10
And Then It All Changed

The Advanced Encryption Standard

When did the Third Millennium begin? On January 1, 2000? A year later? On September 11, 2001? In cryptography, it began on January 2, 1997 with an inconspicuous notice in the *Federal Register* that marked the beginning of a project to replace the aging US Data Encryption Standard (DES).

The contrast with the events that led to the adoption of DES two decades earlier could hardly have been greater. Although the bones of the formal process were the same—a call for proposals in the *Federal Register* and ultimate selection by the Department of Commerce with the advice of the National Security Agency—everything else was different. In 1997, the notice called, not for algorithms, but for comments on proposed algorithm specifications. Whereas the previous standard appeared to have been designed to be just strong enough for non-national-security applications, the new proposal aimed for the highest grade security: a 128-bit block size and keys of 128, 192, or 256 bits.

Two rounds of comments on criteria were followed by a call for algorithms, due June 15, 1998. Twenty-one algorithms were submitted. Of these, fifteen met NIST's complex set of requirements for documentation, implementation, tests, and rationale intended partly to facilitate evaluation and partly to discourage frivolous submissions. At first glance the response was international—ten algorithms submitted by groups outside the US and five by groups within. At second glance it was even more international than that. All but one of the US candidates included non-US nationals on their design teams.[1]

For nearly a year and a half, all fifteen were under study, a process highlighted by two public conferences, one in California and one in Rome. In the late summer of 1999, the number was reduced to five: three American (MARS, from IBM; RC6, designed by Ron Rivest and colleagues from RSA Data Security; and Twofish, designed by Bruce Schneier and his colleagues) and two European (Serpent, designed by cryptographers from the United Kingdom, Israel, and Norway; and Rijndael, designed by two Belgian cryptographers).

Differences between the Data Encryption Standard and its advanced descendant were more than programmatic. The description of DES in FIPS-46 was entirely in engineering terms, speaking of lookup tables and bits and shifts. Over time these structures came to be viewed in more abstract mathematical terms and the mathematical descriptions gave rise to cryptanalytic techniques. The description of Rijndael was given in mathematical terms (Landau 2004). Certainly it had lookup tables, and bits, and shifts, but these followed from rather than preceded the mathematical notions. In a quarter-century, the field had matured. The Advanced Encryption Standard was truly a second-generation cipher.

Why the difference? In the early 1990s two powerful cryptanalytic techniques had been developed in the public research community: differential cryptanalysis, which infers key bits by comparing input and output differences of pairs of encrypted texts, and linear cryptanalysis, which infers them from linear relationships between the input and output bits.[2] At the same time, other researchers were using ideas about algebraic structure to develop block-structured cryptosystems resistant to mathematical attacks. Polynomials proved key to this. Building on approaches developed for error-correcting codes, researchers used algebraic structure, particularly the theory of finite fields, to create a methodology for constructing cryptosystems provably secure against differential and linear cryptanalysis.

Study of the five finalists continued through another year, and another conference before the winner was announced on October 2, 2000. It was the Belgian submission, Rijndael. Bureaucratic processes dragged on for more than a year before the standard received the required signature of the Secretary of Commerce, but on November 26, 2001, the United States

adopted a cryptographic system designed outside its own borders as the "Advanced Encryption Standard."

The adoption of the new standard created an odd situation. AES seemed to be as strong a cryptosystem as anyone could ask for, yet its standing was the same as that of its never-too-strong predecessor. Why couldn't AES be approved as a Type II algorithm, allowing equipment whose functioning was public to be applied to a wide range of government applications? The question hung fire for a year and a half. In June 2003, it was answered in a way that even AES's strongest proponents hadn't dared expect. AES was declared a Type I algorithm, approved for the protection of all levels of classified traffic.

The instrument of this approval was Policy Number 15 of the Committee on National Security Systems (CNSS 15). Curiously, the memorandum was For Official Use Only. It was announced, however, in a virtually identical fact sheet (CNSS15) which appeared in August. The memo is written in a tedious bureaucratic style largely devoted to warning its readers that having an approved algorithm whose workings they know does not entitle them to use anything other than approved implementations for protecting classified information. The important paragraph, however, is perfectly clear:

> The design and strength of all key lengths of the AES algorithm (i.e., 128, 192 and 256) are sufficient to protect classified information up to the SECRET level. TOP SECRET information will require use of either the 192 or 256 key lengths.

Although the Advanced Encryption Standard was now approved for protecting all levels of classified information, this declaration was very much a matter of principle and would remain so until actual equipment was designed, built, approved, and fielded. Nonetheless, the memo had real significance. It showed COMSEC equipment manufacturers a new route to serving their classified markets. Equipment would inevitably follow.

Elliptic Curves, Secure Hash Algorithms, and Suite B

The adoption of AES was not the only radical change working its way through cryptography or through government cryptography in particu-

lar. Advances in computing and discrete mathematics had begun to make the Diffie-Hellman and RSA public-key cryptosystems uncomfortably expensive to operate securely. The computer scientists Arjen Lenstra and Eric Verheul compiled tables of the equivalences of the workfactors of a variety of cryptosystems (Lenstra and Verheul 2000). In order for either of the traditional public-key systems to match the security of AES, they would need to employ keys thousands of bits long and do millions of instructions in each operation.

Fortunately, a solution had been coming to hand since the mid-1980s when two mathematicians, Neal Koblitz from the University of Washington and Victor Miller from IBM,[3] developed a new approach, nearly simultaneously. Put briefly, by using more complicated arithmetic than RSA or Diffie-Hellman, you can make the numbers smaller and get the same level of security. The arithmetic in question grew out of the solutions of algebraic equations of mixed degree. The equations give rise to pretty objects called *elliptic curves* and the new approach was called *elliptic-curve cryptography*. Rather than requiring thousands of bits to implement a secure Diffie-Hellman key negotiation or an Elgamal-type signature,[4] elliptic-curve cryptography requires about twice as many bits as AES to achieve comparable security.

