

Identities and authentication

Fred Piper, Matt J.B. Robshaw and Scarlet Scwidarski-Grosche
Royal Holloway College, University of London

While the Office of Science and Technology commissioned this review, the views are those of the authors, are independent of Government and do not constitute Government policy.

1 INTRODUCTION

As the automation of business and the use of electronic forms of communication increases, we are challenged with finding equivalents to such basic security and crime prevention features as face-to-face recognition and hand written signatures. Although the technology is changing rapidly, when two people communicate electronically, for instance by email, they have usually lost the important facility of face-to-face recognition and need some other means of identifying each other. Similarly, while shoppers in the high street have confidence in the authenticity of the identities of the major stores that they frequent, it is not so easy for Internet shoppers to have confidence in the authenticity of a store's web site.

Our focus in this paper will be the identification and authentication of what we term primary objects whether these be a person, a device or even digital data. This is an area of major concern to government and business and weaknesses in currently used mechanisms allow exploitation by criminals. Many solutions have been proposed. Most are appropriate for certain environments and inappropriate for others. However, it is not always clear that the importance and/or magnitude of the problem are appreciated and there are many situations where systems are insecure and identity theft is a danger. It is perhaps worth noting that in some communities the terms identification and authentication are essentially synonymous. However, this is not true for biometrics where there is a very specific technical distinction (see Section 7).

We also discuss more speculative aspects of the overall authentication mechanism and consider the role of the infrastructure in supporting our applications and security mechanisms. Throughout industry there are islands of progress in securing different aspects of the information infrastructure. However, the issues are complex and, as yet, not particularly well formulated. Nevertheless, we believe that developments in the future will highlight many of these issues and help move the community towards developing a broader understanding of the problem and its solution.

2 AUTHENTICATION AND IDENTIFICATION IN CONTEXT

Before we can discuss the problems associated with identification and authentication in the electronic world we should consider some of the limitations of the techniques used in the pre-electronic age. Suppose, for instance, that you look up someone's telephone number in a directory and dial it. If someone answers and claims to be that person then can you be sure that they are the person you wish to contact? The realistic answer is yes, almost certainly. However, it is worthwhile to stress the assumptions you are making. The first is that your contact is the only person likely to pick up the phone and claim to be them. This, of course, may not be true. Even if the number is correct there may be two people at the same address with identical names, for example, mother and daughter. The phone call may have been re-routed by a criminal to an impostor who is deliberately impersonating the person you wish to contact. The second assumption is that the number in the directory is accurate. This is almost certainly true if you are relying on a paper version of the directory that has been published by, for instance, the telephone company and it would certainly be difficult for fraudsters to change people's entries. However, the same may not be true if you are relying on an electronic copy of the directory where obtaining assurance that the information has not been altered might be much more difficult.

There are three classic ways for a user to authenticate themselves to a system, which may be a computer, a network or another individual. They are: (i) something they own; (ii) something they know; and (iii) something they are (that is, a personal characteristic). Combinations of two or three authentication mechanisms are common, giving what is termed two- or three-factor authentication.

Typically the 'something owned' might be some form of token. If that token has some form of processing capability, for example, a smart card, then the something known might be a password to activate the device. The personal characteristic is likely to be some form of biometric, such as a fingerprint, and this might also be used as an activation process for a smart card. As we will see later, it is common for a smart card to have encryption capabilities and to contain cryptographic keys. The authentication process may then involve sophisticated protocols between the card and the authenticating device.

Before any of these techniques can be used, there must be an identification of the user to ensure that they have, in fact, been given the correct object or knowledge or that the characteristic being associated with them is, in fact, theirs. This process is fundamentally important and is one that is frequently overlooked. The problems associated with establishing someone's identity are significant and this is an area that needs considerable research. Most of the commonly used authentication techniques assume that there has been an initial, accurate identification and rely on that assumption. Authentication techniques that rely on something owned and/or something known cannot authenticate the individual. All that they do is equate the individual with either the knowledge or possession. If the original identification is not conducted properly then obvious disaster looms. Even if the identification process is accurate, there is always the danger that impostors may either obtain the knowledge or capture the token.

3 EXAMPLES AND POTENTIAL PROBLEMS

Your front door key authenticates you as an authorized entrant to the lock and allows you entry to your home. Similarly, a password may authenticate you to your computer and allow you to log on to the system. However, in neither case is the individual being authenticated. Anyone who has possession of the correct key or knows your password will also be accepted. In the case of door keys it is, of course, common for many people to have copies of the same key and, unless some crime is committed, there is no need to be able to determine which of the legitimate key holders opened the door. The situation with passwords is different and, usually, passwords are intended to be unique to specific users.

A common authentication mechanism for shoppers is the use of a credit card with the holder's signature on a white stripe on the back of the card. The card will usually contain the issuer's logo plus a hologram to make counterfeiting difficult. The authentication process is simple. When the user makes a purchase they sign a docket and the retailer compares the signature on the docket with that on the card. When this form of authentication was first introduced the standard way of distributing these cards was to use the postal system and to request the user to sign the stripe as soon as they received the card. This led to an obvious attack. If someone intercepted a card during its postal transmission they could write the cardholder's name (in their own handwriting, of course) on the white stripe. They would then present the card to a retailer and, after completing a purchase, write the cardholder's name on the docket. The two 'signatures', that is, the one on the card and the one on the docket, would agree and the retailer would allow the impostor to walk away with the goods. It is important to note here that the imposter is not actually forging the real user's signature. In fact, they have never seen it. What they are doing is exploiting a weakness in the card distribution infrastructure to get the real user's name written in the imposter's handwriting established as authentic. Having a more secure distribution system easily thwarts the attack. Many modern cards also incorporate the user's photo as an extra security check.

