Topics in Survivable Systems

John C. Knight
A. C. Chapin
Steven Geist
Luis G. Nakano
Matthew C. Elder
Brownell K. Combs
Sean McCulloch
Robert S. Sielken

Computer Science Report No. CS-98-22
August 14, 1998

Topics in Survivable Systems

JOHN C. KNIGHT
Department of Computer Science
University of Virginia
knight@cs.virginia.edu
(804) 982-2216

MATTHEW C. ELDER
Department of Computer Science
University of Virginia
elder@cs.virginia.edu

A. C. CHAPIN
Department of Computer Science
University of Virginia
acc2a@cs.virginia.edu

BROWNELL K. COMBS
Department of Computer Science
University of Virginia
bkc6j@cs.virginia.edu

STEVEN GEIST
Department of Computer Science
University of Virginia
smg9c@cs.virginia.edu

SEAN MCCULLOCH
Department of Computer Science
University of Virginia
stm4e@cs.virginia.edu

LUIS G. NAKANO
Department of Computer Science
University of Virginia
lgn4d@cs.virginia.edu

ROBERT S. SIELKEN
Department of Computer Science
University of Virginia
rss4k@cs.virginia.edu

Table of Contents

Introduction

John C. Knight

In the Spring of 1998, a special topics course in Information Survivability was taught at the University of Virginia. The attendees were graduate students in computer science and the instructor for the course was John C. Knight. This report is a compendium of papers written by students who attended the course.

Information systems are at the heart of many important applications. Banks and other financial institutions operate by transferring messages about financial transactions electronically and storing financial information in large databases. The public switched telephone network is implemented by computers and much of the essential processing in the telephone system is dependent on large databases. In a similar way, many transportation, energy, government, and military systems are dependent on information systems.

An important characteristic of these information systems is that they are distributed, in most cases widely so. For example, the world's banking system is, in fact, a single very large distributed system. Central banks are connected to regional banks, regional banks are connected to local banks, regional banks are connected to other financial institutions such as mortgage companies, central banks from different countries are connected together, and so on. All kinds of financial services are implemented by moving information between computers, storing different types of information on different computers, and processing a wide variety of transactions on different computers. For example, depositing a check or using a credit card for a retail purchase each involve several computers and many network transmissions.

The dependence of these applications on information systems is considerable and not well appreciated. In fact, in many cases the normal activities of society depend upon the continued operation of the information systems. The loss of some of the critical information systems in the telephone network has lead to widespread and protracted failures. The loss of such systems in other applications could have a devastating effect.

Complicating an already complicated situation is the interdependence of some of these applications. For example, although limited protection against loss of power is afforded for some information systems, service following a power loss is usually severely reduced. Thus, for example, management of transportation systems will be affected significantly if there is a widespread loss of power. Similarly, loss of communication service will disrupt many other information systems such as finance, electronic commerce, and transportation.

The dependability requirements that arise with many critical information systems are quite extraordinary. For example, many current systems and others that are being planned are required to operate on widespread networks and require twenty-four-hour-per-day, seven-day-per-week operation. In addition, these systems have to support combinations of dependability requirements. For example, they have to maintain very high levels of availability whilst also ensuring network-wide security.

Dealing with the effects of faults in information systems leads to the notion of survivability. Informally, by survivability we mean the ability of the system to continue to provide service (possibly degraded) when various changes occur in the operating environment. For example, when events such as hardware failure, software failure, operator error, or malicious attack occur, a critical subset of normal functionality or some alternative functionality might be needed to mitigate the consequences of the event.

The topics covered in this report are not comprehensive by any means; the survivability area is too broad for that. The papers do, however, span quite a wide technical range. The first paper, by Robert S. Sielken, is entitled "The Public Switched Telephone Network". This is a short summary of many aspects of the telephone system and provides some background information about how one of the critical infrastructure systems works. The second paper, by Sean McCulloch, is entitled "An Analysis of Non-Security Failures of the Electric, Phone, and Air Traffic Systems". This is an examination of some of the incidents that have occurred in these infrastructure application domains. The limitation to non-security incidents is important because it helps to point out the many sources of failure to which these systems are subject.

The general concern of the community about security is significant, and four of the papers cover security-related topics. The first, by Matthew C. Elder, is entitled "Major Security Attacks on

Critical Infrastructure Systems". This paper discusses recent security attacks on critical information systems, and presents extensive information about the severity of the security problem. The second security paper, by Brownell K. Combs, is entitled "Hacking Information Available on the Internet". This paper examines the material available to hackers on the Internet. The author sought out sources of information so as to determine what is available and how easy it is to obtain. The results are surprising. The third security paper, by Steven Geist, is entitled "Firewalls". This paper presents a summary report on firewalls, an important technology designed to help improve the security of existing systems. The fourth paper, by Luis Nakano, is entitled "State of the Art in Computer Virus Prevention". Viruses are growing in number at an extraordinary rate and many new techniques are being developed by their authors to combat existing detection systems. In this paper, the different types of virus are summarized and approaches to their elimination discussed.

The final paper in this report is by A. C. Chapin and is entitled "Smart Cards: Security in the New Transaction Cards". The field of critical information systems is changing rapidly and this final paper is about an emerging technology that will have a large impact on all aspects of the field in the future.

These papers have been compiled into this report to provide a ready source of information on a variety of topics for the interested reader. In each case, the papers include an extensive bibliography that can be used to obtain more details about the subject of the paper. For more information about the field of survivability, please visit the Web site of the survivability architectures project at the University of Virginia:

http://www.cs.virginia.edu/~survive

The Public Switched Telephone Network

Robert S. Sielken

1. History

1.1 Advances in technology

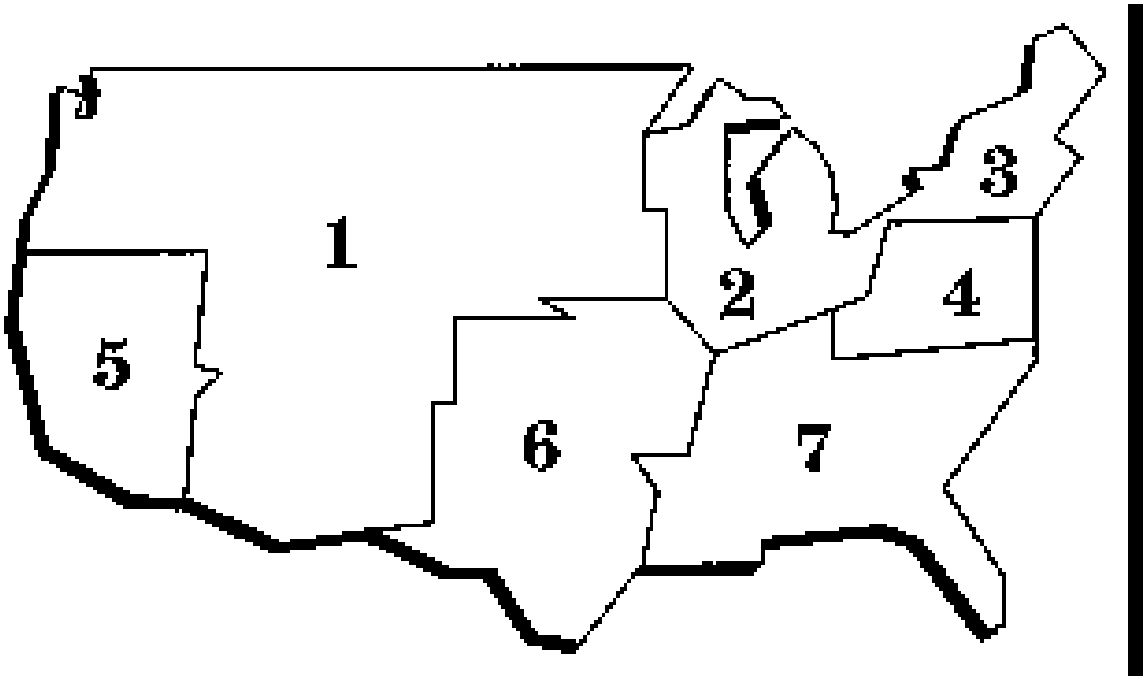
The first working telegraph, the predecessor to the phone, was developed in 1830. About thirty years later, Antonio Meucci designed and built the first transmitter and receiver for a telephone set. At the same time, Philipp Reis developed the first instrument that could transmit music over a wire. However, it was Alexander Graham Bell, a Canadian, who finally discovered that speech could be sent across a wire. In March 1876, Bell received the patent for an electrical speaking telephone [8].

The first telephone switchboard was installed in Boston in 1877. In 1878, New Haven, Connecticut, was the first city to have a commercial telephone exchange. The first commercial telephone service was established in 1884 between Boston and New York City. Almon Strowger of Missouri invented the automatic line selector in 1889; it utilized step-by-step switching to automatically connect two subscribers. The first automatic exchange was installed in Indiana in 1892. By 1896, the pulse dial telephone started to replace the older type of telephone that required the caller to press the button *n* times to dial the digit '*n*' [8].

The first transcontinental line was finished in 1915 and connected New York City to San Francisco. The first coaxial cable linked New York to Philadelphia in 1936. In 1947, microwave radio relay was put into use. The first transatlantic cable was completed between Canada and Scotland in 1956, while the first transpacific line was not completed until 1963 between Canada and Australia. The 1960s saw the first use of satellites to relay signals between continents [8]. By 1970, Corning Glass Works was able to produce a glass fiber with sufficient purity to support telecommunications [16].

1.2 Current Structure

Up until 1982, AT&T had a virtual monopoly on the telecommunications industry. It provided 85% of all local telephone service and 97% of the long distance service. Judge Harold H. Greene from the United States Justice Department created the antitrust settlement that divided the country into 160 local access and transport areas (LATAs). The Bell operating companies were only allowed to provide service within the LATAs, not between them. Each of the twenty-two companies was incorporated into one of seven regional Bell operating companies (see Figure 1) [16].



Number	Name	Constituents
1	US West	Mountain Bell Northwestern Bell Pacific Northwest Bell
2	Ameritech	Illinois Bell Indiana Bell Michigan Bell Ohio Bell Wisconsin Bell

Number	Name	Constituents
3	Nynex	New England Bell New York Telephone
4	Bell Atlantic	Bell of Pennsylvania Diamond State Tel New Jersey Bell The Chesapeake & Potomac Companies
5	Pacific Telesis	Pacific Bell Nevada Bell
6	Southwestern Bell Corporation	Southwestern Bell
7	Bell South	South Central Bell Southern Bell

Figure 1. Regional Bell Operating System [16]

2. Incidents

Despite the high availability of the PSTN (Public Switched Telephone Network) in general (more specifics on the availability later), the PSTN is still subject to incidents that cause the PSTN to function incorrectly. Many of these are physical reasons and some are related to software. There are other reasons why the PSTN might fail, but those are not as well publicized.

2.1 Physical

Physical attacks on the network can be either related to the weather or to people.

2.1.1 Weather-related

The weather is a frequent source of problems in the PSTN. In January 1998, more than 100,000 people had either degraded or no service because of the heavy rains, ice, and variable temperatures in the greater Chicago area (Illinois, Indiana, Michigan, Ohio, and Wisconsin) [3]. At about the same time, snow and ice were hammering the East Coast of the U.S. More than 100,000 customers were without power due to downed lines. The phone system was hampered by downed lines as well, but it was also hampered by the loss of more than a dozen generators from its switching

substations that were stolen to help alleviate the power supply problem [9]. Earlier in the month,rain and melting snow caused flooding in the western half of the country that caused evacuations of people and phone systems not to work in areas that were underwater [24].

Weather related incidents have not happened only recently, they have been around for many years (ever since the beginning). An incident from the past was the San Francisco Bay Area earthquake of 1989. The earthquake only caused a few lines to be downed, but the biggest problem was the flood of long distance calls that followed - 140 million in 24 hours. The AT&T network was forced to give outgoing calls priority, and the switches themselves were able to handle this [7].

2.1.2 Non-weather-related

Of the physical attacks on the PSTN, not all of them are caused by the weather. People cause some of them. The most frequent incident involving people is when they cut a cable. On January 4, 1991, AT&T employees accidentally cut a cable which disrupted long distance service in and out of New York City for five hours. The outage shut down the airports and the financial exchanges. In June of the same year, a sliced cable blocked some long distance customers from calling between Washington and New York [7].

2.2 Software

Besides incidents related to the weather and people, incidents are caused by software errors. An error in a new version of a network computer program led to the famous AT&T outage of 1990. The massive service disruption began at 2:30 PM EST when the error caused the New York City computer to send out alarm messages throughout the network. The cascading alarms caused the switches to refuse to accept any new connections. The company estimated that more than half of the long distance calls placed were not completed. The problem was located that evening and was fixed by midnight. Later that same year, eight million customers in Washington, D.C., Maryland, Virginia, West Virginia, and California were unable to complete calls unless they were nearby local calls or using long distance carriers. The source of the problem was traced to a software problem in the version of Signaling System 7 that was being used to route calls [7].

3. How it works

3.1 Switching

3.1.1 Switches

The first switch was the one invented by Strowger in 1889, and it was commonly called the step-by-step switch. The crossbar switch was introduced in 1938. However, the biggest breakthrough in switching occurred in 1965 with the introduction of the No. 1 ESS that was the first electronic switch. The switch was the first to be controlled by a computer. It had switching for 10,000-65,000 lines and could handle up to 100,000 calls per hour.

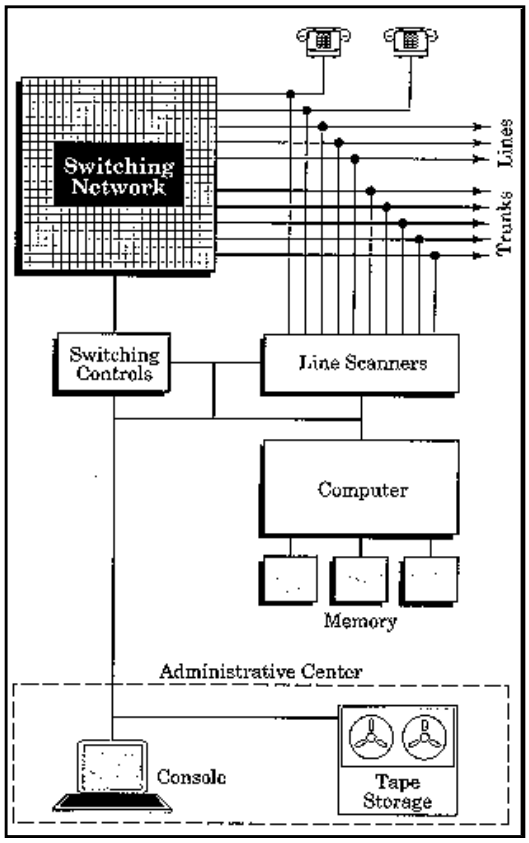


Figure 2. Electronic Switching System [16]

These first generation electronic switches had an analog switching matrix (see Figure 2). The first digital switch was the No. 4 ESS installed by AT&T in 1976. The future of the digital switch is becoming shorter with each passing day as optical switches are being designed to handle fiber optic

traffic. This futuristic telephone switching system could support 10,000 channels operating at 150 Mb/s. In theory, such a system would be able to handle all of North America's voice traffic at once [16].

3.1.2 Computing devices on the PSTN

The PSTN can supply a temporary method of providing a physical connection between two devices called DTE (an acronym for Data Terminal Equipment). A device called a DCE (data circuit-terminating equipment) is used to convert the DTE output into a signal that is suitable to be transmitted over the standard voice circuits [6].

3.1.3 Switching Offices

The North American switching plan classifies each office according to the function it performs (see Figure 3 and Table 1).

Class 5 (CL5, end office). Subscribers are normally connected to these. The main function is the interconnection of any two subscribers.
Class 4 (CL4, TP (toll point)). The main function of this switch is to interface with the class 5 offices and with the intertoll network. They provide the beginning and final stages of concentration and expansion for toll traffic to and from the end offices.
Class 4 (CL4, TC (toll center)). This office provides operator assistance for toll traffic.
Class 3 (PC (Primary Center)). This office interconnects the class 4 and 5 offices with larger geographical areas.
Class 2 (SC (Section Center)). This office interconnects the class 4 and 5 offices with larger geographical areas.
Class 1 (RC (Region Center)). This office interconnects the class 4 and 5 offices with larger geographical areas.

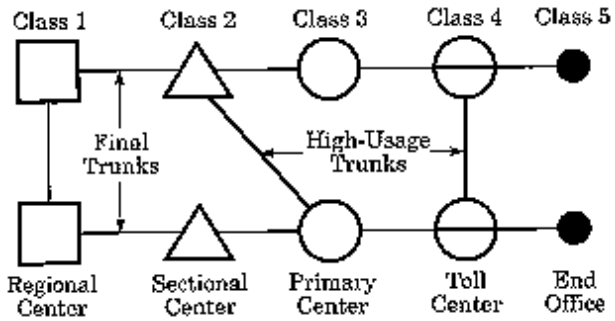


Figure 3. Switching Structure [16]

The North American plan consists of a homing pattern where each lower office (lower offices have the higher numbers) connects to a higher office [8].

Switching System	Class	Switching Function Performed	Homes on:
RC	1	CL1,2,3,4	All other RCs
SC	2	CL2,3,4	CL1
PC	3	CL2,3,4	CL1
TC or TP	4	CL4,5	CL3,2,1
End Office	5	CL5	CL4,3,2,1

Table 1: Switching Functions and Homing Arrangements

Prior to the 1982 divestiture, there were approximately 20,000 end offices, 1,300 toll centers, 265 primary centers, 75 sectional centers, and 12 regional centers. Traffic is always routed through the lowest level (highest numbered) available level in the hierarchy [16].

3.2 Spectrum

The PSTN was originally designed for voice communications. By limiting the sound spectrum to just those frequencies of human speech (not a wide range), engineers could reduce the bandwidth used for each call which would permit more simultaneous calls in the given spectrum. However, while this works well for voice communications, it is a serious bottleneck to data communications [6].

3.3 Power

Originally, phones were powered from the subscriber side. Eventually, this became less practical, so now the power source originates from the central office. Telephones are powered by a -42 to -52 V dc (typically -48) current supplied from the central office [8]. This is why the phone system can remain functional when the AC power goes out. However, this may no longer be the case when fiber optic cables become prevalent since they carry light, not electric current.

3.4 Wires

3.4.1 Coaxial

There are two types of coaxial cables used for transmission of broadband signals with low attenuation: rigid (or air dielectric) and flexible (or solid dielectric). The rigid cables contain an inner conductor insulated from the outer conductor by insulating spacers, while the flexible cables have solid and continuous dielectric. The key advantage to coaxial cable is its ability to minimize noise pickup [8].

3.4.2 Fiber

Fiber optic cables are the way of the future. Although seemingly simple, the wire is actually composed of three parts: core, cladding, and coating. The core is the center where the light travels. The cladding surrounds the core and keeps light from escaping. The coating is a soft layer of acrylate that surrounds and protects the fiber. Light can travel up to 100 miles in the cable before it needs to be boosted. A single laser can be turned on and off up to a billion times per second. A strand of glass can carry multiple wavelengths (colors) of light simultaneously. Therefore, light can handle billions of bits of information per second, far more than any other technology. While fiber optic cables are made from glass, they are by no means brittle. Fiber's theoretical strength is two million pounds per square inch (psi), and its typical strength is about 600,000 psi [10].

3.5 P(A)BX

The private (automated) branch exchange system was developed to meet the needs of private switching in medium to large-sized organizations. The private part of the name is because it is used for within the organization, and the branch exchange part of the name is used to designate the switching system as a branch of the local central office. The PABX switching system is used to transfer from one subscriber in an organization to another subscriber in the same organization through the central office. The PABX performs the same functions as a switch in the central office, but it also has some additional features: call hold, call forwarding, automatic callback, conference calls, and paging. To dial out of the system, the caller must dial an exit digit first (usually an 8 or 9) [8].

3.6 Rotary vs. Touch-Tone

Rotary telephones use dial pulsing where *n* pulses are made to represent the digit *n*. A small break between the pulses signals the beginning of the next digit. When the circuit is closed, it is a make; when the circuit is open, it is a break. The touch-tone telephone uses semiconductors to generate an audio signaling tone. Pressing a button causes the rotation of a set of two rods resulting in two audio tones that represent the digit [8]. Table 2 shows a comparison of these two technologies.

Rotary	Touch-Tone
Dialing: make-break technique	Dialing: combination of two frequencies
Operation with mostly direct control switching equipment	Operation with mostly common control switching equipment
Slow dialing	Fast dialing
Heavy electromechanical device	Light, mostly electronic device

Table 2: Telephone System Differences

3.7 Backbone

Microwave radio has been the backbone transmission system of the long distance telephone network for the past thirty-five years. Radio beams follow a line-of-sight path between towers. High frequencies (4, 6, or 11 GHz) are conducted through metal pipes called waveguides that couple the radio with the antenna. At each tower, the signal is received, amplified, and transmitted to the next tower. On the 4 GHz system, two DS-3 rate signals (each corresponding to 1,344 voice channels) are transmitted in each radio channel [16].

3.8 AT&T Picturephone

AT&T introduced the Picturephone at the World's Fair in New York in 1964. While economically infeasible at the time, current videophone applications are both convenient and economical on the current PSTN. They typically use 112 kb/s using two Switched 56 lines [16].

4. Capacity of Numbers

4.1 NPA

The number system currently used in North America is called the Number Plan Area (NPA) [8]. More and more numbers are needed all the time because of the increasing number of subscribers. The question is not whether or not to add more numbers, but rather how to add these numbers. There are currently seventeen dialing plans in use around the country [4].

