# SEISE tool uses semantic gaps to detect website promotional attacks

May 19 2016


Georgia Tech Ph.D. student Xiaojing Liao and Professor Raheem Beyah are shown with a typical promotional infection, this one advertising essays for sale. Credit: Credit: John Toon, Georgia Tech

By detecting semantic inconsistencies in content, researchers have

developed a new technique for identifying promotional infections of websites operated by government and educational organizations. Such attacks use code embedded in highly-ranked sites to drive traffic to sketchy websites selling fake drugs, counterfeit handbags and plagiarized term papers - or installing drive-by malware.

The new technique, known as Semantic Inconsistency Search (SEISE), uses natural language processing to spot the differences between a compromised site's expected content and the malicious advertising and promotional code. Using SEISE, the researchers found 11,000 infected sites among non-commercial top-level sponsored .edu, .gov and .mil domains worldwide, and are working to extend the method to other domains.

The research was supported by the U.S. National Science Foundation and Natural Science Foundation of China. It will be described in a presentation May 25, 2016 at the IEEE Symposium on Security and Privacy in San Jose, California. SEISE was developed by researchers from the Georgia Institute of Technology, Indiana University and Tsinghua University in China.

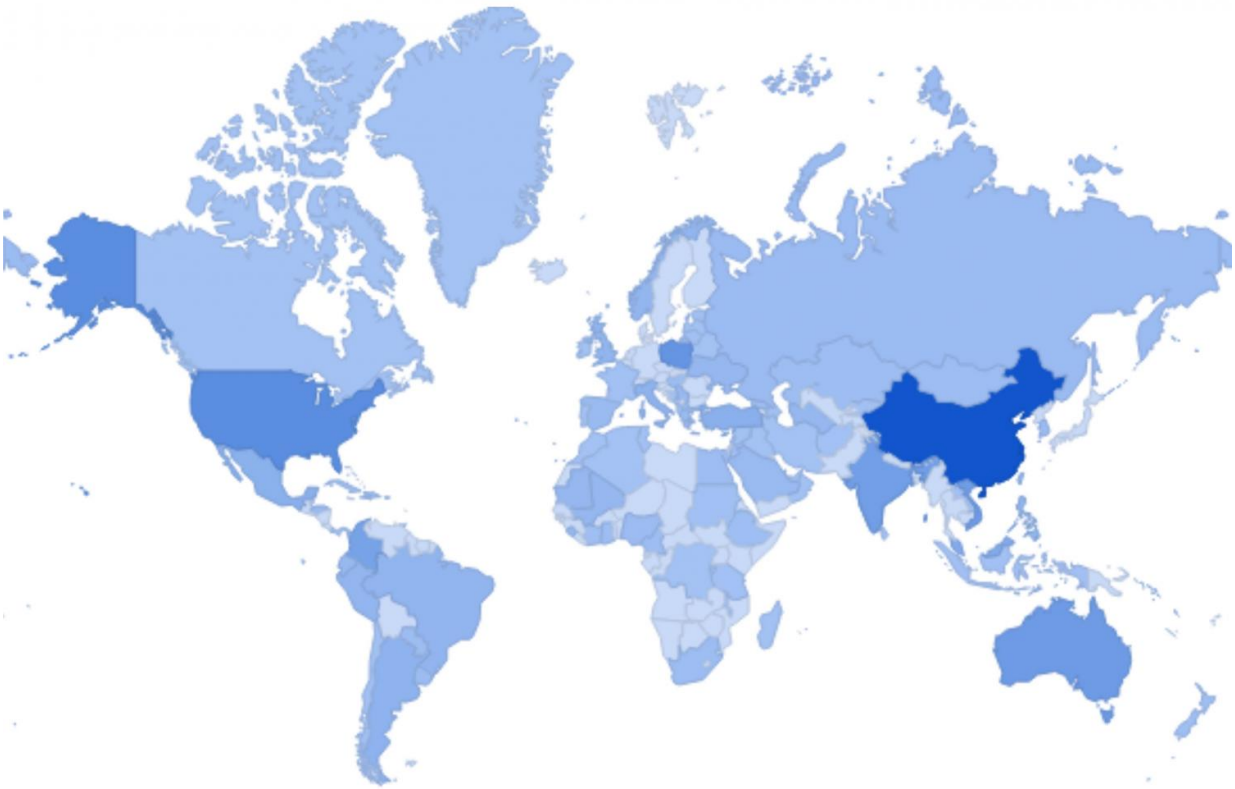
"The basic idea behind promotional infection is to attack websites that are highly-ranked and to leverage their importance to promote various things, most of them illegal," explained Raheem Beyah, who is the Motorola Foundation Professor in Georgia Tech's School of Electrical and Computer Engineering. "The bad content is nested into the prominent site to leverage the traffic of that domain. That gives the attackers a doorway to whatever they are promoting."