An example of a combination of something known and something owned is the process associated with withdrawing cash from Automatic Teller Machines (ATMs). When debit cards are used at ATMs, the card is inserted into the ATM and the customer enters a (secret) Personal Identification Number (PIN). The system then checks that the PIN entered is the one allocated to the account identified by the details encoded on the card's magnetic stripe. Because the human is taken out of the authentication process the logo and hologram are not relevant and attackers can make 'white cards' which, to the human eye, are clearly identifiable as forgeries but which contain the correct magnetic stripe details and are, therefore, accepted by the network. The security of the authentication system depends on the user keeping their PIN secret. It is worth noting that while the user is authenticated by the network, the user has no way of authenticating the system. This provides an example of one-way authentication and can lead to problems involving 'dummy' ATMs (Anderson 1994).

In the next few sections we introduce and discuss some of the fundamental technologies used for identification and authentication. We will begin with the simple password and then move through the use of cryptographic technologies to biometrics.

Throughout though we should keep in mind why we are doing this. In the traditional non-digital world we are constantly performing very sophisticated identification and authentication decisions without realizing it. Our eyes and minds are remarkably adept at recognizing visual clues – from people to objects – and the calculations we perform are remarkably complex to replicate on a

computer. Yet, in cyberspace, we are robbed of this remarkable skill and it is for this reason that we find the need to introduce a replacement set of complex calculations to prove our identity to one another.

4 PASSWORDS

The aim of identification (which is sometimes described as entity authentication within the cryptographic community (Menezes et al. 1996)) is to provide real-time assurance that some entity (a person or a device) is the one claimed. Suppose that a user wishes to authenticate him or herself to some authenticating server so that he or she can receive some service. In the subsequent text we sometimes refer to that user as the claimant, and we use the notion of 'authenticating server' in a general way to indicate any device that authenticates them.

The fixed password is perhaps the most common way of authenticating a user to a device such as a Personal Computer (PC). At the time of authentication, the user is identified by a user-name and prompted to enter a password. The system compares the entered password against the expected response and reacts accordingly.

It is clear that, at the very least, a fixed password system is vulnerable to interception and replay. The extent of the risk to which the system is exposed will depend on the deployment. If passwords are being transmitted across an unprotected network the risk is greater than with a closed system where there are limited opportunities for eavesdropping.

Since user-chosen passwords are memorable, they are likely to contain some inherent structure. Widely available password crackers, perhaps based on a dictionary search, appear to achieve a surprisingly high success rate when attacking user-chosen passwords. To help provide additional protection, some proposals deploy machine-generated passwords, but these are not especially popular. In fact, such approaches can sometimes be counter-productive since a password that is difficult to remember¹ is sometimes written down, which might degrade the overall security of the system.

A third point of weakness with password-based entity authentication is that a user's password must be stored within the system. Depending on the system and the deployment, this can be done in different ways. To reduce the risks of compromise, it is common practice to store, not the password, but the image of the password after it has been processed through a one-way function. A one-way function is a computational process which is easy to perform in one direction, but difficult to reverse (see Menezes et al. 1996). Thus, the image of the user-supplied password can be compared with the stored image of the password. So, even if the password file is stolen or compromised in some way, it remains difficult to reverse the images of the passwords to recover the original passwords.

While there are substantial problems with password-based authentication – and these problems mean that passwords are considered a weak form of authentication – it should be noted that passwords are very familiar and offer a wide-degree of user acceptability and convenience. Added to this, administration safeguards can be used to ensure that user-chosen passwords satisfy certain criteria to help set a minimum level of password acceptability. Users can also be forced to change their passwords at regular intervals, and systems often lock-down after a specific number of unsuccessful login attempts.

As a particular form of password, we have already mentioned the PIN. We are very familiar with this mechanism from the banking industry, but the PIN is little more than a short, restricted password. As far as password-based authentication is concerned, the PIN would appear to offer very little security. However, such PIN-based authentication does not depend on the PIN alone. The PIN is typically used in a two-factor authentication system and it is used in conjunction with the bank (or ATM) card. The user is only authenticated if they have the right card and they know the correct PIN.

Certainly fixed passwords have many good attributes – most particularly the simplicity and cost of administration – but the risk of password discovery, interception, and/or replay might be too great in some deployments. The one-time password is a move towards a stronger means of authentication.

In a one-time password scheme, a user's password may only be valid for a short time frame,

perhaps for 30 seconds or one minute. After this time the password changes. Thus, the window of opportunity for an attacker is greatly reduced since an intercepted password is unlikely to be of use in the future. All that we require is that the sequence of passwords should not be easy to predict after witnessing or intercepting a (potentially large) set of past passwords.

Such a one-time password scheme will require a moderate level of computational complexity and, to provide this, the user will typically be given a token. There are a variety of schemes available, but one of the largest deployments is probably RSA SecurID.² The RSA SecurID technology can be provided in a variety of form-factors, but most typical deployments will involve a tamper-resistant card. The card is issued to a specific user and each card contains a secret quantity, which is also held at the authenticating server. The one-time password is computed as a complex function of the physical time, the unknown secret stored in the card and, optionally, a user-supplied PIN. The password on the token display should then match the password anticipated by the server.