4.2 Area code splits vs. overlays

Despite the large number of dialing plans available, the debate on how to add more numbers boils down to two options: split an area code or add an overlaying area code. Irrespective of how numbers are to be added, it takes the phone company a minimum of six months from the date of the NPA code assignment to reprogram its machines to accept the new number and allow permissive dialing for a short period [19]. (See Figure 4.)

4.2.1 Area code splits

One way to add more numbers is to split an area code. This is being done more and more in the big cities with a doughnut style where the main part of the city keeps the old area code (the inner part) while the outer parts of the city get a new area code. With the split, the dialer may have to dial seven, ten, or eleven digits for some local calls [4]. The advantage of this plan is that it keeps the current geographical separation that people are used to in the numbering system. However, many people must change their existing numbers. This might be expensive given how many places may have a phone number listed. An alternative, though less popular, way to split the old area code is in half with one half keeping the old number and the other half getting a new are code [19].

4.2.2 Overlays

The other main option is to use an overlay. In this system, the old area code would remain and all new numbers in that area code would have the new area code. With the overlay, all dialers dial ten digits (eleven depending on the city [19]) regardless of whether it is billed as a local call or not [4]. This scheme has the advantage that it does not require the current subscribers to change their

numbers. The disadvantage is that everybody always dials ten digits. Estimates are that dialing the extra three digits in Houston alone will result in $26.5 million in lost time (1.5 seconds to dial the extra three digits, $5 hour for an individual's time). The other disadvantage is that the same house may have two different area codes. If the parents had a line before the overlay, it would have the old area code. If they decided to add a second line for a computer, a fax, or the children, it would have the new area code. This would cause people to lose some geographical association with the area code system. Outside of the city, the difference would be negligible since the outside dialers can still associate the old area code number with the same region and just add the new area code to also represent that region [26].

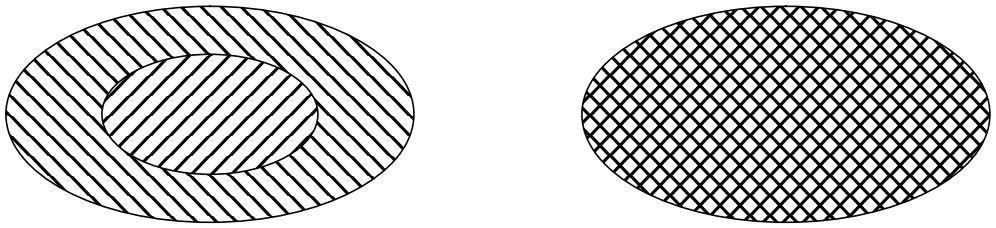


Figure 4. Area Code Split vs. Overlay

4.3 Toll free numbers

A similar phenomenon is taking place with the toll free numbers. Toll free numbers are part of the Toll Free Interexchangeable Numbering Plan (INPA). In April 1998, 877 joined 800 and 888 as the next toll free number [13].

5. Requirements

5.1 Power

Severe weather such as lightning may cause commercial power outages, but it will not immediately affect the phone network. Companies have backup power to last for a few hours. Ameritech has

backup batteries that can last four or five hours before being recharged. However, people must remember that cordless phones, fax machines, answering machines, modems, and some business phone systems will still not work because they may require AC power [20].

5.2 SNR

A signal-to-noise (SNR) ratio is used to measure the performance of a link in the system. The ratio is computed by dividing signal power by noise power. The higher the ratio, the clearer the connection and the more data capacity it possesses [6].

5.3 Capacity

Having the capacity to support all of the subscribers being on the phone at the same time would be a tremendous waste of resources. The typical telephone system will only accommodate 10% of the subscribers as the originators of calls. In areas where business phones are used frequently, capacity might be 15-20%. The number of supported active dialers varies from system to system but is around 1% [15]. An additional 15-25% of that 1% must be added to account for false attempts and prematurely abandoned calls. The network handles over 600 million messages a day over more than 20,000 switching systems [16].

Besides handling normal calling, the network can also be adjusted to accommodate unusual behavior. In 1997, Ameritech noticed about a 25% increase in the number of calls made from Wisconsin after the NFC Championship and Super Bowl XXXI involving the Green Bay Packers. For 1998, Ameritech made sure that the network was ready just in case the Packers won Super Bowl XXXII [23], but the Packers lost.

5.4 Availability

Availability of the PSTN is of the utmost concern for the phone companies and their customers. Bellcore's availability requirement is 99.93%. But the PSTN averaged over 99.999% availability in the early 1990's. This was attributed to reliable software, dynamic rerouting, loose coupling, and human intervention [14].

5.4.1 Kuhn's results

Many studies have been done about the reliability of the PSTN. In 1997, D. Richard Kuhn published the findings from his study of the PSTN from April 1992 to March 1994 [14]. The following table (Table 3) shows his results.

Category and Source	# of outages	Avg. # of customers affected	Avg. outage duration (minutes)	Customer Minutes (millions)
Human Error - Company	64	193,240	143.9	2,160.1
Human Error - Other	73	83,936	360.1	2,415.8
Acts of Nature	32	159,000	828.2	3,124.0
Hardware Failures	56	95,690	159.8	1,210.8
Software Failures	57	118,130	130.2	544.7
Overloads	18	276,760	1,123.7	7,527.2
Vandalism	3	85,930	456.0	110.5

Table 3: Failures in the PSTN

Kuhn defined the categories as the following:

Human error - company: errors made by telephone company personnel
Human error - other: errors made by people other than telephone company personnel
Acts of nature: major and minor natural events and disasters
Hardware failures: hardware component failures
Software failures: internal errors in the software (it should be noted that the software versions (mismatches) error was reclassified as a software failure and not as a human error - company for this paper)
Overloads: situation where demand exceeds supply
Vandalism: sabotage or any other intentional damage

The following pie charts (Figures 5 and 6) represent this same data.

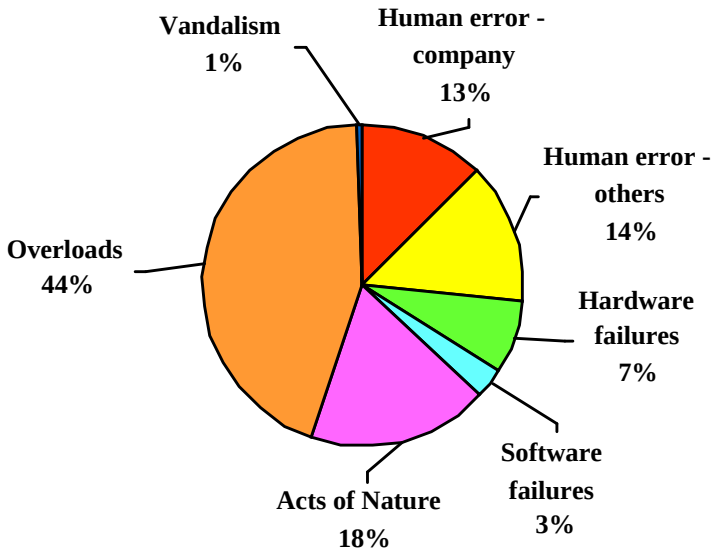


Figure 5. Outages by Category

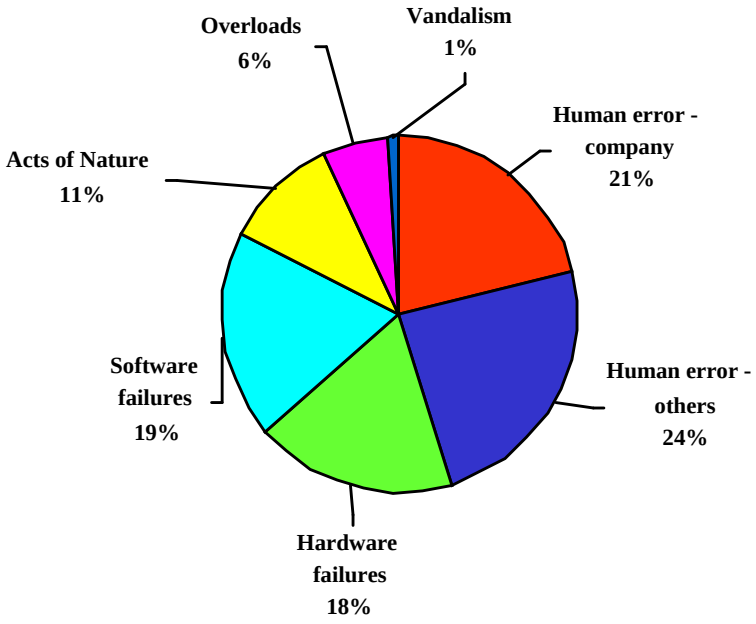


Figure 6. Downtime Percentages in Customer Minutes

5.4.2 Snow's results

Andrew Paul Snow received his Ph.D. from the University of Pittsburgh in 1997. His dissertation, *A Reliability Assessment of the Public Switched Telephone Infrastructure*, analyzed the PSTN in great detail. Some of his more general results will be presented in this section (Tables 4, 5, 6, 7, 8, 9, and 10, and Figure 7) [22].

5.4.2.1 Fault Allocation

Carrier Segment	Faults	Percentage
Major Local Exchange	531	72
Top 3 Interexchange	148	20
Competitive Access Provider/ Other	59	8
Total	738	100

Table 4: Fault Allocation

5.4.2.2 Fault Category Allocation

NRSC Category	Failures	Percentage
Facility	364	49.3
Local Switch	119	16.1
CCS - Local	71	9.6
Tandem Switch	69	9.3
Central Office (CO) Bulk Power	50	6.8
CCS - Net	21	2.8
Natural Disaster	16	2.2
Natural Disaster - Local	11	1.5
Overload	9	1.2
Other	8	1.1
Facility	364	49.3
Switching	188	25.5
All Other	94	12.7
CCS (Common Channel Signaling)	92	12.5
Total	738	100.0

Table 5: Fault Category Allocation

The NRSC is the Network Reliability Steering Committee that defines the fault categories as:

Local Switch: the failure of any element of a local switch that renders the switch or major portions of it unusable to subscribers
CCS - Local: the failure of any element of a local switch that isolates the switch from the CCS network
Natural Disaster - Local: a single local switch failure induced by a natural disaster
Facility: inter-switch transmission failure
CCS - Network: the failure of a CCS network due to causes other than local switch CCS isolation
Overload: a network traffic condition that results in blocked cells
CO Bulk Power: the failure of external or service provided backup CO power
Natural Disaster: a failure induced by an act of god and not preventable
Tandem Switch: the failure of any element of a tandem switch that renders the switch or major portions of the switch unusable to the network and/or the subscribers
Other: failures not covered by the other categories such as water leakages and internal building environment

5.4.2.3 PSTN Fault Determinant Involvement

Fault Determinant	# of Times Involved	Percentage
Cable Cuts	262	28.6
Hardware Problems	204	22.2
Human Error	142	15.5
Software Problems	107	11.7
Power Failure	65	7.1
SS7 Signaling Problems	47	5.1
Weather	42	4.6
Natural Disaster	29	3.2
Traffic Overload	19	2.1
Total	917	100.0

Table 6: PSTN Fault Determinant Involvement

5.4.2.4 Regional Fault Allocation

The telephone companies have long claimed that the faults in the system are geographically independent. However, the faults per million population differences shown in the next table were found to be statistically significant, which contradicts the telephone companies' claims of geographical independence.

Region	Composition	Faults	Percentage	Per Mil. People
1	CT, ME, MA, NH, RI, VT	32	4.3	2.4
2	NJ, NY	53	7.1	2.1
3	DC, DE, MD, PA, VA, WV	57	7.7	2.2
4	AL, FL, GA, KY, MS, NC, SC, TN	122	16.4	2.7
5	IL, IN, MI, MN, OH, WI	117	15.7	2.5
6	AR, LA, NM, OK, TX	130	17.5	4.6
7	IA, KS, MO, NE	45	6.1	3.8
8	CO, MT, ND, SD, UT, WY	37	5.0	4.8
9	AZ, CA, NV	94	12.7	2.7
10	ID, OR, WA	38	5.1	4.3
11	AK, HI	6	0.8	3.6
PR	Puerto Rico	12	1.6	3.4
All	All	743	100.0	2.9

Table 7: Regional Fault Allocation

5.4.2.5 Time

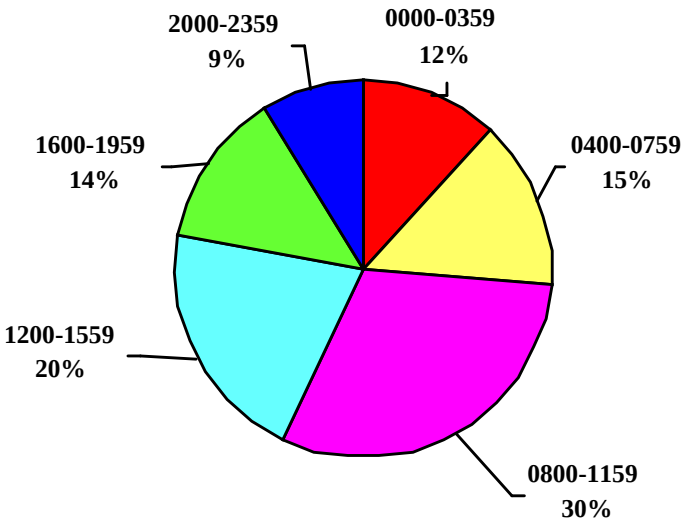


Figure 7. Fault Allocation Time

5.4.2.6 Number of Affected Lines

# of Lines Affected	Faults	Percentage
< 50 K	331	44.9
> 50 K and < 100 K	236	32.0
> 100 K	171	23.1
Total	738	100.0

Table 8: Failure Size

5.4.2.7 Local Switches

Code	Description	Number	Percentage
1	Scheduled	10,234	54.4
2	Procedural Error (Telco install./maint.)	650	3.5
3	Procedural Error (Telco non-install./non-maint.)	388	2.1
4	Procedural Error (System vendor)	304	1.6
5	Procedural Error (Other vendor)	244	1.3
6	Software Design	1,625	8.6
7	Hardware Design	227	1.2
8	Hardware Failure	1,900	10.1
9	Acts of god	403	2.1
10	Traffic Overload	42	0.2
11	Environmental	77	0.4
12	External Power Failure	97	0.5
13	Massive Line Outage, Cable Cut, Other	103	0.5
14	Remote - loss of facilities between host/ remote	233	1.2
15	Other/unknown	2,281	12.1
	Total	18,808	100.0

Table 9: Local Switch Outage and Failure Cause Distribution

Codes	Failure Category	Number	Percentage
2,3,4,5	Human Error	1,586	18.5
6,7	Design Error	1,852	21.6
8	Hardware	1,900	22.2
9,10,11, 12,13,14	External Circumstances	955	11.1
15	Other/Unknown	2,281	26.6
2-15	Total	8,574	100.0

Table 10: Local Switch Failure Cause Distribution

5.4.2.8 Long Distance

The long distance companies are always in fierce competition with each other. However, they do not provide the same reliability of service. MCI and AT&T each account for about 8% of the total FCC reportable failures, while Sprint only accounts for about 5%. But on a failure per toll minute basis, AT&T was five times more reliable than MCI and eight times more reliable than Sprint over a four-year period in the early 1990s.

5.4.3 AT&T

The AT&T network consists of 135 4ESS Switches that can each handle over a million calls an hour. Each call is routed among the most efficient of 134 possible routes [17]. The AT&T network handles more than 230 million calls on an average business day. In 1996, AT&T handled more than 68 billion calls with 99.99% going through on the first attempt [18].

5.4.3.1 Network Redundancy

AT&T believes that redundancy is the best way to avoid network problems. Each 4ESS switch has dual processors. The Signal Transfer Points (STPs) consist of a pair of AT&T 3B computers; STPs are used to route network inquiries over the signaling network. Network Control Points (NCPs), the customer database for advanced services, has dual processors as well as a backup NCP. Digital Interface Frames (DIFs) provide access to and from the 4ESS switches to process calls. Spare DIFs

are maintained in case of a failure. Power relies on the commercial power company, but battery and generator backups are available in the event of a commercial power outage [1].

5.4.3.2 Real Time Network Routing

AT&T has patented a Real Time Network Routing (RTNR) system that causes the 4ESS switches to communicate with one another every five to ten seconds to determine the most efficient routes. Every switch knows the status of every other switch. As an example, RTNR will route calls in the morning on the East Coast through the still sleeping West Coast to avoid the congestion on the east coast. As the day progresses, this trend is reversed as the East Coast becomes less busy with the end of the work day while the West Coast is still hard at work [21].

5.4.3.3 FASTAR

AT&T has the FASTAR (FAST Automatic Restoration) system to provide automated restoration services. FASTAR is fully automated and is designed to restore traffic in less than five minutes [1]. In 1995 in Alabama, a cable cut disrupted service to 82,000 circuits. The FASTAR system was able to restore full service in less than one minute. The San Francisco Bay Area also experienced a cable cut that day with service being disrupted to over 120,000 circuits. Of those 120,000 circuits, 97% had service restored to them in less than one minute [17].

5.4.3.4 Worldwide Intelligent Network

AT&T has a similar network structure around the world. Their reliability and availability allow them to competitively offer guarantees that all outgoing calls, incoming toll-free calls, domestic faxes, international faxes, and long distance cellular calls will all get through. The Data Communications User Survey gave AT&T the highest ratings for network reliability [2].

6. Future

Telecommunications is changing quickly. However, the PSTN has not become obsolete but rather a partner in the telecommunications industry. Each of these services could warrant their own paper, so this will be just a brief overview of the technologies involved.

6.1 Wireless/Cellular

6.1.1 General

Cellular phones can be divided into two groups: analog and digital. Up until now, analog has been the dominant choice, but this is starting to change with the additional capabilities of the digital service. Instead of one large base station, cellular phones systems are divided into many small areas called "cells". The main reason for using cells is frequency reuse. This increases efficiency of channel use, which allows more calls in the system. The Mobile Telephone Switching Office (MTSO) handles connections between the cellular network and the PSTN. The MTSO also controls the cell sites and manages all of the phones via a control channel [12]. The mobile units can tune to any of the 832 FM channels in the 800-900 MHz range [16].

6.1.2 UltraPhone

A company call InterDigital has developed a new wireless system called UltraPhone. It can support up to 95 full-duplex voice circuits per 1.2 MHz of spectrum (over 9.5 times as much as the traditional cellular). The system is based on the concept referred to as the Radio Carrier Serving Area (RCSA). A RCSA can include all area within a 60 km radius. The system is composed of two sub-systems: a Central Office Terminal (COT) and a Radio Carrier Station (RSC). The UltraPhone system interconnects with the PSTN at the central office [25].

6.2 ISDN

ISDN (Integrated Services Digital Network) originally emerged as a viable technology in the 1980s. However, limited coverage, high tariffs, a lack of standards, and a lack of demand all kept it from emerging from obscurity to prominence. Although it caught on faster in Europe, the Internet and the increasing demand for bandwidth have brought it to the forefront recently. ISDN is essentially a digital phone call capable of simultaneously carrying voice and data over existing twisted-pair phone cabling systems. Many of the telecommunications companies have switched to the fully digital network, but the conversion of the local loop has been slow. It is the increase in the number of local loops supporting ISDN that have made it grow [6]. Despite its growing popularity, the ISDN industry still suffers to some degree (but less than before) from tariffs. In 1996, the Maryland

Public Service Commission had a complaint filed by the Consumer Project on Technology against Bell Atlantic contesting Bell Atlantic's ISDN tariff [5]. (See Table 11.)

Benefit	Description
Integration of Multiple Services	ISDN can handle voice, data, image, video, and sound over the same link
Higher Speed	ISDN delivers up to 128 Kbps uncompressed (512 Kbps compressed using 4:1)
More Cost Effective	ISDN may be cheaper than PSTN on a cost per MB basis when you consider the cost of having a voice, data, and fax line
Same Wiring Infrastructure	ISDN uses the same local loop wiring with the installation of a network termination unit (NT1) at the remote client's site
Simultaneous Voice and Data	ISDN has this, PSTN does not
Up to 8 devices on a single ISDN circuit	The number of devices depends on the support of the switch at the central office
Use of Existing Analog Devices	Existing analog devices can be used along with digital devices
On-Demand-Pay-As-You-Use Service	May be cheaper than leased-line services
More Accuracy	Lower error rate for data transmission

Table 11: Benefits of ISDN over PSTN

6.3 ADSL

ADSL (Asymmetric Digital Subscriber Line) follows on the heels of HDSL (High Data Rate Digital Subscriber Line). As the name implies, the data transmission rates downstream (to the end user) and upstream are not the same. Because of the twisted-pair wiring needed to avoid signal coupling, symmetric transmission would significantly limit the data rate that a line could attain. This is the reason for using asymmetric transmission. Also, the PSTN system usually has many more connections coming out of a central office than between the central office and the end subscriber. However, this is not a major problem because most of the target applications have a large need for high data rates downstream but not upstream (see Table 12). These applications include video-on-demand, home shopping, Internet access, remote LAN access, multimedia access, and other specialized PC services [6].

Distance	Data Rate (downstream)
Up to 18,000 feet	1.544 Mbps (T1)
16,000 feet	2.048 Mbps (E1)
12,000 feet	6.312 Mbps (DS2)
9,000 feet	8.448 Mbps

Table 12: ADSL Downstream Capacity Depends on Distance

6.4 Cable modems

The PSTN is not the only path to the Internet. The cable industry is trying to compete with ADSL using cable modems. Cable modems are capable of up to 10 Mbps. The cables themselves are capable of up to 27 Mbps, but, unlike ADSL, this must be shared between the subscribers. Unlike the phone systems, cable modems will not require a dial-up or login process since the services will be continuously available. They will be essentially configured as wide area networks (WANs) [6].