To all appearances, elliptic-curve cryptography has been developed at least as much in the public world as in the secret. Although research and development in the area have been done at many places, one company, Certicom of Mississauga, Ontario, has been completely focused on the field and very vocal in claiming the subject as its own. The company has a large patent portfolio—and alleges a larger portfolio of patent applications—which it has been brandishing at other would-be practitioners. Certicom's claims received a substantial boost in October 2003 when the US National Security Agency paid Certicom $25 million for a very broad license to its technology. The license allowed NSA to sublicense its rights broadly for the development of products to support US national security (Certicom 2006).

Practical application of digital signatures, public-key cryptography's less surprising but perhaps more important facet, requires a new "conventional" cryptographic component, called a *message digest function* or *secure hash function*. Of the two terms, "message digest" gives a better

picture of what is going on, but the name "secure hash" is used for some of the most important standards, and we will use the two phrases interchangeably.

A message digest function begins with an arbitrarily large data object and produces a small (at most a few hundred bits) one that is inextricably tied to the larger, much as the digest of a book or article is tied to the larger work. A message digest is an example of what is called a *one-way* function, a function that is easy to compute forward but difficult to invert, so that the original message cannot be recovered from the digest. One-wayness, however, is not sufficient for a message digest; it also needs to resist all attempts to produce two messages with the same digest.[5] The latter property is difficult to achieve and secure hash functions have had a troubled history. For many years, an algorithm called MD5, designed by RSA inventor Ron Rivest, has been a mainstay of commercial computing. MD5 was designed to have a workfactor only a little greater than that of DES, or 2^{56}. In the early 1990s, NIST acting as the public face of NSA, put forth a standard (FIPS-180) intended to have a workfactor of 2^{80}. Within a few years, NSA had broken its own algorithm[6] and replaced it with a variation called SHA-1. This algorithm stood untroubled until the summer of 2005, when Xianyun Wang and Hongbo Yu (professors at Shandong University in Beijing) and Yiqun Lisa Yin (an independent security consultant) showed that it could be broken with significantly less effort.

The attack by Wang et al. on SHA-1 will probably not become a practical threat for several years, sufficient time to move to new algorithms. Although a movement is underway to intensify the study of hash algorithms with a view to designing some with an entirely different architecture, the government had earlier put forth standards corresponding to the various key sizes of AES.[7] The very size of the new algorithms would seem to mean that the attacks seen so far will have no practical impact on their functioning.

Cryptographic security depends not only on the quality of the individual algorithms used but on careful coordination of the strengths of algorithms used in combination. A set of cryptographic algorithms all selected to support the same workfactor is called a *suite*. In 2005, NSA extended the CNSS15 approach and announced a full suite of public

algorithms that, like AES, were approved for the protection of all levels of classified information. The new construct was called Suite B, apparently by contrast with a previous Suite A, a collection of secret algorithms with colorful names like Juniper and Mayfly. Suite B is made up primarily, though not entirely, of Federal Information Processing Standards. In addition to AES, it contains two secure hash algorithms, SHA-256 and SHA-384, an elliptic-curve version of the Digital Signature Standard, and two key negotiation algorithms elliptic-curve Diffie-Hellman, and MQV, an elliptic-curve algorithm named for Menezes, Qu, and Vanstone from the University of Waterloo in Ontario (NSA 2005). MQV is preferred because it provides authentication at little additional cost.

Suite B is intended to serve a number of objectives. By employing unclassified algorithms, NSA began a convergence between commercial encryption equipment and that used to protect classified information—perhaps eventually making them identical. In so doing, the NSA hopes to draw the major computer and communications manufacturers into a market now dominated by more specialized producers, thereby lowering the cost of acquiring security equipment. These financial objectives of Suite B are bolstered by two interoperability goals.

One is national. A consequence of the increasing interconnectedness of the world and particularly the ever declining distinction between internal and external is the need for greater interoperability among communication systems: those of the military and the intelligence community, those of the police, and those of other "first-responders" like fire departments and ambulance services. Because these systems are intended for responding to terrorist attacks as well as natural disasters, they need to be secured. The two requirements suggest and may actually demand a uniform cryptographic methodology throughout.

The other interoperability requirement is international. Wars are increasingly being fought by ad-hoc coalitions assembled for the occasion. Long standing coalitions like NATO have achieved some degree of cryptographic interoperability among their members.[8] A coalition assembled in weeks can have interoperability only if it plans for it in advance. The central element in this planning is to adopt common cryptosystems, a course of action that makes the traditional secrecy about cryptography meaningless. The only practical solution is to move toward the use of civilian systems and standards by the military wherever possible. Ulti-

mately, this will lead to secure, open, standards accessible throughout the world, another advantage of the Suite B approach.

Export Control

Throughout the late 1990s, the government's rhetoric in regard to cryptography was completely intransigent. Nonetheless, the same period saw a diverse sequence of events coming together to force a change. Every year, almost like clockwork, the government was confronted with a new problem.

In 1996, Daniel Bernstein, a graduate student at the University of California in Berkeley, decided that rather than ignore the export-control regulations as most researchers had, he would assert a free-speech right to publish the code of a new cryptographic algorithm electronically. Bernstein did not apply for an export license, maintaining that export control was a constitutionally impermissible infringement of his First Amendment rights. Instead, he sought injunctive relief from the federal courts. Bernstein won in both the district court[9] and the Appeals Court for the Ninth Circuit.[10] Unfortunately for the free-speech viewpoint the opinion of the appeals court was withdrawn in preparation for an *en banc* review—a review by a larger panel of Ninth Circuit judges—that never took place. With the appearance of new regulations, the government was able to ask the court to declare the case moot, which it did. This indefinitely postponed what the government perceived as the danger that the Supreme Court would strike down export controls on cryptographic source code as an illegal prior restraint of speech.

