You’ve Got Spam.

Some Notes on the Reliability of E-mail Message Filtering

Christian K. Hansen, PhD

Eastern Washington University

cdotkdothansenatieeedotorg

Abstract

Since e-mail became widely adopted by the general public in the mid 90’s there has been significant concerns regarding the reliability of e-mail viewed in terms of the probability that a transmitted message is correctly received by the intended recipient. While the reliability of the network over which the message is transmitted has always been a factor, over the last decade the strain on network capacity due to the massive broadcasts of spam messages has become a dominating factor in this. E-mail reliability is affected when a legitimate message is being falsely classified as a spam message (false positive) as well as when a spam message is being falsely classified as legitimate (false negative) the latter allowing potentially harmful content to pass on to the user’s inbox along with a large number of unwanted messages drawing away attention to legitimate messages. In this article we look at some of the recent developments in spam filter design and its impact on e-mail reliability and we present the results of a case study in which two spam filters are tested in terms of “false negative” and “false positive” classifications.

Definition and Detection of Spam

Spamhaus (www.spamhaus.org), an international consortium providing leading efforts in the fight against spam since 1998, defines spam as “Unsolicited Bulk E-mail.” Their definition emphasizes “consent” rather than “content” as the primary factor in classifying a particular message as spam or legitimate. By their definition it is not a factor whether the particular message contains offensive or potentially harmful material or not, neither is it a factor whether the message contains material of potential interest to the recipient. If a message is distributed in bulk to users who have not explicitly given their consent, it is a spam message, and the distribution of such is in violation with the terms of most Internet Service Providers.

Under this definition, it would be the chief goal of any spam filter to classify each message according to those two characteristics (solicited or not, bulk mail or not). Because these characteristics may be impossible or difficult to establish, instead filters seek to determine the classification based on measures believed to correlate with these two characteristics, including:

- Content analysis (identifying keywords and other content that occur frequently in spam, but not in legitimate messages as well as keywords that occur frequently in legitimate messages, but not in spam.)
- Sender analysis (identifying e-mail addresses and/or domain names who have been “black listed” as spammers or “white listed” as legitimate senders, either using public databases or the individual users preferences.)
- Transmission analysis (considering the path of IP addresses through which the message is transmitted from sender to recipient providing possible evidence of the message being part of a bulk mailing, but not whether the mailing is solicited).
- Similarity analysis (a possible indicator of bulk mail is if a number of identical or similar messages are identified on one or more mail servers for which the filter is operating.)

As far as determining “consent” there is no reliable method by which a computer program can establish the degree to which the recipient has provided consent to receive the message. Even when the recipient has explicitly black listed or white listed a particular sender, consent cannot be established with 100% accuracy because spammers often attempt to falsify their true identity (Sanchez et al 2010). When the identity has not been falsified and no black or white list identification, valuable information can be found by considering the recipients inbox and address book. When a particular sender’s address exists in the recipient’s inbox or address book, it makes it more likely that the transmission is consensual.

The biggest difficulty is in handling a message that has been identified as bulk mail and where the sender is unknown to the recipient. It may indeed be a legitimate message if the recipient has signed up to receive updates or newsletters from a particular organization or business. Often a business or organization will add a user to its distribution list without explicit consent when the user has interacted with the organization or business in one of the following ways:

- Signed up for membership of an organization
- Purchased a product from a business
- Made an inquiry to or requested an estimate from a business
- Entered a drawing at a trade show or other event for which the user’s e-mail address is revealed
- Sharing/sale of e-mail addresses of potential clients between affiliated organizations and businesses.

Unsubscribing from Bulk Mail

US laws require that all bulk e-mail be accompanied by an option to unsubscribe from the sender’s distribution list. Many such unsubscribe options are ineffective even when the sender has full intention of honoring all such unsubscribe requests for reasons including:

- The recipient uses several aliases that all forward to a particular e-mail address. When an unsubscribe request is submitted it may not identify the particular alias being used in the sender's distribution list.
- The sender's e-mail address may already have been shared with a number of other distribution lists.

For a typical spammer who has no intent to respect a user's privacy, submitting an unsubscribe request may adversely impact a user's likelihood of being targeted for spam or being subjected to harmful content:

- When an unsubscribe request is received, it signifies that the replying user represents a valid e-mail address from someone who actually opened and read the spam message. This increases the value of the e-mail address when being sold to other distribution lists.
- Clicking an unsubscribe link may result in further exploitation to spam content or downloading of harmful content such as a virus, spyware or phishing attack.