Unlike the PSTN, the cable network was not originally designed to be a switched network. This causes two problems. First, cable was designed for unidirectional broadcast, so travel in the other direction is usually noisy. Second, as already mentioned, bandwidth must be shared among users, just as it is on a local area network (LAN) [6].

6.5 Internet

The Internet has the capability to connect with the PSTN to support long distance calling. VocalTec Communications has developed some next generation Internet Phone IP telephony software to accomplish this task. The following diagram (Figure 8) explains the basics of how it works [11].

Figure 8. Internet Call Routing

6.6 Satellite

While cellular is ideal for urban areas, it is not as practical for very remote or rural areas. It is in this case that satellites can be of use. Many communications companies have formed a consortium to work on a project called Iridium that will allow customers to call anywhere around the world using handheld wireless telephones. The system is composed of sixty-six low Earth orbiting satellites (LEOS). The low orbit allows the system to use less power and have less delay than with geosynchronous satellites (round-trip delay of about 540 ms) [16].

6.7 Comparisons

Service	Speed (Upstream/ Downstream)	Availability Now?	Best Application	Pros and Cons
PSTN	1200 to 56 KBPS (U&D)	Yes	Remote access from customers	Widest availability; low speed
ISDN	128 Kbps (U&D)	Mostly Yes	Internet and vid-eoconference	High speed, low error rate; relative high costs
ADSL	1.54 to 6 Mbps / 64 to 640 Kbps	Limited	Internet	No new cabling required
Cable	10 to 40 Mbps / 28 Kbps to 15 Mbps	Somewhat	Consumers and telecommuters, corporate LANs	Limited cable not suited for video conferencing
Satellite	400 Kbps to 30 Mbps/ 28 Kbps	Yes	Video broadcast	Limited interactivity

Table 13: Comparisons

7. Conclusions

The PSTN has come a long way since Alexander Graham Bell's first patent in 1876. Although incidents and failures of the network are well documented, they are relatively uncommon. The PSTN's availability is over 99.999% and seemingly getting better with modern network surveillance and dynamic reconfiguration capabilities. However, the PSTN must maintain its high level of service while at the same time undergoing changes in its structure. The other threat to the PSTN is the growth of competitors in the telecommunications/information industry. However, with the PSTN's tradition of excellence and the ability to continue to upgrade its infrastructure to better

technologies, the PSTN will remain a vital, strong infrastructure in North America as well as around the world for many years to come.

8. Bibliography

[1] "AT&T network service reliability and restoration, a white paper". AT&T News Release, January 25, 1994, http://www.att.com/press/0194/940125.chb.html.

[2] "The AT&T Worldwide Intelligent Network: Guarantees". AT&T. http://www.att.com/reliability/guarantee.html.

[3] "Ameritech crews come through with herculean response to phone trouble". Ameritech Press Release, January 12, 1998. http://www.ameritech.com:1080/news/releases/jan_1998/12_01.html.

[4] J. J. Carpenter. "Re: Area code splits vs. overlays - pros and cons?". http://www.cabl.com/telecomreg/9705/187.html.

[5] "The Consumer Project on Technology's Comments on Bell Atlantic's ISN Tariff and Request for Joint Evidentiary Proceeding". Maryland Public Service Commission, June 3, 1996. http://www.cptech.org/isdn/mdcomm.txt.

[6] C. Dhawan. Remote Access Networks: PSTN, ISDN, ADSL, Internet, and Wireless. Washington, D.C.: McGraw-Hill, 1998.

[7] Events in Telecommunications History. AT&T Archives, 1992.

[8] M. Harb. Modern Telephony. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1989.

[9] "Hard-hit Northeast braces for snow atop ice". CNN Interactive. http://www.cnn.com/WEATHER/9801/16/weather.wrap/index.html.

[10] "How Fiber Works". Corning. http://www.corningfiber.com/index.html and handouts/brochures.

[11] "How Internet Calls Work" Internet Collect. http://www.internetcollect.com.

[12] "How It Works: Cellular Phones". Radio Design Group, Inc. http://www.radiodesign.com/cellworks.htm.

[13] "Information Announcement for the 877 Toll Free Code". US West Network Disclosure #404, February 16, 1998. http://www.uswest.com/disclosures/netdisclosures404.html.

[14] D. R. Kuhn. "Sources of Failure in the Public Switched Telephone Network." Computer, Vol. 30, No. 4, April 1997, pp. 31-36.

[15] J. E. McNamara. Technical Aspects of Data Communication. Massachusetts: Digital Equipment Corporation Press, Second Ed., 1982.

[16] J. G. Nellist. Understanding Telecommunications and Lightwave Systems: An Entry Level Guide. New York: IEEE Press, Second Ed., 1996.

[17] "Network Facts: Real Time Network Routing". AT&T. http://www.att.com/reliability/facts.html#restored.

[18] "Network Management". AT&T. http://www.att.com/reliability/managers.html.

[19] "Petition for Approval of an NPA Relief Plan for the 847 NPA". State of Illinois, Illinois Commerce Commission, October 10, 1997. http://www.state.il.us/ICC/libdocs/selected/97/101097.847area.code.htm.

[20] "Power outages won't affect phone network". Ameritech News Release, June 16, 1997. http://www.ameritech.com:1080/news/releases/jun_1997/16_02.html.

[21] "Real Time Network Routing". AT&T. http://www.att.com/reliability/rthr.html.

[22] A. P. Snow. "A Reliability Assessment of the Public Switched Telephone Network Infrastructure." Dissertation, University of Pittsburgh, Department of Information Science and Telecommunications, July 1997.

[23] "Super Bowl phone calls a 'snap': Ameritech's network ready to handle calls from joyous Packers fans". Ameritech News Release, January 22, 1998. http://www.ameritech.com:1080/news/releases/jan_1998/22_02.html.

[24] "Thousands flee as Western rivers overflow: Clinton declares Nevada disaster area". CNN Interactive. http://www.cnn.com/WEATHER/9701/03/western.weather/index.html.

[25] "UltraPhone 110 Wireless Digital Loop Carrier System". http://www.interdigital.com/ult10_2.html.

[26] K. Walt. "Critics say plan for area codes doesn't ring true: Houston hearing on overlay plan Monday". http://www.chron.com/content/chronicle/page1/96/01/06/areacode.html.

An Analysis of Non-Security Failures of the Electric, Phone, and Air Traffic Control Systems

Sean McCulloch

1. Introduction

The critical infrastructure applications of the United States need to be made to survive many different adverse conditions. Some of these conditions are security issues- what happens if someone tries to break a system maliciously? Much more frequently, however, failures of infrastructure systems are due to non-malicious means. Usually these involve the weather, or human error, or bad system design. It is important that we analyze these types of failures of the system, both so that we can see what effects an outage has on the customers of a given service, but also to see if any existing problems can be fixed, making our system more reliable. The three systems to be analyzed are the Electric system, the Phone system, and the Air Traffic Control system.

Regrettably, it is difficult to find reliable information about failures of the infrastructure. It seems that people are reluctant to advertise when critical systems go wrong for some reason. However, the government does require some failures of some infrastructures to be reported, and then that information is made public. This is where the bulk of information was obtained. It was a conscious decision to use only that information, and also to ignore information from other sources, whose reliability could not be proven. This includes failures posted to places like comp.risks, and the like, and also includes lists of infrastructure failures produced by private organizations, where the validity of the lists could not be checked.

2. The Telecommunications Industry

In the telecommunications industry, the Federal Communications Commission (FCC) requires that phone companies report outages of at least thirty minutes that affected at least 30,000 customers. In addition, the phone companies were required to report any outages that affected "critical" systems (airports, 911, military installations, etc.). The Network Reliability Steering Committee was created

to analyze the data reported in this way. Each year they produce an Annual Report that discusses the type and severity of outages in the past year. The 1997 Annual Report [5] is the most recent, and it also discusses how the most recent outages relate to the outages of previous years.

One of the main difficulties in reading the Annual Report is that it measures the severity of outages using a metric called the "Outage Index", that has no explained derivation. The NRSC says that the index is dependent upon the number of customers affected, the time the outage began, the duration of the outage, and the severity of the services affected. It is not an intuitive measure, however, and they admit as much.

The Outage Index is used to measure the severity of each outage, and to measure the aggregate severity of a period of time. These severity numbers are then placed in either the "Green" (acceptable) range, the "Yellow" (warning) range, or the "Red" (immediate action required) range. The severity numbers for 1997 are all well within the Green range.

The Annual Report also categorizes outages by type of outage. They define several types of outages:

Facility outages are outages that happen because the wires themselves break.
Common Channel Signaling (CCS) outages arise from faults in the signaling system.
Local Switch outages happen on the local switches.
Tandem Switch outages happen on the bigger switches that connect networks.
Central Office Power (CO Power) result from power outages at the central office
Natural Disaster outages result from the weather.

The following table shows the number of outages of each type in 1997:

Facility	CCS	Local Switch	Tandem Switch	CO Power	Natural Disaster	Other
85	11	37	21	13	0	3

Facility outages are by far the most common, accounting for about 50% of all outages. Since these are failures of the lines themselves, many of these outages are actually caused by private citizens doing excavation, and digging up the phone lines. The next highest outages, after facility outages, are switch failures.

The Annual Report also measures the duration of outages in minutes, as compared with previous years:

Percentage	Baseline Year	Year 1	Year 2	Year 3	Year 4
95th	931	899	959	1113	791
75th	307	400	342	362	248
Median	133	148	171	194	156
25th	57	70	83	97	68
5th	46	36	39	35	35

The “years” for this table vary from July of one year, to July of the next, so Year 4 is from July 1 1996 to June 20, 1997.

This table shows us that the time for the longest outages in the most recent year are the smallest of any recorded year. It also shows that over 95% of the outages are significantly longer (by at least 5 minutes) than the minimum outage duration required by the FCC. What this means is that if there is an outage serious enough to require reporting, it is usually going to last for a long time.

The Annual Report is not the only job of the NRSC. They also look at the causes of outages, and potential causes of future outages. One document that discusses future outages is the Internet Study Team Report [6]. This report analyzes the growth of the Internet, and the effects it will have on the Telecommunications system in the future.

The first conclusion drawn from the study is that Internet access is redefining the “busy hour”, when telephone usage is at its peak. When before, there were many business calls during the business day, now there are many Internet calls being made at night. They also note that while most voice calls are short (about 2-3 minutes on average), the average Internet call lasts 4-7 times as long. Also, they note that the times of peak usage for Internet calls is also the times of peak usage for 911 calls, so there is some concern that there will be congestion on the 911 lines. The report also goes into some technical suggestions of how to solve these issues.

3. The Electric Infrastructure

In contrast to the telecommunications system, the Electric system has outages much more frequently, at least over small areas. As a result, the Department of Energy (DOE) requires outages to be reported only if they affect over 50,000 people for over 3 hours in most cases, or less severe outages of specific systems for shorter periods of time. These reports contain the duration of the outage, the number of customers affected, and the reason for the outage (if known). The DOE sends these reports to the North American Reliability Council (NERC), a group of electric companies across the country, overseen by the DOE. A part of NERC is the Disturbances Analysis Working Group (DAWG), which contains a database of these reports [7].

From 1990 to the present, the DAWG Database had 152 outages in it with "complete" information. "Complete" means that the outage had a start time and a time when the repairs were finished, and it mentions the number of customers affected, or had some explanation as to why no customers were affected. Most also had a reason for the outage as well.

The DOE requires that outages ("Blackouts"), load reductions ("Brownouts"), and cases where power was transferred in from another generating facility, to be reported. This means that some analysis of the data was needed to determine which of the reported outages were actually blackouts, and which merely resulted in a reduction in load for customers, and which were handled by backup systems, resulting no change in service for the customer at all. Of the 152 outages, 86 (56.6%) resulted in no actual power interruption to the customer. This is a sign that while we all know that the power system has many outages, there are in reality much less than there could be.

Of the 66 outages that did cause blackout, they were caused by many different factors:

Reason	Weather	Hardware Failure	Sabotage	Human Error	Overload	Nature
# of Outages	33	17	4	6	3	3
Avg. #of customers	380,000	276,000	107,500	79,000	2.53 Mill.	81,000
Avg. Duration	2.95 days	2.28 hours	52.5 min	3.33 hours	2 hours	1.5 days

The different categories of outages are:

Weather- Any abnormal weather condition (usually rain or snow, but 2 outages are due to hurricanes) that causes an outage.
Hardware Failure- Any failure of the equipment anywhere in the system to do what it was supposed to do.
Sabotage- Any deliberate, malicious human action.
Human Error- An accidental mistake by a human, working for the power companies.
Overload- An outage caused by a component of the system taking more load than it was designed for.
Nature- An outage caused by a non-weather-related natural incident (such as a fire, or a tree falling on a line).

The first thing to realize about the chart is that the statistics for "Overload" are skewed by a situation On August 10, 1996, where a couple of lines going down on the West Coast caused a cascading overload that took out about 7.5 Million People. Since the report didn't say when the power came back on for those people, the average duration of 2 hours is only for the other two events.

From the table, we see that more than half of the outages are caused by Weather and Nature, two things that cannot be controlled. It can be argued that the goal of a survivable system is to make as many of the outages as possible due to only these causes (since nothing can be done if a tornado destroys a generating station, for instance). The next goal would be to minimize the number and impact of the remaining outages. By this metric, the power system is doing well, better than the telecommunications system. Of course, outages of the power system are still far more frequent.

4. The Air Traffic Control System

The third infrastructure area being analyzed, the Air Traffic Control system, is slightly different from the other two. This is because the Federal Aviation Administration (FAA) requires that every accident be reported, no matter what the cause. These reports take several years to complete, so there are only 28 incidents involving major aircraft from 1990-1996 that have complete reports [3]. Of these 28 accidents, only three were caused by problems with the Air Traffic Control System. Two were caused by errors by the Air Traffic Control Operator, and one was caused by the Air Traffic Control Procedure not covering a certain situation. The most recent of these was February of 1991. None of these were caused by failures of the computers in the Air Traffic Control System.

By way of comparison, there were 12 incidents caused by pilot error, 7 caused by the malfunction of the plane, 3 by adverse weather, and 2 by hijacking.

Incidents with smaller aircraft were much more frequent [4]. There were 60 incidents in 1997 alone that resulted in fatalities, but none of them had anything to do with failures in the Air Traffic Control System. Of those whose cause could be determined, 39 were due to pilot error, 9 were due to plane defects, and 11 were due to the weather.

5. Two Examples of Large Failures

In addition to looking at general statistics for the various outages of the different infrastructures, it is also instructive to look at specific large failures of the infrastructure, to see why they happened, and whether anything has changed to prevent them from happening again. Two such failures in particular are worth noting- the East Coast Blackout of 1965, and the AT&T failure of 1990.

On November 9, 1965, a blackout occurred in the Northeastern US, blacking out 30 million people for as long as 13 hours [2]. The cause of this blackout was a single transmission line overflowing, which caused its load to be redistributed to other lines, causing them to overflow. This chain reaction quickly overloaded many lines, causing each generating station to become isolated from the network. Within five minutes, each of the islands became imbalanced, because it wasn't getting the input it was expecting, and overloaded.

As a result of this accident, the NERC, and other agencies were created to analyze the system to find the cause and prevent a similar situation from happening again. One way the system was changed was to add more redundancy to the system, which means that if one line fails, there are many more lines to share the excess load. Thus, now we have very few outages due to overload (as the previous analysis of the DAWG database shows).

The AT&T outage differs from the Blackout in that it was caused by a problem in software, not in hardware [1]. A bug in the software of the switching system was introduced that had the effect of causing the switch to be unable to handle two messages from another switch in rapid succession. When one switch crashed because of this bug, its backup came online, with the same faulty code, and also crashed. The messages that were sent from the crashing switches were received by other

switches, which could not handle them and crashed. This quickly led to the entire system going down for 9 hours, causing AT&T to lose $60 million in calls, and other businesses losing untold millions in lost revenue.

The bug was caused by a break being placed in the wrong place in a switch statement. It was a simple one-line fix. As a result of this outage, we have a better understanding of the seriousness that software bugs can cause. Of course, there is still no good way to assure that software is bug-free.

6. Conclusion

While there have been many outages in some critical infrastructure applications, and some serious ones, in the normal case, the applications perform fairly well. We see that the phone system is getting better every year, the power system can avoid more than 50% of all potential blackouts, and the Air Traffic Control system has yet to fail and cause a fatality. However, we also realize that this is still only the normal case. The examples of the 1965 Blackout and the AT&T failure show that the system still is very vulnerable, and that software bugs or bad hardware design can cripple the infrastructure. Also, there is no data for an explicit, malicious attack meant to take down an infrastructure. Even the cases listed as "sabotage" before were very small scale. We must prepare these systems to be able to withstand such assaults, or eventually, surely someone will exploit the vulnerabilities.

7. Bibliography

[1] Burke, Dennis, "All Circuits are Busy Now: The 1990 AT&T Long Distance Network Collapse". http://cobra.csc.calpoly.edu/~dbutler/papers/att_collapse.html

[2] Central Maine Power Company, "The Great Northeast Blackout of 1965". http://www.cmpco.com/aboutCMP/powersystem/blackout.html

[3] Lisk, David, "Major Commercial Airline Disasters". http://www.d-n-a.net/users/dnetGOjg

[4] National Transportation Safety Board, "Accident Synopses". http://www.ntsb.gov/Aviation/months.htm

[5] Network Reliability Steering Committee, "1997 Annual Report". http://www.atis.org/atis/nrsc/view.htm

[6] Network Reliability Steering Committee, "NSRC Internet Study Team Report". http://www.atic.org/atis/nrsc/intrept.htm

[7] North American Electric Reliability Council, “DAWG Database”. http://www.nerc.com/dawg/database.html

Major Security Attacks on Critical Infrastructure Systems

Matthew C. Elder

1. Introduction

This paper presents a catalogue, or historical review, of major security attacks on critical infrastructure systems. The details of these security attacks, to the extent that they are public knowledge, will be presented, including when, what, where, who, and with what effect.

The first question to consider is what constitutes a security attack? A system is considered secure if its resources are utilized and accessed as intended under all circumstances [27]. Given this definition, a security attack would be an attempt to utilize or access system resources in an unintended manner. More detail would then be required specifying intended usage of systems in order to utilize this definition practically. Alternatively, a security attack could be defined in terms of security violations and security policies. A security violation is "a violation of a system's security policy," where the security policy defines appropriate behavior regarding the system and its information [1]. Examples of security policies include defining intrusion and misuse - which users are permitted within a system and considered authorized to perform given operations. A "security attack" is then defined as any attempt to perpetrate a security violation. Of course, the security attacks that succeed in causing a security violation are the most interesting in the context of this work.

Security attacks and violations have varying degrees of "seriousness"; that is, the impact of the security attacks range from minimal to catastrophic. This work deals with "major" security attacks on the critical infrastructure; the second question to consider then is what constitutes a "major" security attack? Given that the purpose of critical infrastructure systems is to provide services upon which society depends [23], a "major" security attack would be one that could impair the critical infrastructure system's ability to provide its critical functionality or services. The extent of misuse

or disruption providing critical services (or the extent of damage to the system) distinguishes "major" security attacks from others.

The third question to consider is what is considered the infrastructure? There have been many reports recently exploring this nation's dependence on proper information system operation, including the Presidential Commission for Critical Infrastructure Protection's report [23] and the Defense Science Board report [5]. The critical infrastructure is defined as those systems and applications upon which the society depends for normal functioning or operation.

The national security of the United States is increasingly dependent on U.S. and international infrastructures. Previously, national security was dependent primarily on just the services the military provided. Now, military services are dependent on economic and political interests. Society requires large amounts of infrastructure, such as information, financial, and power infrastructure, to function normally, and these infrastructures are highly interdependent. It is clear that economic and security interests have become inseparable [5].

The PCCIP determined five categories for the critical infrastructure systems:

- Telecommunications: Information and Communications (PSTN, Internet, computers)
- Banking and Finance
- Energy (Electrical Power, Oil and Gas Production and Storage)
- Transportation/Physical Distribution
- Vital Human Services (Water Supply; Government Services: Social Security, records management, and other programs; Emergency Services: police, fire, rescue, EMS)

Those domains cited and explored in the PCCIP report will be taken to constitute the critical infrastructure. In particular, the domains of the critical infrastructure explored in this work are decomposed into the following categories:

- Military services
- Government services
- Emergency services
- Water
- Power
- Gas and Oil Production and Storage
- Air Traffic Control
- Rail Transportation

- Telecommunications
- Banking and Finance

Having outlined the original scope of this work in the preceding paragraphs, it should be stated that incidents not falling strictly into the above categories are included and outlined in this work as well. Primarily, all security attacks against the critical infrastructure are examined. There are various reasons why all security attacks against the infrastructure are, in some sense, "major." For example, regardless of the extent of actual disruption the security attack might cause, any attack on the critical infrastructure of society undermines public confidence and trust. In other words, the implications of a security attack are, in some ways, sometimes, as important as the actual consequences. Proper functioning of society rests in part on the public's confidence in the systems upon which society depends, in addition to the actual functioning of those systems. In addition, security attacks on the infrastructure lead to additional exposure of those systems to attacks (i.e. publicized successful penetrations might make these systems more of a target for attack).

An additional extension to the original scope of this work regards what is considered the critical infrastructure. The PCCIP report focuses on the dependence of the United States on the aforementioned application domains; this report considers security incidents pertaining to these application domains worldwide.