In 1998, a US intercept network called ECHELON became embarrassingly public (Campbell 1999). The ECHELON system is a product of the UK-USA agreement, an intelligence association among the US, the UK, Canada, Australia, and New Zealand. published earlier (Hager 1996), a 1999 report prepared for the European Parliament (EP) stated that ECHELON was targeting major commercial communication channels, particularly satellite systems. This caught Europe's attention. The implication was that the system's purpose was commercial espionage, a view confirmed, at least in part, by former CIA Director James Woolsey's article "Why We Spy on Our Allies" (2000).[11]

The natural European reaction to this evidence that they were being

spied on was an interest in improving the security of European communications but strict regulations on cryptography were an impediment to any such program. The European governments responded by relaxing their rules on the use, manufacture, sale, and export of cryptography, thereby putting the US under pressure to relax its own export rules.

Export control regulations also created direct problems for the government in its role as a major software customer. The military was trying to stretch its budget by using more *commercial off-the-shelf* hardware and software. Economic realities meant that restrictions on cryptography often forced companies to omit security features altogether rather than supporting distinct domestic and foreign versions of the same product. As long as export regulations discouraged the computer industry from producing products that met the government's security needs, the government would have to continue to buy more expensive *government off-the-shelf* equipment for its own use. This was becoming uneconomical to the point of infeasibility. The only way to induce the manufacturers to include sufficiently-strong encryption in domestic products was to allow them to put it in the products they exported as well.

From 1997 to 1999, the US government attempted to bribe the computer industry by allowing the export of products containing 56-bit DES by companies that had made plans to implement key escrow in their products and were making satisfactory progress on their plans. Because development by the companies involved was a commercial secret held in confidence between the companies and the government, the success or lack of success of this program cannot be determined but in the fall of 1999, as noted in the previous chapter, it was abruptly abandoned.

Earlier in 1999, a bill called SAFE (Security And Freedom through Encryption), which would have forced the administration to change the export regulations, passed the five committees with jurisdiction and was headed to the floor of the House of Representatives, when the administration announced that the regulations would be revised to similar effect. By capitulating, the White House avoided the loss of control that would have resulted from a change in the law.

On September 16, 1999, US Vice President and presidential candidate Albert Gore Jr.[12] announced that the government was changing its policies. Beginning with regulations announced for December—and actually

promulgated on January 14, 2000—keylength would no longer be a major factor in determining the exportability of cryptographic products.

The new regulations shifted away from the strength of cryptography as the principal determiner of its exportability. In its place, the new regulations considered two issues: whether the hardware or software was for a commercial or government customer and whether the product was off-the-shelf (the term used in the regulations was "retail") or whether it was adapted to the needs of individual customers. The object was to provide the cryptography needed by industry and commerce while making it difficult for military organizations with special requirements and installed bases of cryptographic equipment to make use of the newly available products. For most purposes, the export control portion of the "crypto wars" had been won by industry.

The key-escrow battle ended more quietly, with most (perhaps all) of the government's programs being canceled. In June of 1998, the Skipjack algorithm that underlay Clipper was declassified so that the military could use it in software for secure email. Despite being an elegant algorithm and strong enough for many purposes, Skipjack was tainted with its key escrow past and has not been widely employed.

The notion of key escrow has not died, however. It has been built into products for commercial use where employers are in a better position to require their employees to submit to potential spying than the government was to impose the same requirement on the citizenry in general. Key escrow, better called key recovery for this purpose, is also essential when cryptography is used to protect stored information rather than communications.[13]

Cryptography after Deregulation

Contrary to the expectation of many of its fans, the deregulation of cryptography did not produce any immediate explosion in either the number of available cryptographic products or the frequency of their use. Anyone who expected most email and phone calls to be encrypted overnight was surely disappointed.

The reasons for slow growth in the cryptographic business, however, seem fairly clear. Foremost, cryptography, at least communications cryp-

tography, is a phenomenon of the interaction among people. To engage in encrypted communication you must make an investment in hardware or software. The value of your investment is proportional to the number of other people similarly equipped and so the market has natural exponential growth.

The benefits of exponential growth are well known in successful fields and the term is applied willy-nilly to anything that is growing quickly. The downside (more precisely the slow upside) of exponential growth is exhibited by public-key infrastructure. Keys, certificates, directory servers, and revocation lists are of very limited use until most people have PKI-enabled products. Consequently, their up-front costs are not offset by a robust revenue stream and must be supported by an investment that is slow to produce returns. This burden has been borne by the US military who have built themselves an electronic key-management system at a cost of tens of millions of dollars but a similar commitment is difficult for the commercial sector.[14]

Other phenomena act to advance or retard the basic exponential progress. Most conspicuous of technical problems is the lack of uniform standards. Several non-interoperable suites of cryptographic algorithms and formats for keys, certificates, and cryptograms have significant shares of the commercial market. Exponential growth is thereby fragmented and must occur independently within each sector.

Regulation, which in the nineties acted primarily to decrease the use of cryptography, may come to play a supportive role as the value of cryptography for protecting private data against the compromises that are so frequently in the news becomes more widely recognized. Similarly, the popularity of laptops and the ease of laptop theft has created a market for cryptographic protection of their file systems. Many companies have imposed requirements—similar in effect to regulations—that the laptops carried by employees be so protected. In June 2006 the US Office of Management and Budget put into place a recommendation that all sensitive data on laptops be encrypted unless exempted by the department's Deputy Secretary (or his designee) (Johnson 2006).[15] Cryptography may get a big boost from regulations giving *safe haven* in the event of compromises of personal data when the data are properly encrypted.

