal of Human-Computer Studies, Volume 63, Issue 1-2, (July 2005), 102-127.*

[10] *Moshe Zviran and William J. Haga, Cognitive passwords: The key to easy access control, Computers & Security (1990) 9(8), 723-736, cited by Karen Renaud and Antonella De Angeli, 'My password is here! An investigation into visuo-spatial authentication mechanisms', Interacting with Computers, Volume 16, Issue 6, December 2004, 1017-1041.*

[11] *Harry P. Bahrick, Semantic memory content in permastore: Fifty years of memory for Spanish learned in school, Journal of Experimental Psychology: General, Volume 113(1), March 1984, 1-29 cited by Susan Wiedenbeck, Jim Waters, Jean-Camille Birget, Alex Brodskiy and Nasir Memon, 'Authentication using graphical passwords: effects of tolerance and image choice', Proceedings of the 2005 symposium on usable privacy and security.*

[12] *Anne Adams and Martina Angela Sasse, 'Users are not the enemy', Communications of the ACM.*

[13] *Martina Angela Sasse, Sacha Brostoff and Dirk Weirich, 'Transforming the weakest link - a human/computer interaction approach to usable and effective security', BT Technology Journal, Volume 19, Issue 3, 2001, 122-131.*

although the real percentage is likely to be higher. Some users even view it as desirable that they share their authentication details.[14] The card holder authorises the other person to act within the scope of the authority grated by the person whose password is used. For instance, the woman manages the finances in 70 per cent of households, thus it may not be acceptable for her to refuse to tell her partner the PIN for the family bank account. Disabled users may be unable to obtain access to the ATM because of cognitive or physical limitations. By necessity, they may need to divulge their PIN to enable a helper to obtain access to the ATM with their authority. Using knowledge-based authentication mechanisms that are easily communicated verbally or in writing means it is difficult to prevent passwords from being shared.

People posing as legitimate users can exploit poor usability. When organisations routinely ask users for their password in order to resolve usability difficulties with the security mechanism, they open the way for malign users to trick legitimate users into divulging their passwords. The poor design of a security system often requires users to share their passwords with an administrator. This in turn enables an attacker to use various social engineering techniques to trick legitimate users into divulging their passwords. An ATM with a mouse or clearly displayed fixed position keys, facilitates shoulder-surfing, sometimes known as 'observer attack'.[15] Software applications that help thieves are common, but the threat is poorly understood by legitimate users. Multilingual dictionary attack software can break 85 per cent of passwords through a simple exhaustive search. Rule-based attacks extend this capability further, altering known words according to a number of rules, for example searching on words written backwards.[16]

Design for ATM users

Characteristics of ATM users

There is no such thing as a 'typical' user at the ATM. Young, old, able-bodied and disabled may all wish to use ATMs. Yet ability and disability are not distinct categories. A wide range of physical, cognitive and sensory abilities may be displayed by users. Abilities may fluctuate, affected by circumstance at any given point in time. This makes it difficult to provide a universal design. For example, consider an elderly person who is holding a number of bags of shopping. Without the bags of shopping, they are normally able-bodied and alert, yet when carrying the bags of shopping, they are temporarily 'disabled'. Their physical function is hampered by tangible external constraints, as they juggle shopping bags whilst entering their PIN. Furthermore, older adults generally find it more difficult to use computer software than younger people.[17]

Usability of current mechanisms

As previously discussed, the prevailing knowledge-based authentication mechanisms are not very good. Users are generally not very satisfied with them, and there are high error rates. Further problems include difficulty in memorising passwords and learning how to use authentication systems.

Knowledge-based authentication mechanisms place a requirement on people that passwords be recalled correctly, otherwise the user will fail to obtain access to the service or system. People find it difficult to remember passwords, because not enough consideration is given to the difficulty that people experience in recalling passwords that are not frequently used. The cognitive ability to remember imprecisely is not taken into account. The policy of many knowledge-based mechanisms is that if the user fails to key in the correct password on the third attempt, they are denied access to the service or system, regardless of a realistic security appraisal of security requirements.[18] When an account holder wants to check the balance on a seldom-used account without withdrawing cash, for instance, it is a matter of debate as to whether their bank card should be revoked if they get one digit wrong on the PIN number.

When a user chooses a new password, they find it very difficult to learn the new password if it is created in the random way that is usually required. The lack of understanding of what constitutes a strong password can lead to the creation of weak passwords that are vulnerable to attack. It is rare for training to be provided on how to create a secure password. Without

[14] *Rachna Dhamija and Adrian Perrig, 'Deja vu: A user study using images for authentication', Proceedings of the 9th conference on USENIX Security Symposium.*

[15] *Volker Roth, Kai Richter and Rene Freidinger, 'A PIN-entry method resilient against shoulder surfing', Proceedings of the 11th ACM conference on Computer and communications security, (Washington DC, USA) (ACM, 2004), 236-245.*

[16] *Rachna Dhamija and Adrian Perrig, 'Deja vu: A user study using images for authentication', Proceedings of the 9th conference on USENIX Security Symposium.*

[17] *Ann Chadwick-Dias, Michelle McNulty and Tom Tullis, 'Web usability and age: how design changes can improve performance', Proceedings of the 2003 Conference on Universal Usability, (ACM Conference on Universal Usability, 2003), 30-37.*

[18] *Sacha Brostoff and Angela Sasse, '"Ten strikes and you're out": Increasing the number of login attempts can improve password usability', Workshop on Human-Computer Interaction and Security Systems part of Conference on Human Factors in Computing Systems 2003, 5-10 April 2003, Fort Lauderdale, Florida.*

appropriate help and advice when a weak password is selected, the user's inaccurate understanding of security remains uncorrected, and security is undermined. This means that the utility of the mechanism that is designed to help provide for security is poor. This in turn reflects on the inability of those people responsible for security to understand that the very measures they have implemented are not secure, and illustrates their failure to design truly secure systems.

Conflict between security and usability

The need to provide for security whilst also providing a system that is relatively easy to use can be in conflict. For an authentication mechanism to be secure, it must use passwords that are secret, and there must be a very wide range of possible passwords available for users to use. For example, if a password can only be two characters in length, it will be far easier to guess than one that is six characters long. Further, the security of a password should be rated for the ability of a potentially malign user to observe the code, guess the code and record the code. Users can be perceived by administrators as undesirable, because they undermine secure systems. They allow their passwords to be observed. They create weak, guessable passwords. They record them in obvious places. Users feel justified in adopting such insecure, apparently careless, behaviour by systems that are poor to use, and that seem inadequately matched to user needs. In contrast, part of the success of hackers can be attributed to the attention that they pay to users and their behaviour.[19]

Users do have a shared purpose with the system and its administrators, but a perceptual divide exists. It is in the interests of both administrators and users to protect data. Unfortunately, there is little dialogue or shared understanding between the two groups regarding what security needs really exist. System designers follow a simplistic approach to security. Users are thus subjected to an impossible burden to memorise complex passwords, forced to remember multiple passwords, and often given no guidance on what makes a password secure.[20]

Reducing conflict

If authentication mechanisms are to become more usable, they must incorporate usability considerations from the start. Knowledge-based systems could use a single sign-on for a whole system and also reduce forced changes, thus reducing the burden on memory. Alternative approaches to authentication, such as graphical or biometric authentication, are also possible.

A shift of emphasis needs to occur. Rather than users and security designers being in conflict, a partnership might be promoted, with an emphasis on shared aims. Communication is essential. By involving users from the start, security system designers can understand users' needs and develop systems that are compatible with their requirements. The degree of security can be appropriate to the risk. Systems can be evaluated for usability with users before being implemented. Cost savings can be made by paying attention to the needs and abilities of users. As passwords are forgotten less, there is less need to spend corporate time and effort resetting them. Moreover, satisfied users are more likely to abide by the security rules.

The value of training users to create strong passwords is disputed: Sasse and her colleagues found that user training increased usability of security systems,[21] yet Yan and others found that this did not significantly improve the strength of the passwords that were created.[22] Online guidance during the process of creating a password may be a better approach. This method is now used by some online service providers. For example, GoogleMail provides real-time feedback to show if a password is weak, fair or strong. The use of colour coding and a bar chart helps to reinforce the comments offered in respect to a password used by a customer. The security software reviews the proposed password as the user types it in. For example, in the process of typing the password 'secret12', the following comments are provided:

'secret1' is reported as 'too short'

'secret12' is reported as 'fair'

'Secret12' is reported as 'strong', because of to the inclusion of an upper case letter

Further advice is given to users on creating a strong password if the user clicks on a link on the same web page, entitled 'Password strength'. Whilst such online guidance forces users to choose a strong password, it

[19] *Anne Adams and Martina Angela Sasse, 'Users are not the enemy', Communications of the ACM.*

[20] *Martina Angela Sasse, Sacha Brostoff and Dirk Weirich, 'Transforming the weakest link - a human/computer interaction approach to usable and effective security', BT Technology Journal.*

[21] *Anne Adams and Martina Angela Sasse, 'Users are not the enemy', Communications of the ACM; Martina Angela Sasse, Sacha Brostoff and Dirk Weirich, 'Transforming the weakest link - a human/computer interaction approach to usable and effective security', BT Technology Journal.*

[22] *Jianxin Yan, Alan Blackwell, Ross Anderson and Alan Grant, 'The Memorability and Security of Passwords – Some Empirical Results', Technical report No. 500 (Cambridge University Computer Laboratory, 2000).*

can also be used to provide the rationale for security, and enable the user to understand what makes up a strong password.

How we remember

Limitations in the way people remember make it difficult to recall alphanumeric passwords. The 'Power Law of Forgetting'[23] describes how an individual may experience rapid forgetting immediately after learning, followed by a further gradual decay. Over time, recall becomes progressively more inaccurate. This inaccuracy is a particular problem when password authentication demands total accuracy. Retroactive interference, where new additions to memory disrupt existing memories, adds to the problem. It may be inferred that multiple passwords are particularly prone to retroactive interference.

Recall is easier when distinct items are familiar, and when they are associated with each other. Mnemonics are useful in exploiting this feature of memory. Amongst the mnemonic systems of memorization, two commonly used techniques are loci and PegWords. Using the loci technique, locations serve as retrieval cues for the information being recalled. PegWords entail learning a series of words that serve as 'pegs' on which memories can be hung. Both techniques have been used in trials of authentication mechanisms.[24]

In contrast to their imperfect ability to recall, 'humans have a vast, almost limitless memory for pictures' – an ability known as the 'Picture Superiority Effect'.[25] Unlike recall, picture recognition appears to be relatively unaffected by the process of ageing. There remains some debate as to why pictures are significantly easier to remember than words. The Dual-code theory suggests that the advantage springs from the brain remembering pictures simultaneously in two different ways, using an image code and a semantic code. Another possibility is that pictures generate a more detailed memory, and are thus easier to extract from long term memory.[26] Regardless of this debate, the brain has a proven greater capacity to store and recognise pictures over letters and numbers. This makes images an excellent candidate for authentication mechanisms, especially given the acknowledged failings of current knowledge based authentication mechanisms.