The structure of this work is the following: the next section outlines the major security attacks against the critical infrastructure. That section is organized according to application domain, and incidents within an application domain are presented chronologically. The following section of this work provides some analysis of the incidents as a whole, including information on how the data was obtained, trends, and general observations. The final section of the report presents conclusions. Finally, a bibliography of sources is given.

2. Security Incidents

The security attacks presented are organized according to the area of the infrastructure that was under attack. In each application domain, an outline of attacks is given first and then the major attacks will be presented in their own subsections.

2.1 Military Services

The military services are a common target for security attacks. There are many reports that confirm this observation. One Pentagon study, reported on in October, 1997, disclosed that over 250 Defense Department computer systems were broken into in 1996 [22]. These Pentagon estimates were used in a GAO study of computer attacks against the Department of Defense, released in May, 1996 [32], [33]. The GAO report stated that unauthorized attempts to enter Defense computer systems might have reached 250,000 in 1995. That document also reported that only one in every 150 attacks is detected and reported, and 65% of internal, tiger team attacks against its own Defense systems are successful.

A more recent military exercise conducted in June 1997 was reported on in April, 1998 [25]. A tiger team of fifty National Security Agency (NSA) hackers tried to penetrate U.S. military and civilian networks. The cyber war game, code-named "Eligible Receiver," demonstrated the vulnerability of computer systems to security attacks. The hackers gained access to many computer systems nationwide, including the U.S. Pacific Command in Hawaii, responsible for 100,000 troops in Asia. The FBI tracked only one unit posing as North Korean hackers.

Many incidents of security attacks against military computers reported in the press are security attacks on Web pages. On December 30, 1996, the Air Force had its Web page hacked, replacing the normal page of aviation statistics with a pornographic picture [4], [15], [18]. The U.S. military shut down access to over eighty of their sites in response. No classified information was accessible from the Web sites though. The Pentagon's U.S. Army Artificial Intelligence Center's Web page was hacked by "Chameleon" on October 4, 1997. Finally, both the U.S. Army and the U.S. Navy had sites hacked by the "No|D Crew" on March 8, 1998 and March 9, 1998, respectively [21].

Other security incidents include the Defense News magazine reporting that the U.S. Army in Bosnia was dealing with many computer virus infections, including those by the Monkey, AntiEXE, and Prank Macro viruses [15]. In addition, Anderson Air Force Base in Guam was broken into by a 15-year-old Croatian hacker using hacking tools available on the Internet in February, 1997 [16]. The attack was detected and no files were accessed. The GAO report presented a few more incidents of attacks against the military, including hackers from the Netherlands penetrating 34 defense sites in 1990 and 1991 [33].

The remainder of this section describes in detail more serious attacks against military systems.

Rome Laboratories Air Force Base: March-April, 1994 [15], [16], [32]

The first major successful security attack that is documented occurred in March and April of 1994. A 16-year-old teenager in the United Kingdom, Richard Pryce, known as "Datastream Cowboy," broke into the Air Force command and control research facility, Rome Laboratories, in New York. A second hacker, "Kuji," also penetrated the computer systems. There were over 150 incidents of the hackers using Trojan horse programs and sniffers to gain access to the lab's systems. The hackers took control of the lab's network, disabling all 33 subnetworks for several days. In addition, sensitive air tasking order research data was stolen and additional military, government, and civilian systems were accessed illegally from Rome Laboratories. Among those other sites penetrated were NASA Goddard Flight Center, Wright-Patterson Air Force Base, Griffiss Air Base, and a Lockheed computer network in California.

Pryce was charged in May 1996 and convicted in March 1997. He was fined $1915. The second hacker, "Kuji," was never caught. No one knows what happened with the stolen data. It is estimated that the cost to detect and recover from the attacks at Rome Labs was over half a million dollars. This does not take into account the value of the data or the cost of the attacks at other sites.

Pentagon: February, 1998 [2], [12], [19], [20]

Pentagon computers were attacked and successfully penetrated for a two-week period in February 1998. Two Cloverdale, CA teenagers, an 18-year-old Israeli master hacker, and two other Israeli teens were responsible for the attacks on non-classified Pentagon computers and other government-related networks. At the time, the Deputy Defense Secretary characterized the events as "the most organized and systematic attack the Pentagon has seen to date."

The attacks included attempts to set up electronic "trap doors" in software systems in order to obtain information illegally. It does not appear, however, that much damage was done or much sensitive information was obtained.

The alleged mastermind behind the attacks was 18-year-old Ehud Tenenbaum, known as "Analyzer." All three Israelis were placed under house arrest on March 18, 1998. The two California teens had their homes and computers searched by the FBI.

Tenenbaum is a member of the computer group, the Enforcers. In retaliation for the house arrest of Tenenbaum, the Enforcers perpetrated a week-long assault on Internet Web pages during March 1998.

Navy and NASA: March, 1998 [9], [35], [36]

On March 4, 1998, unknown hackers executed a denial-of-service attack on thousands of computers at Navy installations, NASA centers, and various universities. Computers running Windows NT and Windows 95 were crashed using the attack called "New Tear," "Bonk," or "Boink." The attack exploits vulnerabilities in the Microsoft implementation of the TCP/IP stack to crash computers that must then be rebooted to fix the problem. No further damage is caused by the attack.

Navy computers at Point Loma, CA, Charleston, SC, and Norfolk, VA were crashed, among others. Nine of the ten NASA major field offices were attacked; the NASA sites reporting the attack were the following: NASA Headquarters (Washington, D.C.), Ames Research Center (California), Dryden Flight Research Center (California), Goddard Space Flight Center (Maryland), Independent Validation and Verification Facility (West Virginia), Jet Propulsion Laboratory (California), Kennedy Space Center (Florida), Langley Research Center (Virginia), Lewis Research Center (Ohio), Marshall Flight Center (Alabama), Moffett Federal Airfield (California), Stennis Space Center (Mississippi), Wallops Flight Facility (Virginia), and White Sands Test Facility (New Mexico). The extent of the attacks at the various sites, however, was unclear: at NASA Ames Research Center, for example, only 50 of the 3,000 computers were affected.

After the attacks, Microsoft posted an update pointing to the patches fixing the problem proposed in January 1998. The attack utilized a fragmented UDP network packet to crash the system.

Defense Information Systems Agency: April, 1998 [10], [12]

In possibly the most major successful security attack to date, reported on April 21, 1998, a hacking group called the Masters of Downloading (MOD) claim to have stolen a suite of programs that run classified U.S. military networks and satellites. The software, the Defense Information Systems Network (DISN) Equipment Manager (DEM), was allegedly stolen from the Defense Information Systems Agency. The DISN is described as the telecommunications backbone of the U.S. military.

MOD released a statement claiming to have stolen the software, detailing its capabilities, and providing images of the software's user interface screens. A copy of the software was made available to John Vranesevich of AntiOnline, a computer security site, who verified its authenticity. The DEM software remotely monitors and manages military computer-related equipment, primarily network devices including routers, repeaters, and switches. The software can control military communications networks and monitor GPS satellites and receivers. The software cannot control the GPS satellites, but it can be used to pinpoint their exact locations. Allegedly, the stolen DEM software could be used to shut down the entire Defense Information Systems Network.

MOD is a group of fifteen hackers worldwide, including eight Americans, five Britons, and two Russians. A MOD member conducted two interviews with Vranesevich in mid-April. The group claims to have stolen the software in October 1997, from a Windows NT server. They state that they have no hostile intentions regarding usage of the software, but thirty individuals have copies of the software worldwide.

2.2 Government Services

The most common form of attack against government computers is the hacking of Web pages. The Justice Department's Web page was hacked on August 24, 1996 in protest of the Communications Decency Act [15]. On September 19, 1996, the CIA's home page was attacked by Swedish hackers of the group "Power Through Resistance" in protest of telecommunications companies pressing charges against hackers [13], [15]. The Florida Supreme Court's home page was hacked on October 25, 1996 [15]. The NASA Web page has been hacked multiple times. On December 23 and 30, 1996, it was hacked by "\\StOrM\\". On March 5, 1997, the group H4GIS hacked the page as a

political statement for Kevin Mitnick [15]. Most recently, the Department of Commerce home page was hacked by H4GIS in retaliation for the arrests of the Pentagon hackers [3].

Other incidents include an attack on the White House e-mail system in March, 1996 [15]. The denial-of-service attack consisted of the e-mail system being flooded with fraudulent, unwanted requests to Internet mailing lists. Compounding the congestion problem was the auto-responder at whitehouse.gov responding to the incoming mailing list traffic. On November 6, 1996, the computer networks of the Environmental Protection Agency's mid-Atlantic region had to be shut down due to a virus; 15% of the region's computers were infected [15]. Finally, as described previously, NASA computers were attacked on March 4, 1998, along with other Navy and university sites [9], [35], [36].

Probably the most serious security issues regarding government services pertain to the State Department [2], [30]. On March 23, 1998, the State Department announced that the General Accounting Office (GAO) had conducted a study of the State Department's computer systems and found vulnerabilities. However, an earlier USA Today story indicated that security attacks on two overseas diplomatic posts during October, 1997 had breached the agency's network, causing it to be partially shut down; the State Department denied that occurred and that the incidents were a part of the GAO report. The GAO report was not released: the State Department designated parts of the report "secret" and the rest as "for official use only."

2.3 Emergency Services

The only reported incident of a security attack in the emergency services sector occurred on April 19, 1996, when the New York Police Department voice mail system was hacked [15]. The voice mail system was repaired quickly and no effects of the hack were reported.

2.4 Water

No security attacks were discovered in the water supply domain.

2.5 Power

The Information Assurance Task Force of the President's National Security Telecommunications Advisory Committee (NSTAC) conducted a risk assessment of the electric power industry. The study found that the most significant information security vulnerability in the power grid laid in the substations; automated devices monitor and control equipment within substations but are not well protected against intrusion. Physical destruction was found to still be the greatest threat to the electric power infrastructure, though electronic intrusion is an emerging threat. However, the study found no evidence of disruption of electric power caused by an electronic intrusion [24].

In contrast to that report on the minimal history of security attacks against the power grid to date, the aforementioned Pentagon Tiger Team exercise in June 1997 underscored the vulnerability of the power infrastructure to security attacks. The report on the Pentagon exercise stated that intruders perpetrated a security attack in which the electric power grid could have been sabotaged, disrupting power distribution to the nation [25].

2.6 Gas and Oil

No security attacks were discovered in the gas and oil production and storage domain.

2.7 Air Traffic Control

There has only been one major security attack against the air traffic control system. On March 10, 1997, a teenage hacker disabled telecommunications at a regional airport in Worchester, MA [6], [8]. That same day he also disrupted telephone service in Rutland, MA. The attack on the airport wiped out telephone access to the airport's control tower, fire department, airport security, and weather service for four hours. In addition, the airport's main radio transmitter and runway lighting control were disabled.

On March 18, 1998, the Justice Department charged the teenager with computer crimes, the first time a juvenile had been brought up on such charges. That same day the hacker accepted a plea bargain in which the juvenile must serve two years probation, be barred from employment at a

computer company, perform 250 hours of community service, pay $5,000 damages to Bell Atlantic, and forfeit the computer hardware and software used in the attack.

2.8 Rail Transportation

No security attacks were discovered in the rail transportation domain.

2.9 Telecommunications

A major area of the critical infrastructure under attack is the telecommunications system, the information and communication infrastructure, which includes the Public Switched Telephone Network (PSTN), the Internet, and computers that connect to it. There are a variety of forms in which attacks on the telecommunications infrastructure can come, for example, toll fraud (including phreaking) and attacks against PBXs, voice mail systems, and Internet Service Providers (ISPs). A study by Telecommunications Advisors, Inc. estimates that the total losses to the economy due to toll fraud and telabuse range from $2-8 billion a year [17].

Many examples exist of Internet Service Providers being the target of security attacks. BerkshireNet in Pittsfield, MA was attacked on February 27, 1996 [15]. A hacker gained administrator privileges and proceeded to vandalize the system, erase data on two computers, and shut down the system. The ISP was down for approximately twelve hours. PANIX, another ISP in New York City, was the victim of a denial-of-service attack in early September, 1996 [15]. The security attack utilized was the “SYN-flooding attack,” where fraudulent TCP/IP requests for connections to non-existent Internet addresses overwhelm a server. The ISP subsequently went out of business when it was the target of this attack over an extended period of time. A disgruntled ex-employee sabotaged Digital Technologies Group, an ISP in Hartford, CT, in October 1996, causing a week-long shutdown and other direct costs controlling the damage. The suspect was arrested in December, 1996 [15]. Finally, WebCom of Santa Cruz, CA was the victim of another SYN-flooding denial-of-service attack on December 14, 1996 [15]. Access to the Web pages of hundreds of businesses was blocked for forty hours by an unknown hacker in British Columbia.

There are also incidents of PBX and voice mail systems being attacked. In San Francisco, it was reported on July 10, 1996 that high school students attacked the PBX of a manufacturing firm and

hacked into the voice-mail system, crashing the system and costing $40,000 in incident response [15]. In August 1996, the PBX at Scotland Yard was hacked and approximately $1.5 million worth of fraudulent calls were made [15]. In September, 1996, Pacific Bell reported that phone phreaks tapping into residential and home telephone lines by clipping onto circuit boxes was costing them possibly several million dollars a year [15].

Cellular phone fraud is a major area of security attacks. In June 1996, the U.S. Secret Service announced the arrest of 259 suspects of cellular phone fraud whose violations were estimated at more than $7 million [15]. Soon after in July 1996, two people were arrested in Brooklyn, New York for stealing 80,000 cellular phone numbers, capable of generating $80 million worth of stolen phone services on the black market [15]. One of the most celebrated hackers this decade, Kevin Mitnick, was indicted in Los Angeles in September, 1996, on counts of stealing software, damaging computers at USC, using passwords without authorization, and using stolen cellular phone codes [15]. He later pleaded guilty.

An incident of security attacks against phone company computer systems occurred when a 19-year-old hacker, Christopher Schanot (“N00gz”), obtained unauthorized access to Southwestern Bell, Bellcore, Sprint, and SRI computers. In April 1996, he was indicted on computer fraud charges; in November he pleaded guilty to two counts of computer fraud and one count of illegal wiretapping [15].

The Internet Worm: November 1988 [28], [29]

Perhaps the most famous security attack against the infrastructure occurred in 1988 when a Cornell University graduate student, Robert Morris, released a worm on the Internet that infected and crashed thousands of computers. Morris, the son of the chief scientist at the National Computer Security Center (part of the National Security Agency), exploited weak passwords and security holes in the UNIX programs sendmail and fingerd. As interesting as the extent of the attack is the aftermath of the attack. DARPA soon after established the Computer Emergency Response Team (CERT) at the Software Engineering Institute at Carnegie Mellon University to coordinate response for and communicate information on computer security attacks.

2.10 Banking and Finance

Banks are frequent and favorite targets of security attacks. The Central Bank of Russia reportedly was the target of approximately 500 intrusion attempts from 1994 to 1996. The ITAR-Tass News agency reported that hackers stole $4.7M in successful security attacks in 1995 alone [15]. In October, 1996, various security attacks against Czech banks were reported, including intrusions where $1.9 million were stolen and an incident where Czech citizens' personal information was stolen and posted to electronic bulletin boards [15]. The next month, November of 1996, seven men in London pleaded guilty to defrauding British banks by tapping communication lines between ATMs and bank computers and using the stolen data to manufacture illegal bank cards [15].

An interesting twist on security attacks involves the threat of security attacks against financial institutions used for extortion [26]. On June 3, 1996, the London Times reported that hackers had been paid 400 million pounds sterling to remain silent about security attacks and logic bombs placed in New York and London financial institutions. The article reported that banks were concerned the public would lose confidence in the security of their systems if the security attacks were publicized. While this proved to be a hoax, in June of 1997, Newsday published a cover story stating that "COMPUTER HACKERS have successfully forced financial institutions in the United States, Europe and Asia to pay millions of dollars in ransom by threatening the companies' computer networks. The payouts were confirmed by law enforcement officials, banking insiders and security experts interviewed over the past several weeks. When most successful, the sources said, the crimes have linked disgruntled insiders with computer experts recruited throughout the world - including the former Soviet Union, India and southeast Asia - by organized crime groups." [16]

Citibank: Summer 1994 [11], [28]

In the summer of 1994, a Russian hacker, Vladamir Levin, led a gang of hackers that broke into Citibank Corporation's computer systems and made unauthorized transfers from customer accounts totaling more than $10 million. Citibank recovered all but approximately $400,000. Levin, a graduate of St. Petersburg Technology University, was accused of using his office computer at AO Saturn, a St. Petersburg computer firm, to break into the Citibank Cash Management System that allows Citibank customers to transfer funds from their Citibank accounts to accounts at other financial institutions over their computer network.

Levin, the mastermind of the scheme, was arrested by Interpol at Heathrow Airport in April 1995. The U.S. Justice Department accused Levin of stealing $2.8 million from Citibank accounts in New York and directing the Citibank computers to send the money to bank accounts in Finland, Israel, and San Francisco (Bank of America). An accomplice of Levin's, Alexei Lashmanov, pleaded guilty in a U.S. court to charges against him in January 1996. Lashmanov admitted to executing unauthorized wire transfers from customer accounts to his accounts in five Tel Aviv, Israel banks and to trying to withdraw $940,000 from these accounts. The maximum sentence Lashmanov can receive is five years in prison and a $250,000 fine. Three others have already previously pleaded guilty for their roles in this incident.

Hong Kong Investment Banks: November, 1996 [15], [16]

Five Hong Kong investment banks were brought down by a disgruntled computer technician at Reuters, November 29, 1996. According to Peter Neumann, Wilson Chan Chi-kong, 29, a former employee of Reuters financial information agency who sabotaged the dealing-room systems, was motivated by revenge after a dispute with his superior. The hacker detonated logic bombs, causing 36 hours of downtime in networks providing market information for trading. Fortunately, alternative services were available to switch to and no serious effects were reported. Damage control took over 1700 man-hours, at a cost of HK$1.3 million [16].

Japan: January, 1998 [14]

Sakura Bank, Ltd. in Japan reported on January 5, 1998 that confidential computer records had been stolen. The information stolen was primarily information on approximately 20,000 of its 15 million customers, such as names, addresses, and telephone numbers. The leaked information for at least 37 of those customers was subsequently leaked to a mailing list vendor in Tokyo. However, no customer accounts were accessed and no money was stolen in the security attack. Sakura Bank believes the data was stolen when bank affiliate Sakura Information Systems Co. upgraded the software for the computer system in 1997.

GAO report on Electronic Banking: January 1998 [34]

The GAO conducted a survey of banks concerning electronic banking and their possible experiences offering electronic banking services. While no specific results and security incidents were presented

in that report, the statistics concerning security were interesting in the context of this work. Of the 93 banks that were offering electronic banking services, significant percentages lacked basic security mechanisms such as firewalls (ten percent) or virus protection (eleven percent). Many of the banks that reported security problems had experienced attempts at unauthorized access, and one bank reported a successful unauthorized access attempt. In addition, a large number of banks had connections from their on-line systems to other computer systems, including Fedwire and other clearing house institutions.

3. Analysis

Incident information in this paper was obtained from many sources, including the general media (news agencies), government reports, the computer science literature, security-specific Web sites, and hacker Web sites. Reliable information on security attacks is difficult to obtain in many situations because the attacked parties do not want negative publicity, and the attacking parties tend to exaggerate the effects of their work. Information in this paper is cited appropriately and the source of the information can be judged on an individual basis for credibility.

News articles were obtained from the following Web sites: CNN, CNET, MS-NBC, Wired, the San Francisco Examiner, and the St. Petersburg Press. These news sources either used their own staff writers or reproduced articles from news wire services such as Reuters, Associated Press, or UPI.

An excellent source of information are reports generated by the United States General Accounting Office (GAO). The GAO conducts studies of various government operations and institutions, including the particularly relevant analyses of information security at the Department of Defense, the Department of State, and financial institutions conducting electronic banking.

There exist many general security Web sites, for a variety of purposes. There are those sites for incident reporting, such as CERT and the Department of Energy's Computer Incident Advisory Capability. However, for reasons of confidentiality these places do not make their databases of incidents available and no information for this report could be obtained there. The two most useful security Web sites were the International Computer Security Association (ICSA) and AntiOnline. ICSA publishes many white papers on security, including Year in Review papers. AntiOnline has a strong hacker focus and its founder, John Vranesevich, often interviews hackers who take

responsibility for security attacks. The information from both sites, however, is sometimes difficult to verify. The ICSA information does not always cite its sources (and sometimes its sources are possibly subject to hearsay, such as new stories from comp.risks). AntiOnline seems closer to the actual perpetrators of the attacks, but the objectivity of information then comes into question.

An interesting Web site on general security can be found at Discovery On-line, the Hacker Hall of Fame. This Web page presents a high-level history of hacking, focussing more on presentation than content. However, an overview of the subject is helpful in as much (or as little) detail as it offers. It should be noted that this Web site makes the distinction between "hackers" and "crackers": hackers are people who have done extraordinary things with computers. A subset of this group are crackers, who utilize their talents in an illegal or subversive manner.

Finally, there are an abundance of hacker Web sites, such as 2600.com. There is a site with an archive of hacked Web pages, including but not limited to those pertaining to the critical infrastructure. Limited information in this report was obtained from these sites.

An analysis of the information in this work begs the question: what is reported and what is not? It is well-known that most security attacks are not detected, and many of those that are detected are not reported at all. Of those that are reported, only a fraction will come to the attention of the press and the public in general. Therefore, it is plausible to assume that there are countless security attacks, major or other, that are unknown and not catalogued here. In fact, it could be argued that the most "major" security attacks against the infrastructure are possibly those that are not detected, and the losses and ramifications of such attacks are subtle or yet to be exposed.

