WHY X.509-STYLE PKI IS A POOR SOLUTION FOR ACCESS CONTROL AND DATA AUTHENTICATION

Dr. Stefan Brands
brands@credentica.com
May 2002

Abstract: Digital certificates were invented a quarter of a century ago, for the specific purpose of facilitating the encryption of messages using public key cryptography. They were never intended to be the basis for access control and data authentication. Consequently, today's commercial PKI offerings (which are all based on digital identity certificates and attribute certificates that are linked to identity certificates) do not take into account any of the scientific advancements in access control and data authentication made in the past 25 years. They offer no protection against the lending of access rights and a myriad of other attacks, do not respect any of the privacy principles that are commonly codified in law, and suffer from a host of performance drawbacks (which certificates were intended to eliminate in the first place) due to their heavy reliance on trusted online repositories. This paper describes in detail the many problems of PKI based on X.509-style certificates when used for access control.

1. Introduction

Digital certificates are widely perceived to be the most secure technique for providing authorization, authentication, and accountability in electronic environments. Since they are just sequences of zeros and ones, they can be verified with 100 percent accuracy by computers and can be transferred electronically and instantaneously without human intervention. Owing to their special mathematical structure, it would take millions of years to forge a digital certificate, even when using all of the world's computing power.

An infrastructure that revolves around the distribution and management of public keys and digital certificates is commonly referred to as a Public Key Infrastructure (PKI). Today's prevailing PKI methods rely on digital *identity* certificates. An identity certificate is a digital signature of a trusted entity, called the *Certificate Authority* (CA), that binds a *public key* of an individual to his or her name. The name could be the individual's true name, but it could also be a Social Security number or any other data that at least the CA can readily associate with the individual. Identity certificates were invented in 1978, at the dawn of modern cryptography, for a very specific purpose: to enable the sender of a message to encrypt that message under the public key of the intended recipient, by binding that public key to the recipient's identity. Over the course of time, the usage of digital identity certificates has been stretched all the way into the realm of digitally signed access control and attribute authentication, without any serious attempts to assess the security, privacy, and performance consequences. As an unfortunate consequence, digital identity certificates are being used today as authenticated pointers into all sorts of online and offline databases, allowing verifiers to look up any information about the certificate holders they are interested in. This is similar to the way organizations use Social Security Numbers, but with stronger authentication and much broader exposure.

