Published on **O'Reilly** (http://www.oreilly.com/)
http://www.oreillynet.com/pub/a/sysadmin/2007/01/05/fingerprinting-mail-servers.html
See this if you're having trouble printing code examples

Fingerprinting the World's Mail Servers

by Ken Simpson and Stas Bekman
01/05/2007

This summer, the sales staff at MailChannels came to the dev team with an urgent request: "Can you tell us which companies are running Sendmail? If we could know that, it would be so much easier to sell our Sendmail-compatible product."

For those of us who understand the SMTP protocol, the answer was, of course, a resounding "Yes." Most mail servers announce their identity when you connect to them on TCP port 25. The dev team decided that this was a summer science project they just had to get on top of. We even gave the science project a name: PingedIn, and we hope to provide more dynamic content on our skeletal website.

Stats Porn Teaser

Before I get into the nuts and bolts of querying millions of mail servers and fingerprinting them, I want to review some of the interesting results of the survey.

Our Survey Approach

First, a note on our survey approach.

One idea we had when we began this project was to survey all the mail servers in the world, using various domain databases as our source material. We actually tried this by downloading the dot-com and dot-net registries from Verisign and pinging the first few million domains.

The problem with this approach is that all domains are not created equal. It may surprise you to learn that speculators and fraudsters own most of the world's domains. When these people own a domain, it's rare that they provide email service on it (we found that only 10 percent of domains provided MX records). The more significant issue is that parked domains and phishing domains are not really a useful sample set for us to examine. Remember, our goal with this project was to find leads for our sales team.

Surveying all the mail servers in the world is also a daunting technical task. There are tens of millions of dot-com domains--never mind dot-net, dot-org, and all the other TLDs. The data from a survey that large would fill terabytes of disk space, eat up monster bandwidth, and starve an already lean and mean startup of much-needed capital resources--not to mention taking many weeks to complete.

Rather than surveying all the domains in the world--which would have been a fun project--we chose instead to survey only those domains that have a real company behind them. We partnered with an old-school company data firm to get a list of 400,000 companies worldwide as our source material. It's not the perfect solution, but it suits our needs, and we can survey all 400,000 in a couple of hours without so much as raising a single abuse complaint.

With this little detail aside, where's the stats porn?

Open Source Still Dominates

Open source still dominates the global mail server software market. The changing nature of email threats such as spam and viruses are causing many companies to install an extra layer of protection at the network edge: witness Postini's rise to nearly 10 percent market share (Figure 1).