What promotes the use of cryptography more than anything else is

its default inclusion in products and its automatic operation without the need for user action. The Secure Socket Layer protocol that comes built-in to all browsers may be the most widely deployed cryptosecurity system of all time. A plausible competitor for this title is the A5 algorithm used in GSM telephony. Another is the use of cryptography in smart cards. A comer in this category is the automatic encryption of phone calls by Skype, a popular Voice over IP system. All of these are commercial products, which have far outrun their military ancestors in deployment.[16]

DRM, DMCA, and TCG

If cryptography has exhibited a growth area, it is one rather different from what cryptography's early pundits anticipated: protecting the interests of purveyors of intellectual property.

The incredible advantage of information in digital form—it can be readily and inexpensively moved and copied—is a disadvantage from the conventional marketing viewpoint, which, roughly speaking, knows how to charge for what is scarce. Digital products, unlike antibiotics for example,[17] can readily be copied. If you have one copy of a program or a digital copy of a picture, you can readily and inexpensively have another but one antibiotic tablet is of little help in producing more.

Attempts to prevent ready copying of digital products have come to be called *Digital Rights Management*. In essence, DRM is a regime in which the goods are kept in encrypted form everywhere except in controlled environments in which the digital products can be viewed, listened to, or otherwise used. An approach of this sort is proactive and imposes a prior restraint on would-be users of digital products. Another approach, more in line with conventional enforcement of copyright is to label each copy of a product in a way that cannot readily be altered, an approach called *watermarking*. When an unauthorized copy of a watermarked work is discovered, it is possible to examine the label and trace the copy back to its source. The tamper resistant, and often covert, labeling of digital objects uses *steganography*, a cryptographic technology that hides messages rather than merely making them unreadable.

One widespread cryptography-based copy protection system is the Content Scrambling System (CSS) used to encrypt the contents of DVDs.

Although intended to prevent the copying of DVDs, particularly onto other media like hard drives, CSS had the side effect of preventing the implementation of DVD players on computers running open-source operating systems, particularly Linux. CSS is administered by the DVD Content Control Association,[18] which is unwilling to license its technology for use in to open-source software. In October 1999, however, *deCSS*, an independent implementation of CSS developed in Norway by Jon Lech Johansen and unknown associates, became available over the Internet. As it turned out, cracking copy-protection systems is not illegal in Norway, and attempts to prosecute Johansen failed.

When NSA director Bobby Ray Inman tried to acquire the legal power to control cryptographic publication in the early 1980s, the idea was widely condemned. The American Council on Education panel that was created in hopes that it would recommend the idea came nowhere close. Ironically, a legal system of censorship of cryptographic research has since grown up to serve commercial ends with no comparable condemnation.

The Digital Millennium Copyright Act shepherded through Congress by the entertainment industry gives legal protection to technical systems used in protecting copyrighted material. Its functioning is comparable to laws against breaking and entering: if you lock your door, even with a very poor lock, you acquire a measure of legal protection that is lacking if you leave your door unlocked. Anyone who breaks your lock and enters your home is guilty of breaking and entering, a crime more serious than mere trespass. In a similar way, the DMCA created both a tort and a crime of defeating copyright protection measures. The law made it a crime to defeat such measures, even when the objective was to use the material in a manner permitted under copyright law. The objective was to criminalize attempts to defeat copyright protection mechanisms independent of any issue of the protection of particular copyrights. An exception was created for research but it was painfully narrow, requiring the researchers to give notice in advance to the owners of the system under study.

This time the crypto community was remarkably docile. Two cases set the tone.

In 2000 the Secure Digital Music Initiative (SDMI), an industry group

consisting of about 150 companies and organizations, put together a challenge to researchers to break the audio watermark they had developed. The contest rules were stringent: three weeks to remove the watermark without badly degrading audio quality (the latter was not an announced contest rule). Felten and his colleagues decided to participate without actually officially joining the contest, thus preventing them from competing for the prize but also allowing them to avoid signing the required confidentiality agreement.

Instead of competing for the cash award, Felten, his students, and fellow researchers at Rice University wrote a technical paper showing how to defeat the SDMI technology. Felten et al. intended to present the research at the Fourth Annual Information Hiding Workshop, held in Pittsburgh on April 25–27, 2001. Through a particularly foolish action on the part of the SDMI, the Recording Industry Association of America (RIAA), and Verance Corporation, Felten and his colleagues were threatened with legal action if they presented their work. The argument was that the Princeton and Rice University computer scientists had violated the anti-circumvention aspects of the DMCA. No matter that DMCA has an escape clause for research—§1201 (g)(2) (B), which permits circumvention if the "act is necessary to conduct such encryption research"—or that the ensuing publicity was likely to cost SDMI and RIAA far more than permitting Felten and his colleagues to go public with their research.

Princeton University declined to provide counsel to defend the scientists and so the researchers withdrew their paper from the meeting. Instead the Electronic Frontier Foundation (EFF), a civil-liberties group focused on citizens' rights in the digital world, stepped in. The researchers filed suit in federal court, seeking a "declaratory judgement" that publication of the research paper would fall within the plaintiffs' First Amendment rights (*Felten v. RIAA*, US DC NJ Case #CV-01-2669). The recording industry backed down, the scientists presented their paper at a different—and more widely attended—venue, the USENIX Security Symposium (Craver et al. 2001), and the case was dismissed for lack of standing.

With Niels Ferguson, a case of interest to cryptographers, the situation worked out differently. Ferguson, an established cryptography researcher and consultant, claimed to have broken the High Bandwidth Digital Content Protection (HDCP) system, an Intel cryptographic system for

encrypting digital video communications between cameras and players (HDTV, etc.). Licensing for the system was available through Digital Content Protection LLC, a subsidiary of Intel.