Of the security attacks that are public knowledge, however, it is easy to see a trend from those that are less recent and severe to those that are more recent and severe. Especially in the military services domain, previous attempts have been more likely to be vandalism (i.e. Web page hacking). This year, however, the incidents effect widespread disruption and heightened levels of misuse and intrusion. The most recent attack on DISA was characterized as a "new level of security breach" by John Vranesevich [10].

There is the question of how one measures severity? There are many ways of characterizing an attack, from the extent of damage to the amount of time lost (inconvenience) to the exposure and

negative publicity generated. In some ways, many measures can be cast in financial terms, for example damage and time. However, it is often difficult to put a price on the value of information, in particular information lost in a security attack. Other difficulties in measuring severity include determining the extent of the information lost and the spread of that information after being obtained in an unauthorized incident. Measuring the effects of exposure and negative publicity is difficult as well.

Finally, an analysis of certain incidents highlights the interdependence of the varying infrastructure systems. For example, in the air traffic control incident, the security violations were primarily telecommunications vulnerabilities exploited. The effects, however, were widespread in both telecommunications and air traffic control (in addition to having emergency services implications).

4. Conclusion

This report has presented a catalogue of security attacks against the critical infrastructure. The infrastructure is vulnerable to a variety of security attacks, from a variety of sources, for a variety of reasons and motives (stated or otherwise). While all infrastructure domains are critical and subject to attack, historically the military, government, telecommunications, and banking infrastructures have been attractive targets of security attacks. In some domains, it is readily apparent that the intensity and severity of security attacks are increasing.

Information was obtained from a variety of sources, but for many reasons is difficult to come by regarding security attacks. The validity and objectivity of information must also be considered when reporting on security attacks.

The severity of attacks was discussed in this paper, but it is clear that there has never been so severe a security attack against the infrastructure such that society was incapacitated and no longer capable of functioning. An event of the magnitude of “Pearl Harbor” has yet to occur in the context of security attacks against the critical infrastructure.

Finally, an obvious question to ask in conclusion is what should be done? While that discussion is beyond the scope of the this work, a succinct and preliminary answer has been provided by a hacker involved in one of the most severe attacks presented. In an interview with AntiOnline, a hacker

from the Masters Of Downloading offered this advice to security professionals: “It’s simple: take all [classified] military systems off the Internet, place only [unclassified] Web servers on the Internet [and] keep the rest on a purely internal network.” [10]

5. Bibliography

[1] Bishop, M., S. Cheung, and C. Wee. “The Threat from the Net,” IEEE Spectrum, August 1997.

[2] Boyle, A. “Backing off from hacker backlash,” http://www.msnbc.com/news/151963.asp, March 27, 1998.

[3] “The Commerce Hackers,” Anti-Online, http://www.antionline.com/SpecialReports/CommerceHack/Story1.html, April, 1998.

[4] “Computer hacker plants porno on Air Force Web page,” http://cnn.com/TECH/9612/30/air.force.porn/index.html, December 30, 1996.

[5] Defense Science Board, Office of the Secretary of Defense. “Report of the Defense Science Board Task Force on Information Warfare - Defense (IW-D),” November 1996.

[6] Festa, P. “Airport hack raises flags,” http://www.news.com/News/Item/0,4,20278,00.html, March 19, 1998.

[7] Festa, P. “Computer security problems growing,” http://www.news.com/News/Item/0,4,19765.html, March 5, 1998.

[8] Festa, P. “DOJ charges youth in hack attacks,” http://www.news.com/News/Item/0,4,20226,00.html, March 18, 1998.

[9] Festa, P. “Hackers attack NASA, Navy,” http://www.news.com/News/Item/0,4,19674,00.html, March 4, 1998.

[10] Glave, J. “Have Crackers Found Military’s Achilles’ Heel?” http://www.wired.com/news/news/technology/story/11811.html, April 21, 1998.

[11] “Guilty plea by alleged Levin aide,” http://www.times.spb.ru/archive/sppress/141/guilty.html, January 9-15, 1996.

[12] “Hackers claim major U.S. defense system cracked,” http://cnn.com/TECH/computing/9804/21/pentagon.hack.reut/index.html, April 21, 1998.

[13] “Hackers vandalize CIA home page,” http://cnn.com/TECH/9609/19/cia.hacker/index.html, September 19, 1996.

[14] “Japan reports cyber bank heist,” http://www3.zdnet.com/zdnn/content/reut/0105/268074.html, January 5, 1998.

[15] Kabay, M. “The InfoSec Year in Review: 1996,” http://www.ncsa.com/knowledge/research/isecyir.htm, 1997.

[16] Kabay, M. “The InfoSec Year in Review: 1997,” http://www.ncsa.com/knowledge/research/iyir.htm, 1997.

[17] Kabay, M. "Totem and Taboo in Cyberspace," http://www.ncsa.com/knowledge/research/g.htm, 1994.

[18] Kornblum, J. "Pentagon reopens sites," http://www.news.com/News/Item/0,4,6577,00.html, December 31, 1996.

[19] "Pentagon, FBI probe latest hacks," http://www.news.com/News/Item/0,4,19472,00.html, February 25, 1998.

[20] "The Pentagon Hackers," Anti-Online, http://www.antionline.com/PentagonHacker/, April, 1998.

[21] "The New>? Pentagon Hackers," Anti-Online, http://www.antionline.com/NoidCrew/, April, 1998.

[22] "Pentagon had 100s of Net break-ins," http://www.news.com/News/Item/0,4,15624,00.html, October 24, 1997.

[23] Presidential Commission on Critical Infrastructure Protection. "Critical Foundations: Protecting America's Infrastructures The Report of the President's Commission on Critical Infrastructure Protection," October 1997.

[24] "Risk Assessment of the Electric Power Industry," http://www.ncsa.com/knowledge/research/elecpower.htm, March 1997.

[25] "Security team finds Pentagon computers unsecured," http://cnn.com/TECH/computing/9804/16/cyberwar.ap/index.html, April 16, 1998.

[26] Shelton, D. "Banks appease online terrorists," http://www.news.com/News/Item/0,4,1465,00.html, June 3, 1996.

[27] Silberschatz, A. and P. Galvin. Operating System Concepts, 4th Edition, 1994, pg. 460.

[28] Slatalla, M. "Hackers Hall of Fame," http://www.discovery.com/area/technology/hackers/hackers.html, April 28, 1998.

[29] Spafford, E. "The Internet Worm: Crisis and Aftermath," Communications of the ACM Vol. 32, No. 6, June 1989.

[30] "State Department admits to flaws in computer system," http://cnn.com/TECH/computing/9803/23/state.dept.hacking/index.html, March 23, 1998.

[31] "Survey says tech crime rising," http://www.news.com/News/Item/0,4,19697,00.html, March 4, 1998.

[32] United States General Accounting Office, "(Testimony) Information Security: Computer Attacks at Department of Defense Pose Increasing Risks," GAO/T-AIMD-96-92, http://www.gao.gov/AIndexFY96/abstracts/ai96092t.htm.

[33] United States General Accounting Office, "Information Security: Computer Attacks at Department of Defense Pose Increasing Risks," GAO/AIMD-96-84, http://www.gao.gov/AIndexFY96/abstracts/ai96084.htm.

[34] United States General Accounting Office, "Electronic Banking: Experiences Reported by Banks in Implementing On-line Banking," GAO-GGD-98-34, http://www.gao.gov/AIndexFY98/abstracts/gg98034.htm.

[35] "Update on Network Denial of Service Attacks (Teardrop/NewTear/Bonk/Boink)," http://www.microsoft.com/security/netdos.html, March 6, 1998.

[36] Zajac, A. "Hackers network to hone tools, skills," http://www.examiner.com/daily/0305hackers.hmtl, March 5, 1998.

Hacking Information Available on the Internet

Brownell K. Combs

1. Introduction to the Problem

Recent reports from the Department of Commerce indicate that electronic commerce is increasing dramatically [5]. Ten million people conducted electronic commerce business in 1997. Business to business purchases are expected to reach 300 billion dollars per year by 2002. Furthermore, the Internet population is increasing at an unprecedented rate. It took only four years for the Internet to reach 50 million users, and it currently has 62 million Americans connected. Internet traffic is reported to be doubling every 100 days.

The most cited risk to electronic commerce is that of the electronic criminal. The skilled computer thief who intercepts credit card or other personal information looms large in the mind of perspective electronic shoppers. Recent trends indicate, however, that the true threat to electronic commerce may lie elsewhere. With the widespread use of automatic encryption between secure web pages and easy-to-use commercial web browsers, the chance of electronic interception of credit cards or other information is decreasing.

On the other hand, if an electronic commerce web site can be disrupted or taken out of commission, then much more overall damage can be done than intercepting credit cards. While there is no individual loss, the company may be losing millions of dollars in lost sales. This may be especially important in the future, as electronic commerce becomes a very big industry. If almost every electronic commerce web site went down today it would not be that big of a crisis. But what about when electronic commerce is a 300 billion dollar a year industry? Not only will electronic commerce be more critical to the US economy, but there will be many more important web sites that can be targeted for attack. It seems logical that the more sites there are, the more likely it is that not all of those sites will be setup to adequately withstand attack.

Who might be doing these attacks? With more and more Internet uses, there is a likelier possibility of those that may choose to do damage. Becoming a skilled hacker is a very time consuming and difficult process. But what if there are tools and information easily available on the Internet that could turn a novice hacker into a dangerous one? Such information and tools could create a legion of hackers capable of doing serious damage to companies and the economy.

2. Terminology and Exclusions

It is common to read about the difference between hackers and crackers. When this split is used, a hacker is someone who has the ability to obtain outside, unauthorized access to electronic data or services [2]. This type of person is certainly capable of 'hacking' their way into a system, but does not do so. A cracker is a hacker who actually uses their skills to obtain unauthorized access. While elements of the hacking community have for some time tried to emphasize the difference between the two, this paper will not. A hacker is someone who has certain abilities. This paper does not seek to make any value judgements about possession of this knowledge. Just because someone does not choose to commit illegal actions does not mean they do not have the skills to do so. For example, a sharpshooter is someone who is extremely skilled with a firearm. One sharpshooter may work for the police or military and help to protect the citizens. Another sharpshooter may belong to a terrorist organization and seek to do wrong. But both are still sharpshooters. Likewise, this paper simply uses the word hacker and hacking to illustrate someone's ability to hack into a system.

Denial-of-service attacks are those that do not attempt to hack into the system, only prevent it from communicating with the network. These attacks include those that flood the machine with connection requests so that no other machine can request a connection; and, attacks that fool other computers into coming to a different location instead of that of the true host machine (IP address stealing). Although there are some applications that make denial-of-service attacks very easy, these types of attacks are not within the scope of this paper.

3. Search Methods

The following search methods were used to locate hacking information and applications on the Internet. The particular methods were chosen to simulate the possible activities of a relatively new

member of the Internet community who was searching for hacking info. All Internet searches were done through Yahoo. While Yahoo may or may not be the most comprehensive search directory, it is one of the easiest to use and appears to be very popular among new Internet users. An attempt was made to not spend more than an hour in any one location, and no site was explored more than 3 links deep. These methods prove that the information detailed in this report is not only accessible, but that it is easily accessible, even to a new Internet user.

A search on Yahoo for "hacking" and "tools" returned 7 different categories. A search for "hacker" returned 254 different entries. Within the first few pages of these results were several pages that were 'Hacking Resource Links Pages' that contained nothing but links to hacking resources. The majority of links used in this page were obtained in less than 30 minutes.

4. Sources

Sources for hacking methods have been broken up into three main categories: publications, web sites, and newsgroups. Some members of one category may be connected to members of another category. For example, 2600, the Hacker Quarterly, has both a published magazine and a web site. The two aspects are separated in different categories because the content of the magazine may not always match the content of the web page. Other sources may be conceivable be connected to two different categories, but were not since the majority of the source could be classified in a single category. A web page solely devoted to selling publications deserves to be mentioned only in publications category because that is from where the hacking information is obtained.

4.1 Publications

There are several different kinds of publications that have roots in the Internet. Perhaps some of the oldest are electronic magazines. These are magazines that are usually e-mailed to subscribers, but sometimes can be viewed on the magazine's web site. These electronic magazines are not included in the web site category, because their home pages often contain nothing but the content of the magazine (no web page only aspects). Phrack is an example of an electronic magazine that provides hacking information. It is published four times a year and has been in existence since 1985. Prior to the existence of graphical web pages, Phrack was only e-mailed to subscribers. Now it can also be seen and downloaded at the Phrack web site (www.phrack.com). Phrack is targeted

towards a more competent hacker, someone that is already fairly familiar with computers and the Internet. For example, some of the topics from the most recent issue of Phrack [9] are: “Weakening the Linux kernel,” “Piercing firewalls,” “Everything a hacker needs to know about getting busted.”

There are also several print magazines dedicated to hacking. These are listed in this report because the web sites that advertise these magazines are prominently displayed in any search for hacking information. While they can be found at many major booksellers, one often learns about the magazine from on-line information. One example is 2600 - The Hacker Quarterly. Each issue is approximately 60 pages and is published 4 times a year. The first issue was published in 1984. This type of publication often targets a level of hacker slightly higher than ‘novice.’ It is, however, not as complex as many of the electronic magazines. Unlike an electronic magazine, the cost of publication forces an actual print magazine to attempt to interest a large as possible audience.

The last type of publication is books and applications that are mailed to buyers. There are a number of companies that sell books that are supposed to teach one the basics of hacking. Often these companies have web sites and sometimes they advertise in the hacking magazines. One such company is Spectre Press (www.spectre-press.com). Although this company has a web site, it is described in this section since the web site is merely advertisement for selling books. One such book is the Computer Hacker’s Bible. It cost 20 dollars and covers the basics of many topics ranging from hacking particular pieces of hardware to password cracking. The contents of the book will be further detailed later.

4.2 Web Sites

There are two basic types of web pages that introduce one to hacking: amateur and professional. Professional pages are those that are hosted by companies that are either advertising for something or trying to sell directly from the web site. The majority of these companies are ones like Spectre Press. These web sites will not be detailed since there is very little information actually on the site. The books and applications for sale are what contains all the hacking information. As previously mentioned, several of the hacking magazines maintain official web pages. The site for 2600 magazine is one example (www.2600.com). The 2600 web site is specifically mentioned separately from the print magazine, because it often contains information not available in the magazine. One example is the section on web pages that have been hacked. The magazine posts the original page

and then what it looked like after it was hacked. Sometimes there is a story accompanying the display. The web site also provides a more timely voice for the hacking community. One of the issues on the web site is the "unfair" imprisonment of Kevin Mitnick [1]. One other category of professional web pages is the home pages of several high profile applications. These sites, like the one for the Satan Utility (www.fish.com/satan/), merely provide information about the application.

Amateur web sites are those put up by hacking groups or individuals interested in hacking. This type of site normally contains a small to large collection of text files introducing hacking or amateur case studies on past hacks. Several examples are the sites of Genocide2600 (www.aracnet.com/~gen2600/) and the Spider's Den (www.vicon.net/~bhoover). Sometimes these sites border on being classified as either professional or amateur. The L0pht (www.l0pht.com) is a hacking group that started to offer its security evaluation services professionally.

Another amateur site, Rootshell (www.rootshell.com) is a web site that is host to many different texts and applications on hacking. It is a particularly good resource for the novice hacker since it has a search engine that searches its collection of hacks for particular machines and software. Users of Rootshell can either choose to browse hacking text and applications chronologically or they can search for information on particular systems. For example, entering "Windows NT" returns several results. One is an application that when placed in a particular directory of an NT server collects passwords of users that log into that machine. Searching for "Solaris" (Sun Operating System) returned several explanations of overflowing buffers on systems without particular patches. Another entry was a C program that would overwrite a buffer for the user, giving them root access to that machine.

4.3 Newsgroups

There are several hacking oriented newsgroups. Among these are alt.2600, alt.hackers, and alt.cracks. These are, however, rarely a source of good information for the novice hacker. Actual techniques seem to be rarely discussed. These newsgroups seem to just be a forum for people that claim to be hackers to discuss random issues. There are, however, newsgroups that can be useful once a novice hacker gains a little experience. Namely alt.security and the CERT newsgroup contain information about newly discovered bugs. Even a novice hacker might be able to employ these techniques against sites where a very new bug has yet to be fixed.

5. Information

The following will be an examination of some of the information found during the search. The majority of information encountered can be classified as introductory. The most common example is some kind of FAQ that defines an area of interest to hackers. Other times a document may discuss strategies, but rarely complex techniques. That harder to find information is, however, not really in the scope of this paper.

One example is the PC Hacking FAQ obtained on a hacking group web page [8]. This document contained information about defeating security measures on a PC. It explained what a boot password was and several simply ways to get around the password (like removing the battery to reset the CMOS). The FAQ also discussed issues like ways to get access to DOS from Windows when a systems administrator had disabled the normal methods.

Another FAQ obtained from the 'newbie' area of a hacking group web site was entitled "Internet Cracking: Firewalls." While not a very in-depth document, it is perfect for the new hacker. It first explained what a firewall was and then detailed several different types of firewalls: dual-homed gateway, screened host gateway, and other vocabulary like proxies and bastion hosts. While no part of the FAQ is particularly technical, there are discussions on basic attack strategies, and tips on how to not get caught once a hacker gets in to the system. There are also pointers to ftp servers that have tools and applications that can help one secure their firewall, or break through someone else's firewall.

There is a widely available FAQ called the Hack FAQ. It contains sections on UNIX hacking, telephone hacking (phreaking), and hacking resources. The information is specifically presented in a format that can be understood by novices. For example, the section on UNIX password files explains where to find the files, what format they are in (with some examples), and what shadow password files are and how to get them. The FAQ also goes into some detailed about what can be logged and on which systems.

6. Tools

There are several tools that can be used to help hack into a system. The Security Administrator's Tool for Analyzing Networks (SATAN) is perhaps one of the best known of this class of application. It is an application that was designed to enable system administrators to examine their systems for 11 basic categories of security vulnerabilities [10]. Upon finding any of these vulnerabilities, the application gives the user a short tutorial explaining the problem and likely ways to fix the problem. None of the categories of vulnerabilities required revolutionary or advanced techniques to be exploited. All had been the subject of advisors from CERT or other similar organizations. The danger of SATAN in the hands of a novice hacker is, however, limited for two reasons. The only version publicly available is fairly old (1995) and is difficult to install. It requires an old version of the Mosaic web browser, which a novice hacker may not even be aware exists. Furthermore, SATAN can only run on UNIX compatible machines and requires root access. Some novice users may not even know what UNIX is, let along have root access to a UNIX machine. There are applications obtainable from Rootshell can accomplish tasks similar to SATAN with a much more user-friendly interface and installation procedure.

There are some other applications that fall under the category of password crackers. Two particular applications are called Crack and Brute. The purpose of these applications is to attempt to crack passwords in a UNIX password file. These files work by making a guess at the password (either systematic or from some list provided by the user), encrypting the guess, and comparing the result of that encryption with all the passwords in the file. If there are any matches, then the user now knows the password of that particular user. Most password checkers return the user name and any other information about the user that is contained in the password file along with the cracked password. As previously mentioned, this requires the hacker to understand what a UNIX machine is, etc. But this is not as hard a problem as SATAN, since the hacker needs only to understand how to log into the machine and move around (not install complicated programs). Especially since there are many tutorials and some applications that will help one download the password file from a UNIX machine if it is not adequately protected. Some examples of password crackers were found in the Computer Hacker's Bible as well as hacking group web sites.

Yet another class of applications allows the user to disguise from where their packets originated. Packets on the network contain the IP address of the computer that created and sent the packet. But

several applications can change this information in the header of the packet. While not something that helps to hack into a system, it is an application that can help a novice hacker to be harder to track.

7. Internet Vulnerability

So, what is the resulting threat of having these tools and information easily accessible? Now that we know what can be found, can anyone do anything serious with this stuff? It seems an incredulous theory that someone with very little serious computer knowledge could break into a system run by trained system administrators. Could someone who barely knows how an Operating System works pose any serious threat? There is evidence to suggest that may be the case. Dan Farmer, an Internet security consultant, used his application SATAN to scan a sampling of electronic commerce web sites. The web sites selected belonged to Banks, Credit Unions, Newspapers, and Federal government interests. This survey of over 1700 sites found that more than 60 percent had some security deficiency that would have allowed him to break into or destroy that host [3]. While it is true that Farmer is by no means a 'novice' hacker, the tool used was SATAN. Once installed, SATAN searches any host it is directed to, and reports any security deficiency found. Furthermore, the application provides a tutorial on the deficiency that could be understood by a novice hacker.

Another recent test by the Department of Defense conducted almost 9000 attacks on government systems [7]. Almost 90 percent of the attacks succeeded in breaking in, but only a total of 390 of those attacks were detected by the agencies running these systems. Both the number of successful attacks and the small number of attacks detected are astounding.

8. Case Studies

Two case studies will be briefly examined in order to give the reader an example of some break-ins that could have been perpetrated by novice hackers. The first case study involves the Department of Justice's web server [4]. In November of 1996 the official web site was replaced with a home page for the "Department of Injustice" by a hacker unhappy with the Communications Decency Act. This hacker gained unauthorized access to the Department of Justice's web server in order to make the switch. An investigation after the switch revealed that only one part time person was responsible for maintaining the security of the web site. The firewall was not configured properly and the most recent patches for multiple services were not installed. The intruder could have gained

access through well-known holes in FTP, SMPT or Telnet. In the wake of the incident, the Department of Justice created several full time security positions. Those employees spend all their time tracking security incidents and current patches.

