Calhoun

Institutional Archive of the Naval Postgraduate School

Calhoun: The NPS Institutional Archive

Faculty and Researcher Publications

Faculty and Researcher Publications

2010-07

Towards Reversible Cyberattacks

Rowe, Neil C.

Monterey, California. Naval Postgraduate School

Towards Reversible Cyberattacks

Neil C. Rowe
U.S. Naval Postgraduate School, Monterey, California, USA
ncrowe at nps dot edu

Abstract

Warfare without damage has always been a dream of military planners. Traditional warfare usually leaves persistent side effects in the form of dead and injured people and damaged infrastructure. An appealing feature of cyberwarfare is that it could be more ethical than traditional warfare because its damage could be less and more easily repairable. Damage to data and programs (albeit not physical hardware) can be repaired by rewriting over damaged bits with correct data. However, there are practical difficulties in ensuring that cyberattacks minimize unreversible collateral damage while still being easily repairable by the attacker and not by the victim. We discuss four techniques by which cyberattacks can be potentially reversible. One technique is reversible cryptography, where the attacker encrypts data or programs to prevent their use, then decrypts them after hostilities have ceased. A second technique is to obfuscate the victim's computer systems in a reversible way. A third technique to withhold key data from the victim, while caching it to enable quick restoration on cessation of hostilities. A fourth technique is to deceive the victim so that think they mistakenly think they are being hurt, then reveal the deception at the conclusion of hostilities. We also discuss incentives to use reversible attacks such as legality, better proportionality, lower reparations, and easier ability to use third parties. As an example, we discuss aspects of the recent cyberattacks on Georgia.

1 Introduction

The U.S. military is concerned about defending against cyberattacks. Little progress has been made in the last few years in this defense, due to increasing dependence on networking and the spread of unnecessarily complex software products. Attackers have grown increasingly sophisticated, and are now predominantly members of large organized groups, either organized crime or government espionage. Militaries are increasingly wondering to what extent these cyberattack techniques could be used as nonlethal weapons (Rattray, 2001).

An emerging consensus is that many existing "laws of war" do apply to cyberconflict (Schmitt, 2002). This includes the Hague Conventions of 1899 and 1907 and Geneva Conventions of 1949 and 1977 (ICRC, 2007). Article 51 of the 1977 Additional Protocols of the Geneva Conventions prohibits attacks that employ methods and means of combat whose effects cannot be controlled or whose damage to civilians is disproportionate, and Article 57 says "Constant care shall be taken to spare the civilian population, civilians, and civilian objects". So it is important to ensure this with cyberweapons.

These laws matter because of the practical difficulties in using cyberattacks effectively in warfare. As we have discussed elsewhere (Rowe, 2010), cyberattacks are less reliable than conventional attacks because they exploit bugs and flaws in software, and bugs and flaws are constantly being fixed. Cyberattacks thus tend to be unnecessarily strong in an effort to ensure their success. This risks high collateral damage. The good news is that some cyberattacks are better than others in that the damage they cause is more easily reversible. Lack of reversibility is an important argument for banning weapons, as with the recent initiatives to ban land mines in warfare. This suggests we draw analogies to outlaw irreversible cyberattacks in international law and treaties.

1.1 The Georgia attacks

The cyberattacks on Georgia in August 2008 suggest a possible future of cyberwarfare; (USCCU, 2009) provides an excellent summary. Attacks were launched to coincide with a military invasion of Georgia by Russia (the "South Ossetia War"), and appeared to be well planned and timed (Markoff, 2008). They primarily involved denial of service against key Georgia Web sites, with some malware involved in support of this, plus some Web-site defacement. Some of attacking

sites were known malware sites, some were new sites created specifically for the attack, some were "botnets" of otherwise innocent computers, and some were computers of people recruited to attack from social-networking sites. None of these were government or military sites.

The targets of the attack were government and business organizations in Georgia that contributed to its ability to withstand the conventional military attack which followed shortly thereafter. They included government agencies associated with communications as well as news-media organizations, apparently with the goal of making it difficult for Georgians to determine what was happening. Later attacks broadened the scope to financial and educational institutions, as well as businesses associated with particular kinds of infrastructure. These cyberattacks were clearly targeted at civilians, and were targeted quite precisely. A wide range of techniques were used in generally nonattributable ways in a well planned campaign. Apparently the attacks were designed (unlike the attacks on Estonia previously) to avoid international outcry.