Ferguson, a Dutch citizen living in Holland, believed that were he to publish his results, he would be subject to arrest for violation of the DMCA, anytime he was in the United States.[19] Ferguson did not make his work public; instead he submitted a letter to the chair of the 2001 ACM Workshop on Security and Privacy in Digital Rights Management describing the chilling effect of the DMCA. He also submitted an affidavit in the *Felten v. RIAA* case.

Other researchers studying HDCP did not react to the chill in quite the same way. In particular, a group of researchers from the University of California at Berkeley, Carnegie Mellon University, and the Canadian company Zero-Knowledge Systems also found an attack on the HDCP system (Crosby et al. 2001). Wanting to publish but realizing the possible conflict with DMCA, the researchers proceeded with caution. One of them, Berkeley professor David Wagner, consulted with University of California lawyers, who made clear they would be defended in any civil suit.[20] Next, in accordance with the requirement of the DMCA, the researchers "made a good faith effort to obtain authorization before the circumvention"[21] and met with engineers from Digital Content Protection LLC, who appeared to appreciate the advance notice they received about problems with their technology.[22] Then the security researchers published their work at the same ACM workshop that Ferguson had avoided.

Trusted Computing Technology

In the late 1990s five major computer companies, AMD, HP, IBM, Intel, and Microsoft, formed a consortium called the Trusted Computing Platform Alliance to develop standards for a broad new approach to copy protection and a variety of other computer security problems. The idea was to add dedicated security hardware called *Trusted Platform Modules* (TPMs), hardware capable of monitoring and controlling all activity, to computers. The work of the TCPA was widely perceived as an attempt to reduce personal computers to the status of such consumer-electronic devices as television sets, devices on which the owner's control over what

programs could be run would become comparable to the viewer's control over what TV programs were available to watch.

Partly in response to the criticism, the TCPA later reorganized and reincorporated itself as the Trusted Computing Group. It also expanded its core membership (the Promoters) from five to seven, adding Sun and Sony, and making provision for adding more as time went on.

The basic objective in adding the sort of security hardware for which TCG is developing standards is to achieve tighter control over the software running on computers. This technology has many possible applications. It can, for example, make it far more difficult for viruses and worms to infect a machine. It can enable system administrators to ensure that only approved programs or only licensed copies of programs or only the latest versions of programs can be run on a system.

Conceptually, trusted computing technology begins with control of what operating system can be run. The process is called *secure boot*: the microcode built into the computer checks a signature on the operating system it is loading and will let the system run only if it bears the correct signature. The most influential developers of secure boot technology were William Arbaugh, Dave Farber, and Jonathan Smith at the University of Pennsylvania.

This approach may be very secure, but it is also very inflexible. For many purposes, it is desirable to allow a computer to run a variety of operating systems. In these cases, it may still be valuable for one system to be able to determine with certainty what set of programs another system is running. This technique is called *attestation*.[23] A computer may be capable of running any of the popular operating systems—Linux, Solaris, Windows—and any applications that those operating systems support. In interaction with other computers, however, it may be asked to attest to its configuration, i.e., to get a signed message from the tamper-resistant TPM describing the configuration of its hardware and software. This allows the computer that has demanded the attestation to decide whether it will allow interaction to proceed.

In some networks, for example those running the electrical power grid, TCG technology is entirely appropriate. Computers are not connected to that network by rights but to serve the interests of the power companies and their customers by managing the country's electric power. A similar

argument can be made for enterprise networks in which all the computers are owned by the enterprise.[24] On the other hand, individual computer owners take the reasonable attitude that they should be able to run whatever programs they wish and fear that if trusted-computing technology becomes widespread, this freedom will be denied them. Other critics, with a more entrepreneurial view fear that trusted-computing technology will stifle innovation by allowing ISPs to discriminate against programs —browsers, for example—that were not provided by their preferred commercial partners. This can be done entirely without malicious intent, for example, by an ISP that is trying to limit the burden of maintaining compatibility with an ever-growing number of versions of a popular program.[25]

The Bigger Picture

The 1990s will be remembered long after the roaring nineties a century earlier, the roaring twenties, or the 1960s have been forgotten. Not only were the 1990s a boom period dotted with great feats and great fortunes, the 1990s changed the foundations of society in a way that may not be appreciated for decades.

The technologists will remember the era for the World Wide Web. Invented in 1989 at CERN (the European laboratory for particle physics), the Web began to take hold about 1993 and was flourishing by three or four years later. The Web made delivery of information over the Internet easy and natural rather than tedious and geeky. Within a decade of its birth, the Web was being used by businesses and governments as the primary way that they should be getting information out to and acquiring information from their customers. Great fortunes were amassed by those who got on the Web bandwagon early and such names as Amazon, eBay, Yahoo, and Google became household words. 'Google' even became a verb.

Deeper, less flashy developments also drove the move toward a society based on digital communications. For its whole previous history, the cost of telephony had been driven by the cost of long distance transmission. In the 1990s, optical fiber technology came of age. By placing fiber through conduits previously occupied by copper, communications

companies multiplied their bandwidth by a factor of 1000 and sometimes made money on the deal by selling the copper they had replaced. The result was the development of a vast overcapacity that still exerts a profound effect on the communications business. Fiber is the bedrock on which the current high-speed Internet and plans for future higher-speed internets are built.

It is hard to think of a dot-com that better captured the positive energy of the 1990s Internet more readily than Google, with its philosophy "Do no evil." But from a security perspective there is a dark side to the enabling technology of search engines, which give the ability to discover targets with known security vulnerabilities. Such information was, of course, public before the Web, but like county court records, it was largely inaccessible, leaving these systems relatively safe. However, now through the use of search engines, it has become a rather trivial job to discover and access public systems with known vulnerabilities (Landau 2006, p. 433).

