The Internet Mail System from the Network's Perspective

Dominik Schatzmann*, Wolfgang Muhlbauer, Thrasyvoulos Spyropoulos,
Xenofontas Dimitropoulos, Bernhard Plattner (*student author)
Computer Engineering and Networks Laboratory
ETH Zurich, Switzerland
{schatzmann, muehlbauer, spyropoulos, fontas, plattner}@tik.ee.ethz.ch

Poster Abstract

The growing popularity of electronic mail has not only influenced our communication habits but has also attracted criminals abusing the system to distribute e.g. spam [3, 4, 2]. As a countermeasure, new mail filtering techniques have been proposed, launching an ongoing arms race. Furthermore, new applications such as Ajax based webmail interfaces are developed to improve its usability.

We propose to use flow based network measurement data collected in the Internet backbone to analyze this evolving and important communication system. This approach allows to analyze large host populations using a rich set of mail applications over a long observation duration in a scalable and non intrusive way. However, the use of flow data includes several challenges, since the collected information is reduced to the number of exchanged packets and bytes per session.

In this context, we face the problem of how all e-mail related traffic can be reliably extracted from flow traces. Standard port based or flow-statistics based classification methods, like [1], provide only limited support. For example, a flow of a webmail application shares the same application port (80) and uses the same protocol (HTTP) as a flow of a normal WWW application. Therefore, it is difficult to distinguish these traffic classes using port or per flow based statistics. In addition, the acquisition of flow statistics often requires packet level data that is not available.

Instead of trying to work on per flow based characteristics, we propose to work on flow aggregates. In more detail, we propose to aggregate all flows of a socket [1], e.g. all flows of a possible webmail application, into a single set and analyze its set characteristics. This aggregation allows us to analyze and search for application dependent pattern that are spread over multiple flows and time. As one example, mail delivery traffic is expected to exhibit clear daily and weekly activity patterns biased by when users normally check their e-mail account(s). Furthermore, users check their mailbox several times per day, something that should be detectable as a low frequency pattern in the flow set. In addition of this user initiated actions, we assume to find a high frequency pattern generated by mail clients like Thunderbird, Outlook or mobile devices like IPhone, Blackberry that poll in regular short time intervals for new messages. This bimodal inter-flow time distirubtion could serve as singature for the classifier.

Beside of analyzing per socket characteristics we propose to compare the flow set of the unknown application with other flow sets of already classified applications. This additional step allows to exploit correlations that exist between different applications for the classification process. One such cross set characteristic could be the IP distance between already classified services and the unknown service. For example, the presence of other mail services close to the target flow set is an indicator that the unknown application could be mail related. Preliminary analysis of our flow sets show that, indeed, ordinary mail services and webmail services tend to be hosted in the same subnet or the same host. In addition, more advanced cross set characteristics, like analyzing similarities of the different user graph, can be used to further improve the classification result. For the case of mail traffic, we assume to find e.g. strong similarity between the user graph of the different mail delivery protocols since they share the same common user pool. Therefore their size and the geographical distribution of the clients IP addresses should be very similar.

The preliminary analysis of these correlation features for webmail applications shows that e.g. up to 90% of the webmail applications are indeed hosted in the same subnet with other mail service. Furthermore we found that webmail and IMAP traffic show similar polling frequencies. Based on these preliminary results we estimated that up to 60% of the observed HTTPS flows belong to webmail applications.

[1] identified by IP, protocol and port number

1. REFERENCES

[1] L. Bernaille, R. Teixeira, and K. Salamatian. Early application identification. In *CoNEXT '06*.

[2] D. Schatzmann, M. Burkhart, and T. Spyropoulos. Inferring spammers in the network core. In *PAM '09*.

[3] B. Stone. Breakfast can wait. the days first stop is online. *New York Times*, 2009.

[4] Y. Xie, F. Yu, K. Achan, R. Panigrahy, G. Hulten, and I. Osipkov. Spamming botnets: signatures and characteristics. In *SIGCOMM '08*.