Alternative authentication mechanisms

Despite the prevalence of knowledge-based authentication at the ATM and online, research continues into alternatives, such as graphical authentication and mechanisms that monitor an individual's behaviour and include additional security checks when their behaviour deviates from its normal pattern. The use of biometric measurements is a further option, but is outside of the scope of this article. Further details on biometric measurements can be found in the dissertation from which this paper is drawn.[27] No mechanism is perfect: all have limitations. The ideal password should be easy to remember but hard to guess. As Renaud comments, 'any authentication mechanism teeters between memorability and predictability requirements'.[28]

Personalisation and behaviour

An emergent trend in security mechanisms is the tracking of patterns of user behaviour. If a user who travels relatively little takes a holiday abroad, they may be surprised and inconvenienced to find their card disabled, when security software notices an abnormal transaction. Conversely, the same user may be relieved when the system prevents their account from being used by a thief. Sasse highlights the desirability of determining the method of authentication after taking into account the nature of the task.[29] For instance, a user could be asked for less stringent authentication if they are performing a task that fits their normal behaviour pattern, and a more stringent method of authentication if their behaviour is unusual. For example, a customer consistently withdraws £20 from her local ATM. If she tries to take out £200 from a different ATM, she will be prompted for extra authentication. This approach is consistent with current societal trends that demand speed and an individual approach from technology services. They may provide increased customer loyalty as a result. The approach does not provide

[23] *A term first used by J. R. Anderson and L. J. Schooler, 'Reflections of the environment in memory', Psychological Science, 2, 1991, 396-408.*

[24] *Jianxin Yan, Alan Blackwell, Ross Anderson and Alan Grant, 'The Memorability and Security of Passwords – Some Empirical Results'.*

[25] *Antonella De Angeli, 'Pictorial security at the ATM: the visual identification protocol', Advances 2002: annual research report for NCR Self-service strategic solutions, Volume 1, 2002, 161-169.*

[26] *Antonella De Angeli, Lynne Coventry, Graham Johnson and Karen Renaud, 'Is a picture really worth a thousand words? Exploring the feasibility of graphical authentication systems', International Journal of Human-Computer Studies, Volume 63, Issue 1-2, (July 2005), 128-152.*

[27] *Wendy Moncur, 'Exploring the usability of multiple graphic passwords', MSc Dissertation, retrieved from http://www.csd.abdn.ac.uk/~wmoncur/publications/Exploring%20the%20usability%20of%20multiple%20graphical%20passwords.doc, 2006.*

[28] *Karen Renaud and Antonella De Angeli, 'My password is here! An investigation into visuo-spatial authentication mechanisms', Interacting with Computers, Volume 16, Issue 6, December 2004, 1039.*

[29] *Martina Angela Sasse, Sacha Brostoff and Dirk Weirich, 'Transforming the weakest link - a human/computer interaction approach to usable and effective security', BT Technology Journal.*

The effects of interference on memorability when a user has many separate graphical passwords are small in comparison to that exhibited when a user has multiple knowledge-based passwords.

authentication in itself. It must be combined with other approaches, such as knowledge-based or graphical authentication before it can be assessed for memorability and predictability.

Graphical authentication mechanisms

Graphical authentication mechanisms use picture recognition to authenticate the user. This is based on the understanding that people remember images far better than words.[30] They can remember more images, more accurately, and with less adverse effects from the process of aging. An attractive feature of graphical authentication mechanisms is that they are harder to disclose than PINs and semantic or syntactic passwords. Research into the potential of graphical authentication continues.

Graphical authentication systems have their own requirements relating to usability. It is recognised that images used must be concrete, nameable and distinct.[31] Each image displayed must be visually dissimilar, and from a separate semantic category. For example, if a user is shown ten images, ideally they should each be from a separate category such as transport, mammals, faces, architecture, to prevent the user from mixing images of a similar nature. Conceivably, they should not be shown two pictures of the same item, such as flowers on the same screen, because this can cause confusion. The effects of interference on memorability when a user has many separate graphical passwords are small in comparison to that exhibited when a user has multiple knowledge-based passwords.[32]

Immaturity of authentication systems

Successful authentication mechanisms should provide a balance between security and usability. Yet no current authentication mechanism fits all the requirements: every mechanism has its failings. Providers of secure systems must look to their users, and to understand and accept their innate cognitive and physical limitations, if they are to create authentication mechanisms that are usable, and thus less flawed. In the meantime, users will continue to exhibit insecure behaviours, circumventing the best intentions of secure systems.

Wendy Moncur is a PhD candidate, working between the Universities of Aberdeen and Dundee. She was awarded an MSc with Distinction in Multimedia and Interactive Systems in 2006. She has seventeen years experience of Information Technology in industry, including ten years working as a database design specialist for a leading UK bank and European financial services institutions.

Dr Grégory Leplâtre is a lecturer at Edinburgh Napier University. His research interests lie in interface design and human-computer interaction, and include collaborative work with NCR, a global manufacturer of ATMs and point-of-sale systems.

[30] *Ian Jermyn, Alain Mayer, Fabian Monrose, Michael K. Reiter and Aviel D. Rubin, 'The Design and Analysis of Graphical Passwords', Proceedings of the 8th USENIX Security Symposium, 1999, available at http://www.usenix.org/events/sec99/full_papers/jermyn/jermyn_html/.*

[31] *Antonella De Angeli, Lynne Coventry, Graham Johnson and Karen Renaud, 'Is a picture really worth a thousand words? Exploring the feasibility of graphical authentication systems', International Journal of Human-Computer Studies.*

[32] *Wendy Moncur and Grégory Leplâtre, 'Pictures at the ATM', Proceedings of the SIGCHI conference on Human factors in computing systems, 2007, 887-894.*

ARTICLE:

KNOWN KNOWNS, KNOWN UNKNOWNS AND UNKNOWN UNKNOWNS: ANTI-VIRUS ISSUES, MALICIOUS SOFTWARE AND INTERNET ATTACKS FOR NON-TECHNICAL AUDIENCES

By **Daniel Bilar**

Introduction

The risks associated with the internet have changed significantly. A recent study claims that a typical Microsoft Windows machine is subjected to autonomous infiltration attempts – not just mere pings and probes – from worms and botnets looking for clients once every six minutes.[1] Stealth – not exhibitionism or hubris – characterizes this breed of attacks and concomitantly deployed malicious software. Unbeknownst even to experienced human operators, surreptitious attacks are able to insert malicious code deep within the bowels of individual computers and the wider supporting internet communication and control infrastructure such as wireless access points, home routers, and domain name servers.[2] In addition to stealth, social engineering via e-mail, Instant Messaging, and social networks plays an important part, as well: unsuspecting users are coaxed to initiate actions that infect their computers and usurp their digital identities.

These attacks are powerful because of the havoc that is causes to the owner or user of the computer or computer network. The effects range from mere nuisance, to the appropriation of sufficient information to impersonate an individual, that can, in turn, lead to financial ruin, up to and including criminal charges against the innocent.[3] They are also powerful because, in many instances, neither individual computer owners, nor the sophisticated network controlled by a government can prevent all of the malicious code from penetrating their computers or networks.

In the past, the actions of hobbyists and isolated mischief makers merely caused disruptions. Now organized and highly technically competent criminals with financial incentives as the primary motivator have taken over. In addition, semi-independent state-sponsored groups occasionally launch attacks on another state. The ramifications of this shift are worrisome: that a person may subscribe to an anti-virus software product from one of the many vendors on the market does not mean that their computer is protected from or necessarily free of malicious software.

Modern malicious software has been shown in tests carried out in independent laboratories to be highly resistant to being identified by anti-virus (AV) products. In addition, these empirical results are consistent with theoretical findings, in that detecting complex malicious software is beyond the effective modelling capabilities of current AV products,[4] and as such is becoming

[1] *Gabor Szappanos, 'A Day in the Life of An Average User', Virus Bulletin, January 2009, 10-13, available at http://www.virusbtn.com/.*

[2] *Most users do not bother to change the default passwords on home devices such as routers. Browser vulnerabilities can then be exploited by malicious software to alter the DNS settings of the router, thereby directing any name lookup query to a DNS of the attacker's choice. This may be used to spoof a bank web site, for instance. See Sid Stamm, Zulfikar Ramzan and Markus Jakobsson, 'Drive-By Pharming', Lecture Notes in Computer Science 4861, (Springer, 2007), 495-506 and Hristo Bojinov, Elie Bursztein, Eric Lovett and Dan Boneh, 'Embedded Management Interfaces: Emerging Massive Insecurity', Blackhat Technical Briefing, Blackhat USA 2009 (Las Vegas, USA, August 2009), available at http://www.blackhat.com/presentations/bh-usa-09/BOJINOV/BHUSA09-Bojinov-EmbeddedMgmt-PAPER.pdf.*

[3] *For examples of people charged with offences, see Patrick v Union State Bank, 681 So.2d 1364 (Ala. 1995); Vic Lee, 'ID Theft Puts Innocent Man In San Quentin', 21 February 2007, ABC7News, available at http://abclocal.go.com/kgo/story?section=news/local&id=5052986 and Mary Pat Gallagher, 'Identity-Theft Victims Owed Duty of Care in Bank Fraud Investigations, N.J. Court Says', Law.com, 11 September 2008, available at http://www.law.com/jsp/article.jsp?id=1202424426977.*

[4] *Yingbo Song, Michael E. Locasto, Angelos Stavrou, Angelos D. Keromytis and Salvatore J. Stolfo, 'On the infeasibility of modelling polymorphic shellcode,' Proceedings of the 14th ACM conference on Computer and Communications Security, 2007, 541–551.*

increasingly difficult to detect in practice, and worryingly, also in principle.[5] To put it simply, anti-virus software does not prevent all forms of malicious software from penetrating computers and networks – some malicious software will not be identified by anti-virus software, which is why this is an important topic for lawyers and judges to understand.

The aim of this article is to introduce the technical issues surrounding modern internet attacks, anti-viral software and malicious software to the individual that has no technical knowledge, and who needs a working understanding of the pertinent issues. As such, its primary goal is to raise awareness, not comprehensiveness. The interested reader is referred to a recent book by Markus Jakobsson and Zulfikar Ramzan, *Crimeware. Understanding New Attacks And Defenses*, (Symantec Press, 2008) for further study.

Software vulnerabilities

Coding errors[6] in software can lead to vulnerabilities. Software vulnerabilities are program weaknesses which malicious software can exploit. The relationship between coding errors, vulnerabilities and exploitation is illustrated by the following analogy: the US Tariff Act of 1872 was to include a list of duty-free items: Fruit plants, tropical and semi-tropical. A government clerk duly transcribed the Act, but erroneously moved the comma: Fruit, plants tropical and semi-tropical. Shrewd businessmen argued that the law, as promulgated, exempted all tropical and semitropical plants from duty fees, resulting in $500,000 loss to the US Treasury.[7] For the purposes of this discussion, the erroneous placement of the comma is the equivalent of a software coding error. The vulnerability resulting from this error manifests itself as an opportunity for alternative interpretation, and the exploit is represented by cleverly taking advantage of duty-free imports of tropical and semi-tropical plants.

Since errors in software coding errors permit malicious exploitation, it seems obvious that efforts should concentrate on writing error-free code. Unfortunately, industrial software has exhibited the same code error density for the past twenty years; on average six faults (errors) for every thousand lines of source code.[8] However, the general increases in the amount of code (Windows Vista has an estimated 80 million lines, whereas Windows 2000 had 35 million lines), as well as the complexities of modern software (interactions between components and protocols, as well as very large applications like Adobe Acrobat Reader with 2 million lines of code) have exacerbated the situation.

The survival time of an unpatched Windows system may serve as corroborating evidence.[9] In 2003, an unpatched Windows PC would last approximately 40 minutes on average, before it would succumb to probes from (presumably) malicious software. In 2004, survival time was reduced to 16 minutes and by 2008, the time window had shrunk to mere 4 minutes.[10]

Just as a motor car needs regular tune-ups to keep running smoothly, maintenance of installed software is performed through regular updates. Since software vulnerabilities are the root cause of many malicious software infections, updating (or equivalently patching) minimizes the number and severity of software vulnerabilities that malicious software may exploit. The poor quality of software code explains in part why anti-virus software is required in the first place (another factor is ubiquitous connectivity). Anti-virus software, however, has problems of its own.