Such tokens have an inherent cost of deployment in terms of manufacturing the cards and supporting the infrastructure. User acceptability is reasonably high and products like RSA SecurID are likely to provide a good level of security when compared to fixed passwords. Interestingly, such technology can also be deployed in software and is supported on a variety of platforms including some mobile phones. In this way, the cost of card deployment is mitigated, and the mobile phone can be used as a convenient channel for deployment. In some sense, the phone itself becomes the token.

Even though the window of opportunity for interception and replay might be reduced with a one-time password mechanism, it is still not referred to as strong authentication. For this we require some real-time interaction and we will need to use some cryptographic algorithms.

5 INTERACTIVE AUTHENTICATION

We now move on to consider stronger forms of authentication. Instead of transferring a password (or a short-lived password) as a means of authentication, the authenticating server and the claimant (typically a card or token) perform some protocol or exchange of messages. In general terms, the server sends a challenge to the token and a cryptographic computation takes place within the card or token. The result is sent back to the server for verification. The cryptographic computation can be based on *secret (symmetric) key* or *public (asymmetric) key techniques*.

In classical cryptography, the two participants in a cryptographic exchange share the same secret key. Such algorithms are referred to as secret key, or symmetric, algorithms. While there is now a wide variety of algorithms for achieving both confidentiality and authentication, it can be difficult to guarantee that both participants have the same key.

Public key, or asymmetric, cryptography allows two participants in a cryptographic exchange to possess different keys. Such systems are designed so that knowledge of one key (the public key) does not allow an adversary to recover the other (the private key). This is a very powerful property and permits a range of interesting applications. As well as providing encryption capabilities – where the sender of a message uses the receiver's public key and only the intended receiver can recover the encrypted message³ – public key techniques can be used to provide what are termed digital signatures. Here, the signer of an electronic document performs a computation on the document using their own private key – this is an action only the signer of the document can perform – while the widely-available public key can be used by anyone to verify that signature.

Of course, public key cryptography is not free of its own unique problems. In particular, ensuring the availability of authenticated, valid, public keys is a significant problem and one that has proved to be practically tractable in only a few specific areas of deployment. Such a supporting infrastructure is referred to as a Public Key Infrastructure, or PKI.

Cryptographic algorithms are typically classified as shown in Table 1.

Table.1 Classification of cryptographic algorithms

	Confidentiality	Authentication
Secret Key (Symmetric) Cryptography	Block ciphers Stream ciphers	Message authentication Codes
Public Key (Asymmetric) Cryptography	Public key encryption	Digital signatures

5.1 Using Symmetric Techniques

There are a number of systems where authentication relies on the use of a secret that is shared between the two entities. In these systems the authentication may be one-way but there may also be the possibility of mutual, or two-way, authentication, where each entity authenticates the other. The basic principle is that two entities share a secret key that they believe is known only to them. The integrity of the authentication process is dependent on this key remaining secret. As a consequence, it will typically be delivered to the user in the form of a tamper-resistant token. Each participant regards the use of that secret as identifying the other.

It is worth noting that any mutual authentication scheme that relies on parties sharing keys can only be suitable for use between parties who trust each other. Furthermore, if one-way authentication is being provided then the verifier will need to be sure of the identity of the other party before agreeing to share a secret with them. Thus, for these systems, the original identification is likely to have taken place before the system is established. (For instance it might involve two personal friends who have known each other for a long time.) This is unlikely to be true in an e-commerce situation.

A variety of techniques is available to perform secret-key based authentication. For instance, in order to establish the identity of their partner an entity might issue a random number as a challenge which the entity claiming to be the partner will encrypt using the shared secret as a key. If the result of decrypting the response with the shared secret key gives the original challenge, then the identity of the partner is authenticated. This kind of interaction is referred to as a challenge-response protocol.

One of the most widespread deployments of a version of secret key challenge-response is in GSM phones. The phone's SIM card is a tamper-resistant module that contains a secret. This is used in a challenge-response authentication protocol when the phone authenticates itself to the network.

When compared to passwords and one-time passwords, we can see why challenge-response authentication is a stronger mechanism. Instead of the claimant sending a currently valid password (which might then be reused over a limited period) the server challenges the claimant to perform some cryptographic computation that can only be accomplished with possession of the secret key. It is this element of timeliness that adds additional strength to the authentication process.

Like the move to one-time passwords, the need to deploy tamper-resistant tokens and the need to manage secret keys make the deployment of such mechanisms more expensive than a simple password-based approach. The computational resources for a secret-key based challenge-response protocol are likely to be more substantial than the simpler algorithms that might be deployed in cheaper one-time password tokens. And again, like one-time passwords, it is common to combine such a token-based authentication method with a user-PIN to authenticate the holder of the token.

5.2 Using Asymmetric Techniques

Once we have moved on to more computationally sophisticated tokens, a variety of technologies becomes available. One problem with a secret-key based solution is that, by necessity, there is a direct prior relationship between the authenticating server and the claimant token since they both need to share the same secret key.

When we move to public key techniques we can deploy tokens that contain both the public and private key for that token. To verify some cryptographic operation, the public key can be sent (together with the transformed challenge) to the authenticating server. The entire computation can, therefore, be verified without any direct pre-existing relationship between the verifying server and the claimant token. While this simple description motivates the potential desirability of a public-key based solution, it also exhibits a minor sleight-of-hand! While there need not be a direct link between the claimant token and the authenticating server, we do require an indirect link, which allows the authenticity of the token's public key to be verified by the server. This indirect link is often provided by a PKI.