The decade was also one of growing internationalization. Falling communication costs, falling shipping and travel costs, and falling barriers to both trade and travel increased the tendency of businesses to expand worldwide. Internationalization merged naturally with the growing business trend toward *outsourcing*, using contractors rather than employees for many tasks, ranging from sweeping the floor to programming, accounting, and public relations. Once international communication became adequate to support close business relationships across intercontinental distances, it became apparent that capable well-educated workforces from countries with low labor costs were readily available. This dramatically increased international business communication and consequently the need for its security.

Internationalization and outsourcing have been eagerly embraced by US businesses. Less engaging from a US perspective is a decline in the preeminent position the US has held in world commerce since the end of World War II. Although the United States is the world's third-most-populous nation, it has only about 5 percent of the world's population. As the gross national products and living standards of many parts of the world increase, US influence over high-tech policy issues will decline.

It is worth noting that whereas in the late 1990s 80% of Web content

was in English (Wallraff 2000, p. 61), by 2002 the percentage had declined to less than 50% (Crystal 2004, p. 87). Indeed, by 1998, over half of the newly created websites that year were *not* in English (ibid.).

Conspicuous on the international scene is the rise of China as a major cultural and economic power and a major creditor of the United States. Indeed, among Internet users the second-most-common native language is Chinese (English is first).[26] China has begun developing standards for digital products in many areas. These standards, some of which are cryptographic, are in potential competition with European and American standards.

For a long time, computer communication involving humans consisted primarily of single-fixed-width-font text. In non-text interaction between computers and humans, the images were usually generated locally. This was natural enough. Network bandwidths were low. Fifty-kilobit backbones and modems running at a few thousand bits per second were considered fast.

The Web brought a new paradigm: HTML, the hypertext markup language. HTML is a crude typesetting language, which, however, provides hyperlinking—the possibility of including one document, by reference, in another. HTML is a small and in many ways primitive subset of an ambitious standard called SGML (Standard Generalized Markup Language) a construct so rich that people are forever creating subsets. A subset that has taken on great significance is XML, the Extensible Markup Language.[27]

XML has proved to be the most popular approach, not for human-to-computer communications but for computer-to-computer communications, communications that are not simply about moving "customer data" but communications that involve negotiation between computers about services, inventories, communications, and security.

The explosions in both the raw power (bandwidth) of communications and in the computational capability to manage the communicated material have created a world in which every sort of information, from names and addresses, to historical documents, to satellite photographs, is more readily available. On one hand, the ease with which real estate speculators or would-be home buyers can appraise property remotely makes some homeowners indignant. On the other, details of some government

buildings have been made indecipherable in the aerial photographs most easily found on the Web.

In a world where more and more material is being brought under tighter and tighter control, covered by non-disclosure agreements, and protected by digital rights management systems, there is one major move toward openness: the open-source software movement.

By lowering the cost of both creation and dissemination, modern computer technology has enabled the user to be a creator in unprecedented ways. During Hurricane Katrina, many users were able to put together information from publicly available sources to enable people to discover whether their homes had been damaged. This is one of many examples of technology enabling individuals to provide services for themselves that could once have been provided only by governments or large corporations.

If a communications network is to be flexible and encourage innovation, then it must perforce allow applications unanticipated at the time of the design of the network. The Internet does this is through the *end-to-end* principle, which is the idea that the communication endpoints —the applications—should implement the functions, rather than letting low-level function implementation be part of the underlying communication system. This principle been fundamental to Internet design since the beginning.[28] The Internet concentrates investment (and particularly "smarts") at the edges. The center is a computationally powerful but fundamentally dumb collection of routers and transmission channels. (In fact, the routers and channels are not so much dumb, as neutral to the application they are transmitting.)

This is fine in theory; in practice, as the Internet has evolved, it has departed somewhat from these principles. In particular, it has evolved into a fragmented network in which many nodes are walled off into private isolated networks that cannot communicate with each other.

The key elements of this "walling off" are *network address translators* (*NAT*s) and Firewalls. The current Internet Protocol (version 4, or IPv4) has addresses that are only 32 bits long. There are about 4 billion possible addresses. This is a large space if addresses are chosen at random but a much smaller space if large sets are reserved to be in contiguous chunks. The fact is that address space is in too short a supply to be used

as originally intended: every host computer has a unique address and therefore can be addressed by any other computer on the network.

Beginning in the early 1990s, organizations that could not get enough unused addresses began to allocate their internal addresses independent of the outside world.[29] Such common functions of the Internet as email do not actually make direct use of internet addresses. If you send email to president@whitehouse.gov, the fact that the IP address of the White House is 63.161.169.137 does not even come to your attention, and the individual address of the workstation on the President's desk is something you may not even be able to discover. The email servers will go from the email address president@whitehouse.gov to the IP address 63.161.169.137 and deliver the email there. At that point the "president" component of the address becomes important, and the White House mail server will take over and discover the correct internal address to which to forward the message. It makes no difference whether the internal address is unique; it is only accessible to computers inside the White House. This is network address translation. It is typical of institutional connections to the Internet.

Closely associated with NATs are *firewalls*, computers that manage and filter traffic between an internal network and the Internet. Firewalls serve several purposes but one of the main ones is explicitly to prevent unfettered access to the Internet. Rather than supporting (permitting) any form of access to the Internet, firewalls frequently allow only a limited range of services. The White House firewall described above might support no service other than email.

An attempt to expand the address space of the Internet has been underway for several years. Internet Protocol version 6, IPv6, has 128-bit addresses (256 billion billion billion billion possibilities) and can provide enough address space for the end-to-end connectivity originally envisioned. As noted earlier, however, the protocol that handles addressing is the protocol that defines the network and must be shared by all network users. Even when one protocol has been designed as an extension of another, transition is difficult.

Had IP originally been designed with larger addresses, firewalls might have been designed differently and NATs might never have established themselves. In the existing Internet, however, many organizations are

delighted to have control of their user's communications, and a return to unfettered end-to-end communication seems unlikely.