The problem with anti-virus software

Most commercial AV products rely predominantly on some form of signature matching to identify malicious code. In the context of AV, a signature is the rough software equivalent of a fingerprint – it is a pattern that identifies malicious software. It is possible to derive a pattern from software code, that is, a static snippet of code (or a uniquely reduced version of it, such as a hash). The fragment is taken as the pattern that identifies the code. A static signature is, in its simplest incarnation, a fixed sequence of characters somewhere

[5] *Grégoire Jacob and Eric Filiol and Hervé Debar, 'Malware as interaction machines: a new framework for behavior modelling,' Journal in Computer Virology, Volume 4, Number 3, August 2008, 235-250.*

[6] *For an overview of such errors, see Katrina Tsipenyuk, Brian Chess and Gary McGraw, 'Seven Pernicious Kingdoms: A Taxonomy of Software Security Errors', IEEE Security and Privacy, Volume 3, Issue 6, (November 2005), 81-84.*

[7] *See 'Forty-Third Congress; First Session Feb. 20', New York Times, February 21, 1874, at http://query.nytimes.com/mem/archive-free/pdf?res=9902EFD8173BEF34BC4951DFB466838F669FDE.*

[8] *Compare John Musa, Software Reliability Measurement Prediction Application (McGraw-Hill, 1987) with Parastoo Mohagheghi and Rediar Conradi, 'An empirical investigation of software reuse benefits in a large telecom product', ACM Transactions on Software Engineering Methodology, Volume 17, Issue 3 (June 2008), 1-31.*

[9] *A patched system denotes a computer on which the latest software updates (normally for the Operating System, but also for Office suites and media software) have been installed.*

[10] *See John Leyden, 'Unpatched Windows PCs own3d in less than four minutes', The Register, 15 July 2008 at http://www.theregister.co.uk/2008/07/15/unpatched_pc_survival_drops/ and Survival Time at http://isc.sans.org/survivaltime.html.*

in a file or in memory and may look something like this:

```
C3 7C FD 1D 31 C0 6F OF 96 18 A4
```

The rationale underlying these character patterns is that they are more likely to be encountered when analyzing malicious software rather than innocent programs. Hundreds of thousands of these signatures are stored in local AV databases (AV signature updates are received, hopefully, at least once a week). An AV scanning engine then tries to match pre-defined file areas against this signature database. These areas are typically located at the beginning and the end of the file, and after what is called the executable entry point of a program.

Strict matching of the byte sequence pattern was most popular in the early 1990s. This method has since been augmented, because those responsible for writing malicious code took action to avoid being noticed by the AV products. They approached their evasion in a straightforward way. Because of time constraints (users tend not to wait more than a couple of seconds), it is not usual to scan the whole file. Malicious authors took advantage of this fact and moved the malicious code to locations in the file that would probably not be scanned. Furthermore, they tweaked their malicious code to make the byte pattern mismatch. One way of doing this is by equivalent instruction substitution. An example will illustrate this point. In the signature above, the substring pattern `31 C0` represents Intel machine code `xor ax, ax`. Its purpose is to set register `ax` to `0`. A substitution that preserves this functionality would replace the substring with `29 C0` (which is machine code for `sub ax, ax`) or `B8 C0 00` (which is machine code for `mov ax,0`).

Generic matching was introduced to add some 'fuzziness' to the signature in order to catch malicious software that is slightly altered so as to evade the stricter matching. Using the example above, the second, third, fourth and ninth bytes are replaced with a wildcard (a 'blank', do-not-care byte) denoted by '??':

```
C3 ?? ?? ?? 31 C0 6F OF ?? 18 A4
```

When searching for this pattern, the '??' directs the AV scanner to ignore whatever byte value is present in the second, third, fourth and ninth bytes of character strings it encounters while scanning the file. For example the string below:

```
C3 99 A0 BB 31 C0 6F OF 77 18 A4
```

would match, as well as:

```
C3 A1 22 00 31 C0 6F OF FF 18 A4
```

Hence, wildcards try to lower AV false negative detection rates by 'softening' the signatures to counteract some of the evasive coding tactics that malicious software is programmed to use to avoid detection. The problem with casting a wider net to catch 'bad' programs is that 'innocent' (that is non-malicious) programs may be identified incorrectly; in other words, there is an increase in the false positive rate.

For an accessible overview of more AV signature detection enhancements, the reader is encouraged to peruse chapter 11 of Peter Szor, *The Art of Computer Virus Research and Defense*, (Addison Wesley, 2005).

Static signatures, as we have discussed them so far, are derived from program code, reflecting the byte value make-up of a program. Malicious software detection at the beginning of the twenty-first century started to incorporate *behavioural heuristics* approaches; that is, a notion of how a given software program interacts with its embedded environment. For instance, a program may interact with a file system (by opening, creating or deleting a file), or the network (opening a connection to a server or setting up a receiving server). These and other interactions of the program can be monitored in what is called a 'sandbox'. A sandbox is a controlled, instrumented container in which the program is run and that records how it interacts with its environment. A sample sandbox output is set out below:

```
[ General information ]
* Display message box (sample) : sample, tikkun olam!
* File length: 18523 bytes.
* MD5 hash: 1188f67d48c9f11afb8572977ef74c5e.
```

Here some general information about the file (its length and its hash) is made visible, together with what is displayed on screen (a message box with the caption 'sample' and message 'tikkun olam!'). The next phase is for the malicious software to carry out instructions to delete a file and place a substitute file in place of the file that has been deleted:

```
[ Changes to filesystem ]
* Deletes file C:WINDOWS\SYSTEM32\kern32.exe.
* Creates file C:WINDOWS\SYSTEM32\kern32.exe.
```

Here we see that the first action of the program is to delete a file and recreate one with the same name, kern32.exe. This is suspicious. Then it is necessary to enter the internal Windows database (the Windows registry). This is illustrated below. This entry makes the file kern32.exe run when system startup begins as the computer is switched on:

```
[ Changes to registry ]
* Creates key
"HKLM\Software\Microsoft\Windows\CurrentVersion\RunOnce".
* Sets value "kernel32"="C:WINDOWS\SYSTEM32\kern32.exe -
sys" in key " HKLM\Software\ Microsoft\Windows
\CurrentVersion\RunOnce".
```

This is very suspicious behaviour, in that the system is instructed to intercept the strokes used on the keyboard and pass it on to a custom function:

```
[ Changes to system settings ]
* Creates WindowsHook monitoring keyboard activity.
```

There follows the network activity: the program connects to a server at address 110.156.7.211 on port 6667, a typical port for Internet Relay Chat (IRC) chat server, logs in and joins a chat channel:

```
[ Network services ]
* Connects to "110.156.7.211" on port 6667 (TCP).
* Connects to IRC server.
* IRC: Uses nickname CurrentUser[HBN][05].
* IRC: Uses username BoLOGNA.
* IRC: Joins channel #BaSe_re0T.
```

In the example above, interactions occur with the file system, the Windows registry (the internal Windows database) and the establishment of a TCP network connection to an IRC chat server. Connecting to a chat server is anomalous enough behaviour that it should raise a concern that something is not correct. Taken together, this set of activities is consistent with the suspicious program being a bot, connecting to a botnet through the IRC server.

Thus, behavioural heuristics seek to establish an 'activity' profile. It is also possible to derive a 'behavioural signature' from such an activity profile (as opposed to the byte-value approach discussed earlier).[11]

Just as there are different ways of rewriting instructions (as seen with the `xor ax,ax` example above), there are ways of effecting the same or similar behaviour: a Windows program may open a file by means of user mode API `NtOpenFile()/OpenFile()`, kernel-mode API `ZwOpenFile()` or may even bypass the API completely and directly access the disk driver with `IoCallDriver()` with manually constructed IO packets. How well these signatures approaches work in practice will be discussed below.

Practical AV concerns: false negatives

A number of independent laboratories regularly test updated AV scanners against millions of malicious software specimens. These scanners predominantly use byte-value signature approaches, though almost all of them today incorporate some form of (much slower) behavioural detection. Some empirical data for sixteen well-known, reputable AV products are shown in Table 1.

Report Date	AV Signature Update	MW Corpus Date	False Negative (%)	Scan Speed (MB/sec)
2009/05	Feb. 9th	Feb. 9th -16th	[31-86]	N/A
2009/02	Feb. 9th	Feb. 1st	[0.2-15.1]	[24.0-3.7]
2008/11	Aug. 4th	Aug. 4th -11th	[29-81]	N/A
2008/08	Aug. 4th	Aug. 1st	[0.4-13.5]	[22.2-2.9]
2008/05	Feb. 4th	Feb. 5th -12th	[26-94]	[25.5-1.6]
2008/02	Feb. 4th	Feb. 2nd	[0.2-12.3]	N/A

Table 1: Miss rates of up-to-date scanners.
Table generated by the author from AV-comparatives.org data

The reader is requested to note how quickly AV signature databases go out-of-date. After failing to update signatures for one week, the *best* AV tested missed between 26 and 31 per cent of the new malicious software, the worst missed upwards of 80 per cent. The empirical test results from http://www.av-comparatives.org/comparatives reviews indicate that the claims made by vendors of AV products must be soberly assessed.

There is preliminary hope pinned on 'cloud computing' environments, where vendors promise reactive signature generation times on the order of

[11] *For a review of behavioural based scheme and a recent prototype of a behaviour based signature approach (using a system-call-data flow dependency behaviour graph), see Clemens Kolbitsch, Paolo Milani Comparetti, Christopher Kruegel, Engin Kirda, Xiaoyong Zhou, and Xiaofeng Wang, 'Effective and Efficient Malware Detection at the End Host', in USENIX Security '09, Montreal, Canada, (August 2009), available at http://www.iseclab.org/publications.html.*

The ability to disguise malicious software becomes more subtle and unpredictable in the light of the different methods by which devices now communicate with each other.

minutes, not days, through active internet connections. This remains to be seen, as the race between AV companies and malicious software writers continues.

The problem with modern malicious software

Modern malicious software is interactive, polymorphic and metamorphic. All these terms have to do with the methods used to bypass the approach used by AV products to detect malicious software (and other signature-based defences such as intrusion detection systems). Polymorphism and metamorphism are both techniques to mutate the computer code of the malicious software while keeping its malicious functionality unchanged. The purpose of this is to evade the signatures of AV.

Though the terms are sometimes used interchangeably, there are technically different: polymorphic malware typically uses encryption on parts of its code containing its malicious functionality. This code must be decrypted by a decryptor routine before it can be executed. Typically, both the encryption and decryption loops can be identified (in unencrypted form) in the malicious software, although it is possible to out-source this function to a remote server – called server-side polymorphism. Hence, the main characteristics of truly polymorphic malicious software are the use of encryption and a fixed decryptor routine.

For detection purposes, encrypted code has a distinct general signature (it has high entropy because of the diffusion property of good encryption); as such, AV can discern the existence of encrypted code, if not its purpose or functionality. Benign code may also be encrypted (and commonly is for intellectual property reasons), thus limiting the usefulness of high entropy detection approaches. The fixed decryptor routine of truly polymorphic code can easily be picked up by byte-pattern signature-based AV. It is for this reason that writers of malicious software have devised schemes to generate mutated, but functionally equivalent decryptors in subsequent generations, leading to what is called 'oligomorphic' code. Oligomorphic decryptor mutation approaches in turn lead to the development of 'metamorphic' code.

Metamorphic code strives to change its appearance from generation to generation, whilst ensuring that it continues to function as it was designed to. Metamorphic malicious software typically does not use encryption. Instead, it is written in such a way that it attempts to re-arrange the relative position of its code, substitute certain instructions, register re-assignments, changes sequence permutation and uses other substitution or permutation techniques. Part of the malicious code incorporates a metamorphic engine that performs these alterations, or the malicious software contacts a server for the task. If the latter, it makes detection harder. Similarly to the truly polymorphic case, a transformation engine residing in the code offers more opportunities for detection purposes.

The ability to disguise malicious software becomes more subtle and unpredictable in the light of the different methods by which devices now communicate with each other. In the widest sense, almost any form of external input might cause malicious software to become active or, more distressingly, provide a missing piece of code to turn apparently innocuous fragments of code into malicious software. Time is used as a mechanism to cause malicious software to become active through internal system clocks (the 1992 Michelangelo virus was activated in this way on the anniversary of his birthday, 6 May), and human activity in using the computer, opening a file or browsing a website can also be used to activate malicious software.

As previously noted, the problem of identifying malicious software is also exacerbated because of ubiquitous connectivity. The vast majority of computers are constantly interacting over the network (end users have little choice in the matter, because software licenses tend to be remotely attested), and at any

moment, passive (as in a simple packet) and active (as in code) prompts can be added to the recipient's system with no prior indication of what this single piece of code will induce. A recent example was provided by the fourth generation of the Conficker worm, when millions of people waited to see what the code would do on the 1 April 2009.[12]

Theoretical AV concerns: detection complexity

As the empirical results suggest, meta- and polymorphic coding techniques pose an aggravated detection challenge for AV. In addition, malicious software has become increasingly modular (utilizing ubiquitous connectivity), and exhibits what is called 'staged downloads'. Staged downloads involve an initial compromise in which a small piece of code is installed. This is effected, for instance, by an network worm exploiting an Operating System vulnerability (such as the 2008 Gimmiv.A worm that targeted the Windows MS08-67 vulnerability[13]) and depositing an initial payload. It could also be effected by the user opening e-mail attachments with malicious code attached, and increasingly, through vulnerabilities in web browsers on computers. The initial infection is subsequently followed up with the installation of more malicious code to fulfil one or more of the objectives that the code is designed to carry out (among them spam relay, stealing of personal information, industrial espionage). Almost 80 per cent of potential malicious code infection exhibit these staged downloads.[14]

Malicious code communicates with its environment for the purposes of propagation and to receive instructions and download new binary code. As a result, the AV detection problem becomes much more difficult. It becomes much harder (impossible in the general case) for AV to decide whether fragments of code are malicious, since not all the pieces may have been assembled. The changing dynamics of malicious code and how it is created and disseminated (complete or in small pieces, and then assembled), means that reliable detection cannot realistically be achieved within time constraints of seconds, if it can be done at all.