Attacking primarily civilian enterprises is a clear violation of the laws of war (Walzer, 1977). Civilians are not supposed to be the primary targets of military actions unless they are substantially contributing to a tactical or strategic military asset. Civilian communications and media outlets are not military assets, and in fact can be adversarial to the military. This is different from the bombing of munitions plants during World War II, where the munitions were directly contributing to the war effort. So since the Georgia attacks were clearly correlated to a subsequent military incursion, they were clearly a violation of the laws of war. It is likely that the reason that military sites were not attacked was that they were better hardened against attack, so attacking them would be less cost-effective.

Denial of service and Web site defacement are relatively benign forms of attack compared to other possible cyberattacks. Once the attacks stopped, repair of sites was relatively straightforward; on this the attacks were relatively ethical. However, they still were not ethical. The government of Georgia could not inform citizens of what to do during the attack, and its ability to provide government services, medical services, and humanitarian activities then was greatly impeded. Some of the damage was permanent because of lost opportunities during the denial of service.

2 The damage of a cyberattack

Let us consider the main factors that affect the damage of a cyberattack. Cyberattacks can range from limited and controlled tactical attacks to broad and uncontrolled campaigns. This means their legal and ethical issues can vary considerably as well.

Collateral damage is a major factor in evaluating cyberattacks. This can cause needless suffering even if no one is killed. Concerns of this kind have limited U.S cyberattacks in the past (Markoff and Shanker, 2009). Collateral damage can be minimized by precise targeting. Such precision may not always be possible. Sites can change their names or shift (as in the case of Georgia) to another country, and self-propagating malware such as viruses and worms can autonomously spread beyond the initial targets.

An important difficulty with cyberattacks is in localizing them, something much more difficult than with conventional military attacks. This makes their effects difficult to repair. Computer systems and networks are complex. When something malfunctions, it may be difficult find the part of it that is responsible. Cyberattacks in particular will pick unusual targets within systems and do surprising things to them to obtain the maximum effect. Debugging of computer systems and networks is difficult, and many operators of computer systems lack detailed understanding of what they are running. The Georgia cyberattacks were focused primarily on the input to Web sites, which simplified their localization. But these attacks were clearly intended to be demonstrations. Most attacks will not be as easy to localize.

Another important factor in evaluating damage is its reversibility, the focus of this paper. If an attack can be quickly undone, this can be used to remove collateral damage, and it permits quick repair after the cessation of hostilities. It also helps achieve proportional response, a key aspect of the laws of war for counterattacks (Darnton, 2006), since an inadvertently too-powerful attack can be partially undone. Some proponents of cyberwarfare claim it is more reversible than other forms of warfare (Shulman, 1999), since damage to programs and data can be repaired by copying the original data over the damaged data. However, restoring programs and data by the victim can be time-consuming and requires well-trained staff (Dorf and Johnson, 2007) which is not always available. Cyberattacks also usually create psychological as well as physical effects, and psychological damage to a victim may not be easy to reverse. Also, attacks on time-critical activities may not be reversible. When a patient in a hospital is on a respirator controlled by a computer and the computer delays actions, the patient can die. Similarly, delaying many important activities of a government, and

particularly a military, can cause plenty of damage. Nonetheless, using attacks that are mostly reversible is a step in the right direction towards the responsible conduct of cyberwarfare.

3 Techniques for reversible cyberattacks

Let us consider ways to enable cyberattack damage to be reversed by an attacker more quickly and better than by restoring from backup. These cyberattacks could be traditional outsider-based attacks from the Internet during which the attacker sends malicious packets to a victim that enable them to take over control of the victim's computer, or they could be attacks accomplished by a malicious insider. Note however that cyberattacks are crimes in most countries, as they amount to malicious vandalism of communications technology.

3.1 Cryptographic attacks

Cryptography is a systematic method for concealing information (Mel and Baker, 2000). Since information superiority is a key objective of warfare, concealing it can be an attack (Libicki, 2007). Then the attack could be repaired by restoring availability of the information.