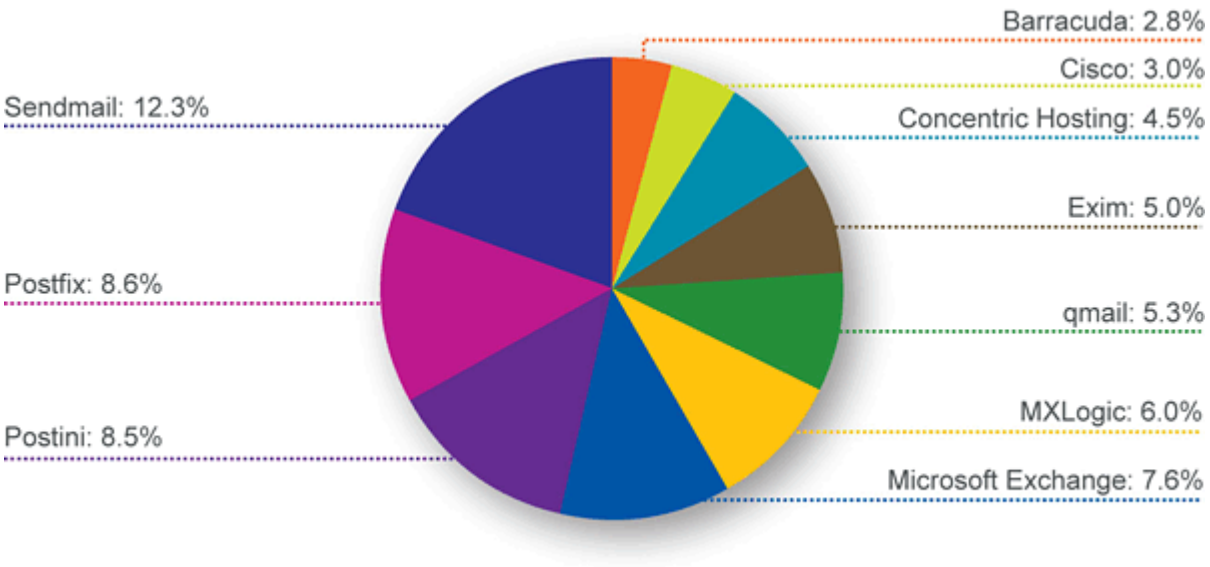

Figure 1. Open source mail server software dominates the market

Of the 400,000 domains we surveyed, 31.2 percent of them (still) receive their email via open source mail server software. Of these, the most popular by far is still the old guard, Sendmail (12.3 percent), with Postfix a relatively close second (8.6 percent). Exim and qmail are roughly tied (5.3 and 5.0 percent, respectively) in third place.

It's hard to tell what email security software lies behind the open source offerings, because generally these are behind the scenes where our pinger can't detect them. But it's a fair bet that many of the open source installations run the venerable SpamAssassin and that the remainder work with open source-friendly vendors such as Sophos, Proofpoint, and Symantec (Brightmail).

Next Up: Hosted Services

A surprising result of our survey has been the emergence of hosted email security services. These services prefilter email traffic for a domain before passing it on to the receiver's destination mail server--supposedly free of spam and viruses. Customers of hosted services pay a premium to get rid of their email security hardware, but the performance of these services is reportedly quite good.

Postini is the clear leader here, having secured 8.5 percent of the domains we surveyed. Next up is MXLogic, with 6.0 percent. Then there is Concentric Hosting with 4.5 percent, Earthlink with 2.7 percent, and Yahoo with 1.0 percent. The rest of the service providers own chunks too tiny to mention here.

The Evil Empire

Microsoft is in fourth place, with Exchange taking 7.6 percent of the domains we surveyed. You might think Microsoft would be doing better, considering that their web server software represents 31 percent (according to Netcraft's November 2006 survey). This weak result shows that Microsoft has a long way to go before it can establish itself credibly as an email boundary vendor.

Toasters, Microwave Ovens, and Other Appliances

There has been a lot of buzz over the past two years surrounding the rise of network appliances. The email space has not been immune to this trend, and now almost everyone seems to be getting into the appliance game. However, despite the buzz, the penetration of appliances remains very small indeed. The most significant appliance vendor in our survey is Barracuda, with a mere 2.8 percent of the market. IronPort is in second place with 0.8 percent. And Ciphertrust (now a division of Secure Computing) has 0.6 percent.

The appliance vendors will probably comment at the end of this article that while they don't own much of the broader market, they do own a significant chunk of specific segments--such as the Fortune 500. I'll get to that later.

The Art of Pinging

Fingerprinting mail servers is a relatively simple process. First, we connect to the server on port 25 (the SMTP port for inbound mail). Then we issue a bunch of commands and record what comes back. Depending on the responses, we can determine the kind of mail server running with a high degree of accuracy.

Banner Advertising

When you connect to a mail server on port 25, the first thing that comes back is the *banner*, a string of text that usually announces the mail server type and certain details about it. There is no standard mandating the kinds of information given by the banner, so banners vary wildly. Here are a few for reference:

```
Sendmail -- the canonical banner:
220 foo.com ESMTP Sendmail 8.13.6/8.13.6; Thu, 23 Nov 2006 13:35:44

Postini -- some banners include anti-spam legal notices:
220 Postini ESMTP 13 y6_8_4c1 ready. CA Business and Professions
 Code Section 1 7538.45 forbids use of this system for unsolicited
 electronic mail advertisement

Hotmail:
220 bay0-mc3-f6.bay0.hotmail.com Sending unsolicited commercial or
 bulk e-mail to Microsoft's computer network is prohibited. Other
 restrictions are found at
 http://privacy.msn.com/Anti-spam/. Violations will result in use of
 equipment located in California and other states. Thu, 23 Nov 2006
 10:38:21 -0800
```

```
Ciphertrust -- some are cagey about the mail server type:
220 SMTP Proxy Server Ready
```