The basic principles involved when using public key cryptography are fairly straightforward. A user is associated with a pair of numbers. The first, the private key, can be used solely by that user and use of that number then effectively identifies that user. Thus, your private key has become, essentially, your electronic identity. The second, the public key, can be used by anyone to verify that the private key has been used. However, it is extremely difficult to determine the private key from the public key. In the real world there are a number of ways in which people can impersonate each other. A recent high profile case involved the theft of a passport by someone who resembled the passport photo of the legitimate holder. When public key cryptography is being used there are at least two different ways in which user A might try to impersonate user B. They might either obtain B's private key or they might try substituting their public key for that of B. The protection of private keys relies on strong algorithms and strong physical protection for the key. The prevention of public key substitution is not so straightforward and the most common current solution involves the use of a trusted third party, called a Certification Authority (CA). Users identify themselves to the CA which then uses its own private key to sign an 'electronic certificate' binding the user to their public key value. Third parties can then check the signature using the CA's public key and be confident that the CA is confirming that they have identified the user and confirmed the value of the user's public key. There are a number of protocols for using public keys to authenticate users. However, clearly, these rely on the accuracy and trustworthiness of the original identification performed by the CA.

Public key techniques can be used to provide encryption and digital signature capabilities and both can be used in a challenge-response protocol. As a consequence, many different proposals are available. The major downside in a move to public key techniques is the increase in computational resources required. Any claimant token is required to perform a computation involving the private key from the key pair since the cryptographic operation provides proof of the possession of the private key. The most widely deployed public key cryptographic algorithm is RSA, but the private key operation for this algorithm is computationally intensive. This means that tokens supporting RSA in this way need to be quite computationally sophisticated and this can put up the cost of deployment.

The most high profile description of a public-key challenge-response authentication protocol is perhaps that contained in the EMV specifications.⁴ These specifications include a range of security techniques for the credit card industry and these are based on digital signatures. While we do not need to look at the details of these techniques, the simplest, Static Data Authentication (SDA) would be used to authenticate data that is stored on the card. However, the remaining two, Dynamic Data Authentication (DDA) and Combined Data Authentication (CDA) would be used to provide assurance that a given card contains a specific private key. In the end, the most appropriate choice of protocol for deployment will be a business decision that depends on the sophistication of the adversary, the level of fraud anticipated, the security required, and the estimated cost of deployment.

Thus, in moving to asymmetric cryptography, we have replaced the management of secret keys (for secret key based challenge-response authentication protocols) by the management of public keys. To some observers this might be viewed as no real gain. However, we have gained in the

form of the deployment, which can now be much more open. A token can be issued by a different administrative entity to the one that performs the authentication. Provided there is an adequate PKI linking administrative domains in an appropriate way, authentication can proceed in a flexible and fluid way.

5.3 Public Key Based Identification Protocols

An alternative to challenge-response protocols based on public key techniques (which are computationally intensive) might be to use what are termed public key based interactive identification protocols (Menezes et al. 1996). However, we will not go into any details on these techniques in this report. Suffice it to say that such techniques offer a very different set of algorithms and protocols to those that have already been discussed. Nevertheless, they can still be used to provide strong authentication, and they typically do so with less computational complexity. This can result in the need for less sophisticated cards or tokens, which, in turn, can lead to cheaper deployment.

Of course, we do not get something for nothing. The algorithms and protocols we use do not provide public key encryption nor do they provide digital signatures. However, they remain public key techniques since the two participants in the interaction possess different keys. So, while they provide entity authentication at a particular instant, interactive identification protocols are much less versatile than the algorithms typically used in challenge-response protocols. Nevertheless, the reduction in computational complexity in these schemes means that interactive identification protocols are used in a range of practical deployments.

5.4 Summary

We can summarize the different identification (entity-authentication) mechanisms we have discussed in Table 2.

Table.2 Summary of identification (entity-authentication) mechanisms

Technique	Pros	Cons
Fixed passwords	Familiar Simple to use Simple admin.	Vulnerable to simple dictionary attacks, interception, replay. Closed-system deployment.
One-time passwords	Simple to use Relatively simple admin. Less vulnerable to replay attacks.	Typically needs a hardware-token plus supporting infrastructure. Closed-system deployment.
Challenge-response (secret key)	Simple to use Relatively simple admin Cryptographically strong	Typically needs a hardware-token plus supporting infrastructure. More complicated interaction than one-time passwords. Closed-system deployment.
Challenge-response (public key)	Simple to use Cryptographically strong. Open-system deployment possible.	Typically needs hardware token plus supporting infrastructure. Administration can be involved; protocols can be computationally intensive.
Identification protocols	Simple to use Cryptographically strong Open-system deployment possible. Computationally cheaper than PK challenge-response.	Typically needs hardware token plus supporting infrastructure. Administration can be involved. Less cryptographically versatile than PK challenge-response.

6 DOCUMENT (DATA) AUTHENTICATION

There are (at least) three easily recognized scenarios where authentication is used: to authenticate a document; to authenticate a device; and to authenticate a person. In this section, we focus on document authentication.