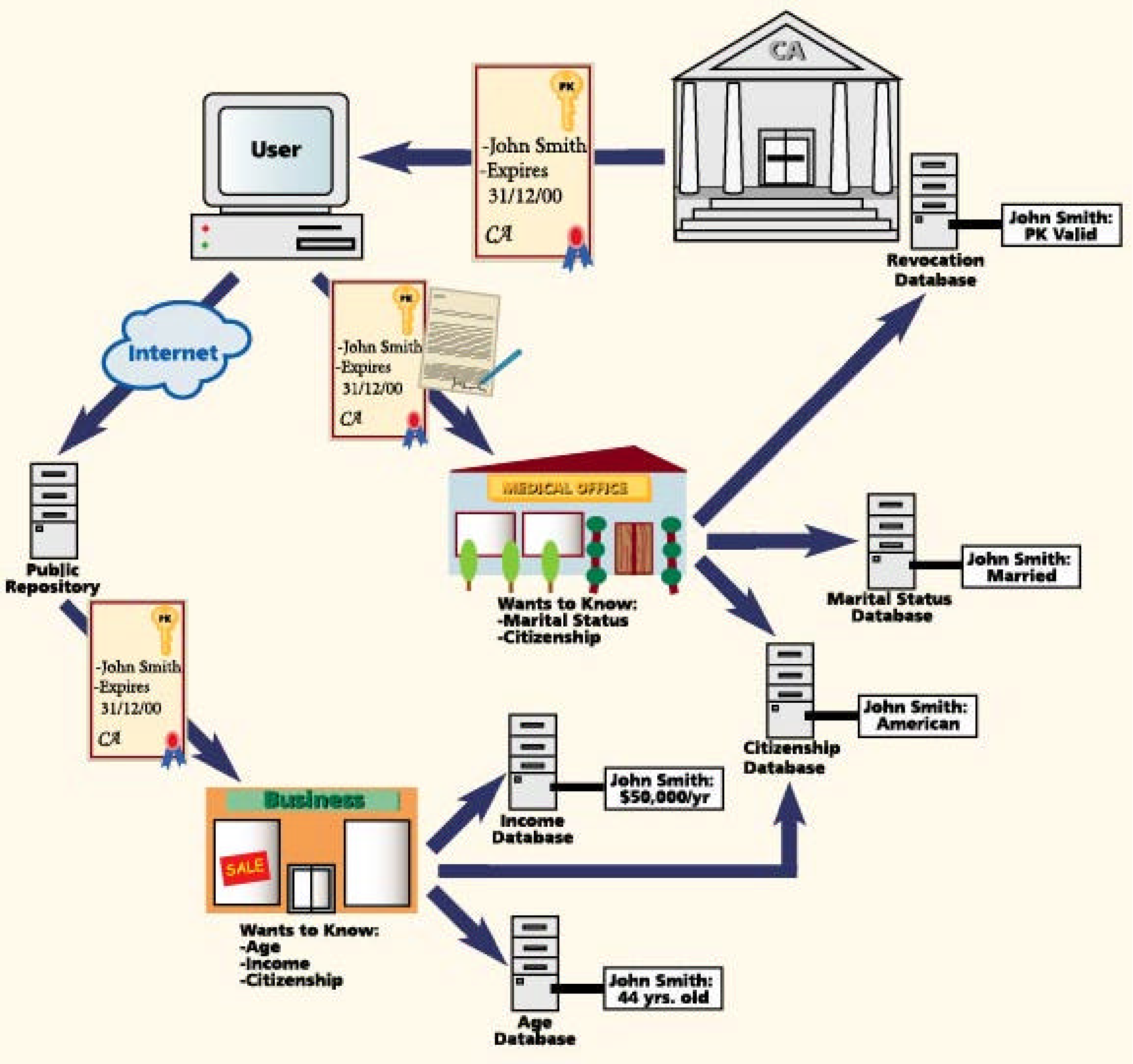

Figure 1: The identity certificate model

Figure 1 illustrates the principle. John Smith receives from a CA an identity certificate that binds his name to his public key. The certificate also specifies an expiry date and possibly other data (not shown here). To engage in a transaction with a verifier, in this case a medical office, John Smith sends his certificate and uses his secret key to authenticate the transaction. John's personal digital signature prevents a replay attack by wiretappers, as well as by the medical office, assuming the message that John signs contains a unique ``challenge." The medical office verifies the binding between the name and the public key specified in the certificate by applying a trusted copy of the CA's public key. Then, it uses the certificate to retrieve John Smith's marital status and citizenship, and any other data it is interested in, by looking into appropriate databases. The medical office also consults an online revocation database to make sure that John Smith's certificate has not been revoked. The alternative would be to use a Certificate Revocation List (CRL). The figure furthermore shows John sending his certificate to an online public repository, from where it is retrieved by a business. Note that the business can look up any data about John Smith without his involvement or awareness.

Security Drawbacks

PKI based on identity and X.509-style attribute certificates does, at best, a mediocre job of protecting electronic security:

1. Systemically relying on user identification more often enables than prevents fraud, as the dramatic rise in identity fraud over the past decade shows. Criminals who manage to steal identity certificates or to assume the identities of unwitting people will be able to misuse certificates in cyberspace on a global scale, while their victims take the blame. Also, CAs will have to establish identities on the basis of legacy paper-based systems, and will thus inherit their insecurity. For example, identities may have been erroneously or maliciously swapped or forged.

2. By relying heavily on central data repositories, identity certificates push the door wide open to devastating abuses of security holes. It is difficult for organizations to protect their online databases against misuse by hackers, let alone by insiders. Furthermore, data records may be outdated, and may be the result of misattributions due to identity theft.

3. Identity certificates offer no intrinsic cryptographic protection to discourage certificate holders from transferring (copies of) their credentials and access rights to other parties: the secret key of a certificate holder is simply a random number, and so revealing it to someone else has no direct negative consequences for that certificate holder. Particularly, when X.509-style identity and attribute certificates are used in closed applications they can only be used in limited ways, so that giving away copies of a secret key will not enable others to misrepresent themselves as the legitimate certificate holder in other applications. This defeats the entire purpose of identity certificates.