An example would be encrypting key programs of a victim with a decryption key that only the attacker knows. Encrypted programs would be unusable by the victim until the attacker is willing to decrypt them. Valuable data such as sensor information can also be encrypted by an attack. Encryption can be accomplished by obtaining administrator ("root") access to a machine by any of a number of methods, and making changes to software and data. Such access can be obtained long in advance and then an attack can be triggered by an external signal. Rather than encrypting an entire piece of software, a simpler alternative is to insert a prolog or "wrapper" to programs that requires a user to enter a password known to the attacker before the program can be used. This is easiest if a program already requires a password, in which case the attacker could modify data to require the attacker's password instead.

Another alternative is to encrypt the data going to a victim, as discussed in section 3.3. Data is usually more vulnerable to modification than programs. Many secure networks encrypt their data in transmission, and this could be changed to use the attacker's key rather than the victim's. This could be done with administrator access to the machines doing the encryption, usually just the first and final ones through which the data passes (with "end-to-end encryption"), and could be accomplished by "rootkits" (Kuhnhauser, 2004). It could also be accomplished by modifying the hardware used for networking, replacing parts or whole computer systems with those designed by the attacker.

The main countermeasure for encryption-based attacks is restoration of the damaged code from backup. However, it may be difficult to determine what to restore because of the attack localization problem discussed above. Even when the attack can be localized, restoration can be difficult when properly trained personnel are unavailable, personnel are unfamiliar with the restoration procedure, or the restoration procedure is complicated (as it can be with complex software like operating systems). Restoration requires valuable time of system administrators that they did not anticipate using, and it might not be fast enough in a "blitzkrieg" cyberattack designed to quickly achieve objectives.

Encryption-based attacks could be easy to detect, since encrypted characters have statistical randomness very different from those of normal programs and software. Attackers who want their attacks to be visible to encourage negotiation would find this property helpful.

3.2 Obfuscating attacks

Computer systems are carefully designed entities. If we can disrupt their organization, they become unusable. So a class of "obfuscating" attacks could rearrange the software and data of a computer system, or map data values to new values in a one-to-one mapping, or insert extra data, in a way known only to the attacker. For instance, we could interchange parts of programs; we could add 13 modulo 256 to a set of designated 8-bit bytes; or we could insert random bytes into designated locations in programs. The plan for how we obfuscate can be arbitrarily elaborate if we record it carefully. Cryptographic methods are a special case of obfuscating methods, but the restrictions of cryptography are not necessary to make a system unusable. For instance, adding many random bytes could greatly increase the size of programs, something cryptographic methods do not do in their attempt to be space-efficient, but which works well as an attack technique given the typically small portion of occupied storage on most computer systems.

Anything can be targeted with obfuscation techniques, but the obfuscation can be more efficient if it targets critical parts of programs and key data. If the attacker wants their attack to be noticed, they can target highly visible parts such as the user interface. Just modifying the appearance of a window can make software ususable with little effort.

Obfuscating attacks can be undone by applying the reverse of the attacking actions in reverse order. So for instance if we interchanged two blocks of code, added 13 to each byte, and then added two bytes of "37" on the end, we can undo the effects by deleting the two end bytes, subtracting 13 from each byte, and then interchanging the same two blocks. Any operation performing a one-to-one mapping (an "isomorphism") on either the contents or location of data can be reversed by an inverse operation. This includes operations that add useless information since they effectively permute locations of the real information.

If the rearrangement is sufficiently complex, it would be virtually impossible for the victim to figure it out and reverse it, although unlike encryption, it could be reversed given sufficient time. Unlike encryption, obfuscation with interchanges and padding can be designed to provide the same statistics as the original system, and thus be difficult to localize. More than encryption, restoration from backup would be difficult and slow because the entire system would need to be restored if the modifications were well dispersed and the victim could not recognize them. Partial restoration could be worse than none at all because it could destroy some of the interchanged halves. Restoration can be made still harder if the attacker combines many techniques.

3.3 Withholding-information attacks

Another potentially reversible attack method directly withholds data. This is similar to blockading in traditional naval warfare and jamming in electronic warfare, both of which share the advantage of relatively reversible damage. Denial of service is an example of indirect denial of data by flooding a resource with false data so there is little time to process the true data. But denial of service is a relatively broad attack which risks much collateral damage.