Document authentication is an important application of cryptographic techniques (Menezes et al. 1996; Stinson 2002). In many situations it is the authenticity of information that is far more important than its confidentiality. In this paper, we use the term 'document' in a particularly broad

way. Not only do we use it to cover the simple electronic representation of physical documents, but also to include other forms of digital information such as that carried on a bank card, or executable code downloaded into a device, or virtual and dynamic documents that might contain links to temporary resources on the Internet or might be generated dynamically using temporary data stored on some server.

Depending on the form of the document being authenticated different techniques are available. If a secret key based infrastructure is viable, then we might use *message authentication codes* for this form of data authentication. Here, the sender and receiver share a secret key and the sender enters the data and the secret key into an authentication algorithm to produce a cryptographic check sum which, clearly, depends on the data and the shared secret key. The receiver performs the same calculation using the received message and compares the value that they calculate with the value sent with the message. If these two values agree then the receiver accepts the message as being authentic in the sense that it is from the partner and has not been illegally changed.

However, given the nature of document distribution (that is, that one copy of a document might need to be authenticated by many recipients) public key techniques – digital signatures – would typically be more useful.

When considering the authentication of a document the complexity of the document can have a significant impact. When we sign a stand-alone electronic document, or some executable code, then it is (reasonably) obvious what we intend the signature to cover and what we intend the signature to mean. However, if a document were to contain links to, or be generated by, other temporary resources, then while the implication behind the signature might be obvious, its execution and continued validity can introduce some significant problems.

7 BIOMETRIC AUTHENTICATION

The only authentication techniques that attempt to authenticate a user directly, as opposed to relying on devices or knowledge assigned to that user, are biometrics. The term biometrics is derived from the Greek words bio (life) and metric (to measure). The field of biometrics is the measurement and statistical analysis of biological data. As we have already noted, the problem associated with authentication methods that are either knowledge based (something that you know) or token based (something that you have) is that they can be passed on to others, or can be lost, stolen or forgotten and therefore do not truly authenticate a person. This is different with biometric authentication methods; they cannot be passed on to others and losing them is pretty difficult (and even if the feature is 'lost', it cannot be used by somebody else). However, impersonation by forgery may be possible.

In biometrics, identification is a statement of who the user is and authentication is the process by which a claimed identity is verified. Identification, therefore, means to find the user in a group of users (a one-to-many comparison) and authentication means to verify the user's identity (a one-to-one comparison). A typical example of identification is matching a fingerprint found at a crime scene to a forensics database, and an example for authentication is withdrawing money from an ATM, where a person claims an identity by inserting a debit or credit card and that claimed identity can be verified by providing a biometric sample (or indeed a PIN).

In a biometric system a personal characteristic such as a fingerprint is used and the basic assumption of the authentication process is that a person's fingerprint identifies them uniquely or, more accurately, that the probability of two people having identical fingerprints is so small that it can be safely assumed to be zero. In a typical biometric system, a user will give a number of copies of the chosen biometric which are converted into bit patterns and stored on a template. When that user wishes to authenticate to the system they provide a copy of the chosen biometric and that copy is compared to the template. If the copy provided is 'close enough' to the template then the user is authenticated. A fundamental problem with applying biometrics is the determination of what is acceptable as close enough. If the demands are too stringent then the likelihood of the correct user being rejected may become too high. If, on the other hand, it is too lax then impostors will be accepted. From the security perspective it must be noted that the template needs protection since, if that can be altered, then the system fails. It is also important to recognize that the person needs to be identified correctly before providing the biometric. Otherwise the wrong person will be identified when that biometric is presented at any later date.

In fact the term biometrics has been generalized slightly from its original concept and we now talk about static and dynamic biometrics. While there are many techniques that are in differing states of technological advance, the main biometric methods in use today are the following:

1. Fingerprint recognition
2. Hand geometry reading
3. Iris scan
4. Retinal scan
5. Face recognition
6. Signature dynamics
7. Speech recognition

Biometric methods 1 to 5 are so-called static biometric methods (also physiological biometric methods); they depend on human features that are always present and which are fixed and do not change (at least in theory). Biometric methods 6 and 7 are so-called dynamic biometric methods (also behavioural biometric methods); they are related to a certain action of the user, to a behavioural habit.

In order to be applicable for authentication, a biometric method must fulfil the general requirements shown in Table 3. Different biometric methods fulfil these requirements to a different extent. However, it should be said that current technology is such that no biometric method fulfils all the requirements to the fullest extent.

Table.3 Requirements for authentication

Universality	Each person should have the characteristic.
Uniqueness	No two persons should have the same characteristic.
Permanence	The characteristic should neither change nor be altered.
Collectability	The characteristic can be measured quantitatively.
Performance	The characteristic can be efficiently measured in terms of accuracy, speed, robustness and resource requirements.
Acceptability	The characteristic should be acceptable to the public.
Circumvention	There should be no easy way to fool the system.

Before a biometric system can be used, the user has to enrol, providing the system with their biometric reference data. These data can be stored in a centralized or distributed database or on a smart card (the latter may be preferable in terms of security). The identification process (deciding whether a user belongs to the set of registered users) or verification process (deciding whether the user is who he/she claims to be) works as follows.

First, the biometric data are captured at the sensor. The biometric features are extracted and then, depending on the biometric method applied, a biometric template is produced. This biometric template is matched with one (in the case of verification) or many (in the case of identification) reference templates and an acceptance decision is made. Only the first two steps take place when the user enrolls in the system.

