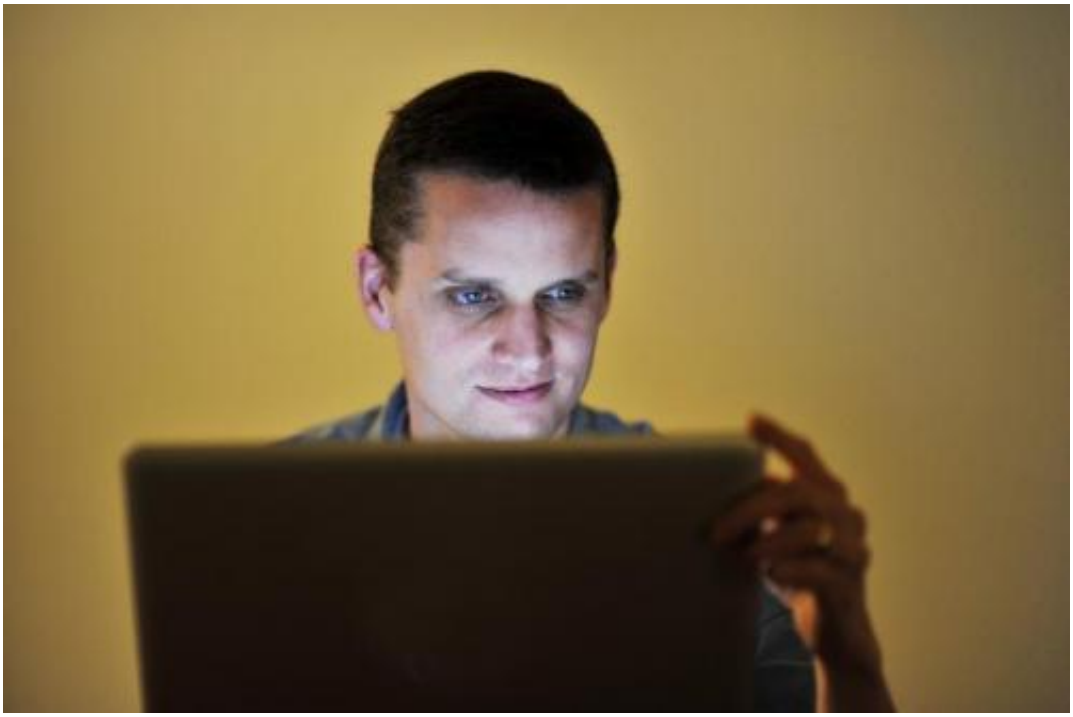
Computer scientist looks for bad guys in cyberspace

February 11 2013



Sandia National Laboratories computer science researcher Jeremy Wendt concentrates on working on a program to find potential targets of nefarious emails. Credit: Randy Montoya

(Phys.org)—The weakest link in many computer networks is a gullible human. With that in mind, Sandia National Laboratories computer science researcher Jeremy Wendt wants to figure out how to recognize potential targets of nefarious emails and put them on their guard.

His goal is to reduce the number of visitors that cyberanalysts have to check as possible bad guys among the tens of thousands who search Sandia websites each day.

Ultimately, he wants to be able to spot spear phishing. Phishing is sending an email to thousands of addresses in hopes a few will follow a link and, for example, fall for a scam offering millions of dollars to help a Nigerian prince wire money out of his country. Spear phishing, on the other hand, targets specific email addresses that have something the sender wants.

Wendt has developed algorithms that separate robotic web crawlers from people using browsers. He believes his work will improve security because it allows analysts to look at groups separately.

Even if an outsider gets into a Sandia machine that doesn't have much information, that access makes it easier to get into another machine that may have something, Wendt said.

"Spear phishing is scary because as long as you have people using computers, they might be fooled into opening something they shouldn't," he said.

