Cognitive Hacking: A Battle for the Mind

Cognitive hackers manipulate a user's perception and rely on his changed actions to carry out the attack. Effective countermeasures must aim at preventing misinformation through authentication, collaborative filtering, and linguistic analysis.

George Cybenko
Annarita Giani
Paul Thompson
Dartmouth College

On 25 August 2000, stockholders were stunned by news that Emulex, a server and storage provider, was revising its earnings from a $.25 per share gain to a $.15 loss and that it was lowering its reported net earnings from the previous quarter as well. The press release, which business news services like Dow Jones, Bloomberg, and CBS MarketWatch were distributing, went on to state that CEO Paul Folino had resigned and that the company was under SEC investigation. Within 16 minutes, Emulex shares plummeted from their previous day's close of approximately $104 per share to $43.

Only none of it was true. A 23-year-old hacker, Mark Jakob, had created the bogus release expressly to lower Emulex stock prices and thus recoup his recent $100,000 loss in a stock short sale. (In a short sale, stock prices must fall for the seller to profit.) Jakob had launched the release via his former employer, Internet Wire, a Los Angeles firm that distributes press releases, and the business news services had picked it up and widely redistributed it without independent verification.

More than three million shares traded hands at the artificially low rates. Jakob earned back his $100,000 nearly three times over, but was subsequently charged with security fraud and is currently facing a 44-month prison term. He also agreed to return his gains—approximately $353,000—and pay a civil penalty of $102,642.

The Emulex case illustrates the speed, scale, and subtlety with which networked information can propagate and how quickly severe consequences can occur. Although Nasdaq halted trading at the artificial price after only an hour, Emulex lost $2.2 billion in market capitalization.

The damage had little to do with penetrating the network infrastructure or technology. It had to do with manipulating perception and waiting for altered reality to produce actions that would complete the attack. Jakob cracked no code, created no Trojan horse, and planted no virus. He merely wrote a convincing press release, used a believable distribution medium, and sat back to watch events unfold.

MANIPULATING PERCEPTION

This manipulation of perception—or cognitive hacking—is outside the domain of classical computer security, which focuses on the technology and network infrastructure. Indeed, the Emulex case is an example of how the variety and complexity of attacks parallel information technologies and the way we use them—with no end in sight for either side.

In 1981, Carl Landwehr observed that "Without a precise definition of what security means and how a computer can behave, it is meaningless to ask whether a particular computer system is secure."[1] Landwehr urged application designers to state in advance the desired system security properties, recognizing that systems with different goals, operating in different environments, will have different security needs.

Twenty years ago, computer security focused on systems for handling military messages. Since then, particularly with the growth of the Internet, computer systems are widely used to disseminate infor-

0018-9162/02/$17.00 © 2002 IEEE

mation of all types to a variety of users. It is precisely this dissemination that has enabled cognitive hacking.

ONE THREAT: MANY FACES

Cognitive hacking can be either *covert*, which includes the subtle manipulation of perceptions and the blatant use of misleading information, or *overt*, which includes defacing or spoofing legitimate forms of communication to influence the user.

In covert cognitive hacking, the attacker tries to disguise the attack; in overt hacking, he does not. At times, the distinction between the two categories is not always obvious. For example, some defacements introduce subtle changes that might be considered covert. Some attacks intended to be covert end up being overt because they're not terribly believable.

Overt cognitive hacking, while more prevalent than covert forms, is more of a nuisance and embarrassment than a serious threat. Covert cognitive hacking, on the other hand, is likely to have more significant and less predictable consequences.

Perception management's dark side

Using information to change behavior is not new. Individuals influence one another through words. Institutions of learning influence students through formal education. Governments and commercial enterprises influence the masses through propaganda and advertising.

Perception management is pervasive in any contemporary society.[2] The intent of a specific attempt to manage perceptions is not always clear. Some management is for the good of the recipient: Education (at least in theory) is benign, for example, but the intent of propaganda is a little fuzzier. The propagator often believes that the message is true, but the presentation is designed to persuade.