Anti-virus: epilogue

Because of the metamorphic and polymorphic dissimulation techniques and the modular staged downloads, current AV is not able to ascertain (within acceptable false negative rates and time limits, and sometimes not in principle) whether code is malicious or not. Worse still, the methods by which an individual can inadvertently download malicious software not only include programs that might be explicitly installed, but code that is installed and executed surreptitiously from a visit to perfectly respectable websites, unbeknownst to the user. The TDSS rootkit serves as an informative case study that demonstrates how malicious software is capable of being installed in seemingly innocuous parts (in the form of a legitimate but maliciously patched DLL) which enables the subsequent downloading and execution of any other arbitrary (malicious) component. Of further interest are the multiple methods of infection used to infect a system (including website vulnerabilities, peer-2-peer networks, video viewing and other software).[15]

The user is faced with stark choices, none of which mitigate the effects of these threats completely. She may disable the functionality that makes web browsing a rich experience to minimize the risk of attack. This means, in effect, reverting back to using the web with 1995 technology.[16] The user might decide to set up a virtual environment that enables the computer to be returned back to a known un-infected state, though this demands a level of discipline that few users are capable of. The Google Chrome web browser is a step in this direction. It incorporates a light-weight virtualized environment called GreenBorder that sets up a protected browser that seeks to shield the computer system from actions originating from browsing the internet.[17] The last choice is the worst and alas, the most common: taking a deep breath, clicking away and trusting anti-virus software to an extent that is not warranted.

Inviting attacks from the internet

There are indications that some safe computing procedures have begun to be understood by end users. For instance, many users now know better than to open e-mail attachments, and they are more mindful of keeping their system patches and AV signatures up-to-date. These measures offer some limited protection.

[12] *For a recent, sophisticated example of binary code updates that is encrypted and electronically signed, see Phillip Porras, Hassen Saidi, and Vinod Yegneswaran, An Analysis of Conficker's Logic and Rendezvous Points, (SRI International Technical Report), 2009 at http://mtc.sri.com/Conficker/ and http://mtc.sri.com/Conficker/addendumC/index.html.*

[13] *See http://www.microsoft.com/technet/security/Bulletin/MS08-067.mspx for the vulnerability and http://www.f-secure.com/v-descs/trojan-spy_w32_gimmiv_a.shtml for a description of the worm.*

[14] *For which, see Symantec's annual Global Internet Threat Report at http://www.symantec.com/business/theme.jsp?themeid=threatreport.*

[15] *Alisa Shevchenko, 'Case Study: The TDSS Rootkit', Virus Bulletin, May 2009, 10-14.*

[16] *One example of a text-only web browser is lynx (http://lynx.isc.org/).*

[17] *GreenBorder was bought by Google in 2007.*

However, the act of browsing the web is more fraught with danger than commonly assumed. Web clients are now increasingly used for banking, health care, governmental services, and retail shopping from the comfort of one's home. Contemporary browsers, such as Internet Explorer, Opera and Firefox incorporate more functions than the mere display of text and images, including rich dynamic content comprising media playback and interactive page elements such as drop-down menus and image roll-overs. These features includes web browser extensions such as Javascript programming language, as well as additional features for the browser, such as application plugins (Acrobat Reader, QuickTime, Flash, Real, and Windows Media Player), and Microsoft-specific enhancements such as Browser Helper Objects and ActiveX (Microsoft Windows's interactive execution framework). Some of these extensions have security vulnerabilities that can maliciously exploited (ActiveX, Flash, and QuickTime make up the vast majority of plug-in vulnerabilities),[18] some are general programming languages or environments that can be tampered with for malicious purposes.

The fundamental issue is one of trust. When the user goes to a website from his browser, he types in a URL, and initiates the connection. Assume the user is visiting an on-line merchant, and assume an encrypted HTTPS connection is established (which is considered 'safe' browsing). The user logs on with his name and password, and a cookie is created. This cookie stores user preferences, such as session information and information about what the customer has purchased, and this cookie is typically placed on the user's computer. Once connected, a relationship of trust is established: the user and the website (the user initiated the connection, and now trusts the page and content display) and conversely, the site and the user (in executing actions from the user's browser). It is this trust, together with the various features incorporated into the browser that attackers try to subvert through what is called Cross-Site Scripting (XSS) and Cross-Site Request Forgery (CSRF) attacks.

Cross-site scripting attacks mostly use legitimate web sites as a conduit, where web sites allow other (malicious) users to upload or post links on to the web site. Such links may contain malicious content (such as Javascript in obfuscated form) within them. They are then presented in an appealing manner ('Click here to view pictures!') to entice the victim to click on them. The malicious script in the link is executed in the victim's browser, and can copy cookie information, change user preferences, write information to files, or (in the form of a CSFR) obtain the data relating to log-ins to merchants and banks to perform actions that purport to be initiated by the customer. It is not only web servers can serve as a conduit: in 2005, a user named Samy Kamkar placed malicious Javascript on his MySpace profile. When a user viewed his profile, an XSS attack would add the user as a friend and place the malicious code in the viewer's profile. In twenty hours, over a million MySpace users were infected.[19]

Prevention of XSS attacks requires both server and client diligence. With respect to the server, software developed for web applications should check links posted by users for potentially malicious content, such as embedded Javascript and HTML code. Since code in such links would be executed in the browser of an innocent user, failure to validate (potentially malicious) input by users represents a software vulnerability that developers should address as a matter of course. Where users are concerned, they should exercise judicious care when clicking on a link. They may also take steps to be much less susceptible to XSS attacks. This can be accomplished by disabling JavaScript, Java, Flash, ActiveX and other dynamic content features in the browser. However, users will incur a severe usability penalty, since many websites depend on these features for to be viewed at their best.

Cross-site Scripting attacks are often used as a stepping stone with more insidious CSRF attacks in which a user's credentials are used for unauthorized transactions. For example, assume the victim is logged into a bank site. There are valid credentials, stored in form of a cookie on the victim's computer. The victim might casually surf a news site where an attacker was allowed to insert code of the sort illustrated below in a

[18] *In May 2009, the web-based Gumblar/JSRedir-R trojan (which accounted for over forty per cent of malicious content found on websites in the first week of May) used obfuscated JavaScript via web browsers to exploit vulnerabilities in Acrobat Reader and Flash Player. See Erik Larkin, 'New Wave of "Gumblar" Hacked Sites Installs Google-targeting Malware', PC World, May 14, 2009 at http://www.pcworld.com/article/164899/new_wave_of_gumblar_hacked_sites_installs_googletargeting_malware.html.*

[19] *See Justin Mann, 'MySpace speaks about Samy Kamkar's sentencing', TechSpot.com, January 31, 2007, where the following was noted: 'Samy Kamkar (aka 'Samy is my Hero') plead guilty yesterday in Los Angeles Superior Court to a violation of Penal Code section 502(c)(8) as a felony and was placed on three years of formal probation, ordered to perform 90 days of community service, pay restitution to MySpace, and had computer restrictions placed on the manner and means he could use a computer – he can only use a computer and access the internet for work related reasons' at http://www.techspot.com/news/24226-myspace-speaks-about-samy-kamkars-sentencing.html.*

It must be emphasized that from the point of view of the user, neither HTTPS (the encrypted channel with the little lock in the browser that denotes 'safety') nor logins protect against XSS or CSRF attacks.

posting or comment on the web site:

```
http://www.bankoflondon.com/transfer.p
hp?account=686868&amount=25000
```

If the attacker succeeds in inducing the victim to click on this link ('Click here to look at Michael Jackson's shroud' might work), a transaction request from the user's browser to the bank would be generated, attempting to transfer $25,000 to (presumably) the attacker's account number 686868.

Sometimes, it is not necessary to click on link on a web site. The news web site might contain HTML code (posted by the attacker) of the sort (purportedly to load an image) as illustrated below:

```
<img src=
"http://bankoflondon.com/transfer.php?
account=686868&amount=25000 width="1"
height="1" border="0">
```

With this code, the browser will try to load a miniscule image. This is a standard procedure to render images in web pages. But the image is not an image: it is actually a HTTP request to the fictional Bank of London, attempting to transfer $25,000 from the victim to the attacker's account number 686868. There is no image available, which means an error (crossed-out) image will be rendered by the browser to the user's screen. The reason for setting the size at 1 pixel by 1 pixel is to suppress this error image, and thus allay any suspicion of the victim.

The nature of transactions that are possible to effect depend on the site for which the credentials are valid. This can range from a posting to a message board with the user's identity; performing bank transactions, to changing the DNS settings of the home router (called drive-by-pharming) and buying stocks. In February 2008, 18 million users of an e-commerce site in Korea were affected by a CSFR attack.[20]

Similar to the XSS example, CSFR attacks can use other conduits (Adobe Acrobat, MS Word, RSS), provided these data formats allow for scripting. It must be emphasized that from the point of view of the user, neither HTTPS (the encrypted channel with the little lock in the browser that denotes 'safety') nor logins protect against XSS or CSRF attacks. In addition, unlike XSS attacks which necessitate user action by clicking on a link, CSFR attacks can be executed without the user's involvement, since they exploit explicit software vulnerabilities on the server. However, it is to be notes that CSFR attacks can be executed without the user's involvement, because they exploit explicit software vulnerabilities (predictable invocation structures) on the server. As such, it is suggested that the onus to prevent CSFR attacks falls squarely on the developers of such applications. Some login and cryptographic token approaches, if conscientiously designed to prevent CSFR attacks, can be of help.[21]

Epilogue

The wide variety of features that are included in everyday programs (such as web browsers and document viewers such as Adobe Acrobat Reader) are a serious concern: almost no user is aware that merely clicking on a URL, handling a PDF document[22] or simply

[20] *For which see 'WHID 2008-10: Chinese hacker steals user information on 18 MILLION online shoppers at Auction.co.kr' at http://www.webappsec.org/projects/whid/byid_id_2008-10.shtml.*

[21] *See the Secret Token scheme reviewed in Adam Barth, Collin Jackson and John C. Mitchell, 'Robust Defenses for Cross-Site Request forgery', in Proceedings of the 15th ACM Conference on Computer and Communications Security (Alexandria, Virginia, USA, October 27-31, 2008). CCS '08. ACM, New York, NY, 75-88, available from http://flyer.sis.smu.edu.sg/srg/ and http://www.adambarth.com/.*

[22] *See http://blog.didierstevens.com/2009/03/04/quickpost-jbig2decode-trigger-trio/ for an example where merely looking at PDF files in Windows Explorer (not opening them by double-clicking) launches the malware.*

surfing on to a webpage[23] may lead to a stealthy compromise and install powerful malicious software. As stated previously, a user has some protection against malicious software by keeping their system conscientiously patched, and maintaining up-to-date AV software. AV performs best if signatures are continuously updated, otherwise the practical detection rate plummets very quickly. Interactive malicious software, as well as user expectations,[24] significantly increases the difficulty of detection for AV software.