4. If the secret key of a certificate is generated and stored on a personal computer or the like, it is virtually impossible to prevent its compromise, loss, disclosure, modification, and unauthorized use. Processing an X.509-style certificate on a smartcard, on the other hand, suffers from numerous drawbacks:

    a. The computational, communication, and storage requirements exceed those that today's simple 8-bit smartcards offer. The addition of complex circuitry and software to a smartcard is expensive, can easily lead to new weaknesses in the internal defense mechanisms, and adversely affects reliability. The ability to protect smartcards against attacks such as differential power analysis hinges on having enough capability and space for a software solution.

    b. Since the goal of smartcards is to shield their internal operations, it is virtually impossible to verify that a card does not leak its card identifier, its access control code, data from other applications running on the same device, and so on; CAs must have strong trust in the honesty of their smartcard suppliers.

    c. The smartcard must be relied on to protect the security interests of its holder. Since standard smartcards do not have their own display and keyboard, user identification data must be entered on a terminal communicating with the card, and this terminal must be trusted not to capture the user's identification data. Likewise, any results that the card wants to communicate to its holder must be displayed on the terminal. The result is that a variety of fake-terminal attacks become possible.

d. There is no way to verify that the secret keys within the smartcards are generated in such a manner that others cannot guess them. In particular, it is very hard to guarantee that the CA or the smartcard supplier cannot reconstruct all the secret keys. This makes the legal status of digital signatures highly doubtful.

5. While revocation is an exceptional circumstance, the task of verifiers to check the revocation status of unexpired certificates is not. They must either have the certificate status validated at the time of the transaction or regularly download a CRL update. This gives the Revocation Authority the power to falsely deny access to targeted certificate holders, by blacklisting their public keys. Worse, it gives an adversary the power to take the system down by breaking into the CRL repository, thereby defeating the main security advantage of off-line certificate verification. Similarly, any uniquely identifying data in a certificate (such as a key holder identifier, the public key, or the CA's signature) can be misused to deny a key holder access to PKI services, and to block his or her communication attempts in real time. For example, blacklists can be built into Internet routers, and transaction-generated data conducted with target public keys can be filtered out by surveillance tools.

6. Any digital signature made by a certificate holder can be used as non-repudiable transaction evidence not only by the legitimate receiver, but also by anyone else who sees the signed statement, including parties that have no right to the information. This leads to all sorts of risks, including conflicts with privacy legislation. It also exposes the CA, the verifier, and other legitimate parties partaking in the transaction to potentially unlimited legal liability.

7. To issue an X.509-style identity or attribute certificate, the CA must know the identity and any other attributes that go into the certificate. This prevents any data separation when that data resides in the databases of other organizations. (This is actually the main reason why the business model of Managed CA Services has never caught on.)

Privacy Dangers

Digital identity certificates are widely touted as a means to protect privacy, since messages can be encrypted with the public key of the intended receiver. Nothing can be further from the truth. Confidentiality (that is, preventing a wiretapper from decrypting intercepted messages) does nothing to prevent all kinds of other privacy threats, such as parties tracing, linking, selling, or misusing the data in whatever manner they see fit. In fact, identity certificates have devastating consequences for privacy:

1. The actions of certificate holders can readily be traced and linked on the basis of the certificates presented, since a digital certificate is a unique bit string. In many cases, it is not only the verifiers who can trace and link the actions of the individuals they interact with, but also the CA. Specifically, this will be the case

when verifiers deposit transcripts of their interactions to the CA (to enable central fraud detection or the computation of visitor behavior statistics), and in any closed application in which verifiers are effectively the same entity as the CA. This enables the compilation and distribution of detailed dossiers about individuals' habits, behavior, movements, preferences, characteristics, and so on. Furthermore, all the dossiers compiled by linking and tracing the actions of participants in one PKI can be tied to the dossiers compiled in other PKIs, and can be linked to a myriad of other sources of personal information.