The dark side comes from using misleading or false statements. Even if the intent is to benefit the recipients, such as to persuade them to vote for a particular issue, some form of manipulation to serve the propagator's personal agenda is involved. If the dark side is dark enough, the propagator may be considered a cognitive hacker.

Advertising is a good example of these degrees of fuzziness. If a Web site advertises a weight loss product in a way that leads the user to believe he will lose weight in one week, but in reality the product is useless, the user may be out $19.95 plus shipping, but it is hardly a cognitive hack. On the other hand, if an employee from the company posts a message strongly implying he has "inside information" about the product line with the goal of making the stock soar for his personal profit, he is considered a cognitive hacker.

What makes the one case advertising and the other cognitive hacking is the propagator's intent *and* the message's context. A person acting on an advertisement should know that the company may be bending the truth. A person reading a personal message posted on the Internet is more likely to expect that the message will be entirely truthful and unbiased.

The Internet's open nature makes it an ideal mechanism for spreading misinformation.

To make this distinction clear, we offer a precise definition of cognitive hacking in the context of perception management: "Gaining access to or breaking into a computer information system to modify certain user behaviors in a way that violates the integrity of the entire user information system." In this definition, system integrity is the correctness or validity of the information the user gets from the system, and the system can be far reaching.

Any Web-based activities that violate the norms of communication or commerce would be cognitive hacking. If someone maintains a Web site that influences its viewers to engage in fraudulent commercial transactions with the site owner, the owner is considered a cognitive hacker because he has compromised the integrity of the information the user receives from the system (the Web).

Misinformation

Misinformation is an intentional distribution or insertion of false or misleading information intended to influence the reader's decisions and activities.

Bruce Schneier provided a definition of semantic attacks that is closely related to our discussion of cognitive hacking.[3] Schneier attributes to Martin Libicki the earliest conceptualization of computer system attacks as physical, syntactic, and semantic. Libicki described semantic attacks in terms of misinformation inserted into interactions among intelligent agents on the Internet.[4] Schneier, by contrast, characterized semantic attacks as targeting "the way we, as humans, assign meaning to content." He went on to note, "Semantic attacks directly target the human-computer interface, the most insecure interface on the Internet."

The Internet's open nature makes it an ideal mechanism for spreading misinformation. In face-to-face interaction, people can more easily evaluate the information being conveyed. They at least know the person's name and can judge the reliability of the information on the basis of the interaction. This type of evaluation isn't possible on the Web.

New York Times Minus a Page

In February 2001, a hacker identified as "splurge" from a group called Sm0ked Crew, replaced a page of the *New York Times* with the page below. A few days earlier, the same group had defaced sites belonging to Hewlett-Packard, Compaq, and Intel. The *Times* defacement included HTML, a MIDI audio file, and graphics. As this example shows, defacements are rarely motivated by anything other than self-aggrandizement.

Pump-and-dump schemes

Pump-and-dump schemes are an example of misinformation. The hacker pumps a stock—presents it in an overly positive way so that its value will increase—thus allowing him to dump the stock at an inflated value. The SEC defines pump-and-dump schemes (also known as hype-and-dump manipulation) as the "touting of a company's stock (typically microcap companies) through false and misleading statements to the marketplace" (http://www.sec.gov/answers/pumpdump.htm). Once these fraudsters dump their shares and stop hyping the stock, the price typically falls, and investors lose their money.

Pump-and-dump schemes often occur on the Internet, where posted messages commonly urge readers to buy a stock quickly or to sell before the price goes down. Telemarketers also use the same kind of pitch. Often, the promoters claim to have inside information about an impending development or how to use an infallible combination of economic and stock market data to pick stocks. In reality, they may be company insiders or paid promoters who stand to gain by selling their shares after the buying frenzy they create pumps up the stock price.

Unauthorized modification

In this type of hacking, the attacker either defaces or spoofs a legitimate communication source. A defacement occurs when a hacker breaks into a legitimate Web site and alters the page. The "*New York Times* Minus a Page" sidebar describes an example of an unauthorized modification. In a Web site spoof, the hacker tries to pass a created counterfeit site as the real site without altering the legitimate site. The "The Day Britney Died" sidebar provides an example of this type of scheme.