Hardware-based malicious code, which can be hidden in underlying integrated circuits (manufactured in China, and possibly compromised in the factory), will cause even more problems. Hardware subversion is not within the ability of AV software to deal with (in fact, it is an open research problem as to how to detect such malicious code in hardware at all). Hence, AV has its limitations, and care must be taken to ensure a digital evidence specialist, when examining a hard disk or live RAM memory, is aware of the various methods by which malicious software can be placed on a computer without the knowledge or authority of the owner or user.[25]

Daniel is an Assistant Professor in the Department of Computer Science, University of New Orleans, Louisiana, United States of America. He is a founding member of ISTS at Dartmouth College (NH, USA), conducting counter-terrorism critical infrastructure research for the US DoJ and US DHS. Active research topics include detection and containment of highly evolved malware and quantitative risk analysis and management of networks.

daniel@cs.uno.edu

[23] *Niels Provos, Dean McNamee, Panayiotis Mavrommatis, Ke Wang, and Nagendra Modadugu, 'The Ghost in the Browser: Analysis of Web-based Malware', Proceedings of the 1st conference on First Workshop on Hot Topics in Understanding Botnets (USENIX Association Berkeley, CA, USA, April 2007), available in electronic format at http://www.usenix.org/event/hotbots07/tech/full_papers/provos/provos.pdf.*

[24] *Whether reasonable or not, users are not willing to wait more than a couple of seconds to ascertain whether they can open, execute or view a file, or safely browse a website; nor are they willing to put in the time or effort to gain reasonable safety proficiency to operate what is increasingly complex hardware and software.*

[25] *By way of addendum, an investigation by Associated Press has found a number of people in the USA where a third party has caused abusive images of children to be downloaded on to their computer, which often results in criminal charges that might or might not be withdrawn: Jordan Robertson, 'AP IMPACT: Framed for child porn — by a PC virus', AP Technology 8 November 2009 at http://tech.yahoo.com/news/ap/20091108/ap_on_hi_te/us_tec_a_virus_framed_me.*

ARTICLE:

REMOTE ELECTRONIC DISCOVERY

By **Gib Sorebo**

Introduction

In the realm of civil discovery ('disclosure' in some jurisdictions) most attorneys in the United States tend to adopt a flexible approach to discovery, because the rules tend to encourage cooperation and give parties significant flexibility to jointly agree on a set of discovery practices. In general, this practice makes sense, because courts only become involved when there is a dispute. Otherwise the parties conduct discovery in a manner that fits the case and the resources they have available. However, the process presumes that attorneys for both sides are qualified to address the numerous issues that arise during discovery. While it is true that clients are ultimately responsible for the competency of their attorneys, it is also true that the legitimacy of our litigation process is undermined each time attorneys conspire, usually unwittingly, to impose unnecessary costs and an unreliable discovery process on their clients due to their lack of understanding of the information they are seeking to discover, the tools that are available to them, and the potential consequences to third parties.

Such a scenario is a daily occurrence in the area of e-discovery where attorneys for each side will negotiate away large swaths of data repositories within a company, accept data without any chain of custody, ignore meta data, and show no concern about the format that the data is produced. All this is not designed to limit discovery to what is important and to control costs. Instead, it is designed to keep matters on a level that they can understand and that their frequently underfunded and litigation support team that does not have the necessary skills can accommodate. Anecdotal comments from judges bear out the fact that such practices are common, and if both attorneys agree, there is little judges can do other than offer advice. Where this behaviour affects only the two parties, it is fair to say that there are more important things to worry about. However, litigation rarely occurs in a vacuum. Third party interests are often implicated. In discovery, all documents in the possession of each party are usually open to being seen by the other side. This could potentially include third party information that may be more valuable to the third party than the litigants. Additionally, privacy laws frequently limit the purposes for which such information can be used and require special authorization for use in litigation, particularly when the subject or data owner is not a party to the case.

As technology evolves, the implications for litigation must also evolve. Traditionally, discovery meant that a requesting party requested information relevant to the litigation with some degree of specificity, and the responding party then set about finding, collecting, and ultimately producing that data after reviewing it for relevancy and privilege. The process was fairly straightforward and limited by how the information was collected. Because the traditional method involved paper documents that were typically in the possession of a single person, usually known as the document custodian, it made sense that attorneys would simply issue a legal hold memorandum to such persons notifying them of the litigation, identifying the kinds of documents that would be relevant, requesting such documents be preserved, and providing a mechanism to deliver the documents or have them photocopied. In most cases, it was the document custodian who searched and delivered the relevant documents. Whether he or she was also responsible for doing the photocopying, sorting, or delivery, the law understood the location of the document custodian and the location of the documents to be synonymous. Discovery rules focused on the fact that a person under the jurisdiction of the court, usually as a function of their being the employee of one of the parties, was under an obligation, and sometimes compelled, to produce the documents. The document reviews, photocopying, Bates stamping, sorting, packaging, and delivery are all

support functions that flow from the obligations of the custodian of the document.

While courts may not always have direct jurisdiction over custodians of documents, particularly if they reside in another state or a foreign country, they still impose the obligation on the custodian indirectly through the jurisdiction they assert on the custodian's employer. The court will require the employer to direct its employees to produce a particular document. In addition to producing a clear chain of responsibility, it also allows any assertions based on privilege, privacy laws, export controls, or a sovereign's outright rejection of the litigation to be heard with respect to the document being requested. Because the same sovereign has immediate jurisdiction over both the custodian of the document and the document itself, it is in a good position to restrict its transfer and eventual production. Until recently, the same concept applied to electronically stored information. While the internet has provided people with instantaneous access to information world-wide, the data most frequently requested in litigation is still modeled after the paper method. The custodian of the data, usually a system administrator or designated data owner, is still requested to produce the information, and significantly, that custodian is usually located within close proximity to where the data is stored. Such proximity may be in another room or another building, but there is a good chance that it is still within the same jurisdiction. Moreover, based on the method typically used to collect the data that is discussed below, the litigation support team, in conjunction with the custodian, usually collects the data directly from the computer that it is stored on or over the network on a computer nearby. Either way, those collecting the data for the purposes of discovery are usually present in the jurisdiction where the data is stored even if it is collected and then loaded onto a repository in another jurisdiction.

What this article seeks to examine is the changing nature of both e-discovery and how data is stored. As new e-discovery technologies are deployed, the potential for widespread collection using remote means not facilitated by a local data custodian is becoming a reality. Because discovery in the United States does not typically involve the court for matters relating to the collection of discoverable material by the producing party, case law is rather limited and discussions about the significance of the location of electronically stored information and any possible restrictions on remote collection are non-existent. In fact, 'no court has squarely addressed where electronic materials are "located" for discovery purposes.'[1] Because jurisdiction and the ability to effectively adjudicate discovery disputes involving both litigants and third parties is generally a product of location in most common law countries, it is important that the law catch up with the technology.

Framing technology issues

The gathering of evidence for litigation is typically directed by counsel whereby the likely locations of relevant information and their custodians are identified. Then legal holds are issued to the custodian who can include both the imputed data owner, which may be a business manager charged with overseeing the business processes that generated or collected the data, and potentially a data custodian, who may be an IT manager or system manager but who is just as likely to be an employee who is simply storing the relevant data on his or her desktop or laptop. Similarly, in the physical world, there are imputed document owners and document custodians. In both situations, both the custodian and the data owner are frequently situated at the same geographic location and usually in the same jurisdiction. In some cases, centralized mainframe computers had required some physical separation. However, in discovery matters, the person physically co-located with the system was usually the person given the task with extracting the data that was required. Additionally, despite the ability to obtain access to the date remotely, there was little question of its location.

The only type of remote discovery that has been considered somewhat routine, is the collection of publicly available information available on the internet. In this case, production is hardly necessary, because the requesting parties can simply search the internet to collect whatever information they choose. Consequently, for the purposes of this article, remote discovery involves the collection and eventual production of non-public information. This includes desktops and laptop computers, servers with directly attached storage, storage area networks, and removable media. Almost by definition, remote access to these storage devices involves some sort of network, either through a traditional circuit-switched telephone network, or dial-up, and, more typically, via a packet-

[1] *Gary B. Born and Peter B. Rutledge, International Civil Litigation in United States Courts (4th edition, Wolters Kluwer Law and Business, 2007), 930.*

based network such as the internet or similar sub-networks within an organization with connectivity being provided locally by the organization or over long distance using internet or private leased line connectivity. In either case, higher level protocols using encryption, circuit virtualization, authentication, and other means can ensure that such connectivity remains private. Within these private networks, the conventional notions of storage and application services are radically changing. No longer is storage tied to a single processing device. It can serve multiple application servers all at once. Moreover, the storage can be distributed across national boundaries as needed. Using sophisticated data mapping technology, what appears to an end user to be a single file or directory at one location could actually be bits stored on multiple devices in several different countries. Moreover, the application retrieving the file for processing and eventual output to the user could also have components residing in multiple locations and potentially owned by a third party. These are called cloud computing applications. While in essence they are a throw-back to mainframe computing concepts of shared processing, cloud computing will probably revolutionize computing and fundamentally alter the notion of electronically stored information. The change is not so much that remote discovery is now possible. In some form, the potential for remote discovery of electronic data has existed as long as there have been computer modems. Instead, the fundamental change is that some discovery can only be achieved remotely, given how some applications and their data are now structured.

Even if the notions of cloud computing and geographically distributed storage are yet to become commonplace, other factors are making remote discovery an all but unavoidable scenario in litigation. Due to the globalization of many corporations and the need to collaborate, the operation of largely autonomous subsidiaries organized by country has largely vanished. High network bandwidth over long distances has meant that information technologies can simultaneously use data stored in multiple places and the need to have 'local' copies of all data needed by the local users has all but vanished in many large organizations. Moreover, enterprise search technologies are being deployed in a way where data owners can authorize data custodians to grant access to appropriate parties and refrain from overseeing such access. They are no longer the go-between in satisfying requests for data. Instead, once access is granted, the data can be available world-wide, subject to export control and privacy laws. Individuals often have no concept of where the data is physically located, nor do they care. They may be aware that a particular document was written by an employee residing in another country, but they have no way of knowing if it was actually drafted in that other country. Also, in the spirit of collaboration, documents are routinely edited and data is supplied from a number of countries. This means that to argue that the data falls under the sovereignty of a particular nation or US state based on the nationality of the author or the location of authorship is a misnomer. While privacy and export control laws may dictate some degree of data segregation by country, such laws are rendered otiose by the vast amount of data involved in commercial litigation that does not fall into those categories. Additionally, with the Safe Harbor provisions,[2] privacy laws, arguably, may no longer require that covered data honour national boundaries.

Traditional legal issues with cross border discovery

Aside from the logistical challenges brought by new technology, legal issues also present challenges of their own. Because there is limited legal precedent for remote discovery, the focus will be on drawing parallels with cross border discovery decisions. From a statutory and administrative perspective, discovery is the same whether the activity is conducted in the United States or abroad. Unless a judge is asked to compel discovery, litigants are free to request and conduct depositions and request production of information wherever it might reside.[3] When a court orders foreign discovery, additional considerations may need to be addressed. As described below, the court jurisdiction will usually be the primary arbiter in deciding whether discovery can be compelled. The potential interests of third parties, when

[2] *Under provisions of Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, OJ L281, 23.11.95, p. 31 (EU Data Protection Directive), and its national implementing laws, countries such as the United States may be allowed to hold private data on European citizens if the country provides for legislation that legally obligates organizations receiving such data to follow the provisions of the Data Protection Data Directive and guidance from individuals nations with respect to dissemination and uses for that data. U.S. Department of Commerce, Safe Harbor Privacy Principles (July 21, 2000), http://www.export.gov/safeharbor/SH_Privacy.asp.*

[3] *Fed. Rule Civ.Proc. Rule 28(a)(1) provides for depositions in a foreign country and may require some notice depending upon how the deposition is procured.*

When used effectively, discovery can be the mechanism that unearths corruption, holds large organizations accountable, and gives litigants with limited means the opportunity to make their case.

no challenge is raised, are usually not considered in the absence of a third party being added to the proceedings. Because most discovery efforts are carried out with little or no public notice, third parties typically have no way of effectively intervening. This presents some interesting privacy and sovereignty considerations that are discuss below.