Modeling Spam Filters using a Bayesian Framework

During the past decade, the use of a Bayesian framework to model and develop spam filters have gained substantial popularity (Samami et al 1998, Yang and Elfayoumy 2007, Yeh and Chiang 2008, Issac et al 2009). This framework is founded upon the theory of conditional probabilities, in particular the well-known identities:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = P(B|A)\frac{P(A)}{P(B)} \qquad (1)$$

For the purpose of quantifying the reliability of a spam filter consider the following events and probabilities (Sahami et al 1998):

M_s=A message is spam

M_l=A message is legitimate

F_s=A message is classified as spam by the filter

F_l=A message is classified as legitimate by the filter

$P(M_s|F_s)$=Spam Precision Rate

$P(M_l|F_l)$=Legitimate Precision Rate

$P(F_s|M_s)$=Spam Recall Rate

$P(F_l|M_l)$=Legitimate Recall Rate

Here the precision rate reflects the likelihood that a message classified as spam (legitimate) is indeed spam (legitimate), whereas the recall rate reflects the likelihood that a spam (legitimate) message will be correctly classified as spam (legitimate). Under ideal circumstances (i.e. when a filter always correctly classifies a message as spam or legitimate), all four of these rates will equal 1 (100%). From a practical perspective, the most desirable situation is one where the spam precision rate is 100% (all messages caught in the spam filter are spam) and the legitimate precision rate close to 100% (the vast majority of the messages delivered to the inbox are legitimate). In this situation the user can delete all messages caught in the spam filter without inspection and at the same time the user is only minimally inconvenienced by having to delete a few spam messages incorrectly passed onto the inbox.

On the other hand the recall rates are more reflective of the spam filter's true capability than the precision rates. That is because the recall rates do not depend on the ratio of spam to legitimate messages, whereas the precision rates are highly affected by this ratio. Even a poorly designed filter may have a high spam precision rate if nearly all messages are spam and a high legitimate precision rate if nearly all messages are legitimate. Conversion between precision rates and recall rates are possible through the conditional probability properties (1) when the overall proportion of e-mail being spam (legitimate) is known.

Further Classification of Spam Messages

Another factor that influences the four rates defined above is the nature of the spam (legitimate) messages received by the user. Some spam (legitimate) messages are more difficult to correctly identify. In particular bulk messages containing unsubscribe options may either be spam or legitimate depending on the user's consent which as mentioned earlier is nearly impossible to establish by the filter. For the purpose of this article we shall consider 4 different types of e-mail messages:

> <u>Legitimate</u>:
> This includes any person-to-person communication as well as messages sent to distribution lists for which the recipient has either deliberately added him or herself or belongs to an organization (employment, professional, sports etc.) that makes such means of communication appropriate.
>
> <u>Type I Spam: Low Risk Spam with an Unsubscribe Option</u>:
> This includes messages sent to distribution lists for which the recipients did not intentionally add him or herself. The sender is a legitimate business or organization that has not attempted to camouflage its identity or the content of the message and who fully respects the privacy of the recipients by offering an option to unsubscribe.

Type II Spam: Low Risk Spam without an Unsubscribe Option:
This includes messages from a questionable business or organization who is attempting to circumvent spam filters and users' preferences by using rotating e-mail addresses or by using a false/disguised identity. Frequently words are intentionally misspelled words and subject headers are not representative of the content of the message. Apart from the annoyance and possible offense, the content of the message itself, however, is considered harmless. There is no intent of the sender other than to market a product (good or bad) or a political or religious view. An unsubscribe option may be listed, but is listed only to disguise the intent of the message, and clicking this option may further exploit the sender's address for spam purposes.

Type III Spam: High Risk Spam:
This includes messages that are sent with a clearly malicious intent containing such content as viruses or malware, solicitations leading to scams, phishing attacks or identity theft for the purpose of fraudulent financial gain.

Not every legitimate e-mail is necessarily desirable for or of interest to the recipient. By definition a message sent without the recipient's consent is not considered spam unless it is also part of a bulk distribution. Such messages, however, can easily be blocked by the recipient to prevent further messages. Except as defined by the user's preferences (custom filters and black and white lists) it is generally the intent of all spam filters to pass all non-bulk messages to the user's inbox unless it is determined to have high risk content as in spam Type III above.