A more precise way to withhold information is to use a "man in the middle" deployment: The attacker inserts their own hardware and/or software (perhaps by address hijacking) between the victim and the victim's intended communicants, or takes control of an existing machine on that route. Only information that the attacker allows will then pass from or to the victim. This could be accomplished by encrypting the information as suggested in the last section, so the victim gets or sends all their normal volume of data but no one can read it. Alternatively, if the volume of data is not large, it could be withheld from the victim and saved for restoration after the cessation of hostilities. Returning the data repairs some damage, since getting it late is usually better than not getting it at all, and facilitates tracking of secondary damage caused by the withholding. However, this will not reverse all damage, like physical damage from the failure of a computer system to prevent a subsequent irreversible physical attack.

If the "middleware" doing the diversion is designed to be highly selective, bandwidth may not be a problem. For instance, an attack on a sensor system might withhold only locations, which are easy to spot in text because they use standard formats. Or it might block the locations of attacker military units while reporting locations of casualties to permit Red Cross activities. Or a messaging system might only block orders down the chain of command, not reports going up, preventing the application of force while keeping the victim informed of what is happening. Man-in-the-middle deployments can be done with hardware, as by inserting a new device on the Internet connection to a computer. Or they can be done with software, by modifying the implementation of Internet protocols with the assistance of a rootkit.

An alternative way to withhold information is to misroute it. So a man-in-the-middle attack could deliberately change destination headers of Internet packets sent through it to that of a site it controls. If the corresponding correct destinations are stored somewhere, and the packets are stored at the new destination, the attack could be reversed. This has the advantage of eliminating storage at the man-in-the-middle device.

A countermeasure for man-in-the-middle attacks is to bypass the middle, much as blockaded countries can find new routes to conduct trade. Attackers can prevent this by attacking a "bottleneck" computer such as a server for a local-area network, or by exploiting key protocols for networking that must be used by a user. Firewall and intrusion-prevention computers, and TCP and HTTP protocols, provide good opportunities for such attacks. Almost as good are file and Web servers for a local-area network.

3.4 Resource-deception attacks

Another approach is to deceive a victim with illusory damage. Then "repairing" the attack means just revealing the truth to the victim. A simple way is to modify the victim's operating system to issue false error messages on any attempts to use the system for something important (Rowe, 2007). Most users take error messages seriously, so false error messages can be quite disruptive to them. Such messages raise no legal problems themselves, unlike man-in-the-middle attacks, since they happen not infrequently when the software lacks sufficient information to accurately diagnose an error. However, the necessary modification of the operating system is still a form of vandalism.

Resource deception can be implemented by modifying the operating system of a computer by a rootkit. One good way is through "software wrappers" on key components of the operating system and applications software (Michael et al, 2002). Normally the wrappers can behave transparently, passing on commands to the operating system and passing back responses. But in specified circumstances recognized by a system monitor, the wrappers can issue false error messages and other confusing information. Alternatively, hardware exceptions can be triggered by the wrappers, which can simulate serious errors.

Error messages can be made more persuasive using techniques from spam and phishing (Spammer X, 2004). Official-looking graphics and verbal manipulation can be used. Claims of authorization from authorities and experts, rewards for compliance, and threats for noncompliance can be cited. Creating an atmosphere of urgency will also help, e.g. flashing red letters saying "Security Violation -- Log out now." However, persuasion works less well when it coincides too closely with the attack, since an intelligent victim will likely infer a common cause. It would be better to begin deception well before the attack. Also, repeated deceptions lose effectiveness as the victim becomes familiar with them, so deceptions should only be used occasionally.

A countermeasure to resource deception is to reinstall the software generating it. But this is time-consuming. It may also be unnecessary because a clever adversary could modify an operating system in superficial ways that do not require reinstallation.

While automated deceptions do not damage data and programs to the extent of the previously mentioned attacks, they do create persistent damage in the form of increased distrust of computer systems by the victim. This can unfairly reduce their ability to use computer technology for a long time, and these effects can extend to a broad range. Trust is built up slowly, but distrust can increase quickly given a single act (Ford, 1996). Thus deception-based attacks may need to be used cautiously.

4 Additional factors contributing to reversibility

Not all cyberattacks following the above methods will be equally feasible or equally reversible. Other considerations are involved.