Identifying malicious intent

Sandia cybersecurity's Roger Suppona said the ability to identify the possible intent to send malicious content might enable security experts to raise awareness in a potential target. "More importantly, we might be able to provide specifics that would be far more helpful in elevating awareness than would a generic admonition to be suspicious of incoming email or other messages," he said.

Wendt, in the final stretch of a two-year Early Career Laboratory

Directed Research and Development grant, presented his work at a Sandia poster session.

Wendt has looked into behaviors of web crawlers vs. browsers to see if that matches how computers identify themselves when asking for a webpage. Browsers generally say they can interpret a particular version of HTML—HyperText Markup Language, the main language for displaying webpages—and often give browser and operating system information. Crawlers identify themselves by program name and version number. A small number Wendt labels "nulls" offer no identification, perhaps because the programmer omitted that information, perhaps because someone wants to hide.

What Wendt is looking for is a computer that doesn't identify itself or said it's one thing but behaves like another and trolls websites in which the average visitor shows little interest.

Going to an Internet site creates a log of the search. Sandia traffic is about evenly divided between web crawlers and browsers. Crawlers tend to go all over; browsers concentrate on one place, such as jobs.

Crawlers, also known as bots, are automated and follow links like Google or Bing do. "When we get crawled by a Google bot, we aren't being crawled by one visitor, we're being crawled by several hundreds or thousands of different IP addresses," Wendt said. An IP or Internet Protocol address is a numerical label assigned to devices on a computer network, identifying the machine and its location.

Distinguishing bots and browsers

Since Wendt wants to distinguish bots from browsers without having to trust they're who they say they are, he looked for ways to measure behavior.

The first measurement deals with the fact bots try to index a website. When you type in search words, the crawler looks for pages associated with those words, disregarding how they're arranged on a page. That means a bot pulls down HTML files far more often.

Wendt first looked at HTML downloads. Bots should have a high percentage. Browsers pull down smaller percentages.

More than 90 percent of the nulls pulled down nothing but HTML—typical bot behavior.

A single measurement wasn't enough, so Wendt devised a second based on another marker of bot behavior: politeness.

Bots could suck down webpages from a server so fast it would shut down the server to anyone else, he said. That might prompt the site administrator to block them.

So bots take turns. "They say, 'Hey, give me a page,' then they may crawl a thousand other sites taking one page from each," Wendt said. "Or they might just sit there spinning their wheels for a second, waiting, and then they'll say, 'Hey, give me another page.'"

Some behavior is 'bursty'

Browsers go after only one page but want all images, code, and layout files for it instantly. "I call that a burst," he said. "A browser is bursty; a crawler is not bursty." Bursts equal a certain number of visits within a certain number of seconds.

Ninety percent of declared bots had no bursts and none had a high burst ratio. Sixty percent of nulls also had no bursts, lending credence to Wendt's identification of them as bots.

But 40 percent showed some bursty behavior, making them hard to separate from browsers. However, normal browser behavior also falls within set parameters. When Wendt combined both metrics, most nulls fell outside those parameters.

That left browsers who behaved like bots. "Now, are all these people lying to me? No. There could be reasons somebody would fall into this category and still be a browser," he said. "But it distinctly increases suspicions."

He also looked at IP addresses. Unlike physical addresses, IP addresses can change. Say you plug your laptop into the Internet at a coffee shop, which assigns you an IP address. After you leave, someone else shows up and gets the same IP address. So an IP address alone doesn't necessarily distinguish users.

There's another identifier: a particular browser on a particular operating system, which leads to what's called a user agent string. There are thousands of distinct strings.

IP addresses and user agent strings can collide, but Wendt said odds are dramatically lower that two people will collide on the same IP address and user agent string within a short period such as a day. That tells him they're probably different people.

Now he needs to bridge the gap between splitting groups and identifying targets of ill-intentioned emails. He has submitted proposals to further his research after the current funding ends this spring.

"The problem is significant," he said. "Humans are one of the best avenues for entering a secure network."

Provided by Sandia National Laboratories