The second case study involves an unauthorized access into over 200 systems of a First Fidelity Bank [7]. This intranet had a software server that was trusted by all of the bank's systems and was connected to the Internet. A hacker was able to gain unauthorized access to that software server, and therefore had access to every computer in the network. Surprisingly this software server had very little protection. Apparently its mission deemed relatively unimportant, the software server was not considered a target by bank personal concerned with protecting critical financial data. Furthermore, the bank had no incident response policies. When a technician accidentally found the hacker logged in (masquerading as a legitimate user), he did not know what to do. For several days those in IT watched the hacker as he or she roamed around the system looking at and possibly changing data at will. An external audit after the incident found that the hacker had used a hacking tool called esniff that is easily accessible on the Internet. Furthermore there were so many basic security flaws in the system that it was impossible to determine which had been used by the hacker.

9. The Future

It is true that most of the cited reports were compiled more than a year ago and it can be argued that many of those problems have been fixed by now. It should also be pointed out that few if any of those systems had critical or private data stored on them. They were merely the hosts for a web site. One must remember, however, that the nature of the environment may change in the future. It is entirely possible that soon keeping a web site secure could mean millions of dollars a month for a company. Furthermore, there will always be 'new' security flaws. The process of keeping up with patches never ends at a particular point.

What can be done? The flow of this information will never be stopped. For one it would never be deemed legal. For another reason it is probably not since not all hackers perpetrate crimes. Third it would not be practical. The flow of information is critical because network administrators must understand these issues. It is particularly critical to have advisories on new security flaws like those posted by CERT. Without this flow of information it is likely that the vast majority of systems would be vulnerable.

What must happen is that network administrators pay more attention. As the case studies illustrated, most attacks can be prevented by utilizing careful measures. Security advisories and associated patches or recommendations must be implemented immediately. Administrators should examine these case studies of past successful attacks since it is unlikely that a novice hacker is going to think of anything revolutionary, safeguarding against the traditional methods of unauthorized access will defeat the vast majority of attacks. Software designers and administrators must keep abreast of new automated hacking tools. New and current systems must be tested against these tools, since it is likely they will come under attack from those tools at some point.

10. Bibliography

[1] http://www.2600.com The Hacker Quarterly Web Site

[2] Blatchford, Clive. "Hacking: Myth or Menace, Part I." Computer Fraud & Security. February 1998. 16-18.

[3] Farmer, Dan. "Shall we dust Moscow?" December, 1996.

[4] "A Hack on the Department of Justice web site." I/S Analyzer. Nov. 1997. 7-11.

[5] "Internet Traffic Doubling Every 100 Days." Lexington Herald Leader. April 16, 1998.

[6] Littman, Jonathan. The Fugitive Game. Little, Brown, and Company: Boston. 1996.

[7] McCarthy, Linda. "Visitors in the Night: A Story from the Trenches."Computer Security Journal. Vol.XIII. No. 2. 1997. 1 - 11.

[8] The PC Hacking FAQ. Version 2.0. July 1996.

[9] Phrack, electronic magazine. First Quarter Issue. 1998.

[10] http://www.fish.com/satan/ SATAN Application Web Site

[11] Sterling, Bruce. The Hacker Crackdown. Bantam: New York. 1992.

[12] Tsutomu, Shimomura. Takedown. Hyperion: New York. 1996.

Firewalls

Steven Geist

1. Introduction

In recent years the Internet has grown by leaps and bounds. Organizations are beginning to realize the benefits of being connected to the Internet. However, it has also become easier for unskilled people to hack into networks disrupting service, corrupting files, or even stealing confidential information. Some organizations are beginning to wonder if connecting to the Internet is worth the risk of opening themselves up to cyber attacks. The implementation of a firewall is a good way of reducing the risks while still gaining the benefits of a connection to the Internet. Note that while firewalls are usually used to protect a network from the Internet, they can be used to protect a network from any other network. This section gives an overview of firewall technology and examines their overall effectiveness.

2. What is a firewall?

There are many varying definitions for firewalls, some better than others. The definition that best characterizes firewalls is given in Lodin and Schuba [12].

> A firewall is a set of mechanisms that collectively *enforce a security policy* on communication traffic entering or leaving a guarded network domain.

Most definitions of firewalls include a statement about the protection of a network from outside communication traffic. This is expected as this protection is indeed the essence of a firewall; however, the “security policy” part of the definition is an important addition. Without a security policy the implementers of a firewall have no benchmark for measuring whether the firewall is doing its job. Those without a policy likely are not aware of which types of attacks are stopped by the firewall and which attacks the system is still vulnerable to. Plugging such holes in a network protected by a firewall is essential to using one effectively [11].

3. Why implement a firewall?

As was stated in the introduction, a firewall is implemented when an organization wants to protect their network from another connected network. A firewall presents a single point of entry and exit between the two networks. In this way the security of the system can be monitored by monitoring the firewall. If a firewall is not implemented then security must be measured instead on a machine-by-machine basis, a more difficult task [14]. An organization may decide not to connect to the Internet because it poses too much of a risk to their network. If this is the case then employees may decide to gain access themselves by connecting through a modem [15]. Now there is a security problem that is not being monitored thus creating an even greater risk. In the case of the Internet, organizations can connect to the Internet while providing a measure of security through the use of a firewall. There are also cases where a company wants to separate one part of their network from the rest of the network. In this case the firewall is called an internal firewall which will be discussed later in this report.

4. Design decisions

Design decisions in firewalls become part of the basic policy of the firewall. There are two main design decisions for firewalls [15].

> That which is not expressly permitted is prohibited.
> That which is not expressly prohibited is permitted.

These two choices represent the trade-off between security and ease of use [15]. The first option provides a greater degree of security, but there may be more problems connecting to the outside from behind the firewall and possibly slowed performance. The second option makes it easier to use the resources outside of the firewall, but it presents more potential security problems. The various firewall architectures discussed below can generally be configured to follow either of these options.

5. Components of firewalls

Firewalls are composed of varying components. The three most commonly seen components are screening routers, proxy servers, and bastion hosts.

5.1 Screening routers

Screening routers are routers that can forward or reject packets on an individual basis [6]. This process is called packet filtering. Packets can be filtered based on port number, destination IP address, or source IP address [6]. This way the administrator can control the traffic moving in and out of the network. The screening router can be set up to block incoming packets from untrusted IP addresses. A screening router can also be set up to block certain services that use a particular port number. For instance, it can reject all packets coming in on TCP port 23 effectively disallowing incoming telnet requests [6].

5.2 Proxy servers

Proxy servers are mechanisms that act as connections between clients on the internal network and services outside of the firewall [15]. Proxy servers are sometimes called application level gateways [6]. When using proxy servers all users on the network must have special versions of the applications that work outside of the internal network [6]. These applications send their requests to the proxy server that then forwards the requests to the outside network and returns any replies to the requester. The benefit of a proxy server is that it has fine-grained control over exactly how applications are interacting with the world outside of the local network [6]. A proxy can be set up to allow users to import files but not to export them or it could allow imported files only from specified hosts [6]. The problem with proxy servers is that an organization must obtain a new proxy version of each application to be used on the network. SOCKS is a tool that simplifies this problem by helping users construct proxy versions of applications from existing non-proxy versions [11].

5.3 Bastion hosts

A bastion host is simply a computer in a firewall that is highly secure [15]. These computers are the cornerstones of many firewall architectures. Generally all communication between the network and computers outside of the firewall has to go through a bastion host. A bastion host must be very secure as access to it usually means access to the entire network behind the firewall [15]. Bastion hosts can have extensive logging of information that can help trace and determine the types of attacks that are occurring [15].

6. Firewall architectures

There are many different firewall architectures that provide varying levels of security. The following architectures will be examined: screening router, dual-homed host, screened host, and screened subnet.

6.1 Screening router

In general screening routers alone are not considered firewall architectures; however, it is helpful to look at the problems inherent in using only a screening router to protect a network. In this architecture all traffic between the internal network and the network outside of the firewall must go through a screening router. Policies can be set up to allow and disallow various types of traffic as was discussed above. One of the weaknesses here is that screening routers cannot control a service at a level lower than "permit" or "deny" like a proxy server can [6]. Another problem is that if a break-in does occur it is difficult to trace and possibly even to discover [15]. A screening router can be defeated by IP spoofing and Trojan horses [11,15]. The screening router can, however, detect packets that are attempting to spoof IP addresses of machines on the internal network. A hacker who finds a way past the screening router has access to the entire network [15].

6.2 Dual-homed host

A dual-homed host architecture is characterized by a bastion host that sits on both the internal network and the outside network [6]. All traffic between the two networks must pass through this host. Access to the outside from the internal network can be provided in one of two ways: allow logins on the bastion host or use proxy servers [6]. Allowing logins on the bastion host presents problems of its own. All internal users must use passwords that are difficult to crack. History has shown that in general this is not a good thing to count on [10]. The second option is the use of proxy servers which causes the extra burden of obtaining or creating proxy versions of all applications that connect to the other side of the firewall [6]. However, a dual-homed host with proxy servers is likely more secure than one allowing user logins. Again, if an attacker is able to break through the bastion host the entire internal network is exposed [15].

6.3 Screened host

A screened host architecture is characterized by a bastion host on the internal network that communicates with the external network through a screening router [6]. All traffic between the two networks must go through both the screening router and the bastion host. A somewhat weaker configuration of this architecture allows some traffic directly between the screening router and the internal network [6]. Using the stronger configuration means that an attacker must get by both the screening router and the bastion host before gaining access to the internal network [15].

6.4 Screened subnet

A screened subnet architecture consists of a perimeter network with a bastion host with screening routers leading to both the external and internal networks [6]. All traffic must pass through the bastion host and both screening routers thus providing a higher level of security. Other computers that require a lesser level of security or a higher level of access to the external network may sit on the perimeter network [6]. In fact, there can be more than one level of perimeter network providing numerous levels of security [6]. If an attacker breaks through the first screening router and the bastion host he still only has access to the perimeter network [15]. Also, unlike other architectures he cannot snoop on the communication lines of the internal network [6]. The attacker would have to break through the bastion host and both routers to gain access to the internal network [15].

The firewall architectures discussed above are not the only architectures, but they are the main ones. There are many variations and combinations of these architectures that have their own levels of security. As Ranum notes, an obscure firewall architecture can actually deter (or at least slow down) an attacker [15].

7. Internal firewalls

There are times when an organization wants to protect one part of its network from the other part. Perhaps a company has a number of computers with very sensitive information and does not want the rest of the network to have access to it [6]. Sometimes an organization has a lab used for training that needs better access to the external network and requires less security [6]. A firewall used for such purposes is referred to as an internal firewall [6]. Another case in which an internal firewall

is useful is when there is a lab researching network applications [6]. The firewall in this case is to keep any "problematic" software from escaping the lab and wrecking havoc on the local network.

8. Effectiveness of firewalls

There are a number of arguments as to why firewalls in general are not as effective as one may think. One of the problems is that they may give a false sense of security [11]. Once a firewall is in place some people begin to think that security is no longer an issue. Such ignorance is a break-in waiting to happen. Another problem with the effectiveness of firewalls relates to the trade-off between transparency and security. The more transparent the firewall is to the internal network the less secure your network becomes. If an organization wants to provide full access to all aspects of the Internet it is likely that the firewall will not provide the level of security it is interested in attaining. As was previously mentioned certain firewall components can be somewhat easily fooled. Despite all of the firewall mechanisms discussed above they are not enough to keep a network one hundred percent secure. As Blakeley puts it [2]:

> All the firewall systems in the world won't prevent the damage done by a disgruntled employee or a truly dedicated and knowledgeable system cracker.

At this point one might be wondering, "Why even bother to implement a firewall to begin with?" Firewalls must be seen for the limited protection they can provide. As more and better tools are becoming widely available to the non-expert hacker it becomes more important to protect your network from them. A firewall can help to protect the internal network from these types of attacks. A firewall can also discourage an expert hacker. Some hackers attack networks simply because they can, and they do not care who they attack. If a system has a firewall perhaps it will be enough to discourage hackers from attacking your network and encourage them to move on to an easier target. In order to make a firewall more effective the holes left in the software must be plugged [11]. New bugs are being found in applications and operating system commands every day. If network administrators have been notified of such a bug in the short term they can alter the firewall policy (through proxies and screening routers) to restrict access to the affected software. When a patch has been made available it can be applied and the firewall policy can be changed back to the original configuration.

9. Firewall products

A number of studies have been done on the effectiveness of commercial firewall products [4, 7, 16]. With the growing need for firewalls there has been an increase in the number of firewall products on the market and some of them are not providing the level of security expected [16]. According to Seachrist, ninety percent or more of firewall failures are due to misconfiguration [16]. This was not a factor in the testing as in all cases the vendors themselves were allowed to configure their own firewall. The following are a few sample firewalls with various statistics [7, 16].

9.1 Altavista Firewall97 version 3.0

This firewall uses packet filtering and proxy servers as the main components of its architecture [7]. It will run on both UNIX and Windows NT based operating systems [7]. It will log information based on service, time, source, and destination of packets [16]. It supports a wide variety of encryption and authentication schemes [7]. Its costs: $3,995 for 50 nodes, $7,995 for 200 nodes, and $14,995 for unlimited nodes [7]. The Altavista Firewall97 was one of the top performers in handling high traffic situations [16].

9.2 Gauntlet Internet Firewall version 3.2

The Gauntlet firewall is produced by Trusted Information Systems and also relies on a combination of packet filtering and proxy servers [7]. It is also supported on both UNIX and Windows NT based operating systems [16]. It provides logging based on service, time, source, and destination [7]. The price for Gauntlet is constant across sizes of 25, 100, and 1000 nodes at $11,500 for the software only and $16,500 for both software and hardware [7]. Gauntlet did not perform very well under high traffic situations [16].

9.3 Sunscreen EFS version 1

This firewall is produced by Sun Microsystems Incorporated and relies on packet filtering and stateful inspection (essentially a smarter version of packet filtering) to provide security [7]. It will only run on Solaris platforms [7]. It supports service, source, and destination based logging [7].

Costs for the Sunscreen firewall are as follows: $1,495 for 25 nodes, $4,995 for 100 nodes, and $14,995 for 1000 nodes [7]. Sun was one of the top performers in handling high traffic situations [7].

These are just a few examples of the many firewall products commercially available today. In one of the tests 19 vendors participated while another 22 declined which gives some idea as to the number of products available [7]. There is quite a variety of products out there and less than half of the vendors were willing to allow their products to be vigorously tested. This implies that possibly over half of these companies do not have confidence in their own products. The testing in [7] was performed with an application called Safesuite produced by Internet Security Systems Incorporated. It was configured to run 100 different attempts to penetrate the firewall [7]. These attempts were on well-known weaknesses of many firewalls and some standard denial-of-service attacks. A number of firewalls failed to defend from all of the low-risk attacks and some even failed in the category of medium-risk [7]. None of the firewalls succumbed to a high-risk attack [7]. Of the three products mentioned above the worst failure was in the low-risk category [7]. Some of the medium risk failures in other firewalls were with SYN flooding (a denial-of-service attack) and TCP sequence prediction [7]. Failure to prevent these types of attacks can have serious consequences for the internal network.

What should you be looking for in a firewall? One of the more important aspects of a firewall is its ease of configuration and management. If ninety percent of all failures of firewalls are due to misconfiguration the first thing to do is make sure that yours is configured correctly. Some vendors include personal help in installation in the price of the firewall [7]. A firewall should have a good logging facility. It is important that your firewall react correctly when a log or disk becomes full [7]. Most firewalls that were tested provide remote notification via e-mail or page when a serious attack is underway [16]. Most firewalls come with a secure version of the operating system they usually run under [16]. As has been mentioned before, plugging the holes in the operating system is an important part of having a firewall. The three firewalls that were mentioned above performed well in the testing and are some of the stronger products available.

10. Conclusion

Firewalls are not a cure-all when it comes to network security, but they are a good step in the right direction. There are a number of different architectures and many different commercially available

firewall products. Any user of a firewall needs to strike a balance between the transparency of the firewall and the security of the system. If data is too sensitive to leak to an outside network then the computers it resides on should not be connected in any way to the outside world. If an organization is looking for a reasonable amount of security with the benefits of being connected to another network then a firewall is a good way to obtain those goals.

11. Bibliography

[1] Anderson, J. P. et al., "Firewalls: An expert roundtable," *IEEE Software*, September/ October 1997, pp. 60-66.

[2] Blakeley, M., "Keeping the Visigoths Out," *PCWeek*, January 23, 1995, pp. N1-N2.

[3] Bryan, J., "Build a firewall," *Byte*, April 1995, pp. 91-96.

[4] "Can Firewalls Take the Heat?," http://www.data.com/Lab_Tests/Firewalls.html (November 21, 1995).

[5] Cobb, S., "Internet firewalls," *Byte*, October 1995, pp. 179-180.

[6] "Firewall Design," http://www.sun.com:80/sunworldonline/swol-01-1996/swol-01-firewall.html (January 1996).

[7] "Firewalls: Don't Get Burned," http://www.data.com/lab_tests/firewalls97.html (March 21, 1997).

[8] "Firewalls in Many Flavors," http://www.sun.com:80/sunworldonline/swol-01-1996/swol-01-security.html (January 1996).

[9] "Internet Firewalls Frequently Asked Questions," ftp.tis.com/pub/firewalls/faq.current.

[10] "It's After Midnight, Do You Know Who Your Modem is Talking To?" http:// www.decus.org:80/decus/pubs/decus94/modem.html (1994).

[11] Kerr, D., "Barbarians at the Firewall," *Byte*, September 1996, pp. 80-82.

[12] Lodin, S. W. and Schuba, C. L., "Firewalls fend off invasions from the Net," *IEEE Spectrum*, February 1998, pp. 26-34.

[13] Loshin, P., "Defending from the unthinkable," *Byte*, December 1997, pp. 67-73.

[14] Oppliger, R., "Internet security: Firewalls and beyond," *Communications of the ACM*, May 1997, pp. 92-102.

[15] Ranum, M., "Thinking About Firewalls," Trusted Information Systems, ftp.tis.com/pub/ firewalls/firewalls.ps.

[16] Seachrist, D. and Holzbaur, H., "Firewall software for NT and Unix," *Byte*, June 1997, pp. 130-134.

[17] "Watch Your Back Door," http://www.sun.com:80/sunworldonline/swol-12-1995/swol-12-security.html (December 1995).

State of the Art in Computer Virus Prevention

Luís G. Nakano

1. Introduction

Computer viruses are programs designed to replicate and spread, generally with the victim being oblivious to its existence. Although there have been confirmed occurrences in Amiga, Atari, Macintosh and some Unix systems [14], IBM-PCs are the most usually infected machines due to their prevalence in the market place. With the exception of macro viruses, most viruses are operating-system specific. Among the operating-systems-specific viruses, DOS is the most prevalent operating system of choice. Although most of the techniques can also be applied to viruses in other environments, in this paper, we will discuss PCs viruses and prevention mechanisms.

Viruses sometimes can carry destructive payloads that can be activated by predefined events chosen by their programmers. This payload can cause arbitrary damage to data or almost arbitrary behavior degradation to a system. For instance, they can be used to modify databases by adding, removing or modifying records. Among the typical behavior degradation that the payload can cause are screen animations and sounds, general processing slowdown, and unexpected writes to the disk.

Even when a virus does not carry destructive payloads, bugs in its code or invalid assumptions about the environment have been reported to cause service disruption and data modification [12]. In either case, viruses cause loss of productivity and confidence in the system. This loss of productivity comes partly from machine slowdown but mainly from the time it takes to remove the virus.

Viruses are costly to remove since there are several tasks involved in the process. First, it is necessary to determine which virus is in the system. Sometimes, bugs are reported as viruses with the converse also being true. As software becomes more complex, it is increasingly difficult to establish beyond a reasonable doubt that the unexpected behavior observed is a virus. Anti-virus products are used in this process with varying degrees of success depending on the virus and product

used. For an analysis of the current anti-virus products, see the Virus Bulletin Comparative Reviews [15].

The second step after finding out that a system is infected by a virus is obtaining information about it such as its behavior, infected areas, and removal procedures. This is not as easy as it would seem because diverse anti-virus products report different names for the same viruses. This happens because most viruses do not have distinctive characteristics that can be used to name them and also due to the dynamic nature of the anti-virus production tool [9]. In some cases, this information gathering step can be ignored if the anti-virus tool can also remove the virus.

The third step is the actual virus removal. This includes scanning all media that could be infected and applying the appropriate fixes. The cost associated with the full process is estimated to be $8,366 per incident with the highest reported cost being $110,000 for a single computer virus incident according to the NCSA Virus Prevalence Survey [8]. The same report indicates that there is a probability of 28 virus encounters per 1000 machines in a year. The average infection includes 107 PCs and 1.8 servers, and takes about 46.6 hours and 10 person-days to recover from [8].

As Figure 1 shows, viruses are growing at least exponentially in numbers, with typical estimates of 400 new viruses being generated every month. The actual number of viruses is not known, mostly because there is not an established way of counting viruses and partially because some anti-virus manufacturers count viruses that are experimental and are not capable of infecting other machines. However, most anti-virus producers and virus research centers agree that the growth in the number of viruses is exponential.