Methods for reversible attacks need precision in targeting to simplify their reversal as well as reduce their chances of collateral damage and disproportionality in attacking. Precision enables an attacker to focus on a few well-chosen military targets and better assess the effect of their attack. If the effect is too small, the attack can be increased; if the effect is too large, reversing the attack can be done even before hostilities cease.

Reversibility may decrease with time. For instance, a victim may close its Internet connections so that systems cannot be reached to repair them. Or an attack may be detected by the victim and repaired ineptly to make reversing impossible, as by loading the wrong backup copies resulting in incompatible software modules that will not work. These contingencies need to be addressed in attack planning. Reversal may also have a latency (time delay) that causes additional harm. For instance, if Georgia had surrendered midway during the campaign against it, it is unlikely that the attacks would have stopped for quite a while, since they were coordinated primarily at the planning stage and not during execution. Then Georgia would have incurred considerable damage after surrendering, a clear violation of the laws of war.

Another important factor is the ability of the defender to identify the attacker ("attribution"). It is desirable that the victim know this for a reversible attack since the attacker knows best how to reverse it. After all, anonymous attacks are typical of terrorism; a responsible country wants their attacks to be attributable to help achieve precise outcomes. A simple form of attribution is to consistently use a well-known source site. (Otherwise, just leaving the sites up for a while after the attack will aid tracing of their location.) More generally, an attacker can attach crytographic signatures to the attack code

or data. Signatures can be embedded in unnecessary instructions in code or in comments in data; steganography (Wayner, 2002) can be used to conceal the signatures if necessary. Signatures can use the private key of a public-private key pair so that only the attacker can attach them.

5 Enforcement of reversible attacks

A question is what incentive a cyberattacker has to use reversible attack methods. A similar question can be asked about many other military technologies, such as conventional weapons instead of nuclear weapons. Some incentives in these cases come from international outcry at using unethical methods and the resulting ostracism of the offending state or organization. But more importantly, nations agree to laws of warfare, and unethical methods can violate those laws. Responses of the international community to such violations include sanctions, boycotts, fines, and legal proceedings (Berman, 2002).

A good incentive for reversible attacks occurs if the attacker must pay to repair the damage. Estimates of repair costs could be an important factor in the amount of reparations required to settle a conflict (Torpey, 2006). Reparations are also enforceable against non-state actors such as factions within a country. If a neutral party can enforce reparations, reversible attacks are advantageous to belligerents. The reversal methods proposed here can be initiated remotely, so territorial-integrity concerns that impede cleanup of damage of conventional warfare are less burdensome with cyberweapons. A related issue is the attacker proving that they have removed all traces of their attack, which is not difficult to do for the relatively simple attack techniques of this paper. For instance, an attacker that used signatures can prove that none remain on a system.

Another incentive to reversible attacks is if a victim is likely to respond in like kind. Then use of reversible attack could encourage an adversary to do the same because otherwise they would appear to be escalating the conflict (Gardam, 2004).

Cryptographic attacks can exploit three-party cryptographic protocols such as key escrow (Mel and Baker, 2000). In this, a neutral third party holds a key for deciphering an attack's encryption scheme. Similarly, in obfuscating attacks the "swap plan" functions like a key and could be held by the third party. A neutral third party like the United Nations could confirm signatures of attacks and assess damage, which could be more acceptable to the belligerents when done by a disinterested party.

A neutral third party could also provide selective or staged repair of reversible cyberweapon damage for belligerents that do not trust one another. The third party could alternate repair between two countries in stages so that none ever has a significant repair advantage over the other. A third party could try to calm crises by enforcing limited sanctions on the belligerents using the methods described above, that would allow food, medicine, and other forms of humanitarian assistance to be arranged across the Internet, while prohibiting activities that led to the conflict such as denial-of-service actions by either party to the conflict. Going further, a third party such as the United Nations could even employ reversible attacks themselves as a form of humanitarian intervention in a conflict to stop it, as for instance for a genocide. Reversible attacks then might be the more ethical than doing nothing.

6 Conclusions

All warfare aims at precise effects on its victims. Reversibility of attacks aids and supports precision. We have discussed four ways for implementing reversible cyberattacks, and some of the secondary factors that affect their reversibility. Reversible cyberattacks are clearly feasible, cost-effective, and some can be made undetectable. Thus reversibility appears to be a desirable property of cyberweapons.

