A new kind of pub crawl

August 24 2012, By Angela Herring

Engin Kirda, an associate professor of information assurance at Northeastern, developed new software for detecting and containing malicious web crawlers. Photo: Dreamstime.

Websites like Facebook, LinkedIn and other social-media networks contain massive amounts of valuable public information. Automated web tools called web crawlers sift through these sites, pulling out information on millions of people in order to tailor search results and create targeted ads or other marketable content.

But what happens when "the bad guys" employ web crawlers? For Engin Kirda, Sy and Laurie Sternberg Interdisciplinary Associate Professor for Information Assurance in the College of Computer and Information Science and the Department of Electrical and Computer Engineering, they then become tools for spamming, phishing or targeted Internet attacks.

"You want to protect the information," Kirda said. "You want people to be able to use it, but you don't want people to be able to automatically download content and abuse it."

Kirda and his colleagues at the University of California–Santa Barbara have developed a new software call PubCrawl to solve this problem. PubCrawl both detects and contains malicious web crawlers without limiting normal browsing capacities. The team joined forces with one of the major social-networking sites to test PubCrawl, which is now being used in the field to protect users' information.

Kirda and his collaborators presented a paper on their novel approach at the 21st USENIX Security Symposium in early August. The article will be published in the proceedings of the conference this fall.

In the cybersecurity arms race, Kirda explained, malicious web crawlers have become increasingly sophisticated in response to stronger protection strategies. In particular, they have become more coordinated: Instead of utilizing a single computer or IP address to crawl the web for valuable information, efforts are distributed across thousands of machines.

"That becomes a tougher problem to solve because it looks similar to benign user traffic," Kirda said. "It's not as straightforward."

Traditional protection mechanisms, like a CAPTCHA, which operates on an individual basis, are still useful, but their deployment comes at a

cost: Users may be annoyed if too many CAPTCHAs are shown. As an alternative, nonintrusive approach, PubCrawl was specifically designed with distributed crawling in mind. By identifying IP addresses with similar behavior patterns, such as connecting at similar intervals and frequencies, PubCrawl detects what it expects to be distributed web-crawling activity.

Once a crawler is detected, the question is whether it is malicious or benign. "You don't want to block it completely until you know for sure it is malicious," Kirda explained. "Instead, PubCrawl essentially keeps an eye on it."

Potentially malicious connections can be rate-limited and a human operator can take a closer look. If the operators decide that the activity is malicious, IPs can also be blocked.

In order to evaluate the approach, Kirda and his colleagues used it to scan logs from a large-scale social network, which then provided feedback on its success. Then, the social network deployed it in real time, for a more robust evaluation. Currently, the social network is using the tool as a part of its production system. Going forward, the team expects to identify areas where the software could be evaded and make it even stronger.

Provided by Northeastern University