AwlWrongWay

Why does the AWL sometimes assign scores the "wrong way"?

It's a common misconception that the AWL should always contribute a positive adjustment to the score of spam, and a negative one to nonspam. However, that's not the case, and it's not a bug or misconfiguration of SA. In fact, the AWL wouldn't work as well as it does if it didn't assign scores in the "wrong way" sometimes.

Fundamentally, the AWL is a score averaging system. You can read the "How does It Work" section in AutoWhitelist for a more detailed explanation, but for now it should be sufficient to know that it works by keeping track of the historical average score for a given sender, and adjusting the score of the message towards their past average.

Now, given that it's an averaging system, it would be impossible for it to always assign negative scores to nonspam messages, unless the pre-AWL scores of those messages were always above the average. In the long run, that would only happen if the email scores from that sender were increasing, meaning that the average from that sender was constantly increasing.

Really the fact that the AWL sometimes assigns scores the "wrong way" for the type of mail isn't a problem, unless the AWL pushes the email across your tagging threshold in the wrong direction, which means that the average for that sender is bad and needs to be fixed. (The --add-to-whitelist and --add-to-blacklist command line switches are currently designed for this use, but check with the man page.)

For example let's say my friend sends me mostly emails in rough ballpark of score=1.0, but one this day sends an email that happens to trip off a negative scoring rule or two and winds up scoring a -4 before the AWL is factored in. The AWL will look at his past average of 1.0 and the current email of -4, and it will average them. The average of these scores is -1.5, so the AWL will add 2.5 points to the message. However, the email is still correctly categorized as not being spam.

However, if that same friend instead sent me an email that had some racy jokes in it, and scored a 7, the AWL would again average 1.0 and 7. The resulting average is 4, so the AWL would knock 3 points off the score in this case.

The same basic effects also happen with spam, where a past average of 10, and a current score of 20 can result in the AWL knocking 5 points off the score, resulting in a final score of 15.

As you can see from these examples, the AWL operates by averaging out the "score spikes" between different emails. By the very nature of averages, this means that it will push the high points down, and the low points up, but it will always push them towards the average for that sender. As long as the average for that sender is on the right side of the spam/nonspam fence, it will do its job nicely.

What can go wrong

Now, with that said, it IS possible for the AWL to be polluted and cause problems. Generally this is the result of past misconfiguration or scoring problems that have since been fixed, but the AWL retains the old average and causes score problems, pushing things onto the wrong side of the spam/ham threshold line.

AWL database entries contain pairs of a sender's e-mail address along with an IP address from which mail entered the site's trusted zone. It is essential that SpamAssassin extracts the correct client's IP address from Received header fields. In order to do so, the following parameters must be correctly configured: trusted_networks, internal_networks (and the more exotic msa_networks). A misconfiguration can cause an incorrect IP address to be stored in AWL records and used in lookups, potentially treating both internal and faked inbound sender addresses as belonging to the same network IP address space.

Another potential problem is that AWL only keeps the first two octets of an IPv4 address (a /16 network block). If a spammer happened to send his message from some zombiized computer in the same /16 network block as a valid correspondent, using his faked e-mail address, then both would share the same AWL record. Consequently, a future legitimate mail could receive inappropriately high spam score from AWL, or vice versa, a future spam could benefit from legitimate mail correspondence from the same /16 network address space. A variation of this same problem is when mail arrives over IPv6 - the AWL as of version 3.2.5 is unable to store such IP address to a database and consequently treats such mail as if it were unable to determine an IP address, using only a sender's e-mail address as a key.

Blind white- or blacklisting can cause large score values to be entered into AWL records. Such values take lots of messages to be eavened out. Make sure that white- or blacklisting or other rules with high scores do not apply to messages for which they were not intended. For whitelisting only use selective methods such as whitelist_from_rcvd, whitelist_auth, whitelist_from_dkim or whitelist_from_spf - never use a plain whitelist_from.

Solution

If you have this problem, you can use `spamassassin --remove-addr-from-whitelist` to remove any prior knowledge about a given address from the AWL database. If you consult the main spamassassin manpage, there are other commands to force an AWL entry towards the black or white, but use these somewhat cautiously.

If your AWL is pulling scores wildly in the wrong direction and you have no idea why, check your configuration for problems. This is especially true if you use add-on rules or custom rules. If an errantly scored email got averaged into the AWL once, it can happen again, and tweaking the AWL is only going to provide a short-term solution to a greater problem.

(by MattKettler)