An issue is whether the availability of reversible cyberattacks will encourage attacks. Reversibility, even if partial and delayed, lowers the cost to the attacker as broadly measured. However, any kind of attack introduces risks for the attacker, such as international outrage, sanctions, and counterattacks. Warfare still remains difficult to justify.

References

Berman P. (2002) "The Globalization of Jurisdiction", *University of Pennsylvania Law Review,* Vol. 151 No. 2, pp. 311-545.
Darnton, G. (2006) "Information Warfare and the Laws of War", in Halpin, E., Trovorrow, P., Webb, D., and Wright, S.

(eds.), *Cyberwar, Netwar, and the Revolution in Military Affairs*, Palgrave Macmillan, Houndsmills, UK, pp. 139-156.

Dorf, J., and Johnson, M. (2007) "Restoration Component of Business Continuity Planning", in Tipton, H., and Krause, M. (Eds.), *Information Security Management Handbook, Sixth Edition*, CRC Press, pp. 645-1654.

Ford, C. (1996) *Lies! Lies!! Lies!!! The Psychology of Deceit*, American Psychiatric Press, Washington, DC, USA.

Gardam, J. (2004) *Necessity, Proportionality, and the Use of Force by States*, Cambridge University Press, Cambridge UK.

ICRC (International Committee of the Red Cross) (2007) "International Humanitarian Law – Treaties and Documents", retrieved December 1, 2007 from www.icrc.org/icl.nsf.

Kuhnhauser, W. (2004, January) "Root Kits: An Operating Systems Viewpoint", *ACM SIGOPS Operating Systems Review,* Vol. 38, No. 1, pp. 12-23.

Libicki, M. (2007), *Conquest in Cyberspace: National Security and Information Warfare,*: Cambridge University Press, New York, NY, USA.

Markoff, J. (2008, August 13) "Before the Gunfire, Cyberattacks", *New York Times*, p. A1.

Markoff, J., and Shanker, T. (2009, August 1) "Halted '03 Iraq Plan Illustrates U.S. Fear of Cyberwar Risk", *New York Times*, p. A1.

Mel, H., and Baker, D. (2000) *Cryptography Decrypted, 5th edition,* Addison-Wesley Professional, Boston, MA, USA.

Michael, J., Auguston, M., Rowe, N., and Riehle, R. (2002, June) "Software Decoys: Intrusion Detection and Countermeasures", *Proc. of IEEE Information Assurance Workshop*, West Point, New York, pp. 130-138.

Rattray, G. (2001) *Strategic Warfare in Cyberspace*, MIT Press, Cambridge, MA, USA.

Rowe, N. (2007, May) "Finding Logically Consistent Resource-Deception Plans for Defense in Cyberspace", *Proc. of 3rd International Symposium on Security in Networks and Distributed Systems*, Niagara Falls, Ontario, Canada, pp. 563-568.

Rowe, N. (2010) "The Ethics of Cyberweapons in Warfare", *Journal of Techoethics*, Vol. 1, No. 1, pp. 20-31.

Schmitt, M. (2002) "Wired Warfare: Computer Network Attack and J*us in Bello*", *International Review of the Red Cross,* Vol. 84, No. 846, pp. 365-399.

Shulman, M. (1999) "Discrimination in the Laws of Information Warfare", *Columbia Journal of Transnational Law,* Vol. 37, pp. 939-968.

"Spammer X" (2004) *Inside the Spam Cartel*, Syngress, Rockland, MA.

Torpey J. (2006) *Making Whole What Has Been Smashed: On Reparations Politics*, Harvard University Press, Cambridge, MA, USA.

USCCU (United States Cyber Consequences Unit) (2009, August) "Overview by the US-CCU of the Cyber Campaign against Georgia in August of 2008", US-CCU Special Report, downloaded from www.usccu.org, November 2, 2009.

Walzer, D. (1977) *Just and Unjust Wars: A Moral Argument with Historical Illustrations,* Basic Books, New York, NY, USA.

Wayner, P. (2002) *Disappearing Cryptography: Information Hiding: Steganography and Watermarking,* Morgan Kaufmann, San Francisco, CA, USA.

The views expressed are those of the author and do not represent those of any part of the U.S. Government.