

HOW COLLABORATIVE FILTERING CAN STOP FUTURE FORMS OF MESSAGING ABUSE

**Vipul Ved Prakash, Founder
and Chief Scientist, Cloudmark, Inc.**
**Adam J. O'Donnell, Senior Research
Scientist, Cloudmark, Inc.**

ABSTRACT:

Conventional anti-virus software relies on a staff of researchers to isolate and analyze viruses, identify them with a fingerprint, and then write and test code and rules to block them. This process takes up to 24 hours and often blindly blocks many legitimate messages with attached executable code. In contrast, Cloudmark's Collaborative Security Network™ (CCSN) uses a fingerprinting algorithm to identify each incoming message, combined with a reputation-based, trusted community of real-time users to identify malicious viruses. Using the Cloudmark Trust Evaluation System™ (TeS), Cloudmark is able to let trustworthy, credible users identify viruses. Cloudmark's virus fingerprinting algorithm automates the time-intensive reverse engineering analysis of conventional technologies, allowing its system to identify and squelch new worms and virus strains in real time. The Cloudmark technology is language-agnostic, format-agnostic, representation-agnostic, and protocol-agnostic, making it particularly suited to combat all forms of malicious content.

Individuals who use messaging services, such as e-mail, SMS, and MMS, are rather good at discerning the difference between content that is spam and content that is legitimate. When users are pooled together and allowed to “vote” on which content is or is not spam, the quality of their opinions is astonishingly high. The concept of pooling individual opinions regarding a piece of data is known as collaborative filtering.

The principle of community-based collaborative filtering was critical in the design of our message security system, known as the Cloudmark Collaborative Security Network. While the design is quite complex, the concepts behind its operation are simple. Users vote on whether messages are “spam” or “not spam” by reporting the fingerprints of messages to a central system. If enough members of the community agree that the message is spam, then the central system labels the fingerprint as truly representing spam. Each piece of e-mail received is checked against this list to determine if it is spam or not. Just as in a real-world human community, individuals who behave well by quickly and correctly identifying spam become trusted, and are rewarded by having their opinions weighted more heavily in the future.

In practice, the reporting, fingerprinting, and filtering operations are transparent to the service provider. Users report content as being “spammy” by clicking on a button in their e-mail or web mail client, which automatically computes the fingerprints and forwards them to the Cloudmark Network Classifier (CNC). Fingerprints that represent spam are rapidly redistributed to Cloudmark-enabled content filters which sit inside the SMTP stream.

The Cloudmark system is not limited to the collaborative filtering of e-mail. From the outset, the Cloudmark system has been designed to implement a general framework for filtration based on collaborative opinion. The core of the classification system is language-agnostic, format-agnostic, representation-agnostic, and protocol-agnostic. The Cloudmark approach can be rather easily adapted to almost any messaging medium by enabling feedback and filter hooks into the medium. For example, consider the following messaging systems that have the potential for abuse:

E-Mail-to-SMS Gateways

E-mail-to-SMS gateways are frequently used by mobile customers for a variety of legitimate purposes, but these systems have a high potential for abuse. A spammer is able to send content to a handset using the same tools they currently possess for e-mail spam, since the majority of mobile service providers possess e-mail to SMS gateways. Therefore, sending spam to a large number of SMS users is as easy as swapping their mailing lists. However, if handset users are able to report message abuse to the CNC, either by contacting the carrier directly, or by directly forwarding the SMS message to the CNC, the abuse could be prevented. This functionality is available today, since our content filtering technology can sit inside the SMTP stream used by E-mail-to-SMS Gateways.

Handset-to-Handset SMS Messages

Cloudmark's fingerprinting schemes are bearer-protocol agnostic, and can generate fingerprints for content regardless of how it is being delivered. In fact, any content that can be delivered as a single message and can be forwarded to the CNC can be filtered. For example, if a mobile user spammed other handsets the recipients can forward it to the CNC, which would generate a fingerprint for the content. The fingerprint would then be used to filter content sent to the SMS back-end engine.

MMS Messages

The reporting, fingerprinting, and filtration architecture offered by Cloudmark is not limited to text-based content. The technology treats underlying content equally, and as such, any binary data, including sound clips, pictures, and movies sent via MMS can be filtered. If the content is sent to a large number of users, and the users report the content to the CNC as abuse, a fingerprint will be generated and used to filter the content from all messages being sent to handsets.

If there is anything that the reader should take away from this description it is that the collaborative filtering architecture is not limited to combating conventional spam. Viruses can be identified by the community as well. Phishing, spyware, and a whole host of other abusive messaging that is recognizable by the average person can be solved by pooling their opinions together to form a communal opinion.

REFERENCES

1. V. V. Prakash and A. O'Donnell. *Fighting spam with reputation systems*. Queue, 3(9):36–41, 2005.