Ultimately, the issue is often that non-US jurisdictions find the American discovery process unwieldy and fundamentally flawed. They see its stated goal of learning the truth through exhaustive review of all relevant information as simply a charade meant to mask the true intent of the litigants, which is to conduct fishing expeditions designed to increase the other party's costs, expose embarrassing facts that are only tangentially related to the matter at hand, and engage in countless acts of gamesmanship and chest thumping to distract the fact finder from the case and enhance the image of the attorneys. While the statistics that show only a small percentage of cases reaching trial would seem to support the claim that the American litigation system is unwieldy, it is misleading to suggest that such an outcome is based on the system's overwhelming discovery burdens or the publicity sought by attorneys. When used effectively, discovery can be the mechanism that unearths corruption, holds large organizations accountable, and gives litigants with limited means the opportunity to make their case. While true smoking guns are rare, discovery often forces settlement because the information produced by each side provides overwhelmingly evidence that favours one party or the other. The fact that litigants in civil proceedings routinely produce incriminating information that it likely to be used against them is a testament to not only the effectiveness of the discovery process, but of their adherence to ethical conventions and the rule of law. That said, the process is not without its faults. The process can certainly be expensive and unwieldy, with many of its failures not a product of its fundamental principles, but rather adherence to inefficient processes and poor use of technology. Nonetheless, as the processes and revised and the technology is improved, it must be recognized that streamlining processes to more efficiently adhere to these principles may have the unintended effect of denigrating the principles that others hold dear. While compromising principles may not be an option, the methods chosen can be open to compromise.

E-discovery issues

State level

While most legal issues with remote electronic discovery involve the movement of data across national borders, there are a few issues that are relevant between US states. While states are typically obliged to honour requests made by courts of other states and generally show deference to depositions and document productions that originate from litigation in another state, it remains the expectation that some protocol should be followed. For example, where a judge authorizes a party to seize evidence from the other litigant in another state, it is expected that local law enforcement will be engaged and perhaps even local courts will enforce the order.

Recently, some states have passed laws requiring that computer forensics examinations that are part of litigation be performed by licensed private investigators.[4] While this particular requirement is problematic on a number of levels, it does reflect states' desire to assert some quality controls over the process of collecting evidence and to retain oversight over the process. However, the laws are unclear whether they

[4] *2008 American Bar Association Section of Science & Technology Law, Report to the House of Delegates 301, available at http://www.abanet.org/scitech/301.doc (noting specific PI licensure requirements for performing computer forensics in Illinois, Texas, Michigan, Georgia, Rhode Island, South Carolina, North Carolina (pending), Massachusetts, Nevada, New York).*

would apply to remote electronic discovery. While the South Carolina Attorney General has asserted that any computer forensic examinations performed in other states must be conducted by a South Carolina licensed private investigator when the evidence is gathered to be used in a South Carolina court proceeding, there is no such guidance for evidence remotely gathered using a computer forensics process on a device located in a state with such a private investigator requirement where the information is to be used in a matter outside that state.[5] This is despite the fact that some of these states have asserted that licensed private investigators be used when computer forensics is performed in their jurisdiction regardless of where the evidence will ultimately presented and even applies if no litigation is anticipated. It is one of many examples where laws are written too simplistically to resolve a perceived problem rather than to address the true objectives of the situation. The quality of computer forensic examination is certainly an issue that needs to be addressed. However, the solution proposed and implemented is not always the most appropriate. Rather than passing a law that is enforceable in the least costly manner possible but ineffective at accomplishing the objective, states should recognize that the true solution may be to learn more about the problem, seek consensus where possible, and regulate last. Failing to do so simply leads to circumvention and higher costs and ultimately causes more harm to the very people it seeks to protect.

Aside from forensic examinations, the very notion of remotely collecting data in other states raises a number of issues relating to the state's desire to accord privileges to its citizens. Because most state privacy laws target personal data about its citizens without regard to location, the privacy aspects seem not to be implicated. Moreover, constitutional protections of interstate commerce would seem to preclude a state from restricting the flow of such data. However, because this data may be destined for a court, practitioners should be wary of state specific privileges that may arise. Conflict of law principles are far from settled in this area as it is unclear whether privileges apply to data at the point it is generated or in the state where the court is located.

International

By far the most significant legal issues with remote discovery involve data that passes across international borders. Because remote discovery typically does not require anyone in the foreign country to facilitate the data transfer, such transfers can be transferred with relative ease. Typical foreign discovery challenges usually involve conducting depositions in another country or requesting someone in that country to produce a document. For the most part, there is little authority on the issue of whether the fact that no one involved in the discovery need be present in that country raises any concerns. The typical remote discovery scenario would be where relevant information resides on a server in a branch office of a multi-national company that was outside the United States. Assuming that personnel in the United States already have remote access to the data, then a foreign government has limited ability to prevent access, because it cannot sanction anyone under its jurisdiction for the immediate transfer. After all, '[t]he location of the person, not the document, is also a hallmark of discovery under the Federal Rules of Civil Procedure: "Persons resident or found within the United States may have in their possession or under their control evidence located abroad. It has long been recognized that such persons may be required to produce such evidence in courts in the United States."'[6] However, if a local employee of the company is required to grant access or otherwise facilitate the transfer before it can be sent, then foreign law may pose some challenges depending upon the data at issue, because the facilitator risks violating their own law or causing their employer to violate US law.

Beyond the unique circumstances associated with remote discovery, the challenges posed by discovery of information in a foreign country for use in a US proceeding can be daunting. As mentioned above, many countries, particularly those using the civil law system, show a particular distaste for the American discovery process. 'As of 1986, some 15 states had adopted legislation expressly designed to counter United States efforts to secure production of documents situated outside the United States.'[7] These have taken a number of forms, from providing mechanisms for its citizens to

[5] *Deb Radcliff, 'Computer Forensics Faces Private Eye Competition' Baselinemag, January 2, 2008, p 1 http://www.baselinemag.com. ('In April [2007], the state attorney general opined that even if you never set foot in South Carolina, if you're collecting evidence to be used in court here, you still need a South Carolina [PI] license', says Steve Abrams, a licensed independent PI and computer forensic examiner based in Sullivans Island, S.C. 'Licensing authorities in New York, Pennsylvania, Texas and Oregon have opined the same way.')*

[6] *Charles McClellan, America, 'Land of (Extraterritorial) Discovery: Section 1782 Discovery for Foreign Litigants', 17 Transnational Law & Contemporary Problems 809, 822 (2008) (quoting Hans Smit, 'American Assistance to Litigation in Foreign Aid and International Tribunals: Section 1782 Title 28 of the U.S.C. Revisited', 25 Syracuse Journal of International Law and Commerce, 1, 10 n.46 (1998)).*

[7] *Restatement (Third) Foreign Relations Law § 442, Reporters' Note 1 (1987).*

refuse requests, to prohibiting its citizens from cooperating altogether in the case of France.[8] The process generally acceptable to most countries is through the Hague Convention on the Taking of Evidence Abroad in Civil or Commercial Matters.[9] That treaty calls for letters of request to be issued by the court having jurisdiction over the matter and sent to the relevant authority in the country where evidence is being sought. While this process has assisted litigants who previously had no recourse when seeking discovery of witnesses or documents located in a foreign country, the US Supreme Court noted in the seminal case of *Societe Nationale Industrielle Aerospatiale v. District Court*[10] that the use of the treaty is not mandatory in foreign discovery matters and that it does not override the Federal Rules of Civil Procedure.[11] In this case, the Supreme Court took a pragmatic view in suggesting that the treaty was merely in place to protect the rights of foreign litigants and that where both parties are amenable to the discovery request, there is no need to involve foreign authorities in the matter. Justice Blackmun's dissent clearly alludes to this perception and argues that with the civil law system, in particular, the process by which evidence is collected, normally by a judge rather than the litigants, is as much a part of the analysis as the willingness of parties to comply with the request. He notes judges are often entrusted with the role of balancing the rights of the parties as well as the rights of the public as a whole, including affected third parties, when deciding whether to transfer that evidence to a foreign court.[12]

Considering the view of the majority and similar holdings in other courts that effectively suggest that the treaty's procedures should be the last resort rather than the first, it would be safe to conclude that remote discovery would probably be a matter requiring little, if any, consultation with foreign authorities as far as US courts are concerned. In effect, what little related case law on this subject tends to bear this out. For example, in 2002, Vasily Gorshkov was convicted of stealing credit card numbers by a Seattle-based federal court. The evidence in the case was gathered by FBI agents who lured Gorshkov and his accomplice into the United States from Russia, where through an undercover ruse, they asked the two men to type their username and password into a computer the FBI was monitoring. The account credentials were for a computer located in Russia. The agents then used the credentials to gain access to that computer and download the evidence implicating the two in multiple cases of fraud. Because the FBI used its own computer, told the defendants they wanted to watch them, and obtained a search warrant before viewing the downloaded file, the Federal District Court Judge ruled that the evidence was admissible and that the FBI had done nothing wrong. He further asserted that the fact that the agents' action violated the law of the Russian Federation was not relevant because the law of the Russian Federation did not apply.[13] It transpires that the Russian authorities did not agree, and filed a criminal complaint against the agents, while the agents received the Director's Award for Excellence as a result of the successful sting operation.[14]

While such a flagrant flaunting of another nation's laws may result in the exclusion of evidence in civil matters, courts have nonetheless shown that US interests, particularly those of the litigants, come first. In response to blocking statutes, courts have adopted a five factor test for considering whether a party should be compelled to produce information that resides in another jurisdiction, particularly in cases where the party needs to travel to that country to retrieve it or request employees in the foreign country to facilitate its delivery even when the foreign nation specifically forbids it. The factors include: (1) the importance to the litigation of the information requested; (2) the degree of specificity of request; (3) whether the information originated in the United States; (4) the availability of alternative means of securing the information; (5) the extent to which failure to comply would undermine the interests of the United States or compliance with the request would undermine the interests of a foreign sovereign nation.[15] However, case law suggests if the

[8] *James Chalmers, ' The Hague Evidence Convention and Discovery Inter Parties: Trial Court Decisions Post Aerospatiale' , 8 Tulane Journal of International and Comparative Law 189, 213 (2000) (noting that '[i]t is difficult to take the French ' blocking statute' at face value given that, taken literally, it appears to prevent French nationals doing business abroad from taking court action in foreign tribunals. Instead, it appears that the statute was intended to assist French nationals involved in litigation abroad by providing them with a reason for refusing to disclose information.')*

[9] *Opened for signature, 18 March 1970, 23 U.S.T. 2555, T.I.A.S. 7444, 847 U.N.T.S. 231.*

[10] *482 U.S. 522 (1987)*

[11] *482 U.S. 522 (1987) at 544 (declining to hold to hold, as a blanket matter, that comity requires resort to Hague Evidence Convention procedures without prior scrutiny in each case of the particular facts, sovereign interests, and likelihood that resort to those procedures will prove effective).*

[12] *482 U.S. 522 (1987) at 548 (Blackmun, J, dissenting) (' In my view, the Convention provides effective discovery procedures that largely eliminate the conflicts between United States and foreign law on evidence-gathering. I therefore would apply a general presumption that, in most cases, courts should resort first to the Convention.').*

[13] *Mike Brunker, Judge OKs FBI hack of Russian computers, ZDNet, 31 May 2001, http://news.zdnet.com/2100-9595_22-115961.html.*

[14] *Lawyer to challenge FBI in Russian sting, Reuters, 25 August 2002, http://news.cnet.com/Lawyer-to-challenge-FBI-in-Russian-sting/2100-1002_3-955251.html.*

discovery request could be satisfied without leaving the United States or requesting the aid of someone in a foreign country through a means such as remote discovery, courts are not likely to even consider treaty requirements or blocking statutes when deciding whether to grant the request to compel.[16]

While the court's inclination to ignore the wishes of a foreign government in both civil and criminal cases is certainly the most expedient means of resolving a discovery dispute when that foreign government's assistance is not needed, it is nonetheless troubling. Disregard for international comity can arise in other forums that are not directly of interest to the court or the litigants but could have a chilling effect on future cross border litigation and even the transfer of data across borders outside litigation. For example, under provisions of the EU Data Protection Directive[17] and its subsequent enforcement of member nations, the default position is that the United States does not have sufficient data protection laws for the protection of personal data to permit the transfer of such data. However, under the Safe Harbor provisions negotiated with the US Department of Commerce,[18] an organization can voluntary submit to such provisions that the Department of Commerce will then enforce as a condition of receiving personal data on EU citizens. However, the Safe Harbor provisions are problematic within the context of discovery, because the provisions only apply to data transferred to another country but within the same organization. Because the purpose of discovery is to disclose data to another party, the Safe Harbor provisions do not provide adequate protection. Additionally, while consent of the subject of the data is usually sufficient to exempt the application of privacy laws, where the consent is by an employee, EU authorities typically view such consent as coerced and therefore not allowed.[19] As an alternative, the EU Data Protection Directive does allow for transfers outside the Safe Harbor protection where 'the transfer is necessary . . . for the establishment, exercise or defense of legal claims.'[20] However, such transfers must be ordered by a European judicial authority pursuant to a letter of request, such as provided for under the Hague Convention.[21] Based on recent precedent, litigants are not likely to make such a request if the information can readily be obtained by simply obtaining access to a remote computer.

