Lecture 31: Filtering Out Spam

Lecture Notes on "Computer and Network Security"

by Avi Kak (kak@purdue.edu)

April 6, 2017
6:16pm

Goals:

- Spam and computer security
- How I read my email
- The acronyms MTA, MSA, MDA, MUA, etc.
- Structure of email messages
- How spammers alter email headers
- A very brief introduction to regular expressions
- An overview of procmail based spam filtering
- Writing Procmail recipes

CONTENTS

31.1: SPAM AND COMPUTER SECURITY

- Spam is a major source of malware that infects individual computers and, sometimes, entire networks.

- Much spam tries to lure you into clicking on URLs of websites that serve as hosts for viruses, worms, and trojans. Consequences of inadvertently downloading such software into your computer can be deadly — as previously described in Lecture 30.

- In addition to the dangerous spam that may try to steal information from your computer or turn it into a spambot for spreading even more spam, there is also another kind of spam these days: This consists of email generated by legitimate businesses and organizations that you either have no interest in reading or have no time for following up on. [For example, half of my spam consists of unsolicited messages sent to me by marketing companies, public relations houses, government agencies, university departments advertising their activities, and students in various parts of the world seeking to come to Purdue. Even just opening all of these messages would consume a significant portion of each day.]

- I am not much of a believer in spam filters that carry out a statistical analysis of email to decide whether or not it is spam.

These filters are also sometimes called Bayesian filters for blocking spam. A statistical filter with sufficiently low "falses" to suit my tastes would require too many samples of a certain type of spam before blocking such messages in the future. On the other hand, with a regular-expression based filter, once you see a spam message that has leaked through, it is not that difficult to figure out variations on that message that the spammers may use in the future. In many cases, you can design a short regular expression to block the email you just saw and all its variations that the spammer may use in the future in just one single step.

- Based on my personal experience, and in line with my above stated observation, you can design nearly 100% effective spam filters with tools that carry out regular-expression based processing of email messages. [A spam filter is close to 100% effective if it traps close to 100% of what **YOU** consider to be spam and lets through close to 100% of the messages that **YOU** consider legitimate.]

- Spam filter that are close to 100% effective for **your** specific needs in the sense defined above can only be built slowly. My spam filter has evolved over several years. It needs to be tweaked up every once in a while as spammers discover new ways of delivering their unwelcome goods.

31.2: HOW I READ MY EMAIL

- These days most folks read their email through web based mail clients. If you are at Purdue, in all likelihood, you log into Purdue's webmail service to check your email. Or, perhaps, you have it forwarded to your email account at a third party service such as that provided by gmail or yahoomail. **This way of reading email is obviously convenient for, say, English majors. However, if you happen to be a CS or a CompE major, that is not the way to receive and send your email.**

- The web based email tools can only filter out standard spam — this is, the usual spam about fake drugs, about how you can enlarge certain parts of your body, and things of that sort. But nowadays there is another kind of spam that is just as much of a nuisance. As mentioned in the previous section, you have generally well-meaning folks (and organizations) who want to keep you informed of all the great stuff they are engaged in and why you should check out their latest doings. These include local businesses, marketing companies, PR folks, etc. **When you write your own spam filter, you can deal with such email in a much more selective manner than would otherwise**