Carnivore

On July 11, 2000, the *Wall Street Journal* disclosed that the FBI had been "wiretapping" the Internet (King 2000). The FBI had developed a program with the ill-chosen name of *Carnivore* to scan and record network traffic.[30] Despite a public inured to cookies that let web sites track user visits, reports of Carnivore hit a public nerve. What exactly was the FBI recording? Was the Internet versions of pen registers and trap and trace devices recording more information than they would for traditional land-line telephone systems? Was the mail of nontargeted individuals inadvertently being read? Why was an FBI device attached to an ISP, instead of, as had always been done by telephone taps, the Internet Service Provider (ISP) doing the sorting of relevant traffic for the bureau?

The truth turned out to be both more complex, and, on the surface, less threatening, than initial newsreports indicated. Carnivore was a *packet sniffer*, a program that analyzes network traffic; such a program can be configured to search for traffic to or from a particular user. ISPs serve as the local post offices of the internet world, sorting and delivering email and other services to their users. When Carnivore was placed at an ISP, it received all packets that traversed the Ethernet connection on which it was placed (Smith 2000, p. 13). Using filters, Carnivore recorded the traffic that fit pre-determined patterns, generally traffic to or from a particular user. Carnivore could be configured for full wiretap or pen register mode; in the latter, the data content was "X-ed" out (ibid., p. 56).[31] The FBI presented Carnivore (later renamed DCS, or Digital Collection Service 1000) as a natural application of wiretap law to Internet technology. Civil-liberties groups and computer experts disagreed. Because the Internet packet-routed architecture is fundamentally different from the circuit-switched telephone networks, the application of wiretap laws to the Internet is less straightforward than it would appear. Unlike the telephone system, transactional information (number called, number calling) is hard to separate from the content. Consider a targeted user reading a web page. The web page will be received from a website with an IP

address. Carnivore will log the communication between the targeted user and the website. The web page, however, may contain hyperlinks to other websites. These will be resolved into IP addresses and Carnivore will log these as well. Even when the authorization is for a pen register and does not include recording the content of the web page itself, the content will be largely reconstructible from the pattern of IP communications.[32]

Carnivore also contains a mode for analyzing SMTP communications. (SMTP, the Simple Mail Transfer Protocol, is the standard for Internet email transmissions.) In this mode, email header information will be recorded. Email headers contain much more information than phone numbers. Some of this is analogous to information on physical envelopes and can be justified on the grounds that the same information would be obtained in a mail cover. Some, however, is not. If the entire mail header is captured, lists of addressees and copyees that would not be on an envelope will be included. A variety of version and status information about the mail clients and their configuration is also commonplace.

The other violation was not privacy, but engineering. CALEA had already created an odd situation in which the FBI was placed in the role of designing standards for the telephone network. Carnivore went a major step further, placing *the government's own search devices* directly onto the ISP's networks. In at least one case, this was done *over the ISP's objections*.[33]

Attorney General Janet Reno authorized an outside academic review of Carnivore. Several universities with strong research groups in computer security, including Dartmouth College, the Massachusetts Institute of Technology, Purdue University, and the University of California at San Diego, expressed interest in performing such a study but the restrictions the Department of Justice placed on the review—which included confining the study to narrow technical questions, pre-publication review of the study by DoJ, and DoJ approval of the final report—were such that these universities bowed out. Instead, the Illinois Institute of Technology (IIT), an institution not previously known for expertise in computer security, performed the study.

Like the Clipper chip before it, Carnivore suffered from design flaws. In the Clipper case, the design worked as advertised: it securely encrypted data under an 80-bit key. The flaw was that it was possible to spoof

Clipper so that one could use the Clipper chip to encrypt, but do so in such a way as to prevent law-enforcement access (Blaze 1994). In Carnivore's case, the potential problems affected the underlying security of the system. The IIT report found that the filter settings, which determined what traffic could be captured, could be changed by anyone with access to the system password, which was compiled into the Carnivore source code, and thus plausibly to anyone with access to the Carnivore installation, a poor security design.[34] IIT noted that the program had no audit trail, "Incorrectly configured, Carnivore can record any traffic it monitors"(Smith 2000, p. 17)—*any traffic* should not have been an available option—and, "except for FBI procedures and professionalism, there are no assurances against additional copies being made of an inadequately minimized intercept" (ibid., p. 60).[35] Such design runs contrary to the intent of wiretap law, which is very specific about data minimization, requiring that only content appropriate to the warrant be collected.[36]

When various members of Congress expressed concern about the limited nature of the IIT review, legislation restricting Carnivore use seemed possible. The Justice Department promised an internal review of the policy issues regarding Carnivore. But before it was completed, the events of 9/11 intervened, and the USA PATRIOT Act, section 216 of which legalized the application of Carnivore-like systems to packet-routed networks, was passed.

Meanwhile, use of Carnivore is down in favor of commercial systems to do the job (FBI 2003a,b).

Identity and Anonymity in the New World

To casual observation and in casual use, the World Wide Web appears to provide a sort of anonymity similar to that available to a shopper in a big city. You browse through sites and look at their contents. You are not asked for identifying information and you are not aware of providing any. Depending on your communications arrangements, however, you will be providing various sorts of information from which other information about you can be determined.

If you communicate from a fixed IP address, the web pages you visit will be able to recognize your visits as those of a single entity. As a prac-

tical matter, they or may not be able to convert your IP address into any of the usual customer data such as your name. If, as is more common, you communicate through a local ISP and get a different IP address from session to session, it will be far more difficult for the visited websites to track you but easy for the ISP. The latter may or may not be sharing your information with other commercial entities.

There are common and legitimate purposes for which casual anonymity is not sufficient. Investigators of many kinds from academics to reporters to police try to hide the patterns of their inquiries, even when they are consulting open sources. Many individuals value their privacy. For example, they wish to learn how to handle their illnesses without revealing their conditions to medical marketers or colleagues. Companies may be unable to do competitive analysis unless they can conceal their identities as they visit competition's web sites.