As a result, the situation is difficult. As technology advances to the point that multinational corporations can easily resort to these self-help measures without risking sanctions or even awareness by foreign governments that this is going on, the pattern will continue, with the criteria for compliance being US privacy laws that Europeans, in particular, find inadequate. However, it may take one significant and public privacy breach to convince European governments that the Safe Harbor provisions are relatively weak within the US legal system and may be rescinded, making cross company communications problematic across borders. Other implications could be outright refusal to allow US persons direct remote access to the personal data of EU citizens residing on systems within an EU country. Ultimately, distributed storage and cloud computing may either make that discussion moot, or laws could effectively limit the use of such technology across national borders. Given the undesirable consequences that could result from direct regulation of such technology for the regulating country, it is to be hoped that a more efficient solution that preserves international comity and the rights of each country's citizens while satisfying the demands of the US discovery system will come about. The models described above could work, but no one currently has the incentive to implement them.

Gib Sorebo is an information security consultant and assists organizations in managing their information security and privacy risks, and compliance obligations. He speaks on various topics, including information security liability, electronic discovery, and security breach laws. He holds a Juris Doctor from Catholic University.

gibsorebo@hotmail.com

[15] *Restatement (Third) of Foreign Relations § 442(1)(c) (1987)*

[16] *Article 29 Data Protection Working Party, Working Document 1/2009 on pre-trial discovery for cross border civil litigation at 5, 00339/09/EN, WP 158 (Feb. 11, 2009) (noting 'that if the company is subject to US law and possesses, controls, or has custody or even has authorized access to the information from the US territory (via a computer) wherever the data is "physically" located, US law applies without the need to respect any international convention such as the Hague Convention.').*

[17] *EU Data Protection Directive at 31–50.*

[18] *U.S. Department of Commerce, Safe Harbor Privacy Principles (July 21, 2000), http://www.export.gov/safeharbor/SH_Privacy.asp.*

[19] *Carla L. Reyes, 'The US Discovery-EU Privacy Directive Conflict: Constructing a Three-Tiered Compliance Strategy', 19 Duke Journal of Comparative and International Law 357, 374-78 (2009) (discussing limitations of Safe Harbor provisions and consent); but see Stanley W. Crosley, Alan Charles Raul, Edward R. McNicholas and Julie M. Dwyer, 'A Path to Resolving European Data Protection Concerns With U.S. Discovery', 6 Privacy & Security Law Report 1, 5 (Oct. 15, 2007) (suggesting that consent, particularly when obtained in advance of the litigation, may be sufficient).*

[20] *EU Data Protection Directive, art. 26(1)(d).*

[21] *Carla L. Reyes, The US Discovery-EU Privacy Directive Conflict: Constructing a Three-Tiered Compliance Strategy, note 17, at 365-66.*

ARTICLE:

LEGAL PRIVILEGE AND THE HIGH COST OF ELECTRONIC DISCOVERY IN THE UNITED STATES: SHOULD WE BE THINKING LIKE LAWYERS?

By **Daniel R Rizzolo**

The digital era has had profound effects on the practice of litigation in the United States, not the least of which is the paradox that lawyers and their clients have faced trying to protect legal privileges during electronic discovery (e-discovery). The legal gymnastics that must be undertaken in order to protect the relationship between lawyer and client have proven inordinately time-consuming, expensive and fraught with errors. Concerns have been raised by attorneys, who must protect client confidences within the disclosure framework of U.S. discovery law; judges, who have to resolve increasingly complex e-discovery disputes; and clients, who have to pay.

Reform has already occurred. The U.S. Federal Rules of Civil Procedure were updated in 2006 to address the complexities of e-discovery, and the Federal Rules of Evidence were subsequently revised in 2008 in an attempt to reverse the trend toward escalating e-discovery review costs. However, some problems remain, and critics call for additional reform. The U.S. litigation bar, its clients, the courts, and the rule makers have focused their efforts to date on procedural protection in an attempt to mitigate the effect of breaches of legal privilege. It is the heuristic aim of this article to suggest an alternative to the procedural approach.

Legal privileges in the United States

The laws of the U.S. afford special evidentiary protection to information and communications that arise out of the working relationship between a lawyer and client. These protections come in the form of two closely associated legal rules:

The Attorney-Client Privilege preserves the secrecy of communications between client and legal counsel.

The Work-Product Doctrine shields works created by or for counsel in the context of litigation.

The protections under these rules are afforded by excluding privileged information from disclosure and from being introduced into evidence. The exclusions under the U.S. rules are absolute and are exercised without consideration of the materiality or probative value of the underlying information. They are compelling protections, particularly in light of the broad scope of U.S. discovery. Current underlying policy considerations for maintaining the breadth of the legal privileges are that compliance with the law is encouraged by fostering an open relationship between attorney and client based on trust, and the quality and

The complications and uncertainties posed by what has become a byzantine maze of U.S. privilege law forces a disproportionate allocation of costly resources to what is essentially a clerical exercise in procedure.

thoroughness of an attorney's preparations are improved when she need not fear that her work will fall into the hands of adversaries.

Together, the Attorney-Client Privilege and the Work-Product Doctrine bind the lawyer-client relationship. As such, attorneys in the U.S. are cautious to a fault when trying to preserve the legal privileges.

Legal privilege law in the U.S. is complex and well developed. There has been voluminous inspection and interpretation by law makers, courts and commentators.[1] The application and scope of the privilege protections often vary by jurisdiction and are particularly susceptible to the volatilities of judicial interpretation. The volume of information causes additional complications during the e-discovery process. The complications and uncertainties posed by what has become a byzantine maze of U.S. privilege law forces a disproportionate allocation of costly resources to what is essentially a clerical exercise in procedure.

The Attorney-Client Privilege

The Attorney-Client Privilege is designed to protect confidential communications between a client and attorney, and is very broad in scope. There are five commonly recognized elements that must be present to claim that a communication is subject to the Attorney-Client Privilege:

A client – the person or entity asserting the privilege must be a client or must be attempting to become a client at the time of disclosure. Under U.S. law, the definition of a client includes private individuals, corporations, and other private organizations. Governmental bodies and public officers are also protected as clients to the extent that the public interest in open government is not outweighed.

An attorney – the person to whom the communication is made must be a licensed attorney and must be acting as an attorney with regard to the communication. The U.S. definition of an attorney includes outside counsel, and is generally expanded to include in-house attorneys. Communications to agents and subordinates working under the direction of counsel are also generally protected by the privilege.

Confidentiality – the communication must be related to the attorney or subordinate by the client or prospective client in confidentiality and outside the presence of strangers. Confidentiality is the requirement most often contested under privilege law. It takes on new dimensions and must be carefully protected when digital communications are involved. For example, an e-mail or voicemail copied or forwarded to a party outside the attorney-client relationship may be deemed a breach of confidentiality and therefore a waiver of the privilege.

The intent to obtain legal advice – the primary purpose of the communication must be to obtain legal advice or services. The mere fact that an attorney is involved does not automatically make a communication privileged.

A right of claim – the client or prospective client must have asserted the privilege, and the privilege must not have been waived either deliberately or inadvertently.

Once the Attorney-Client Privilege attaches, its

[1] *See generally Edna Selan Epstein, The Attorney-Client Privilege and the Work-Product Doctrine (5th edn, 2007, ABA) for an excellent and comprehensive survey of Attorney-Client Privilege and Work-Product Doctrine law.*

protections are absolute. They cannot be overcome by a showing of need.

The Work-Product Doctrine

The Work-Product Doctrine extends protection to works created in anticipation of litigation by or under the direction of counsel. It is a more recent concept in American jurisprudence than the Attorney-Client Privilege. The Work-Product Doctrine was first recognized by the Supreme Court of the United States in 1947 in *Hickman v. Taylor*,[2] and has subsequently been codified at both the Federal and state levels.[3] There are three threshold questions that must typically be addressed in order for work-product protection to take effect, as discussed below.

Whether the information sought is protected

The most commonly cited formulation of the Work-Product Doctrine is found in Federal Rule of Civil Procedure 26(b)(3) (Rule 26(b)(3)). It applies to 'documents and other tangible things.' This definition has been interpreted to include information committed to a physical format such as hard copy writings, photographs, and diagrams. It has also been extended to digital information. In addition to format, there is a further question of content type, which is weighed on a sliding scale. An attorney's mental impressions and thought processes are afforded an almost absolute level of protection, while at the other end of the spectrum, purely factual information is afforded none.

Whether the work was created in anticipation of litigation

While the Attorney-Client Privilege protects communications regardless of the type of legal work, protection under the Work-Product Doctrine is limited to works prepared in anticipation of litigation or trial.

Whether the work was created by an attorney or an attorney's representative

The common law formulation of the Work-Product Doctrine protects works created by an attorney, members of the attorney's staff, and non-lawyers working under the attorney's direction. It is worth noting that in Federal civil proceedings, Rule 26(b)(3) extends protection to the work-product of non-lawyers working on behalf of the client, whether or not an attorney supervises them. The works of consultants, investigators, insurers, physicians, employees, and others may be afforded protection providing they were created in anticipation of litigation and not in the ordinary course of business. As a matter of practice, this last distinction is subject to interpretation by the courts. This means that supervision of non-lawyers by counsel significantly reduces the likelihood that a work will be found to have been created in the ordinary course of business.

An important distinction between the Attorney-Client Privilege and the Work-Product Doctrine is that protection of work-product is not absolute. An adversary may obtain discovery of attorney work-product upon a showing of substantial need and material hardship in obtaining the information elsewhere. Should a court decide that work product is discoverable, it must still 'protect against disclosure of the mental impressions, conclusions, opinions, or legal theories of a party's attorney or other representative.'[4]

Digital discovery and the increased risk of privilege waiver

The protections of the legal privilege rules are lost through acts that constitute waiver. A waiver may be deliberate or inadvertent. It may be caused by the acts of the attorney, the client, or third parties. The variations and minutiae of U.S. waiver law are seemingly endless.[5] The focus of this article will be the risks posed by inadvertent waiver during discovery.

An inadvertent waiver may occur when counsel accidentally turns over privileged materials in the course of discovery. In such cases, remedies are limited, and the results can be calamitous for both attorney and client. When the question of inadvertent waiver is adjudicated, the U.S. courts will enter into an analysis to determine whether and to what extent privileges have been waived. The jurisdictions are split into three schools of thought on the effect of inadvertent disclosure:[6]

Lenient – a small group hold that there is no waiver when an inadvertent disclosure occurs.

[2] *329 U.S. 495 (1947).*
[3] *For instance, see FED. R. CIV. P. 26(b)(3) for the Federal enactment of the Work-Product Doctrine used in civil proceedings.*
[4] *FED. R. CIV. P. 26(b)(3)(B).*
[5] *See generally Edna Selan Epstein, The Attorney-Client Privilege and the Work-Product Doctrine 390-636 (5th edn, 2007, ABA) for an overview of the multiple waiver variations that may come into play under Attorney-Client Privilege law, and Volume 2, 1027-1122 for an overview of waiver law applicable to the Work-Product Doctrine.*
[6] *Laurie A. Weiss, 'Protection of Attorney-Client Privilege and Work Product in the E-Discovery Era', in Vincent S. Walkowiak ed., Attorney-Client Privilege in Civil Litigation: Protecting and Defending Confidentiality 163, 166-168 (4th edn, 2008, ABA) for additional discussion of the split of authority in U.S. privilege law.*

Moderate – the largest group uses a balancing test to determine whether a privilege waiver has occurred. Factors taken into consideration include the reasonableness of precautions taken to prevent a disclosure, the extent of the disclosure, and the promptness with which remedial actions were taken. If a waiver is found, the court determines the extent of the waiver. It will be typically limited to the disclosed documents but can be extended at the court's discretion.

