Herding Cats:
Practical Security Tips for a Wacky World

March 2008

BrandenWilliams
SECURE BUSINESS GROWTH

Have you ever wanted to see if sensitive data your company protects exists outside of designated areas? Maybe you are looking for Personally Identifiable Information (PII), intellectual property, or cardholder data that might be sitting around in flat files. I suggest turning to Grep[1], a GNU searching tool that is included on most Unix based operating systems (and there are MS ports)!

Grep can use the power of regular expressions to quickly search for patterns in files. The results obtained will help you triage data leakage that may occur through the normal course of business.

I was recently working with a customer who found additional consumer information in batch files they were receiving from a financial institution. I'm not talking about the consumer's hair color or meal preference. I'm talking about their social security number. Our customer had no idea they were receiving this information and yet they would likely be liable for its disclosure if a breach occurred.

The method described below is not foolproof; however, it will find the majority of files that would contain sensitive data. It will not reliably (without additional tweaking) find information encoded inside binaries or database files, but you can use it to look through compressed files[2] and flat files.

Let's say you wanted to take a cursory look at data that could contain social security numbers. Your grep command would look like[3]:

```
grep -rl "[[:digit:]]\{3\}[-,[:space:]]\?[[:digit:]]\{2\}
[-,[:space:]]\?[[:digit:]]\{4\}" / > files-to-triage
```

In this command, we are recursively (-r) looking for any file that has nine digits in a row, with or without delimiters, and list those filenames (-l) into a file called files-to-triage. After doing some test runs with this, I discovered that this will match quite a few files that do not have any of this data, but I think I could filter those out pretty quickly.

What if you wanted to look for stray cardholder data? To find most variations of valid cardholder data (apologies if I missed any) that would occur in a solid string of numbers or blocks of four separated by dashes or spaces, you would type:

```
grep -rl "\(\(4[[:digit:]]\{3\}\)\|\(5[1-5][[:digit:]]\
{2\}\)\|\(6011\)\)[-,[:space:]]\?[[:digit:]]\{4\}
[-,[:space:]]\?[[:digit:]]\{4\}[-,[:space:]]\?[[:digit:]]\
{4\}\|3[4,7][[:digit:]]\{13\}\|3[0,6,8][[:digit:]]\{12\}" / |
mail -s "Sensitive Data Report for `uname -a`" sensitive_data_
reporting@company.com
```

This is what it looks like to write a regular expression that will search for most known types of credit card data. Regular expressions look like a work of art when viewed by the trained eye, but like hieroglyphs for the rest of us.

FOOTNOTES

[1] *http://www.gnu.org/software/grep/*

[2] *To do this, you need to use "zgrep."*

[3] *Note: you may have slight syntax differences in your regular expressions. Non-maintained versions of operating systems need not apply.*

But what is that new fancy piece on the end? Say you wanted to run this regularly and email it to a report gathering mailbox. Provided your server has the mail program installed and its communication with a mail server is not blocked by firewall rules, this may be acceptable.

The automation can get pretty sophisticated from here. You could enhance the regular expression more or further parse the output to remove false positives, have the information dropped into a database or aggregate report on a daily basis with scripts, or simply just run them on a schedule from cron.

Now keep in mind, this is not a foolproof approach. By no means the most efficient way to search for these types of data. There are several commercial packages on the market today to choose from that will do what you see above, and much more. But if you are looking for a starting point to see if you have a problem, using open source tools like Grep can be a cost efficient way to see how deep the hole really goes.

Contact information for requests for permission to reproduce or distribute materials available through this course are listed below.

About the Author:

Branden R. Williams, CISSP, CISM, CPISA/M, has been making a name for himself in the Information Technology and Security arena since 1994, as a high school Junior. Now, a graduate of University of Texas, Arlington earning his BBA in 2000 with a concentration in Marketing and the University of Dallas, where he earned an MBA in Supply Chain Management & Market Logistics, in 2004, Williams is sought after as both an Adjunct Professor and Information Technology & Security Strategy Leader in the corporate world.

Williams regularly assists top global retailers, financial institutions, and multinationals with their information security initiatives. Read his blog at or reach him directly at http://www.brandenwilliams.com/.