2. When digital certificates are implemented in smartcards and other tamper-resistant devices, the privacy of certificate holders is even more at risk. A card can directly leak personal data, for instance by sending along additional data when engaging in a protocol, by encoding information in message fields or random numbers, and so on. Indeed, as Moreno, the inventor of the first generation of smartcards in the early seventies, warned, smartcards have the potential to become ``Big Brother's little helper."

3. CRLs are distributed to all verifiers, and potentially to anyone who requests them; in this manner, entities can collect data about key holders they have never communicated or transacted with. Online certificate validation services are even worse: they allow anyone to verify not only negative data but also positive data (such as the mere fact that one is a participant), and enable the Revocation Authority to learn in real time who communicates with whom. They are also undesirable from the perspective of verifiers, since third parties learn the identities of their visitors, their peak hours, and other data that is either competitive data or that must legally be protected.

These undesirable properties make it impossible for individuals and verifiers alike to control how much data they actually disclose to other parties in the system. Identity certificates that specify a "pseudonym" instead of a real name, an approach that is often proposed in digital signature law, are not a valid solution:

- "Pseudonymous" certificates that can only be obtained by certificate applicants who identify themselves do not prevent tracing; they provide no more privacy than Social Security numbers, credit card numbers, and health registration numbers, all of which a verifier can readily link to an identity by looking into any number of database entries to which the "pseudonyms" point. Moreover, the correlation is at all times known to the CA, typically the most powerful party in the system. Similarly, X.509-style attribute certificates that do not explicitly specify true names can be linked and traced as easily as identity certificates, on the basis of their public key or the signature of the CA. In fact, they worsen the privacy problem, since the dossiers that CAs, verifiers, and wiretappers can compile are even more complete.

- The alternative of not requiring certificate applicants to identify themselves at all offers better privacy, but makes it impossible to ensure accountability, and to protect against lending, copying, discarding, and other misuses of certificates; it is not even possible to contain the damages due to fraud. The approach is also impractical in almost all applications. Even if the CA certifies only personal attributes that do not identify the certificate holder, such as age and marital status, often the only way for the CA to verify the attributes is by establishing the applicant's identity and using this to look up the attributes in a trusted database. Also, registration without identification may be difficult and would prevent applicants from building a long-term relation with the CA. More generally, the idea of issuing an identity certificate to an unidentified party does not make much sense in the first place.

Performance Drawbacks

The central database architecture on which X.509-style digital certificates rely creates numerous performance problems:

1. The transaction process requires a sufficient delay to identify and correct frauds or other undesirable conditions. This may result in organizations not being able to serve as many customers as they could otherwise, or in customers leaving and going elsewhere (especially when browsing on the Internet).

2. Online certificate validation is costly, hard to scale to large communities, and suffers from all the security problems of the central database paradigm. The distribution of CRLs or CRL-updates, on the other hand, requires verifiers to manage their own versions of a CRL and to deal with certificates they are not interested in, and creates a lag between the time a certificate becomes invalid and when it appears on the next CRL update. If validity periods are long, CRLs will grow and additional computing resources are needed for searching and storing them.

3. There is significant uncertainty in the outcome of the transaction process, because the certificate verifier makes its authorization decision on the basis of remotely stored data that may be erroneous or irrelevant, or simply because the online connection fails (e.g., due to peak load or a natural disaster). Also, requests for central database look-up may be dishonored for many reasons and may be expensive (many large databases are operated by commercial organizations such as consumer reporting bureaus).

4. In case the representatives of an organization are spread out geographically, central database verification may be expensive (due to communication costs or the difficulty of dealing with peak load) or may simply not be an option because of the absence of network connections.

These drawbacks can be alleviated by using an X.509-style attribute certificate that contains all the data for the verifier within it. However, this introduces a serious problem due to the fact that all the attributes within a certificate are systematically revealed when showing the certificate. Quoting the authors of the SPKI/SDSI standard, ``because […] certificates will carry information that, taken together over all certificates, might constitute a dossier and therefore a privacy violation, each […] certificate should carry the minimum information necessary to get a job done." This implies that all the data pertaining to a single party must be distributed across many digital certificates that are issued to that party, which introduces serious administrative overhead. Moreover, this means that a single certificate is not suitable to serve its holder in multiple applications.