It is important to note that no two biometric templates match 100 per cent. Instead, the similarity between the two has to be calculated. In order to make a decision, a certain threshold is defined which maximizes the acceptance rate for authorized users and minimizes the acceptance rate for impostors. Two types of error are defined to measure the performance of biometric systems.

Type 1: The system fails to recognize a valid user (false rejections).

Type 2: The system accepts an impostor (false acceptance).

While there is not necessarily a precise link between the two error rates, in practice they are

typically linked. When the false rejection rate is kept small, the false acceptance rate tends to rise and vice-versa. The equal error rate denotes the case where the false rejection rate and the false acceptance rate are the same. This equal error rate is often the manufacturer's default setting of a biometric system. However, the acceptable rate for errors of either type depends on the application, and the acceptability of precise threshold values will vary. For instance, when access control to a high security location is required, it is likely that a relatively high level of Type 1 errors will be acceptable provided that the chances of a Type 2 error are essentially zero. On the other hand, when an application does not require very high security, but when the smooth operation of the system for the valid user is important, then the Type 1 error rate will be kept small and a higher Type 2 error rate may be tolerated.

The application domains for biometric authentication coincide with the applications domains of conventional authentication methods. They include access control to networks, physical access control to sites, entity identification and time and attendance control. Some applications that have attracted recent attention in the media include passports and identity cards. Many airports now issue smart cards with biometric templates to allow speedy checks at immigration. In the US the biometric is typically either hand geometry or fingerprint while at Heathrow Airport in the UK it is iris recognition.

Despite the fact that different authentication methods are frequently adequate for their purpose, they display obvious security limitations. Tokens can be lost or stolen and passwords and PINs can be guessed or copied. The use of biometrics can, at least in theory, remove some of these insecurities. Today we have reached the situation where some biometric authentication techniques have become quite advanced, with hardware and software technology reaching a sophisticated level. This is especially true in the dominant fingerprint recognition market (which accounts for about two-thirds of the overall biometric market). However, it is not clear that there is yet any reliable consistency in biometric products. Indeed, there seem to be other major inhibitors that might prevent the widespread use of this technology, including user privacy concerns and low acceptance by the public. These two issues are related because user privacy concerns are one reason why biometrics experiences a low acceptance by the public. Another reason is a perceived health threat (for example, from scanning of a retina or through direct contact with an unhygienic fingerprint sensor).

As opposed to conventional authentication methods, a biometric feature is bound to the user and cannot be removed without causing physical harm. This is a major advantage of biometrics, but it also leads to user privacy concerns. First, biometric features are often publicly available, such as photographs or fingerprints. Hence, it is not the biometric feature as such that can be used in biometric systems, but the fact that it comes from the live user. Biometric user templates, therefore, need to be secured to the highest degree. Otherwise, there is the risk that the user becomes trackable or that confidential user data are compromised. Depending on the biometric technology, this can include information on the user's health, for example, an image of the user's retina can reveal cardiovascular diseases.

Biometric systems consist of hardware, namely the sensors, and software for recognition. Moreover, biometric systems need to be integrated with existing systems and applications. There is, therefore, a significant research industry built around a variety of problems. With regard to the hardware, there is constant pressure to increase its reliability while maintaining low costs and maintenance. Reliability is also an issue with the software, particularly in terms of keeping the Type 1 and Type 2 error rates low while increasing the performance.

Security issues are essential and it seems that the main research contribution in security relates to system integration. Because the transmission of biometric data between different system components is one of the main weaknesses of a biometric system, much research is devoted to integrating a sensor (generally a fingerprint) onto a smart card and performing the whole recognition process on the card, without any biometric data ever leaving the card. However, although smart cards that contain sensors do exist, it appears that they are not yet sufficiently powerful to perform the whole recognition process.

The infrastructure surrounding biometric deployment such as the storage of biometric templates (in databases or on smart cards) raises many security issues, as does the secure transmission of biometric data during the authentication process. Note that the biometric data are transmitted from

the sensor to the feature extractor and then to the matching module and on to the application. All these transmissions need to be secured. Even though we might be attempting to minimize Type 1 and Type 2 errors, we will need to provide alternatives for users who inadvertently fail or are unable to use a given biometric test.

In an ideal world, where we assume that all the security problems of biometrics are solved, including the privacy issues, the user could use a single biometric authentication mechanism for all authentication purposes. The best solution might be to have the reference data (templates) stored on a smart card or another device that the user can carry. Further, the sensor and all phases of the recognition process might be integrated into this single device (smart card or pen). In this case, the biometric data of the user would never have to leave the secure environment of the user.

Biometric methods have the potential to make our life easier because the information being checked is part of us and always with us. This also has the potential to increase the security although considerable research still needs to be conducted before this is achieved. As a slight justification for this assertion we note that there are very few products on the market where the transmissions between sensor and matching module are secured.

No discussion about the security of biometrics would be complete, however, without reference to the work of Matsumoto who used sweets called gummy fingers to create forged fingerprints. This is an important piece of work that shook the biometrics community and a description of what he did can be found in his presentation 'Impact of Artificial Gummy Fingers on Fingerprint Systems'.⁵ This particularly highlights the issue of liveness detection; it is critical that the template used at both registration and authentication is from a live user!⁶

8 PROCESSES AND INFRASTRUCTURE

In some sense, the issues we have addressed are ones of primary object identification and authentication. A much more subtle, and in many ways, more complex set of issues arises when we consider secondary levels of authenticity and trustworthiness.