Strict – a small minority adhere to a strict liability approach to waiver. They hold that an inadvertent disclosure of privileged materials will constitute a waiver of the privilege with regard to the documents produced and also with regard to the breadth of subject matter covered in those documents (Subject Matter Waiver).

Even in situations where no waiver is found and documents are returned, the result for the client whose privileged materials have been disclosed is not satisfactory. Attorneys describe this situation to trying to put a genie back in the bottle (or as other critics have noted, like trying to un-ring a bell,[7] or 'closing the barn door after the animals have already run away'[8]). The damage has been done, and client confidences or litigation strategy have been exposed to an adversary. The client's position may have been weakened, or the client may have been exposed to new risk outside the pending litigation. The consequences for the disclosing attorney can be catastrophic. Loss of client, fee disputes, malpractice claims, and bar sanctions are all foreseeable results. To make matters worse, the digital era has fundamentally changed the privilege landscape. E-discovery has become a significant problem in the U.S. litigation system because of the volumes and complexities of digital data, a constantly changing landscape, and the fear of breaching client confidentiality. The result is a skittish litigation bar that proceeds with extreme caution during discovery.

Electronic discovery in practice

A brief account of U.S. e-discovery practices is warranted at this point. In the context of commercial litigation, a representative exercise will follow from the issuing of a subpoena to the production of documents as follows:

Stage	Description
Notice	The defendant is served with a subpoena or other request for digital records. Counsel is engaged.
Identification and Preservation	The defendant issues a notice to its employees indicating litigation has begun, and that documents are not to be destroyed or deleted; works with its counsel and consultants to identify potential sources of relevant information, and negotiates discovery terms with opposing counsel. These terms may include a non-waiver agreement, which will be discussed at a later point in this article.
Collection	The potentially relevant information is collected in a forensically sound manner and forwarded to a specialist in litigation data processing.
Processing	The specialist, following specifications provided by counsel, culls the data using software filters, and removes duplicate records.
Document Review	The remaining information is loaded to a tool designed for legal Document Review and is screened by the defendant's counsel.
Production	Privileged records are segregated and logged. The remaining responsive records are prepared to negotiated specifications and produced to the requesting party.

Problems with the current process become apparent when data volumes and costs associated with this kind of discovery exercise are considered. At current pricing and productivity rates, representative estimates for the process described above might be as set out in scenario 1 below.

Scenario 1 – Electronic Discovery in 2009[9]

Stage	Records remaining after Stage	2009 Cost (U.S. $)	Percent of 2009 Cost
Notice	1,696,105,350	19,520	1.0%
Identification and preservation	30,000,600	44,784	2.2%
Collection	20,000,400	100,200	5.0%
Processing	1,038,981	220,550	11.0%
Document Review	42,598	1,515,542	75.8%
Production	42,598	99,855	5.0%
	Total cost	US$ 2,000,451	100.0%

These figures are representative of a moderately complex e-discovery exercise as it would be conducted

[7] *Ashish S. Joshi, 'Clawback in Commercial Litigation Agreements: Can You Unring a Bell?', Michigan Bar Journal, December 2008, 34, 36.*

[8] *Victor Stanley Inc. v. Creative Pipe Inc., 2008 WL 2221841 at 28 (D. Md. May 29, 2008).*

[9] *For the model and detailed analyses used to develop the data presented in the tables in this article, see Daniel R. Rizzolo, Representative Ediscovery Exercise Corporate Response to Discovery in Commercial Litigation (2009), http://www.rizzologroup.com/publications.html.*

in 2009 for a business named as a party in commercial litigation. They assume that forty employees of the business have been identified as witnesses and that archival media (e.g. backup tapes) are not included in the scope of discovery. Because of the size of the case, it is also assumed that junior attorneys in a law firm, rather than contract attorneys would perform a review of documents. The amounts shown include costs for in-house counsel and IT staff, as well as fees from law firms, forensic consultants and e-discovery service providers.

The major share of the expenditure is allocated to the manually intensive process of 'Document Review'. Before the electronic records may be turned over to an opponent, standard practice requires that a party's counsel review them, one item at a time, to determine whether they are relevant to the issues in the dispute and responsive to the requests in the subpoena (Relevance Review), and protected by the Attorney-Client Privilege or the Work-Product Doctrine (Privilege Review).

During the stages of Document Review and Production, privileged records are digitally flagged, segregated from the responsive population, scrutinized by counsel, and recorded at a summary level on a privilege log. This log is provided to the opposing party and the court as part of the eventual document production.

Interestingly, significant cost efficiencies have been realized in e-discovery in recent years. However, they have bypassed the task of Document Review. Had the discovery exercise described above been performed five years earlier using 2004 pricing, the results would have been as set out in scenario 2 below.

Scenario 2 – Electronic Discovery in 2004[10]

Stage	Records remaining after Stage	2004 Cost (U.S. $)	Percent of 2004 Cost
Notice	1,696,105,350	14,792	0.4%
Identification and preservation	30,000,600	33,924	1.0%
Collection	20,000,400	111,444	3.2%
Processing	1,038,981	2,080,957	59.2%
Document Review	42,598	1,194,450	34.0%
Production	42,598	76,944	2.2%
	Total cost	US$ 3,512,511	100.0%

The nominal cost of the Processing stage would have dropped 89 per cent between 2004 and 2009, due primarily to improvements in technology, standardization of techniques, and competition. For the same period, nominal Document Review costs would have increased by 27 per cent.

The high price of electronic Document Review has afflicted the U.S. litigation system. The economic effect is significant. Accurate statistics on the annual U.S. expenditure for discovery related Document Review are not available, however the magnitude of the problem is demonstrated by considering the following data:

> One analysis published in the U.S. projects that US$4 billion will be spent with e-discovery consultants and vendors in 2009.[11] This does not include the costs of Document Review or other fees paid to law firms. Nor does it reflect the investments that U.S. organizations are making in preventive measures.
>
> A 2006 study published by the accounting firm KPMG estimated that attorney Document Review accounts for 58-90 per cent of total expenditure on e-discovery.[12]
>
> Research conducted by the RAND Corporation Institute for Civil Justice in 2008 estimated that the cost of attorney document review is 70-90 per cent of total e-discovery expenditure.[13]

It is possible to extrapolate from this data that the approximate range of annual U.S. expenditure for Document Review is in the region of US$14-20 billion. Whether accurate or not, the estimate provides an illustration of the size of the problem.

There are other effects that result from the costs involved with e-discovery. The RAND Corporation report previously cited suggests that the cost of e-discovery has changed settlement models and negotiating power in U.S. litigation. This is manifest in a variety of ways, including a situation where a party with few digital documents to consider may take a more aggressive stance with an opponent that has a great deal of data. Parties that are prepared for e-discovery also have an advantage over parties that are not. There is also a disparity in cases where e-discovery costs are likely to

[10] *Daniel R. Rizzolo, Representative Ediscovery Exercise Corporate Response to Discovery in Commercial Litigation (2009), http://www.rizzologroup.com/publications.html.*

[11] *George Socha and Tom Gelbmann, 'Mining for Gold', Law Technology News, August 2008, available at http://www.lawtechnews.com/r5/showkiosk.asp?listing_id=2117297, in which the Sixth Annual Socha-Gelbmann Electronic Discovery Survey is discussed.*

[12] *KPMG LLP, A Revolution in e-Discovery: The Persuasive Economics of the Document Analytic Approach, (2006), 10, available at http://www.kpmg.ch/docs/20060812_A_Revolution_in_E-Discovery_Eine_Revolution_im_Bereich_e-discovery.pdf.*

[13] *James N. Dertouzos, Nicholas M. Pace ad Robert H. Anderson, The Legal and Economic Implications of Electronic Discovery Options for Future Research (Rand, 2008), 3, available at http://www.rand.org/pubs/occasional_papers/2008/RAND_OP183.pdf.*

exceed the value of the claim. A recent example of this is the case of *Spieker v. Quest Cherokee, LLC*, where e-discovery costs tripled the total amount at issue.[14] Attorneys are quick to embrace tactical advantage, and it should be no surprise that knowledgeable litigators have begun to use the high cost of e-discovery offensively.

The cost of Privilege Review is also affecting non-parties that are subpoenaed for records. Under U.S. rules, a litigant may not subject a third party to undue burden in complying with a subpoena. Should a third party feel that an onerous burden is being forced upon it, the third party may appeal to the presiding court for relief. U.S. courts tend to be more open to burden objections and cost shifting arguments when the recipient of a subpoena is a non-party. However, survey results published by the Sedona Conference in 2008 indicated that 73 per cent of the respondents had witnessed situations where non-parties were subject to undue burden in complying with a subpoena.[15]

The problems a third party faces in respect of discovery were demonstrated in the recent Federal appellate decision in the matter In re: *Fannie Mae Securities Litigation*. The U.S. Office of Federal Housing Enterprise Oversight (OFHEO), a government agency, failed to object in a timely manner to a third party subpoena for e-discovery. The lower court ordered OFHEO to comply, and the appellate court concurred. While OFHEO was not a party to the action, the agency was required to spend over US$6 million, representing nine percent of its annual budget, to meet the request. The bulk of the expenditure went towards Document Review.[16]

Finally, the high costs associated with Document Review lead to undue weight and consideration being given to what should essentially be mundane procedural exercises. As will be demonstrated in the next section, the strategic components of Document Review lend themselves to enhancement through technology. Unfortunately, the process is currently mired in expensive and time consuming manual tasks designed to avoid privilege waiver. This practice is draining resources that could be allocated to the substantive merits of a case.

The slow advance of automated Document Review

E-discovery has benefitted from significant technological efficiencies over the last five years. It is curious that the figures discussed above show an increase in the costs of Document Review between 2004 and 2009. This is primarily due to the effects of inflation and the increased billing rate of lawyers over that period. The productivity gains become clearer when the fiscal fluctuations are removed from the equation. A comparison of the real costs of Processing and Document Review, rather than the nominal costs, is illustrative. After adjusting for the increases in billing rate and inflation (and converting to equivalent 2009 dollars), the cost model shows as follows:[17]

Stage	2004 costs (in 2009 U.S. $	2009 costs (in U.S. $)	Percent increase /decrease
Processing	2,297,368	220,550	(90.4%)
Document Review	1,564,250	1,515,542	(3.1%)

While significant productivity gains have been realized in Processing, the productivity of Document Review has stagnated. The reason why Document Review has not benefited from the types of efficiencies that have affected e-discovery Processing is because of the nature of the technologies that are available for each task.

The principal purpose of the Processing stage in e-discovery is to reduce the number of documents by using automated filters. Commonly used data culling tools include programs that identify and eliminate duplicate files, text search engines for key word filtering, date extraction and query tools to limit the review population to a defined period, and programs that select or exclude specified file types. While these are powerful and increasingly sophisticated tools, their functionality is limited to a well-defined set of problems (e.g. find all records that are a bit-for-bit match, find all occurrences of a specified text string within the data population). The solutions to these problems, while technically challenging, are essentially mechanistic. The tasks they perform can be precisely defined. They lend themselves to solution through structured computer programs, which have become relatively generic and reusable across different types of data.

Automated Document Review (ADR) tools must solve problems that require complex analysis. The problems are issue and fact specific. They are greatly influenced by the nuances of human thought and language. Even a basic explanation of the rules to be followed in a

[14] *2008 WL 4758604 (D. Kan. Oct. 30, 2008), 2008 U.S. Dist. LEXIS 88103 (D. Kan. Oct. 30, 2008); Spieker v. Quest Cherokee, LLC, "Spieker II", 2009 U.S. Dist. LEXIS 62073 (D. Kan. July 21, 2009).*

[15] *The Sedona Conference, Commentary on Non-Party Production and Rule 45 Subpoenas 9 (2008).*

[16] *In re: Fannie Mae Securities Litigation, 2009 U.S. App. LEXIS 9 (D.C. App. Jan. 6, 2009).*

[17] *Daniel R. Rizzolo, Representative Ediscovery Exercise Corporate Response to Discovery in Commercial Litigation (2009), http://www.rizzologroup.com/publications.html.*

