Spam filter design to benefit from internet routing data

September 12 2006

Internet service providers could better fight unwanted junk email by addressing it at the network level, rather than using currently available message content filters, says Georgia Tech College of Computing Assistant Professor Nick Feamster. Credit: Photo by Gary Meek

A database of more than 10 million spam email messages collected at just one Internet "spam sinkhole" suggests that Internet service providers could better fight unwanted junk email by addressing it at the network level, rather than using currently available message content filters.

Also, the research – conducted at the Georgia Institute of Technology's College of Computing -- identified two additional techniques for combating spam: improving the security of the Internet's routing infrastructure and developing algorithms to identify computers'

membership in "botnets," which are groups of computers that are compromised and controlled remotely to send large volumes of spam. The findings are now directing the researchers' design of new systems to stem spam.

"Content filters are fighting a losing battle because it's easier for spammers to simply change their content than for us to build spam filters.," said Nick Feamster, a Georgia Tech assistant professor of computing. "We need another set of properties, not based on content. So what about network-level properties? It's harder for spammers to change network-level properties."

Feamster and his Ph.D. student Anirudh Ramachandran will present their findings on Sept. 14, 2006 in Pisa, Italy, at the Association for Computing Machinery's annual flagship conference of its Special Interest Group on Data Communication (SIGCOMM).

From 18 months of Internet routing and spam data the researchers collected in one domain, they have learned which network-level properties are most promising for consideration in spam filter design. Specifically, they learned that:

-- Internet routes are being hijacked by spammers;

-- they can identify many narrow ranges within Internet protocol (IP) address space that are generating only spam, and

-- they can identify the Internet service providers (ISP) from which spam is coming.

"We know route hijacking is occurring," Feamster said. "It's being done by a small, but fairly persistent and sophisticated group of spammers, who cannot be traced using conventional methods."

Route hijacking works like this: By exploiting weaknesses in Internet routing protocols, spammers can steal Internet address space by briefly advertising a route for that space to the rest of the Internet's routers. The spammers can then assign any IP address within that address space to their machines. They send their spam from those machines and then withdraw the route by which they sent the spam. By the time a recipient files a complaint related to this IP address, the route is gone and the IP address space is no longer reachable.

"Even if you're watching the hijack take place, it's difficult to tell where it's coming from," Feamster explained. "We can make some good guesses. But Internet routing protocols are insecure, so it's relatively easy for spammers to steal them and hard for us to identify the perpetrators."

Feamster and researchers elsewhere are actively working to improve the security of Internet routing protocols, he added.

Better spam filtering will also result from a system, which Feamster hopes to design, based on collaborative, network-level filtering among ISP operators.

"Within the single domain that we are studying, it's interesting that you don't see the same IP addresses repeatedly being used to send spam to that domain," Feamster said. "So ISP operators need to be able to securely share information about IP addresses associated with spam."

In addition to studying network-level properties of spam, Ramachandran and Feamster compared their lists of IP addresses used to send spam against eight frequently used "blacklists" compiled by network operators to help filter spam.

"We found that these blacklists listed IP addresses for only about half of the spam being sent using route hijacking," Feamster said.

"The best case scenario is that these blacklists are still missing IP addresses from which at least 20 percent of spam is sent…. This 20 percent rate of false negatives is likely to cause a high percentage of false positives, and so this approach may also cause a lot of legitimate email to be mistakenly tagged as spam."

The researchers also plan to use this finding in the spam filter development efforts, Feamster added. Meanwhile, the researchers are continuing to collect Internet routing and spam data.

"It's always nice to have long-term data to help us see trends," Feamster noted. "These are valuable studies that help us see if people's behavior changes over time."

Indeed, it has in this case. The rate of spam has nearly doubled in the past two years in the one domain where the researchers collected their routing data for this study.

Source: Georgia Institute of Technology