An estimated 90 percent of Web page attacks are total page hacks, in which the attacker replaces the entire page of the attacked site. Defacements and spoofs are usually blatant enough for viewers to recognize, and the hacker's main concern seems to be the publicity associated with the nuisance or the satisfaction of being able to do it.

COUNTERMEASURES

Combatting cognitive hacking requires either preventing unauthorized access to information assets (to counter defacements and spoofing) or detecting posted misinformation before it affects user behavior (possibly after dissemination). Detection is difficult because the attack may not involve unauthorized access to information, as in pump-and-dump schemes that use newsgroups and chat rooms. Detection would also involve verifying that user behavior has indeed changed.

Thus, countermeasures aim to detect misinformation and correlate it with user behavior. Detecting the preconditions of cognitive hacking may in effect prevent it. Any information service that claims to add value to information disseminated on the Internet, or that claims to provide reliable information, should be responsible for using these detection techniques. To the extent that cognitive hacking becomes criminal behavior, the government also should become involved in regulating or monitoring Internet transactions.

Single-source hacking

In single-source cognitive hacking, the user has no independent source of information about the same topic and therefore cannot judge if the information is accurate. The most typical example is a single false posting to a newsgroup, where there is no other information source for the story.

Source authentication. Authentication amounts to due diligence in identifying the information source and ascertaining its reliability. In some cases, such as with an individual fake Web site, the user may be responsible for realizing that the source is not authentic. If the source is part of a news service, however, the content provider must protect accuracy.

Mature certification and PKI technologies can detect the spoofing of an information server, for

example. Additionally, the provider can use reliability metrics for an information server or service that score its accuracy over repeated trials and different users.

Clifford Lynch describes a framework that providers can use to establish trust among individual users.[5] Lynch advocates using PKI techniques to verify the identity of an information source and mechanisms such as rating systems to describe the source's behavior. However, authentication will take time and social or corporate consensus to evolve.

Information trajectory modeling. A government agency that aims to regulate Internet transactions or a content provider can use a source model based on statistical historical data or an analytic understanding of how the information relates to the real world. The model could, for example, calibrate weather data coming from a single source (a Web site or environmental sensor) against a historical database or a predictive model extrapolated from previous measurements. A large deviation would give the user a reason to hesitate before committing to a behavior or response.

The model would run as a kind of background filter. The subtler the hack, the more valuable the model becomes. In essence, the model extrapolates an aggregate effect from a series of small deviations, which taken individually would still seem credible. Well-scripted novels or films exploit this idea. The most satisfying and successful stories involve a sequence of small deviations from the expected. Each twist is believable, but the aggregate pushes the audience to a conclusion that is far from reality (the "red herring" in murder mysteries). Accurate modeling, of course, depends on the availability of relevant historical data.

The cognitive hack described in "The Jonathan Lebed Case" sidebar is an example of an attack that information trajectory modeling could have thwarted. If AOL had had an information trajectory model related to pump-and-dump schemes, the model would have shown normal patterns of stock value movement for low-cap stocks. When a particular stock's value seemed to deviate from the model, the model would issue an alert.

Ulam games. In his autobiography, *Adventures of a Mathematician*,[6] Stanislaw Ulam posed this question:

> Someone thinks of a number between one and one million (which is just less than 2^{20}). Another person is allowed to ask up to twenty questions, to which the first person is supposed to answer only yes or no. Obviously, the number can be guessed by asking first: "Is the number in the first half-million?" and then again reduce the reservoir of numbers in the next question by one-half, and so on. Finally, the number is obtained in less than $\log_2(1{,}000{,}000)$. Now suppose one were allowed to lie once or twice, then how many questions would one need to get the right answer?

Of course, if an unbounded number of lies are allowed, no finite number of questions can determine the truth. On the other hand, if k lies are allowed, someone can repeatedly ask each binary search question $2k+1$ times, which is clearly inefficient. Several researchers have investigated this problem, using ideas from error-correcting codes and other areas.[7]

A framework based on Ulam games involves a sequence of questions and a bounded number of

The Day Britney Died

On 7 October 2001, CNN's top-ranked news story was a hoax titled "Singer Britney Spears Killed in Car Accident."