Essentially, said Beyah, the attackers are stealing the site's good name, even if they don't install malware or otherwise inflict harm on web visitors.

"The attackers essentially become part of the prominent website's brand and share in the ranking they have," he added. "It's like setting up operations inside a well-known coffee shop chain. The attacker leverages the brand by becoming co-located with it."

The promotional attacks can be difficult to detect, especially if they don't contain malicious computer code. But the semantic differences between the host site and the attacker's code can tip off the SEISE algorithm. Once it has characterized the content expected on a website - educational information on an .edu page, for example - the pitches for gambling or inexpensive prescription drugs become obvious.

"If you are visiting the website for a prestigious university, you don't expect to see information promoting casino gambling," said Beyah. "If we expect one thing from the website and see something significantly different, there is a huge semantic gap that we can detect."



Map shows the geolocation distributions of the sponsored top-level domains across 141 countries. Credit: Xiaojing Liao, Georgia Tech

SEISE doesn't have to review an entire site to determine what should be there; it can sample the pages to learn context that makes attacker terms stand out. Because their domain purposes are clear and well established, the researchers began with education and government websites. They now hope to extend the automated approach to commercial and other domains whose intended purposes may be less consistent.

"We are trying to figure out how to get the context right for these domains so we can help companies detect these infections," Beyah said. "There's no reason to believe that the commercial domains are any less attractive to attackers than the non-commercial ones."

Beyah and Georgia Tech Ph.D. student Xiaojing Liao began the work by using Google searches to find sites with known "bad words" denoting illicit products. They then utilized natural language processing to find terms associated with these known bad words, which were then used to train the SEISE before it was sent out to analyze 100,000 domains for the presence of the illicit terms. The approach identified 11,000 infected sites with a false detection rate of just 1.5 percent and coverage of more than 90 percent.

SEISE found promotional infections on the websites of top U.S. universities and government agencies, though the problem was truly worldwide, with three percent of .edu and .gov sites infected. Of the infected websites noted, 15 percent were in China and six percent were in the United States.

Sites are infected using proven attack techniques such as SQL injection, URL redirection and phishing to compromise the credentials of users, Beyah said. Though central websites of the organizations may be secure, pages of individual users and units may be more vulnerable - and still provide the prestige of the overall domain.

Existing techniques for detecting promotional infections rely on examining redirects and following links, or observing how sites change over time. But those techniques aren't scalable and can't be automated in the same way as the new semantic gap approach, Beyah said.

The researchers want to share their technique with the larger security community, and are discussing how best to make the algorithm available. "Our study shows that by effective detection of infected sponsored top-level domains (sTLDs), the bar to promotion infections can be substantially raised," the authors wrote in their paper.

About those 11,000 compromised webpages? The researchers are attempting to contact the operators of all 11,000 of them to share the bad news. "We have spent a lot of time contacting those folks and letting them know what we have found," Beyah said. "We're still in the process of doing that because there are so many."

Provided by Georgia Institute of Technology