Dimensions of a Reputation System

Designing a next-generation e-mail reputation system

Julian Mehnle

E-mail abuse has become an almost unmanageable problem: current statistics report about 85% of all e-mail sent as being spam, fraud, mal-ware, or a combination thereof, with a growing tendency. The prevalent approach to e-mail abuse control, *content-based* classification of messages, has a good efficacy record, however it does not facilitate the feeding-back of complaints to the providers whose infrastructure is being abused, and the content filters, which are growing ever more complex, require large amounts of processing power. Thus, *identity-based* AKA *reputation-based* classification of messages is considered "the next big step" in abuse control by the leading e-mail service providers and anti-abuse vendors, as evidenced by recent anti-abuse conferences of the e-mail industry.

There are reputation systems today already, such as *SPEWS*, *SpamHaus*, or *SpamCop*, which *assign binary reputation to individual IP addresses*, or sometimes entire net blocks. IP addresses, however, are a cheap throw-away commodity nowadays, given abusers' common practice of mass-hijacking end-users' PCs and building huge bot-nets. Also, very few of the existing reputation systems accept systematic feedback from end-users or respect end-users' subjective criteria for what is considered abuse and what is not. Due to these and other reasons, the efficacy of those basic reputation services is limited.

Given the growing overall focus on reputation systems, it seems to be a worthwhile effort to review the characteristics of existing reputation systems, take a closer look at possibilities for building on the new e-mail sender authentication methods, to draw parallels to other fields where reputation systems are deployed, such as search engines or community-fed website bookmark services, and to *devise and implement a next-generation reputation system*.

Variable Granularity

As mentioned before, one of the fundamental problems of most current e-mail reputation systems is that they operate merely on IP addresses, which are anonymous as well as cheap to acquire (from an abuser's point of view). At least, some systems escalate an IP address's bad reputation to its superior net blocks under certain conditions, exerting pressure not just on the owner of the system behaving abusively but also on the service provider hosting the system. However, as many such misbehaving systems merely are hijacked end-user PCs, there is little that ISPs can do about it (short of blocking outbound traffic on TCP port 25 flat, a measure that is usually way out of proportion). The approach thus misses the typical structure of today's spam operations.

Using the identity information created by e-mail sender authentication technologies such as *SPF*, *DKIM*, *PGP*, or *S/MIME*, reputation of IP address granularity can be lifted to the level of domains or even individual e-mail addresses. Assigning reputation to yet other identity types such as domains' registrar information (from the *WHOIS* system), and carrying over strong reputation to related identities, is also conceivable, thereby for example matching more closely spammers' strategy of registering their throw-away domains through just a few spammer-friendly registrars. Finding ways to combine existing and new identity types and draw relations among their reputation is a likely key element to the design of an advanced reputation system; for instance, some *statistical classifier* (such as a *Naïve Bayes classifier*[1]) could be applied to the set of (authenticated) sender identities of an e-mail message.

[1] It remains to be seen whether the "naïve" assumption of independent input features conflicts with the actual *lack* of independence of sender identities in an e-mail message. So far, a similar lack of independence of input words has not stopped text-based *Bayes* filters from effectively classifying messages as spam or ham anyway.

Aggregating Reputation Sources

Despite their relative simplicity, the number of existing reputation systems is plenty, as are the sets of objective assessment criteria they employ. Yet, so far there have been virtually no efforts to combine reputation data from multiple sources in non-trivial ways; most mail servers that use sender reputation to classify incoming mail merely use a list of DNS-based reputation systems in a "Three strikes (or even just one), and you're out!" mode. One of the few more sophisticated spam classification systems is *SpamAssassin*, which weights several reputation assessments (in addition to content-based tests) differently according to a profile configured by the system administrator or user. However, SpamAssassin also does not incorporate user feedback on the accuracy of its reputation sources.

A next-gen reputation system could *syndicate reputation data from many 3rd party sources and weight and aggregate assessments* (in addition to its own) according to user feedback and preferences in order to provide users with an partitioning of e-mail senders into "good" and "bad" that is as accurate as possible for their individual needs. Automatic evaluation of *accuracy statistics and user-supplied ratings could further lead to "meta reputation"*, thus facilitating a competition among selections and weightings of reputation sources for the benefit of the system's users. The specific methods by which reputation data, user preferences, and user feedback can best be aggregated, respecting different granularities of identities, would be a matter of research.

Distributed Operation

The fact that customary reputation systems such as *SpamHaus* frequently do not accept outside feedback is not least due to their centralized nature, which on the one hand allows them, and on the other hand makes them dependent on, direct and unmasked access to large "feed" mail streams, which they are usually coupled with. In contrast, decentralized architectures have inherent privacy and reliability problems, as suspect messages, or even just identifying characteristics of them, cannot simply be distributed to other nodes without violating privacy, and anonymized and aggregated results can be difficult to trust.

Some of the features of such an advanced reputation system (such as meta reputation) may allow it a degree of distributedness not feasible for customary reputation systems, whereas other features (such as a sender-identity-based statistical classifier) may make a central design more desirable. Thus it is necessary to find a good compromise between a centralized and a distributed architecture.