The hoax began with a spoof of CNN.com. Then, because of a bug in CNN's software, when people at the spoofed site clicked on "e-mail this article to a friend," the real CNN system distributed a real CNN e-mail to recipients with a link to the spoofed page. At the same time, each click on "e-mail this" at the bogus site incremented the real site's tally of most popular stories.

Allegedly, a researcher started this hoax by sending the spoof to three users via AOL's Instant Messenger. Within 12 hours, more than 150,000 people had viewed the spoofed page.

The Jonathan Lebed Case

Using only AOL accounts with fictitious names, 15-year-old Jonathan Lebed earned between $12,000 and $74,000 daily over six months—for a total gain of $800,000. According to the US Security Exchange Commission, Lebed would buy a block of stock and then post a message like the one below. He did this about 11 times, increasing the daily trading volume from 60,000 shares to more than one million.

> DATE: 2/03/00 3:43pm Pacific Standard Time
> FROM: LebedTG1
>
> FTEC is starting to break out! Next week, this thing will EXPLODE . . .
> Currently FTEC is trading for just $2 1/2. I am expecting to see FTEC at $20
> VERYSOON . . .
> Let me explain why . . .
> Revenues for the year should very conservatively be around $20 million. The average company in the industry trades with a price/sales ratio of 3.45. With 1.57 million shares outstanding, this will value FTEC at . . . $44. It is very possible that FTEC will see $44, but since I would like to remain very conservative . . . my short term price target on FTEC is still $20!
> The FTEC offices are extremely busy . . . I am hearing that a number of HUGE deals are being worked on. Once we get some news from FTEC and the word gets out about the company . . . it will take-off to MUCH HIGHER LEVELS!
> I see little risk when purchasing FTEC at these DIRT-CHEAP PRICES. FTEC is making TREMENDOUS PROFITS and is trading UNDER BOOK VALUE!!!
> This is the #1 INDUSTRY you can POSSIBLY be in RIGHT NOW. There are thousands of schools nationwide who need FTEC to install security systems . . . You can't find a better positioned company than FTEC!
> These prices are GROUND-FLOOR! My prediction is that this will be the #1 performing stock on the NASDAQ in 2000. I am loading up with all of the shares of FTEC I possibly can before it makes a run to $20. Be sure to take the time to do your research on FTEC! You will probably never come across an opportunity this HUGE ever again in your entire life.

Lebed's aim was to influence people to buy the stock, thus pumping up the price. The messages looked credible, and people did not even think to investigate the source of the messages before making a decision.

Initially, the SEC forced Lebed to give up everything, but he fought the ruling and was able to keep part of what he gained.

Richard Walker, the SEC's director of enforcement, referring to similar cases, stated that on the Internet there is no clearly defined border between reliable and unreliable information; investors must exercise extreme caution when they receive investment pitches online.

lies, known a priori. This framework could be useful in preventing a cognitive hack that involves a sequence of interactions between a user and an information service, as in a negotiation or multistage handshake protocol.

The first step in thwarting the Emulex hack, for example, would be to create a sequence of questions to determine whether or not Emulex's situation was dire. The second step would be to have a variety of Web sites or other sources answer those questions. The sequence of questions would follow the 20-questions model, and a solution to Ulam's problem would determine the question ordering or selection. The content of the false release could be seen as a conjunction of predicates, for example, and the solution would then be to seek independent sources to verify each predicate.

The Ulam games technique resembles what Dorothy Denning[4] describes as information warfare—a struggle over an information resource by an offensive and a defensive player. The resource has an exchange and an operational value. The value of the resource to each player can differ depending on factors related to each player's circumstances. Offensive information warfare can have several outcomes: The offense has higher availability to the resource, the defense has lower availability, or the resource integrity decreases.

Viewing the Emulex case in information warfare terms, Jakob is the offensive player; Internet Wire and the other newswire services are the defensive players. The outcome is the decreased integrity of the newswires' content. In cognitive hacking terms, the outcomes involve victims. The main victims in the Emulex case were the misled investors, and an additional outcome was the money they lost.

Multiple source hacking

In multiple source cognitive hacking, multiple information sources are available on the same topic. Examples include financial, political, and other news coverage.