What do we mean by this? Authenticating (or identifying) a specific entity – a human or a computational device – is a very specific problem. Underpinning our solutions to this problem we often made the assumption that the supporting infrastructure would be trustworthy and that it would behave as intended. Without this trust it is difficult to imagine that any solutions would be viable and we tend to ignore, or at best acknowledge and then ignore, this issue.

One of the very obvious places where we are forced to address this issue, and which we have continually alluded to throughout this paper, is the issue of enrolment and registration. It has to be acknowledged that all the very wonderful security mechanisms we have mentioned in this paper could be easily compromised by something as simple as a failure in the administration procedure. We consider this in more detail in Section 10. However, this reliance on additional mechanisms, that often lie outside the scope of our technical solutions, should give us a rather uncomfortable feeling.

Thus, we need to be sure to pinpoint where technology is not enough, and it seems that many security problems occur when the human being directly interfaces to the digital world. This happens at user registration. An obvious second example is when a user is prompted for action by some application. For instance, when certificate verification fails after pointing a web browser to some secure site, how is the user supposed to react if (say) a certificate is rejected because it is being used either before the date of validity or after? How is the uneducated user supposed to react to this? Even better, how is the educated user supposed to react to this?

It would seem to require, therefore, a rather significant leap of faith to assume that the whole system will necessarily work as intended. Indeed, as more rights are managed and conferred by digital means – for instance, with the use of digital identification cards as a way of providing access to services – the stakes are raised and the illicit gains of fraudulent behaviour increase.

These examples are merely prompts for discussion. There are, in fact, many different aspects involved in considering the trustworthiness of the infrastructure. Gradually, safeguards for the infrastructure are being introduced, but in a piecemeal manner. There is a wide range of industry bodies working to protect their particular parts of the digital fabric. Here we list a few diverse issues:

- We are already in a position where our consumer devices – PCs, PDAs (Personal Digital Assistants) and mobile phones – can import code that changes their functionality. To help decide between good and potentially malicious code, initiatives such as code-signing have been developed that allow a device to digitally verify the authenticity of a particular application.
- Smartcard manufacturers spend millions of dollars every year in research on the best ways to provide additional security features on the cards they produce. The whole integrity of a smart-card based solution is dependent on the fact that the smart card offers a secure storage and computation environment.
- Secure computing initiatives such as Microsoft's NGSCB (Next-Generation Secure Computing Base), formerly called Palladium (Microsoft 2003) and TCGA (Trusted Computing Platform Alliance) (TCG 2002) attempt to provide a secure and trusted computing environment.
- There are industry initiatives to promote good engineering and secure coding practices. How can we be sure that good security implementation practices are used within deployments?

It is the control of information that is fast becoming the issue of our times. Ubiquitous computing and ad hoc networking means that our personal information can be increasingly used without our knowledge. We might delegate our devices to authenticate on our behalf as they seek out wireless-based interactions with other devices in the room. Yet, it is not always clear that this is to the user's advantage if the user is concerned about their privacy. In a different field, what is the best way to authenticate dynamic documents that either point to transitory information or use transitory information in their construction?

Thus, the use to which we put information is an increasing concern. One step forward has been the introduction of meta-data where information has a description of its own use and functionality attached. An extension of the very same concern occupies the minds of executives of companies providing information for our entertainment (that is, music and videos). The use and potential misuse of this information drives the whole area of Digital Rights Management (DRM) and, very interestingly, leads us full circle to the issue of registration. Indeed, one DRM solution that is much touted is to effectively 'register' the devices on which information can be accessed. Note that, unlike the case of human registration where there is no digital interface, registering a device is technically rather straightforward, despite the formidable privacy and consumer-acceptance issues involved.

We might also broaden the concept of trustworthiness to include reliability. Catastrophic failures are usually easy to detect and, more often than not, to fix. What can be harder to sort out are the problems that are intermittent, thereby leading to degradation rather than an outright failure in service. In the case of a communications infrastructure it is hard to imagine at what stage a loss of service would become noticeable. Would a network operating at 50 per cent efficiency be noticeably slow to the user? Depending on the network and the application, perhaps not. Would a network operating at 50 per cent efficiency lead to measurable losses if accumulated over a sufficiently long time? Depending on the network and the application, perhaps so.

We feel that it is particularly important to begin to speculate on the continued robustness of the supporting infrastructure. As agent software is increasingly used for workflow, middleware and automatic negotiation, the identification and authentication of software and data objects as well as people will grow in importance.

9 THE PROBLEM OF ORIGINAL IDENTIFICATION

Suppose for the moment that we are confident that we know the identity of Mrs Mary Smith. If her son wanted his identity to be John Smith son of Mrs Mary Smith, then there is only one stage of his life at which we can have total confidence in this claim. That is while the umbilical cord is still joining John to his mother Mary.

As soon as the cord is cut, procedures are required to ensure that there is some form of binding between John and his identity. If these procedures go wrong, for whatever reason, then either someone else will have John's identity or John will have the wrong identity, or both. If we wish to be confident that John Smith has the correct identity for the rest of his life, it could be argued that the binding must take place while the umbilical cord still provides an undeniable physical link

between the two parties. Two obvious options that are often discussed are: taking a DNA sample or the physical insertion into the baby John Smith's body of a microchip containing his identity. If the DNA sample is taken then procedures are still needed to ensure that the record that associates the DNA sample with John Smith is accurate and cannot be altered at any time during John Smith's lifetime. If the chip is inserted into John's body then there need to be assurances that this cannot be removed or replaced by another person and that the information stored on the chip cannot later be changed via remote access.

