Identifying 'anonymous' email authors

March 8 2011

Benjamin Fung, a professor of Information Systems Engineering at Concordia University, has developed an effective new technique to determine the authorship of anonymous emails. Credit: Concordia University

A team of researchers from Concordia University has developed an effective new technique to determine the authorship of anonymous emails. Tests showed their method has a high level of accuracy – and unlike many other methods of ascertaining authorship, it can provide presentable evidence in courts of law. Findings on the new technique are published in the journal *Digital Investigation*.

"In the past few years, we've seen an alarming increase in the number of cybercrimes involving anonymous emails," says study co-author Benjamin Fung, a professor of Information Systems Engineering at Concordia University and an expert in data mining – extracting useful, previously unknown knowledge from a large volume of raw data. "These

emails can transmit threats or child pornography, facilitate communications between criminals or carry viruses."

While police can often use the IP address to locate the house or apartment where an email originated, they may find many people at that address. They need a reliable, effective way to determine which of several suspects has written the emails under investigation.

Fung and his colleagues developed a novel method of authorship attribution to meet this need, based on techniques used in speech recognition and data mining. Their approach relies on the identification of frequent patterns – unique combinations of features that recur in a suspect's emails.

To determine whether a suspect has authored the target email, they first identify the patterns found in emails written by the subject. Then, they filter out any of these patterns which are also found in the emails of other suspects.

The remaining frequent patterns are unique to the author of the emails being analyzed. They constitute the suspect's 'write-print,' a distinctive identifier like a fingerprint. "Let's say the anonymous email contains typos or grammatical mistakes, or is written entirely in lowercase letters," says Fung. "We use those special characteristics to create a write-print. Using this method, we can even determine with a high degree of accuracy who wrote a given email, and infer the gender, nationality and education level of the author."

To test the accuracy of their technique, Fung and his colleagues examined the Enron Email Dataset, a collection which contains over 200,000 real-life emails from 158 employees of the Enron Corporation. Using a sample of 10 emails written by each of 10 subjects – 100 emails in all – they were able to identify authorship with an accuracy of 80

percent to 90 percent.

"Our technique was designed to provide credible evidence that can be presented in a court of law," says Fung. "For evidence to be admissible, investigators need to explain how they have reached their conclusions. Our method allows them to do this."

The new authorship identification technique was developed in collaboration with Mourad Debbabi, a Concordia expert in cyber forensics, and PhD student Farkhund Iqbal. "Our different backgrounds allowed us to apply data mining techniques to real-life problems in cyber forensics," says Fung. "This is an excellent illustration of how effective interdisciplinary research can be."

More information: Cited research: www.dfrws.org/2008/proceedings/p42-iqbal.pdf

Provided by Concordia University