Several aspects of information dissemination through digital network media make cognitive hacking possible and in fact relatively easy to perform. Obviously, there are enormous market pressures on the news media and newsgroups to quickly disseminate as much information as possible. In the area of financial news, in particular, competing news services strive to be the first to provide reliable news about breaking stories that impact the business environment. Such pressures are at odds with the time-consuming process of verifying accuracy.

A compromise between the need to quickly disseminate information and the need to investigate its accuracy is not easy to achieve. Automated tools could be an effective aid in evaluating the veracity of information from multiple networked information systems.

Collaborative filtering and reliability reporting. The reliability, redundancy, pedigree, and authenticity of information are key indicators of its overall trustworthiness.[8] Collaborative filtering and reputation reporting have been receiving more attention recently, especially in online sales. The many online price comparison services commonly use a reliability rating to inform potential customers about vendor reliability. The services use customer reports to compute the reliability rating.

Both filtering and reliability reporting involve user feedback about information received, which builds up a community notion of a resource's reliability and usefulness. The automation in this case is in the processing of the user feedback, not the evaluation of the actual information itself.

Online auction sites, such as eBay.com, use both collaborative filtering and reliability reporting. When sellers misrepresent items to be sold—in effect, a cognitive hack—they usually do it only a few times. eBay aggregates reports from buyers and publishes them. If a seller's profile is bad enough, the company can bar the seller from participation.

Byzantine generals model. K. Mani Chandy and Jayadev Misra define the Byzantine generals problem as follows:[9]

> A message-communicating system has two kinds of processes, *reliable* and *unreliable*. A process, *general*, may or may not be reliable. Each process *x* has a local variable *byz[x]*. It is required to design an algorithm, to be followed by all reliable processes, such that every reliable process *x* eventually sets its local variable *byz[x]*, to a common value. Furthermore, if *general* is reliable, this common value is *d0[g]*, the initial value of one of *general*'s variables. The solution is complicated by the fact that unreliable processes send arbitrary messages. Since reliable processes cannot be distinguished from the unreliable ones, the straightforward algorithm—*general* transmits its initial value to all processes and every reliable process *u* assigns this value to *byz[u]*—does not work, because *general* itself may be unreliable, and hence may transmit different values to different processes.

This problem models a group of generals plotting a coup. Some generals are reliable and intend to go through with the conspiracy while others are feigning cooperation and in fact will support the incumbent ruler when the action starts. The problem is to determine which generals are reliable and which are not.

This model is clearly relevant to the more sophisticated information sources that might arise in the future, such as e-commerce brokers—software agents that negotiate on a user's behalf with a variety of Web sites trying to find a site that best meets the user's criteria. The agent could use a Byzantine generals' model to evaluate which sites were reliable and which were unreliable before ultimately deciding on a transaction.

In a pump-and-dump scheme, a group can agree to post misleading stories on several Web sites and newsgroups.

Linguistic analysis. Stylistic techniques from linguistics are also potential tools for determining the likelihood that the documents being analyzed are authentic. For a set of documents authored by one or more people using multiple pseudonyms, the tool could partition the documents into subsets of papers, all belonging to the same author.

The main idea is to embed the given document into a finite dimensional linear feature space of stylistic language usage. By performing cluster and other types of analyses on the writing and linguistic style, the tool might be able to establish which sections of the documents are stylistically similar and so are likely to be by the same writer.

Linguistic analysis only works with messages of consistent length and sufficient words, but for that type of message, it could determine with high confidence the author's stylistic characteristics or the source.[10]

Linguistic analysis might have prevented the Lebed hack by alerting users that the messages supposedly sent from different people were in reality from one person. After receiving an alert from the information trajectory model, AOL could have used linguistic analysis to examine all postings similar to the Lebed message. If the comparison found common authorship, AOL could have e-mailed a warning to its users to beware that postings about the stock in question might be part of a pump-and-dump scheme.

