Finn Brunton (us)
New York University
Postdoctoral Researcher
finnbr@gmail.com

An Infinite Continuum of Spewage: Bayesian Filtering and the Reinvention of Spam.

```
//    ... it is unreasonable to
//      assume that any finite number of samples can appropriately
//      represent an infinite continuum of spewage, so we can bound
//      the certainty of any meausre [sic] to be in the range:
//
//      limit: [ 1/featurecount+2 , 1 - 1/featurecount+2].
-- crm_markovian.c, crm114-20070810-BlameTheSegfault.src
```

"Norbert Wiener said if you compete with slaves you become a slave, and there is something similarly degrading about competing with spammers." The writer is Paul Graham, the prominent Lisp programmer; the quote is from his 2002 essay, "A Plan for Spam", one of the most influential documents in the the anti-spam movement. (Graham 2002) Influential for three reasons: first, because it suggested a way to get to grips with spam, to turn it into an object; second, because it won, effectively destroying spam as it then existed, sidestepping its social complexities to attack it on a precise technical point; and finally, because it lost, the pure and elegant technical attack being based on a new set of design values and social assumptions, interstices into which spam moved, transforming itself in the process, and accidentally producing a literary experiment on the grandest scale in human history.

Formless

Spam in all its diverse modes – from email campaigns directed at people to bot-generated blogs to affect search engine results – bases its resiliency and strength on two sources, one deep and one shallow. The deep source of spam's vigor as a form is that spammers operate largely by taking existing "good things," technologically and socially, to unforeseen extremes, making it difficult to destroy their capacities without doing damage to much larger constituencies of users and institutions, as well as dearly held values embedded in the design of the Internet and the systems built on it. The shallow source is simply that spam is often very hard to clearly define, whether the goal is legal, political, technical or scientific. Like art or pornography, it has often been a matter of knowing it when you see it, and early projects to filter spam emails before they reached the inbox tried to generalize particular experience, using crude techniques like word blacklists and blocking groups of addresses, with very mixed results. Spammers could fake addresses, and cook up innocuous subject lines and new scams faster than some centrally maintained list could keep up; dull-edged blocking tools tended to result in far too many missed legitimate messages, breaking the open square of email up into small, Balkanized camps, stricken by constant conversational uncertainty (have I missed something important? Did the other receive my message?). Spammers as a group seemed similarly formless in their mores, beyond guilt and shame, their "crime" without adequate legal definition to deter them.

Quantified Language

Graham proposed applying a Bayesian filter to this problem, with a twofold goal: the filter would transform the words in email messages into probabilities of spam or not-spam, attacking the language of spam methods, the only area in which spammers could not hide their intentions. This filtering, done on an individual basis, would not stop all spam, just enough to dramatically raise the cost of a spammer's business, with far more messages needed to get a single response. "Spammers are businessmen," Graham averred, and whether criminal or legitimate would leave if the work stopped paying. The regularity of spam language became its weakness, as the Bayesian system, a very sophisticated method of inferring likelihood from past events, learned from every message marked "spam" and "not-spam." Spam was words like "madam" and "guarantee"; non-spam "although" and "evening." Wiener had worried that workers in competition with the automated production of machines would become no better than machines themselves, slaves; Graham meant to put spammers in the same position, competing with mechanical readers that would filter and discard their messages with relentless, inhuman attention, persistence, and acuity. The method worked well and was widely adopted. Combined with changing perceptions and increasingly effective police action, it effectively killed the 1990s culture of spam, with its limp pretense of respectability, and language from marketing and salesmanship, leaving the field to the smartest and most overt of the criminals.

Litspam

The weak points seized on by the remaining generation of spammers were several, of which the strangest was a direct attack on Bayesian filters with the automated production of seemingly meaningful language. The filters couldn't be too strict, for fear of discarding too many legitimate messages, and enough non-spam words could get the message through the filter to the inbox – but most words, very rarely used, had a spam/non-spam probability of 50/50 and would make no difference. There was source of language in use, however, ready-made for algorithmic processing and statistical analysis, letters of transit to get a spam message before a person's eyes: the text files of public domain literature. Thus historical archives, Sinclair Lewis, pirated e-books, epic poems, and a thousand forgotten authors were pressed into the service of getting credit card numbers, producing a ceaselessly refined corpus of messages that suggest the clumsily mechanized avatars of Burroughs and Brion Gysin, Tristan Tzara and Louis Zukofsky, spam's high modernism.

References

Graham, Paul (2002). A Plan for Spam [Online]. Retrieved from: http://www.paulgraham.com/spam.html [Accessed 4 June 2010]