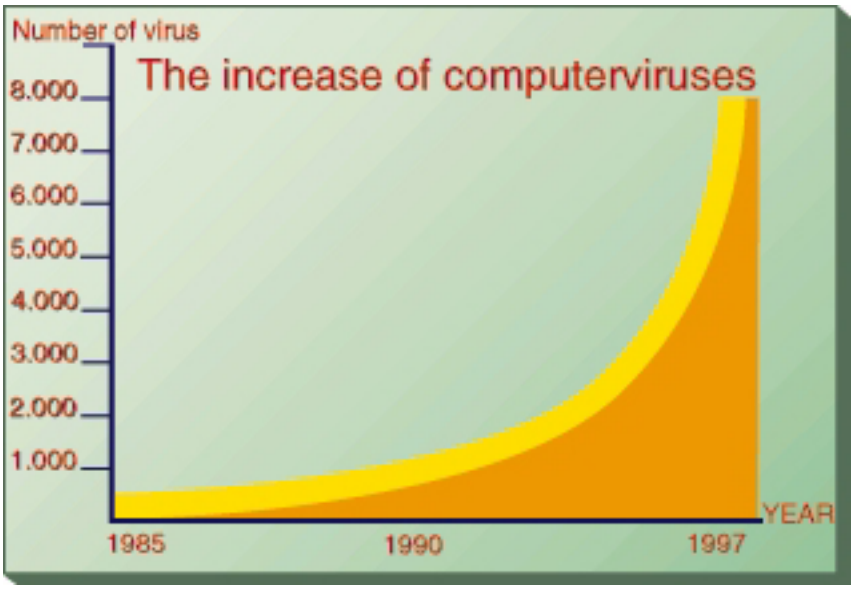


Figure 1: Computer virus growth [5].

This growth is partially due to the increased number of machines and the wide distribution of virus creation tools (some of which are reported to have easy instructions and user-friendly graphical user interfaces), but mostly due to the widespread communication facilities of the Internet and the large proportion of machines with no anti-virus protection tools.

Macro viruses have been the most prevalent viruses since the Fall of 1995. These viruses can propagate easily through e-mail attachments since they are carried by infected documents. Most of these viruses are Word Macros, with Word.concept being the most prevalent. An infected document will cause the macro to execute when the document is opened by Word. The infection is capable of spreading fast in part because many users have their machines set so that Word documents embedded in e-mail messages are opened automatically, thus contaminating the machine right away if the document contains macro viruses.

Given the widespread infection, viruses are clearly problematic and not likely to go away soon. However, there are steps that can be taken to reduce its occurrence. NCSA reports that “it has been shown that if as few as 30% of the world’s PCs used a relatively current, full-time anti-virus protection method, that the effect of ‘herd immunity’ would nearly eliminate the world-wide computer virus problem.” [8]

To better understand the virus infection process, section 2 will discuss the life cycle of viruses. Section 3 will detail the diverse types of viruses known today, as well as techniques to prevent infections for each type. Finally, section 4 will provide conclusions about the current state of the art in virus protection techniques.

2. Life Cycle of Viruses

2.1 Creation

Computer viruses used to be the work of technically trained programmers without employment [12]. Now, there are several authoring kits that enable even individuals with no programming knowledge to generate viruses. Usually, viruses are created by misguided individuals who wish to cause widespread, random damage to computers. Unfortunately, there are no international laws that can be used to prosecute virus writers, and even if there were, establishing the author of a virus is in most cases non-trivial. Therefore, it is doubtful that viruses are going to be stopped in this stage.

2.2 Gestation

After creating a virus, the writer copies it and makes sure that it will spread. Common techniques used are to infect a popular program on a BBS or to distribute copies through offices, schools and other large organizations. However, with the introduction of macro viruses, even documents attached to e-mail messages can transmit viruses if the documents are allowed to open automatically. It is impractical to prevent gestation of viruses because information transfer over networks, such as e-mail, and general file repositories such as BBSs are unlikely to disappear without similar substitutes. The existence of a large number of BBSs and the fact that misguided individuals write viruses using techniques that prevent tools from locating the virus make it unlikely that gestation can be prevented.

2.3 Replication

Viruses are designed to replicate and propagate. The first viruses were somewhat naive in that they would only propagate in very specific hardware (for instance, only on 360KB floppies). With time, more complex viruses capable of infecting diverse media started to appear. The worst types of

viruses have a high replication factor that enables them to spread quickly. To prevent the same machine from being infected several times and thereby provide clear signals that something is not normal, most viruses have sequences of instructions that verify if there is an active instance of themselves on the machine. To achieve this identification, some viruses use signatures that can also be used by anti-virus to identify them.

For some viruses, such as pure boot ones, replication can be prevented by forcing the machine to boot from a hard disk and not from a floppy unit that might be contaminated. Macro viruses can also be made ineffective by disabling automatic macro execution when a document is opened, thus eliminating a potential security problem. While these techniques cannot prevent replication in all cases, at least they would reduce the degree of infection.

2.4 Activation

Essentially, any event or condition can be used by a virus to activate its payload. Some viruses have no damage routines, only distractive characteristics. However, there are reported cases of viruses that were intended to be inoffensive, but had a bug that caused real damage to the system. Since it is unknown in advance what are the events that can trigger an unknown virus, there is no way to prevent the activation of all viruses.

2.5 Discovery

This phase might come before activation, but it usually comes afterwards. It usually requires some technically trained person to detect and isolate the virus and then sends it to the International Computer Security Association in Washington, D.C., to be documented and distributed to anti-virus developers. Some techniques that have been introduced by anti-virus manufactures can identify suspicious behavior and help reduce the spread of the infection in the early stages [7]. Although there is no doubt that virus writers will try to circumvent them, the application of such techniques can help anticipate discovery time and thus reduce the possibility of infection spreading farther. The sooner an infection is detected, the less it can spread.

2.6 Assimilation

After receiving new virus information for ICSA, anti-virus companies start to work to update their products so that they can detect and safely remove the new viruses. This can take anywhere from a few hours to several months, depending on the complexity of the virus. Assimilation is very much a damage control process where the only gain is to reduce the spread of a virus. However, it is unclear how to make assimilation faster, with improvements only in the deployment of new viruses definitions for certain anti-virus tools.

2.7 Eradication

Although no viruses have been known to disappear completely, this could happen if enough users install up-to-date anti-virus software. A list of the viruses in the wild, i.e., viruses that have been known to infect at least two different sites in the last two years, is kept on the Web [16]. Some viruses have ceased to be a major threat and are no longer in the wild list, but they can become active again if newer anti-virus packages no longer search for them. As mentioned before, if 30% of the world's machines were using a relatively recent anti-virus tool, computer viruses would not be a problem. However, anti-virus tools are relatively expensive and considered by most as an unnecessary addition to their systems. This probably happens because most of the campaign towards an increased use of anti-virus comes from anti-virus companies who are understandably unwilling to lose profit by giving away anti-virus tools.

3. Virus Types and Prevention Techniques

There are three main types of viruses: file infectors, boot viruses, and macro viruses. These viruses can be distinguished by the infection process that they use to propagate. Besides these three main classes, there are three more orthogonal dimensions to classify viruses: whether they use stealth techniques, are multipartite, or use polymorphic techniques.

3.1 File Infectors

File infectors [6] attach themselves to ordinary program files, usually infecting arbitrary COM and/or EXE programs although some can infect any program for which execution or interpretation is requested, such as SYS, OVL, OBJ, PRG, MNU and BAT files.

Infection characteristics divide this class of viruses into two:

- Direct Action: A virus that infects one or more files when it is executed. Therefore, they only infect certain files, such as COMMAND.COM, and are somewhat easy to remove, once the particular virus is known.
- Memory Resident: A virus that uses terminate and stay resident techniques to stay in the computer memory, infecting any executable file that is executed after it is loaded in memory.

It has been shown [1] that it is not possible to decide if a program is a virus or not in all cases. Consequently, the best approach to prevent file infectors is to perform signature-based detection to search for known viruses in all executable files to be installed or run in the system. Signatures are sequences of instructions that can be used to detect the presence of a specific virus. The best way to implement this policy is to have full-time anti-virus tools running in the background of the system scanning every executable that is copied into memory for execution. Checksumming files to detect changes can also be used, but their effectiveness is limited since some viruses will infect only floppies (which are not usually subject to a checksum) and only infect hard disks when they are copied from the floppies. Of course, it is always recommended that only software from reliable sources be installed in the system, but even commercial distributions of software have been known to distribute viruses by mistake.

3.2 Boot Viruses

Boot viruses infect executable code found in certain system areas on a disk. There are ordinary boot-sector viruses which only infect the DOS boot sector, master boot record (MBR) viruses that infect the master boot record on fixed disks and the DOS boot sector on diskettes, and partition table viruses that infect the partition table of hard disks and either the boot sector or the master boot record. None of these types can infect a system if they are not loaded during boot time unless there is user intent (for instance, by making a disk image of a contaminated boot sector, sending it over

to someone, and having him/her unpack it to the boot sector, the virus can infect the system without a boot). For a long while, this class was responsible for most infections.

To prevent this type of viruses, most computers allow users to set boot sequences so that the machine is going to boot only from the hard disk and not from any floppies that the user has left inside the boot drive. Another solution is to have the machine boot only from a read-only media that cannot be infected. If used properly, the safe boot techniques described above can prevent most boot viruses from infecting a system. Other techniques include signature-based detection of known viruses on the boot sector, master boot record, and partition tables as well as write-protecting floppies whenever possible.

3.3 Macro Viruses

A macro virus is a piece of self-replicating code written in an application's macro language [3]. Macro viruses that work alone can only be written in languages that support auto-execute macros. Auto-execute macros respond to certain events and do not have to be initiated by an explicit user command. While executing, a macro can copy itself to other documents, remove files, substitute commands of the user interface, and cause all sorts of disruption to the system. In particular, if a macro has the same name as an internal command in Word, the macro will be executed instead of the command.

Macro viruses propagate when they are executed because either a document was opened or a user executed a substituted command. Although there are no viruses that do this, it is not unreasonable to consider file infectors or attack applets that can copy a macro to the normal.dot template of Word and thus contaminate all documents from there on.

Since macro viruses are application specific, not platform specific, some of them can execute in several different systems. In some cases, however, the same viruses might fail to execute properly on some of the platforms where the application runs.

Another feature in Word that makes the behavior of macro viruses hard to predict is the password protection for macros. This feature enables companies to sell macros and users to execute them without knowing their content.

With the advent of Microsoft Office 97, Word Basic is being replaced by Visual Basic. However, old macros written in Word Basic are automatically converted to Visual Basic and executed seamless in the new version, thus providing automatic portability for most macro viruses. Visual Basic macros can execute in the other programs that constitute Office 97, thus potentially providing virus infections to Access databases, Excel spreadsheets and PowerPoint presentations. The last one is frequently used by sales people to transport presentations with the use of local machines being common practice. If the presentation is infected, it will probably infect the machine and all future presenters might become unwilling infection carriers.

The potential for infection is enormous, and this is coherent with the fact that more than 50% of all viruses encounters in the last year are due to Word macro viruses with the different types of macro viruses exceeding 200 [4]. They do not need executable exchanges or boot-ups, only document transfer, and thus are ideally suited for transfer over the Internet and telecommuting.

To prevent macro viruses, most virus scanners have been updated to detect infected documents and many can disinfect them [3]. Microsoft also made macros for detecting macro viruses in Word and Excel, but they are not activated if the user opens a document by double clicking on it. Furthermore, some anti-virus companies have updated versions of firewall software that scans all e-mail attachments for viruses and rejects those that cannot be decoded or that contain unauthorized macros. Another approach used is to keep a database of checksum values for the macros that are allowed to be executed in the system and consider any macro that fails this test suspicious until the operator has decided if they are viruses or not. This last technique is the only one that can prevent new macro viruses from infecting the system and is particularly important when the knowledge necessary for producing a virus is so low that even unsophisticated users can generate their own viruses. Also, macro viruses are probably going to become more common, since virus-generating programs with complete manuals and graphical user interfaces are being distributed over the Internet.

3.4 Stealth Techniques

Stealth techniques can be used by viruses to avoid detection using active procedures. These techniques can be added to any of the virus classes described previously, producing a stronger category of viruses. For instance, some viruses make infected files appear normal when they are

accessed by DIR or anti-virus. Stealth boot viruses typically keep a copy of the original boot sector and master boot record elsewhere and provide this copy to any program that asks to read them. Stealth techniques require viruses to stay in memory to be able to hide themselves.

To prevent against stealth viruses, use the techniques for the main virus type. For instance, a stealth boot virus cannot infect a machine that never boots from any read-write media. Stealth viruses must be detected while in the memory of the system then deactivated before disk-based components can be corrected.

3.5 Multipartite viruses

Multipartite viruses are combinations of two or more of the pure types described previously: file infectors, boot viruses and macro viruses. However, they are typically boot viruses with file infectors associated. In some cases, one of the parts is used to transfer the other. Thus, a file infector that contaminates the boot sector with a boot virus is a multipartite virus.

To prevent against multipartite viruses, the techniques used for the viruses of each of its parts should be employed. Extra care must be taken to ensure that none of the parts is in memory.

3.6 Polymorphic Techniques

As anti-virus tools based on signatures became more powerful, virus writers devised a way to prevent their creations from being detected reliably using polymorphic techniques. These techniques are used by many modern viruses, and, at the very least, they can make virus detection an unreliable process.

The first polymorphic viruses used variable encryption of the virus and sometimes NoOps to change their signatures. However, the description routine was constant in all the different infections. This fact was exploited by the anti-virus tools so virus writers introduced an extra component: a mutation engine that generates new encrypting/decrypting routines. Some of these mutation engines are capable of generating billions of combinations, and this makes it impractical to search for all possible decrypting routines as signatures. Using this technique, signatures became very short, and the probability of misdetection of a virus increased. Also, there is always some chance that an anti-

virus will not be able to identify all mutations that are infecting a machine, and thus report the infection as removed while it continues on the background. Some mutation engines have been distributed to virus authors, increasing the probability of new viruses using this technique being created.

To prevent infections caused by polymorphic viruses, programs can be written to detect sequences of code used by mutation engines in the decrypting code. However, every time a new polymorphic virus appears, a new program has to be generated, and this is much more costly than updating a parameter file that contains signatures of viruses.

A second approach is to execute each program in a virtual machine and scan the memory where the program was loaded from time to time until it can be assumed that no viruses are present. This method has an inherent problem of deciding when it has executed enough of the program in the simulated virtual machine. Also, executing all programs in this virtual machine slows down processing.

Another approach is to create heuristics to decide when a program should be executed longer in the virtual machine environment. For instance, if a program computes values and discards the resulting register before using it, it might be an infected program trying to pass as a normal program. The problem with this approach is that every time a new profile is added to the system, exhaustive regression testing is necessary to guarantee that older viruses are still going to be detected.

The two last approaches are being combined in the Striker system [7], but with a difference: each virus is profiled, and the profile is compared with the behavior of the program when executing in the virtual machine. If the profile could match a virus, then only the viruses that have that characteristics are scanned. This provides speed, for files that do not have executing profiles close to any virus are loaded in normal memory and are allowed to execute at full speed. Also, this approach eliminates the need for exhaustive regression testing since the rules to detect old viruses are not modified. Although this is the most complete approach so far, a virus writer could design a virus that only deciphers the body half of the time and thus prevents detection. Also, a virus could be made to infect a machine only if some external event, such as a predefined keystroke, happens in the system. Since the virtual machine would not have this event, the virus would not be deciphered and thus would escape detection.

4. Conclusion

If past is any indication of the future, it is not likely that computer viruses are going to cease being a problem in the near future. The main reason behind this is that virus writers are always willing to spend more time creating viruses that can defeat the existing detection techniques.

There has also been a great deal of social engineering among virus creators [10,11], and this has made it possible for virus writers to share their most recent techniques. However, this is not the case for anti-virus tool developers as their products use proprietary techniques.

Another important factor to consider is the requirement for backward compatibility that plagues the PC industry: old programs must be able to execute in new operating systems. For this reason, some old viruses are still able to infect machines running under Windows 95, with varying degrees of success, even though they were designed for DOS. Furthermore, one Windows 95 specific virus has been reported [17], and it is likely that its distributed source code will pave the way for other virus writers.

It is very difficult to make predictions about future trends in viruses, but it seems safe to assume that polymorphic viruses will continue to spread for a long time. One of the main reasons is that polymorphic viruses cannot be detected reliably without slowing down the machine, making infection by polymorphic viruses very unlikely to disappear.

Macro viruses are a growing problem for which the solution seems to be simply disabling autoexecuting macros present in documents. Indeed, some tools do exactly that [13]. In summary, protection against viruses is likely to be a problem requiring resources best used in other tasks in the foreseeable future.

5. Bibliography

[1] F. B. Cohen. Computer Viruses: Theory and Experiments. In: IFIP-TC11 Conference. Toronto, 1984.

[2] F. B. Cohen. Protection and security on the information superhighway. New York, Wiley, 1995

[3] D. L. Crawford; W. J. Orvis. CIAC Virus Information Update. Technical Report CIAC-2301. Also available online as: URL: http://ciac.llnl.gov/ciac/documents/CIAC-2301_Virus_Information_Update_9-97.pdf, Sept. 28, 1997. (visited on March 28, 1998).

[4] Datafellows. Press Release: Number of Macro Viruses Now Over 2000. URL: http://www.datafellows.com/news/pr/f-prot/macr2000.htm (visited on March 28, 1998).

[5] Dr. Solomon. Virus Statistics. URL: http://www.drsolomon.com/vircen/stats.cfm (visited March 27, 1998).

[6] N. FitzGerald. (Maintainer, Virus-L/comp.virus FAQ sheet). Frequently Asked Questions on Virus-L/comp.virus. URL: http://www.bocklabs.wisc.edu/~janda/virl_faq.html. October 9, 1995 (visited on March 29, 1998).

[7] C. Nachenberg. Understanding and managing polymorphic viruses. The Symantec Enterprise Papers, Vol. XXX, 1996.

[8] NCSA. NCSA 1997 Computer Virus Prevalence Survey. Carlisle, PA, NCSA, 1997.

[9] Project VGrep. The Project VGrep Home Page. URL: http://www.virusbtn.com/VGrep/ (visited on March 29, 1998).

[10] R. M. Slade. History of Computer Viruses. URL: http://www.bocklabs.wisc.edu/~janda/sladehis.html (visited on March 27, 1998).

[11] Solomon, A. A Brief History of PC Viruses. URL: http://www.bocklabs.wisc.edu/~janda/solomhis.html (visited on March 27, 1998).

[12] Symantec. Computer Viruses: An Executive Brief. URL: http://www.symantec.com/avcenter/reference/corpst.html (visited on April 12, 1998).

[13] TrendMicro. Taking Virus Protection into the 21st.Century (White paper). Trend Micro, Cupertino, CA, November 1997.

[14] US Department of Energy Computer Incident Advisory Capability. CIAC Virus Database September 1997. URL: http://ciac.llnl.gov/ciac/CIACVirusDatabase.html (visited on March 27, 1998).

[15] Virus Bulletin Ltd. Virus Bulletin Comparative Review. URL: http://www.virusbtn.com/Comparatives/ (visited on April 19, 1998).

[16] J. Wells. Wild List March 1998. http://www.virusbtn.com/WildLists/199803.html. March 1998. Visited on March 27, 1998).

[17] I. Whalley. Viruses in Chicago: The Threat to Windows 95. In: Proceedings of the National Computer Security Association IVPC’96, Washington, DC, April 1-2, 1996.

Smart Cards: Security in the New Transaction Cards

A. C. Chapin

1. Introduction - What Are Smart Cards?

There is an increasing demand for applications involving information transactions that can be carried out on the move. The use of smart cards can increase the security of such transactions in an off-line mode by restricting access to the information on the cards themselves and improving the security measures with which the card, the card holder, the terminal by which the card is connected to a network, and the transaction itself are checked for authorization.

Previous security techniques for transaction cards are proving unable to withstand criminal ingenuity, and on-line security measures are suffering for their reliance on communications networks and centralized operation. At the same time, the technologies involved in the creation of smart cards are becoming more affordable. This report discusses types of smart cards, issues in their design, and the security measures smart cards can provide.

A smart card is a way of carrying around information that controls access to itself, both to effect some application behavior, and for security. It is essentially a computer -- with limited data storage and processing power -- packaged in a form that is convenient to carry, typically inside a rigid plastic envelope the size of a credit card. A terminal of some kind with which the smart card can communicate is also usually needed.

In smart cards, the need for portable, secure data storage and access meets the decreased size and cost that hardware advancements have made possible. High memory density (3-8 kilobytes for AT&T's 1994 Smart Card [11]) allows vastly more information to be held on a smart card than on magnetic strip cards, while a microprocessor allows in-card control of how this memory is accessed, allowing offline verification as well as increased security for transactions.

Billions of dollars are lost yearly due to failure of network access security systems, but such systems are in increasing demand [11]. Since the mid-eighties, smart cards have seen increasing use in Europe, especially in France, to combat this fraud [13]. In addition to their obvious use as replacements for magnetic-strip ATM cards, smart cards have been used as prepaid cards for telecommunications, transportation, and utilities. They are also used for information tracking and journaling, for medical histories and prescription records, and for business records.

2. Background

Smart cards developed as an improvement upon previous transaction card technologies, taking advantage of computer hardware advances. It is important to understand the strengths and weaknesses of these previous card types, in order to see the contributions that smart cards can make. A brief outline of the early development history of smart cards is also given.

2.1 Other Card Technologies

Magnetic strip cards are commonplace through most of the world, especially when used for financial transactions, as in ATM or credit cards.

Magnetic strip cards are used for

- Identification and physical access: ID badges, door key cards.
- Information access: library cards, account cards.
- Financial transactions: ATM, credit/debit cards, prepaid telecommunication and transportation cards.

On a standard magnetic strip card is a piece of magnetic tape divided into three parallel tracks for different applications. Track 1 can hold an account number and a name, and has historically been used only by airlines. Track 2 can hold an account number, and is the track mainly used for financial applications. Track 3 is rewritten on each use and holds a PIN verification value that a card terminal can use to verify the entered PIN offline without going to the central computer. Even with all three tracks fully loaded, only 226 characters will fit on a magnetic strip [2].