As a new threat, cognitive hacking requires new countermeasures. Source authentication, information trajectory modeling, Ulam games, the Byzantine generals model, collaborative filtering, and linguistic analysis are relatively mature technologies in the context of applications such as e-commerce. However, these measures are immature as applied to preventing misinformation and detecting user behavior. In applying information

Linguistic analysis could determine that messages supposedly sent from different people were in reality from one person.

trajectory modeling, for example, how can we distinguish the stock price fluctuations that result from legitimate stock analysts' reports from those caused by a cognitive hack?

Legal issues are another concern. Users currently have little protection against the consequences of attacks. Often, the penalties for spoofing and defacement are light or nonexistent. Relevant laws that apply to other media should apply to the Internet, but the application of the law to cognitive hacking and other Internet-related areas is volatile.

The events of 11 September 2001 have only made the situation more unpredictable, as the balance between privacy and security has shifted toward security. More legislation affecting this area must be enacted, and the associated case law will continue to evolve. Until then, users must be constantly alert. ■

Acknowledgments

Support for this research was provided by a Department of Defense Critical Infrastructure Protection Fellowship grant with the Air Force Office of Scientific Research, F49620-01-1-0272; Defense Advanced Research Projects Agency projects F30602-00-2-0585 and F30602-98-2-0107; and the Office of Justice Programs, National Institute of Justice, Department of Justice award 2000-DT-CX-K001 (S-1).

The views in this document are those of the authors and do not necessarily represent the official position of the sponsoring agencies or of the US government.

References

1. C.E. Landwehr, "Formal Models of Computer Security," *Computing Surveys*, vol. 13, no. 3, 1981, pp. 247-278.
2. D. Denning, *Information Warfare and Security,* Addison-Wesley, Reading, Mass., 1999.
3. B. Schneier, "Semantic Attacks: The Third Wave of Network Attacks," *Crypto-Gram Newsletter*, 14 Oct. 2000; http://www.counterpane.com/crypto-gram-0010.html.
4. M. Libicki, "The Mesh and the Net: Speculations on Armed Conflict in an Age of Free Silicon," Nat'l Defense University McNair Paper 28, 1994; http://www.ndu.edu/u/inss/macnair/mcnair28/m028cont.html.
5. C. Lynch, "When Documents Deceive: Trust and Provenance as New Factors for Information Retrieval in a Tangled Web," *J. Am. Soc. Information Science & Technology*, vol. 52, no. 1, 2001, pp. 12-17.
6. S.M. Ulam, *Adventures of a Mathematician,* University of California Press, Berkeley, Calif., 1991.
7. D. Mundici and A. Trombetta, "Optimal Comparison Strategies in Ulam's Searching Game with Two Errors," *Theoretical Computer Science*, vol. 182, nos. 1-2, 15 Aug. 1997, pp. 217-232.
8. R. Yahalom, B. Klein, and T. Beth, "Trust Relationships in Secure Systems—A Distributed Authentication Perspective," *Proc. IEEE Symp. Research in Security and Privacy*, IEEE Press, New York, 1993, pp. 156-164.
9. K.M. Chandy and J. Misra, *Parallel Program Design: A Foundation*, Addison Wesley, Reading, Mass., 1988.
10. J.R. Rao and P. Rohatgi, "Can Pseudonymity Really Guarantee Privacy?" *Proc. 9th Usenix Security Symp.*, Usenix, Berkeley, Calif., 2000; http://www.usenix.org/events/sec2000/technical.html.

George Cybenko *is the Dorothy and Walter Gramm Professor of Engineering at Dartmouth College. His research interests include signal processing, distributed information systems, and computational intelligence. He received a PhD in mathematics from Princeton University. Contact him at george.cybenko@dartmouth.edu.*

Annarita Giani *is pursuing a PhD in computer engineering at the Institute for Security Technology Studies and Thayer School of Engineering at Dartmouth College. Her research interests include communication theory, signal analysis, and computer security. She received a Laurea in applied mathematics from the University of Pisa. Contact her at annarita.giani@dartmouth.edu.*

Paul Thompson *is a senior research engineer at the Institute for Security Technology Studies and Thayer School of Engineering at Dartmouth College. His research interests include document retrieval, information extraction, and computer security. He received a PhD in library and information studies from the University of California, Berkeley. Contact him at paul.thompson@dartmouth.edu.*