If either of these two options is to be given serious consideration then it will be necessary to explore people's attitudes to such perceived intrusive measures as taking DNA samples or inserting chips at birth. Another research area is the likely physical consequences of implanting chips in someone's body for life plus, of course, the durability of the chip.

The example of John and Mary Smith plus the proposal of two somewhat extreme solutions are included to illustrate a point rather than to discuss a specific situation. Indeed, many would argue that there are very few occasions when anyone would deliberately steal a baby's identity and that less severe solutions are sufficient to prevent accidental mix-ups. Be that as it may, there is no doubt that the general problem of identifying 'the original' is difficult and frequently overlooked. One area where it is particularly relevant concerns the concept of digital evidence. If, for example, a digital image is to be produced as evidence then it will be necessary to protect it from alteration. However, if the digital image is obtained using a digital camera then a question that needs to be answered is how do we know that the image for which the protection was provided is the original? If the protection, for example, a digital signature, is constructed and attached inside the camera then we need assurances about the tamper-resistance of the camera. If, on the other hand, it is applied using another device then we will need to rely on procedures to ensure that it was not changed before the protection was applied.

The problems associated with establishing identity are frequently ignored in many discussions relating to, for instance, the issuance of passports, digital certificates and all the authentication techniques that rely on biometrics. Most of the current methods of establishing identity seem to depend on the fact that that person's identity has already been established somewhere else. Each new process is merely endorsing the old one. For instance, a passport application requires the production of a birth certificate. If the birth certificate is not the correct one for the person claiming the identity, then the passport may be issued to the wrong person. Similarly, any record associating a biometric template with a specific individual has a built-in assumption that the identity of the person was correct at the time that the biometric samples were taken. We could go on and produce numerous examples where the ability to impersonate someone at some point in the registration stage implies the ability to steal their identity and impersonate them for life.

In this paper we have looked at a number of identification and authentication techniques. If they are to be trusted then it is vital that the process of what we might call original identification is adequate. This is an area where much research is needed. In addition to the necessary technical research into the security of the technology, more basic research is needed into the effectiveness of the identification processes used for such important, everyday processes such as passport applications and bank account or credit card applications, including their costs and failure rates.

10 CONCLUSION

With regard to the identification and authentication of primary subjects cryptography is, at present, the only 'strong' mechanism available to us. It is now in widespread use and helps to support a wide range of identification and authentication mechanisms. However, we should not be in complete thrall to the technology since there may well be situations where it can be subverted or used as a tool for denial of service.

As with all security solutions, an outstanding question is how much 'security' or 'strength' is enough, and in what parts of the identification and authentication process is it essential to use 'very strong' methods? We also note the requirement that procedural approaches and architectural solutions (separation of duties) be used to significantly reduce the risk of social engineering vulnerabilities in what might otherwise be 'trustworthy' processes.

Finally, we observe that the increased reliance on the automatic creation and distribution of information in all its guises, places interesting and novel requirements on the trustworthiness and

reliability of the supporting infrastructure.

NOTES

- 1 Some software is claimed to generate high-quality yet pronounceable passwords.
- 2 See <http://www.rsasecurity.com/products/secured/>, accessed 17 Apr. 04; note RSA is the acronym for Rivest, Shamir & Adleman.
- 3 In a little more detail, it is typical to use encryption with a public key cryptosystem when it is necessary to agree on a symmetric key between users since this is a faster way of encrypting bulk data.
- 4 See EMVCo., The EMV Specifications at <http://www.emvco.org>, accessed 17 Apr. 04.
- 5 Available via <http://www.itu.int/itudoc/itu-t/workshop/security/present/s5p4.pdf>, accessed 17 Apr. 04.
- 6 For more detailed information on biometrics, see Jain et al. (2003) and Woodward (2003). Also, there are several independent organizations focusing on biometrics, namely the International Biometric Group, <http://www.biometricgroup.com/>, The Biometric Consortium, <http://www.biometrics.org/> and the Association for Biometrics (see <http://www.afb.org.uk/>, all accessed 17 Apr. 04.

REFERENCES

- Anderson, R. (1994), 'Why Cryptosystems Fail', *Communications of the ACM* 37(11): 32-40.
- EMVCo (2000), The EMV4.0 Specifications, www.emvco.org, accessed 17 Apr. 04.
- Jain, A.K, Maltoni, D. and Maio, D. (2003), *Handbook of Fingerprint Recognition*, New York: Springer-Verlag.
- Menezes, A., Van Oorschot, P. and Vanstone, S. (1996), *The Handbook of Applied Cryptography*, London: CRC Press.
- Microsoft (2003), *NGSCB: Trusted Computing Base and Software Authentication*, Microsoft Corporation, Windows Platform Design Notes.
- Stinson, D. (2002), *Cryptography: Theory and Practice*, 2nd edn, London: CRC Press.
- TCG (2002), *Trusted Computing Platform Alliance (TCPA) Main Specification*, Version 1.1b, The Trusted Computing Group, Portland, February, <http://www.trustedcomputing.org/home>, accessed 17 Apr. 04.
- Woodward, J. (2003), *Biometrics and Strong Authentication*, Emeryville CA: Osborne/McGraw-Hill.