Only reciprocal agreements between groups of card issuers have allowed the cards of one issuer to be usable in the terminals of another, and in most countries this type of interchangeability is rare or nonexistent [2].

Because magnetic strip cards lack any onboard processing ability, their security relies entirely on the terminal -- the machine used to read them. Equipment to read or change the strips is relatively easy and cheap to obtain, and although watermarking techniques and other security enhancements have been added to deter forgery, costs from magnetic card fraud continue to rise. Also, to combat bad debts on the part of authorized card holders, magnetic strip cards are often given a time-to-expiration of only three years [13], although the cards could last much longer; magnetic strip cards have no way of monitoring those transactions they are used in for level of risk.

However, magnetic strip cards have the advantage of being inexpensive (average $0.88 compared to $1.50 - $15 for smart cards [13]), and their use is well established worldwide.

Memory Cards consist of a high-density memory inside a plastic envelope. In the LaserCard (TM) invented by Jerome Drexler [2] memory is implemented with a laser technique like that used in compact discs, while infrared scanning and capacitive and inductive coupling are used in other cards.

Memory cards are primarily used for portable data storage and bulk storage, in applications such as medical records and machine tool control.

While memory cards are usually cheaper than smart cards, the absence of processing ability gives them no real advantage over magnetic strip cards in the area of security, and most of the information storage techniques used do not allow information to be erased and rewritten.

2.2 Development of Smart Cards

Smart cards were invented in the early 1970's by Roland Moreno in France, and simultaneously by Kunitaka Arimura in Japan. Moreno's cards are the basis for smart cards as we know them today, and his Innovatron company holds the international patent on 'a card with a self-protected integrated

memory,' although Arimura holds the earliest patent (his filing was limited to Japan) and was the first developer of contactless smart cards in 1978 [2].

Since the most widespread intended use for smart cards was always as a replacement for current types of financial transaction cards, smart card designers have been constrained by the need to make smart cards match the ISO standard for such cards, including physical dimensions and magnetic strip. This has required an emphasis on miniaturization before cost or performance.

The first three licensees of the Innovatron patents were Honeywell Bull, Flonic Schulmberger, and Philips. Both Honeywell Bull and Philips developed microprocessor-based smart cards; Flonic Schlumberger chose a hard-wired logic approach that allowed them to keep prices down and achieve the ISO standard for magnetic strip card compatibility (a thickness of 0.76 mm) from the outset, at the cost of multi-functionality. Through the 1980's, Schlumberger was the largest supplier of smart cards in the world [2].

The microprocessor cards both used non-volatile memory, but neither supported reprogrammability, so that the cards had to be replaced when their memories were exhausted (at this time EEPROM technology was still in development and very expensive). At their first introduction in 1979 these cards were still thicker than the ISO standard for magnetic strip cards, but at this time all three companies began to develop and market terminals for communicating with smart cards. In 1981, the microprocessor cards reached ISO standard thickness [2].

In 1982 and 1983 the first French public trials of smart cards were carried out, with cards issued to consumers to be used as retail payment cards. The trial met positive reaction from the retailers involved, but ambivalence on the part of the card holders, who felt there was not enough emphasis on convenience. In 1984, hybrid cards with both a magnetic strip and smart card features were release by both Visa and MasterCard in France [2].

Since then smart cards have been used increasingly in France for railway payment, mail-order, retail payment, and especially pay phones, and have been successfully introduced into many other European countries.

3. Design of Smart Cards

3.1 The Physical Card

To meet the ISO standard for size, smart card chips are generally smaller than 2mm square [12].

Smart cards are generally equipped with an 8-bit microprocessor (Motorola uses its 6805 microprocessor while Siemans uses the Intel 8051 microcontroller [9]). This processor need not be very powerful, because it need only be able to handle data transactions for one or a few applications and the requisite security for those applications. For the most part, the behavior of the smart card is to generate an Answer to Reset (the information returned by the card when power is applied), and then send and receive bytes through a specific serial protocol [8].

Smart card memories are typically on the same chip as the processor, and variously use ROM, EPROM, and/or EEPROM.

In order to communicate with the outside world and effect transactions, smart cards generally need terminals to communicate with, and contacts through which to interface to the terminal. The ISO layout for these contacts is in two rows of four; the fourth and fifth contacts are generally not used and often omitted; the seventh is the main data line [7].

The ISO standard for smart cards is number 7816, and has four parts. Part one concerns physical characteristics. Part two involves dimensions and locations of the contacts. Part three has to do with electronic signals and exchange protocols, and part 4 involves the minimal set of commands. This standard has passed its first two parts, but has recently failed a vote on part three [9].

3.2 Types of Smart Cards

There are several major design issues along which smart cards are classified. The most important of these are whether the card is to be microprocessor based or hardwired, whether the card will require physical contact with its terminal, whether the card can carry out transactions in the absence of a terminal, and whether the card is equipped with a magnetic strip.

Microprocessor or Hardwired

In order to provide functionality quickly and at low cost within the ISO specification size for magnetic strip cards, Flonic Schlumberger, one of the first manufacturers of smart cards, chose to give their cards hardwired logic, rather than the microprocessors being developed elsewhere. The advantages of this were so pronounced that hardwired cards were the most-produced type of smart card through the 1980's and even into the early 1990's [2]. However, hardwired cards lack multi-functionality, and cannot be programmed after manufacture.

More expensive microprocessor cards have become the current standard, now that miniaturization and falling prices have brought them to the same size and closer to the price range of magnetic strip cards.

Contact or Contactless

Contact cards must be in actual physical contact with the terminal in order to communicate with it. Contactless cards, also known as remote or close coupling cards can transfer data and get power from a terminal simply by being placed close to it. RF identification cards can transfer information between card and terminal from long distances using RF (radio frequency) techniques; these are particularly useful for toll applications.

Contactless strategies are more popular for general use. They allow the card to be completely sealed, with no contacts to let in contaminants, and they avoid the potential stress on the card of constantly being run through readers. Also, contactless cards can be any size and shape since they do not have to conform to a standard terminal aperture.

For transportation and other uses, consumers prefer to simply have the card do its work without their having to go to the trouble of running it through a terminal, or even of getting the card out of a bag or pocket; this also increases customer throughput [6]. It is therefore important that the transfer of data employed by contactless cards be secure, and that there be a clear and obvious way for card users to determine whether their transactions have gone through correctly without inconveniencing them. Contactless cards also tend to be more expensive.

Active or Passive

Active cards have on-board input and output, so that they can be used in the absence of a terminal; passive cards can only be accessed through a terminal. Active cards are most useful for keeping track of information such as medical history - prescriptions and a journal of pills taken could be kept on a card - or schedule. It would also be convenient to be able to check the current status of a financial account on the card itself, without plugging into a terminal.

Strip or No Strip

Some 'hybrid cards' have a magnetic strip conforming to standards for magnetic strip cards. This allows them to be used in terminals that do not accept smart cards, and is a popular approach to encouraging the adaptation from magnetic strip to smart cards. However, hybrid cards must be shaped to be able to pass through a card reader even if they are contactless.

3.3 Physical Survivability Design

Smart cards are computers that will be subjected to the same elements as traditional magnetic strip cards. For this reason, the smart card must be rugged enough to protect its chip, while still maintaining a size and shape that fits the standard.

The housing of the card must protect the chip from mechanical, chemical and electrical damage. The plastic packaging must protect the chip from all twisting, bending, and striking of the card. This may involve locating the chip at the point of least stress.

The contacts, the electronic conduits from the chip to the terminal, must have contaminant resistant surfaces. It is also important to assign high voltage contacts so that high voltage does not wipe across the other contacts when the card is inserted or removed from a terminal.

A plastic envelope is usually used inside the plastic housing, to protect the chip from corrosives and, to some extent, from electricity.

The issuing of the card is the most potentially destructive phase in a smart card's existence. Embossing is generally the greatest stress a card is put under, so it would be best to avoid embossing

the card once it has the chip inside; this is not, however, the standard manufacturing method, because embossing is considered part of the later, personalizing process. Recording any magnetic stripe that the card may have for backwards compatibility, and loading information into the card also put the card under physical stress. After this, the card must also survive being mailed to the card-holder, and then there must be some simple way for the recipient to check whether the card is in working order.

If smart cards are going to be used widely, it is essential that their designers and manufacturers understand the stresses the cards will be subjected to during their usage life, and be willing to invest in survivability design.

4. Security and Smart Cards

When carrying out a transaction using a card, there are a number of points where fraud can be attempted. A secure transaction system protects these points. This section explains the traditional on-line security techniques used with magnetic strip cards, gives an overview of the points in a transaction system that must be protected from security breach, and then explains how smart cards address these security issues.

4.1 Requirements for Transaction Security

On-line vs. Off-line Security

A transaction system is described as off-line if most activities require communication with some remote center.

On-line systems have many advantages due to the immediacy with which they allow the center to manage transactions. The center can react instantaneously to queries, and can rapidly identify breaches of security and invoke counter-measures, while in an off-line system, the center will not even know of a transaction until after it has been completed.

However, since on line systems rely on communications between the remote terminals and the center, rising networking costs are a problem. Also, the high number of communication links means

a high probability of service interruptions. A central malfunction can cause widespread problems, as can overloads.

Also, although on-line validation catches certain types of frauds, experience from ATM systems indicates that the weakest link in the security of the system as it now stands is the easy counterfeit and alteration of magnetic strips [2].

The problems of on line systems have been addressed by replacing them with off-line security, sometimes with no attempt to provide security beyond the dubious inviolability of the magnetic strip cards themselves, but also sometimes with the use of authorization telephone calls, which are inconvenient and time consuming, or with verification terminals that do a quick on-line check through a dedicated line; these terminals tend to be too expensive for most retailers to use.

Smart cards also encourage an off-line security paradigm for transaction systems. Again, off-line security lacks the tight central control of an on-line model -- the responsibilities of security are transferred to the periphery of the system and are held inside the smart card chips themselves. But in the case of smart cards, there is considerably more reason to believe that the cards can protect themselves from unauthorized intervention and amendment. Also transactions outside certain limits can still be automatically referred to the central system when necessary.

This is why the 'smart' part of the smart card is so important in security. A smart card can be programmed, as a magnetic strip or other chipless card cannot, to ensure the protection of a transaction system under off-line operation. Balances can be checked against spending limits, security algorithms can be run, and the authorization information held in the cards can be protected from outside view or tampering.

The elements of security for a system involving transaction cards are:

- Authentication of the Card.

It is possible (although expensive) to simulate a smart card, either with a faked or adjusted card, or with a larger computer that has a smart-card-shaped input/output device. It is also possible that a valid card might be substituted by an invalid one in mid-transaction. It must be possible to determine

whether the card presented is valid - that is, the card is considered valid for use by a trusted card issuer.

- Authentication of the Card-Holder.

Smart Cards provide no new safeguards against being physically stolen or abused. It must be possible to determine that the person presenting the card is authorized to be using the card in the requested way.

- Authentication of the Terminal.

False terminals could be set up, either to make unauthorized transactions directly when a card-holder tried to use one, or to store information about cards for later use. It must be possible to determine that the terminal with which the card is to be used is authorized for use by the card issuer.

- Authentication of the Transaction.

It must be possible for both the sender and receiver of the transaction to determine that the elements of the received transaction are genuine, that is, that this combination of card, holder, and terminal is authorized to carry out the action it is attempting.

- Security of the Message.

The flow of information between the smart card and the terminal is tappable, and it might be possible to send external signals in order to provide false responses to the card. It must be possible to encrypt and decrypt the message in order to provide security and privacy between the sender and the receiver.

Before smart cards, some attempts to address these points included:

- Randomly doing an on-line check of a certain percentage of transactions usually qualifying for off line authorization (e.g. because the transaction value is below a limit).
- Doing an online check if the same card number has been used very recently.
- Keeping a file of cards previously reported as stolen or abused.

4.2 Addressing Elements of Transaction Security with Smart Cards

Authentication of the Card

In a typical use of a smart card with a terminal, the terminal issues a pseudo-random number as a challenge to the card. The card then computes a result using its matching algorithm. The card returns this as a response that the terminal compares against its own result, using the algorithm for the card type (that is, for the card's issuer). This ensures that the card has the algorithm specified by the card issuer. The algorithm may also involve use of a secret identification number associated with this particular card's public-identification number, in order to ensure that the card has a valid number as well as the correct algorithm.

Clearly this form of check cannot be carried out with a chipless card.

It is also important that if the terminal goes out of communication with the card (it is removed in the contact case or a remote link is lost in the contactless case) it immediately destroy all record of the transaction so that an invalid chip cannot be substituted.

Authentication of the Card Holder

The conventional technique of a Personal Identification Number, as used with magnetic strip cards, is also used with smart cards to check the authorization of the card holder. However, in the smart card case, the PIN is checked within the card's protected memory, which can only be accessed by the card's own chip. Only a positive or negative result is ever passed outside of the card. The card may even lock itself permanently if a certain number of false PINs are attempted in a row.

There is some feeling that PINs are too easily compromised by card-holder carelessness. For this reason, various biometric identification techniques such as signature verification, hand geometry, finger prints, voice recognition, retina scan, and so on, have been suggested. The relevant data for the authorized card holder is kept in the smart card's protected memory -- a magnetic strip card would typically not have enough space on the strip to hold this information. A terminal can then run the physical check and send the data to the card for verification. A combination between one or more of these techniques and PIN requirements could prove a superlative security measure, although at some cost.

In the case of active cards used without a terminal, only that information that can be read through the card's on-board I/O can be used for authentication (and we must rely on the card holder's taking care that no miscreant has substituted a counterfeit for the correct card). In many cases, contactless cards, when they are used to increase speed of use, perform no check at all that the person carrying the card is authorized to do so.

Authorization of the Terminal

Each terminal operator should have a smart card of a special type that can interrogate the terminal's authorization, load the secret challenge-response used to check transaction cards, and ensure that this information is removed from the terminal when it is left unsecured. It can also be used to verify that the terminal operator is authorized to use the terminal. Naturally, the measures used for authorizing card and card holder in the case of a regular transaction card are also required in this case.

Authorization of the Transaction

Smart cards can include and verify electronic signatures in all messages sent to effect transactions, to guarantee that the messages come from an authorized source. They also have the ability to check dates and spending limits internally, and can themselves disallow transactions.

Security of the Message Content

Smart cards can also encrypt and decrypt their own messages (using, say, RSA or Fiat-Shamir cryptography), to protect their security and privacy between card and terminal. Again, this is clearly impossible to do in a chipless card. It is also a good idea to design terminals so that there is no exposed tappable wire between the card reader and the rest of the machine.

5. Using Smart Cards

5.1 Some Criticisms of Smart Cards

Although their use is widespread in several European countries, and their supporters are teeming with a lot of news about their usefulness with many cheerful facts about the improved security they may provide, smart cards have never quite caught on in the United States. Is this only another

example of the same obduracy that met the metric system, or are there bigger problems with smart cards?

A common criticism of smart cards is that they are only a status symbol for the technologically literate - the equivalent of a personal Web page, nice to have, but unnecessary. As crisp as the critiques of magnetic strip card security are by supporters of smart cards, it may be that the level of security they provide is acceptable to most businesses and consumers. Even those who are a little uncomfortable with the current level of security may find themselves willing to live with it for some time rather than go to the effort to migrate to smart cards.

Smart cards encourage the spread of and reliance upon applications that involve card transactions, but do not address in any way the protection and providence of services to those whose cards have been lost or stolen. If, as many smart card proponents are declaring, smart cards will entirely replace physical currency (even if only in some arenas) it is essential that the machinery be in place to efficiently provide replacement cards - it is one thing to spend a night stranded without cash due to a lost wallet, and quite another to spend, say, six weeks unable to use public transportation or telephones while waiting for processing of a new card. One approach to this might be public kiosks that read biometric information to check user authorization and can both issue temporary cards (magnetic strip, to cut down costs) and report the lost or stolen cards.

So why not just do all transactions based on biometric identification and account numbers? This certainly makes forging or stealing authorization messy (and painful for somebody), and there are no cards to lose. But consider what a transaction would involve. At the terminal, each user would provide biometric identification, but unless an account number is provided, we have the nontrivial problem of searching a database of bioprints, so users would most likely have to memorize many account numbers - longer than the PINs that most people find difficult to remember now. There is also the question of whether the identification information should be held in the terminal (off-line) or at a remote center (on-line). Biometric techniques are also still expensive and still provide many false negatives and false positives.

Smart cards can provide more anonymity in convenient distance payment systems than was possible before, which the US Justice Department Computer Crime Unit finds unacceptable. They propose requiring that payments using smart cards be trackable. This has only added to the widespread

complaint that smart cards, if tracked, could also make the activities and movements of private citizens more open to government scrutiny than ever before [1].

5.2 Examples of Smart Cards

Cards that are prepaid and then decremented with use are in wide usage for public telephone and public transport in many parts of Europe, especially France. Utilities such as water companies in the UK have been using prepaid cards to cut down on bad debts [14].

In America, the Washington D. C. metro tested a prepaid smart card called the GoCard in 1996-1997, that could be used for the metro bus system and connections to trains, as well as on the metro itself [6]. MicroCard Technologies' business journaling cards for peanut farmers are also in wide use [13].

Microcard has also developed a Smart Shopper Card that acts as an electronic purse, holding information for several checking and credit accounts, and also storing information such as dates and clothing sizes. Microcard will be testing a card combining medical records and medical payment information [5].

Visa made an early push for active cards. Visa's Super Smart Card that has a keyboard, display and magnetic strip emulator, and was tested at the Olympic Games in Atlanta [3].

The AT&T SmartCard supports multiple applications, possibly from several vendors, and has a processor-supported OS with a variety of security techniques and levels of security. It has 3-8kb of nonvolatile memory, an 8bit microprocessor with ROM, EEPROM, and a small amount of RAM, and contactless reader/writer capacitive plates and inductive power transfer coil. It also meets ISO standards for magnetic strip cards. It can communicate up to 19200 bits per second [11].

The Philips 83W858 Smart Card Microcontroller has an 8051 microcontroller and a crypto-processor that can compute an RSA signature in less than 400 ms [10].

6. Conclusions

Information transactions that can be carried out on the move -- including access to resources such as restricted areas or money, identification, and access to information -- are in increasing demand. Smart cards are a new technology that will increase the security of such transactions; smart cards consist of information and processing ability for this information, packaged in a form that is convenient to carry.

Magnetic strip cards are currently the most popular type of transaction card, but they are easy to alter and counterfeit. Smart cards overcome security problems of magnetic cards by restricting access to the information on the card, and by being expensive and difficult to counterfeit and alter.

Smart cards typically consist of a processor with onboard memory, packaged in a plastic casing. A terminal that can communicate with the card is also generally necessary.

Cards can be made more cheaply if they have hardwired logic rather than micro-processing capability, but this type of card lacks multi-functionality. Contactless cards, which do not have to be in physical contact with terminals, are more convenient and safer from contamination, but may result in less secure transactions. Active cards do not require a terminal for all transactions because they have some on-card I/O. Some 'hybrid cards' are equipped with magnetic strips for backwards compatibility.

Secure transactions require authorization of the card, the card holder, the terminal, and the transaction, as well as security of the transaction. Providing security on-line allows a remote center to act on transactions as they happen, but this is increasingly costly, and does not catch most of the fraud now carried out with magnetic strip cards. Smart cards can provide security off-line because their processing power allows them to take an active role in keeping the information in their memories secure. This is not possible with chipless cards.

Because of low prices for processors and memory, smart cards are now a viable and more secure alternative to magnetic strip cards for mobile applications requiring information transactions.

7. Bibliography

[1] R. Aguilar, "Government Worries About Smart Card Security," CNET NEWS.COM (http://ne2.news.com/News/Item/0,4,5166,00.html) May 1996.

[2] R. Bright, *Smart Cards: Principles Practice Applications,* Halsted Press, New York, 1988.

[3] T. Clark, "Smart Cards To GO," CNET NEWS.COM (http://ne2.news.com/News/Item/0,4,4894,00.html?st.ne.ni.rel) October 1996

[4] J. de Wilde, "The Philips PCs, CD-ROM, Smart Card and Laservision," North-Holland Computers in Industry 7, 1986

[5] F. V. Diest, "Mobility and Information A La Card," Electronic Design, January 1997.

[6] D. Fleishman, N. Shaw, A. Joshi, R. Freeze, R. Oram, *Fare Policies, Structures and Technologies,* National Academy Press, Washington D. C. 1996

[7] P. Gueulle, "Simple PC Smart Card Reader," Electronic Design, October 1997.

[8] P. Gueulle, "Tiny Smart Card OS for PIC16C84," Electronic Design, April 1997.

[9] S. L. Martin,"Smart card development expands as standard nears final approval," Computer Design, September 1988.

[10] S. Pecot, "Ashling introduces New Development System for Philips High-Security Smart Card Microcontrollers," Ashling Microsystems Ltd (http://www.ashling.com/pr83w858.html), May 1997.

[11] S. Sherman, R. Skibo, R. Murray, "Secure Network Access Using Multiple Applications of AT&T's Smart Card," AT&T Technical Journal, Sept-Oct 1994.

[12] Stargenix Corporation Home Page, Stargenix Corporation (http://www.starcard.com/story/), 1998.

[13] J. Svigals, *Smart Cards: the new bank cards,* Macmillan Publishing Company, New York, 1987

[14] V. Wyman, "Water firms try out smart card metering to combat bad debt," The Engineer (London, England) v. 274, Mar. 1992.