Over the course of this project, we developed more than 300 rules that characterize the nearly 200 mail server types that make up the 400,000 domains we surveyed. The first rules we developed were simple regexes that matched on patterns in the banner. These allowed us to fingerprint about 50 percent of the domains in our survey set.

The other 50 percent remained mysterious to us because their banners revealed nothing useful. For example, as shown previously, Ciphertrust's banner is just "220 SMTP Proxy Server Ready." This banner is present in many other mail servers. Cisco PIX firewalls hide the banner completely, returning responses like the informative:

```
220 **********************
```

DNA Analysis

Because the banner is easy to spoof and is not always very revealing, we had to dig deeper to fingerprint about half the domains in our survey. Our first idea was to analyze the responses to the EHLO and HELP commands.

EHLO stands for "extended HELO" (and HELO stands for "Hello"). Usually at the start of an SMTP session, the client will send EHLO followed by its domain name:

```
EHLO foo.com
```

The server then responds with a list of SMTP extensions it supports. Extensions support operations such as sending email in an encrypted SSL session. Here is a sample response to EHLO as generated by a Sendmail system:

```
250-mx1.foo.com Hello client.bar.com [192.168.0.1], pleased to meet you
250-STARTTLS
250 SIZE 83886080
```

Here is the response generated by Hotmail's ultra-customized MTA:

```
250-bay0-mc10-f14.bay0.hotmail.com (3.3.0.19) Hello [x.x.x.x]
250-SIZE 29696000
250-PIPELINING
250-8bitmime
250-BINARYMIME
250-CHUNKING
250-AUTH LOGIN
250-AUTH=LOGIN
250 OK
```

We sent EHLO to a bunch of well-known MTAs and built up a catalog of their responses. To fingerprint a host using EHLO, all we had to do was analyze the response to EHLO and compare it against each of the known responses. Whichever known response matched most closely was the best candidate.

Unfortunately, this approach didn't work so well. It turns out that EHLO responses depend strongly on the specific configuration of the system in question and that system administrators often change these configurations.

For example, two different Sendmail sites might return completely different EHLO responses if one of them, say, supports TLS-encrypted sessions, whereas the other supports size limitations but does not support TLS.

EHLO wasn't a complete loss. We found that we could achieve slightly better results than the banner alone by combining EHLO and the banner in computing a "best guess" of the target MTA. With plenty of tweaking, we increased our knowledge of the surveyed domains up from 50 percent to about 70 percent.

Switching to the Dark Side

After looking at EHLO, we added the HELP command in the hopes that we could improve our results significantly; it didn't help much (no pun intended).

What did help tremendously was to query for negative responses--that is, errors. MTAs may be cagey in their positive responses, but when errors occur, their responses are quite revealing.

For example, here is how Sendmail responds to a NULL character in a command:

```
HE\0LO
500 5.5.1 Command unrecognized: "HE"
```

In other words, it scans the line up to the null termination character and then can't understand what you meant by "HE." Ciphertrust appliances, on the other hand, go a bit farther:

```
HE\0LO
500 Command unrecognized:  he\0lo
```

We found dozens of patterns like this for each invalid command we tried sending and found the responses very accurately matched with particular MTAs.

Conclusions

By combining all the techniques in our arsenal--and particularly by looking at negative responses--we improved our fingerprinting ability to 85 percent. The remaining 15 percent of MTA types that we still cannot identify will soon fall as our sales guys call these customers and ask, "So, what MTA are you running?"

Ken Simpson founded and directs MailChannels.

Stas Bekman co-authored two books on mod_perl: mod_perl2 User's Guide and Practical mod_perl.

Return to O'Reilly SysAdmin