As AOL's indiscretion in releasing query logs for thousands of its customers revealed (Barbaro and Zeller 2006), a knowledge of the information a person seeks imparts a vast amount about the person's activities and plans and it is also likely to disclose the person's identity, even if that identity is not directly contained in the query record.

In response to the danger of Internet users being tracked and identified, various commercial entities have arisen to provide services that shield browsers' identities more effectively from the websites they visit. Such services, which include the late Zero-Knowledge Systems and Anonymizer.com of San Diego, find much of their customer base in commercial entities that want to survey their competitors websites in confidence that they are seeing the material that would be shown to typical browsers and not a show that has been put on especially for them.

In the 1980s David Chaum proposed several anonymity systems that wrap communications between the sender and receiver in layers of public-key cryptography. Central to these systems is the appropriately named *mix*. Each time the communication reaches a new mix on its journey from sender to receiver, the mix unwraps one layer of the cryptography, mixes up the order of the messages, and sends them all on. One descendent of Chaum's research is onion routing, the current version of which is called Tor (The onion routing) (Onion Routing 2006). Tor modifies mixnets by using anonymizing proxies (proxy servers sit between a client and a

server fulfilling requests for the client if it is able—otherwise it ignores the request) (Reed et al. 1996).

When an application (perhaps a Web browser, perhaps an IM client), connects through the Tor network, a client proxy chooses a route for the traffic using server nodes called *onion routers*. The Tor proxy establishes a circuit using public keys belonging to onion routers in the selected path. Application data is passed using keys determined by the proxy and each onion router as the circuit was established. While each of the onion routers will know its predecessor and successor on the path, only the client proxy is aware of all nodes on the path which the communication traverses.

The Tor network is an overlay on the public Internet.[37] Widely used applications include anonymous web browsing (http and https), instant messaging (IM), and Internet Relay Chat (IRC). Tor has been running since October 2003. As of August 2006, the Tor network had about 750 Tor servers and the number was doubling approximately every 6–8 months. How many users does Tor have? Tor deliberately does not keep track.[38]

Initial work on onion routing began at the Naval Research Laboratory in 1995. Tor was begun in 2002 and was jointly designed by scientists at NRL and at the Free Haven project, working under contract to NRL. Most of the funding for onion routing, including Tor, has come from the Office of Naval Research and the Defense Advanced Research Projects Agency.[39] Given current law-enforcement concerns about tracking network users,[40] it might come as a surprise that the original funding sources for a public anonymizing network came from the Department of Defense. It should not. A group in the Navy, for example, who, during the last half decade, have been periodically stationed in the mid East have found Tor an excellent way to disguise their communication patterns. No one watching their ISP connection locally in country can learn their affiliation through tracking with which agency in the United States the Naval personnel contact and no one watching their communications in the United States can learn to which country—and *which house*—their communications are going.[41]

Naturally, this naval unit which had taken such pains to conceal itself using anonymizing technology wanted neither to be named nor to have

its location identified. This illustrates a difficulty of explaining the value of anonymizing technologies: when an anonymity system is successfully used for a "good" task, its usage does not generally become public. Publicity only attends those cases when there is a problem with the anonymizing technology or where the anonymizing technology is used for some nefarious purpose. That the funding support for Tor came from a variety of agencies of the US Department of Defense makes it clear that the technology is beneficial in a wide variety of government situations.

Why might the Department of Defense fund an anonymizing system for the general public rather than build one just for military use? As the developers of Tor point out, "anonymity loves company" (Dingledine 2006). The more users an anonymity system has, the easier it is to hide the traffic. Thus a widely used anonymity system provides Department of Defense users the best protection from prying eyes.

At the same time that work was occurring on anonymizing technologies, there was also great effort underway to develop electronic identification systems of various types.

With the opening of the Internet to commercial traffic in the early 1990s, electronic commerce—ordering a book from Amazon, selling collectibles on eBay, making travel arrangements with Travelocity or borrowing money through eLoan—caught on faster with consumers than anyone but a few (now wealthy) backers expected. There was, however, a serious irritant: the constant need to reenter your name and password, not to mention credit-card information, billing and shipping addresses, and phone numbers each time you made a transaction over the Internet. Out of this difficulty was born the notion of single sign-on, a way for the user to sign on and authenticate herself to a single site and have that authentication carry over to multiple sites.

Microsoft developed the Passport system to which Hotmail users had immediate access (thus giving Passport a large installed base). Microsoft's approach centralized customer data and there were immediate objections from civil-liberties groups over the threat to privacy this entailed. Meanwhile, in 2001, in concert with American Airlines, Bank of America, Cisco, Nokia, Sony, and a number of other companies, Sun Microsystems proposed a federated system called the Liberty Alliance,[42] in which identity information (name and authentication) could reside with an "Identity

Provider," while site-specific information—the dates of car rental, the type of car desired—would be with the "Service Provider." The goal of the Liberty Alliance was a set of interoperable specifications for federated network identity allowing users to authenticate once—single sign-on—and link elements of their identities without centrally storing their data (Liberty Alliance). Liberty protocols provide pseudoanonymity in the communications between the Identity Provider and Service Provider, a featured designed to prevent data aggregation. The Liberty protocols have attracted wide interest, and various governments as well as large commercial enterprises are participating in the process.[43]

Anonymity and identity are among the many threads in human culture that have existed in uneasy harmony for millennia. The revolutionary changes of the 1990s—globalization, mobility, greater availability of information—brought many of these threads into open conflict and a new balance among them has yet to be found.

At a moment in human history, however, when reflection and tolerance might have served us best, the events pushed everyone in a direction that, by maximizing security, minimized privacy and individual liberty.