Low risk spam with unsubscribe options are difficult to handle by a spam filter because what might be considered spam by one user may be considered a legitimate message by another. Generally, it will be up to the recipient to filter such messages e.g. by submitting an unsubscribe request or by black listing the sender.

Enterprise and User Level Filtering

E-mail filtering usually occurs at two levels, which we shall refer to as "enterprise level" and "user level". Enterprise level filtering is generally invisible to the user and operates directly on messages arriving at the incoming enterprise mail server. Many Type II and III spam messages are eliminated at the enterprise level, which is desirable because it reduces the cost of storage, however, a "False Positive" that occurs at this level is particularly critical because the message is deleted for good. Individual users generally have no access to or influence on the operation of the enterprise level filter which are administered by system administrators. Messages not determined to be spam with high confidence are directed to either a spam message quarantine for user inspection or to the user's inbox. It is for these messages, and these alone, that user-level filtering is possible. For the message quarantine, the user can browse through messages, and choose to direct messages to the user's inbox or delete them permanently. The user may also be able to define expiration dates for messages such that messages are automatically deleted if not released by the user by that date.

A Case Study

As a practical demonstration of spam filter reliability, two separate e-mail accounts subjected to message filtering, both maintained by the author of this article, were observed for 30 days. One of these e-mail accounts is hosted by a university and has been used for nearly 17 years. Apart from being widely exposed to potential spammers through the many years of open usage, also this address appears in many unprotected directories making it an easy target for collection of e-mail addresses collected and traded for spam usage. All incoming messages are filtered at the enterprise level through the Microsoft Forefront Spam Filter with an optional secondary filtering available through the built-in customizable Microsoft Outlook message filter(s) as well as a user-defined black-list and white-list.

The other e-mail account, hosted by Yahoo, has been used by the author for 7 years in numerous on-line purchases, on-line travel reservations and similar e-commerce transactions as well as recipient for various special interest e-mail discussion boards and distribution lists, thus also a natural target for bulk e-mail solicitations. Like Microsoft Forefront, Yahoo filters messages at the enterprise level with many messages being rejected at the enterprise level without possible intervention of the end user. Because no data is available regarding the number of spam messages being rejected at the enterprise level, the true spam precision and recall rates cannot be measured, thus we consider only precision and recall rates based on messages either passed to the end-users inbox or the spam filter (quarantine) available for user-inspection. The secondary filter for this account consists of user-defined filtering rules as well as black and white lists.

For the period of 30 days all e-mail messages received in either an inbox or a designated filter quarantine were manually examined to determine its true nature (Legitimate or Spam Type I,II or III as defined above). Also, when applicable (mostly for Type I spam) messages were followed up by an unsubscribe request or a black or white listing of the sender to determine whether these user interventions had a significant impact on the number of spam messages passed on to the end user or legitimate messages that failed to be delivered to the user.

Consider the data summarized in Tables 1 and 2. As expected, the university hosted e-mail account being the user's primary e-mail address and the one that has been in use the longest attracted the heaviest e-mail traffic during the 30 day test period (634 messages) compared to the secondary Yahoo hosted e-mail account (110 messages). Also, not surprisingly, the distribution of message types varies greatly between the two e-mail accounts; see Figures 1 and 2, with the Yahoo based e-mail account drawing a much larger percentage of spam messages. Because the two e-mail accounts are used for very different purposes, and thus attract different types of e-mail messages, it is not meaningful to compare the reliability of the different spam filters being used, but it appears that at least two of the three filters tested are performing quite well. Both the

Microsoft Forefront and Yahoo spam filters exhibited a 100% legitimate precision rate and recall rate with no legitimate e-mail being caught in these spam filters. When considering spam message recall rates, at first glance these rates appear not to be too impressive. For example, only about 62% of actual spam messages sent to the university hosted e-mail account (and not filtered at the enterprise level) were correctly placed in the Forefront spam filter. However, a large percentage of these messages were Type I spam which by nature are difficult to classify because of the similarity with solicited (legitimate) bulk messages. If considering only Type II and Type III spam, the recall rate goes up to 84% and if considering only Type III spam the recall rate is 100%, i.e. every Type III spam message were correctly caught by this spam filter. Similar results were observed for the Yahoo based filter's spam recall rates. Considering the secondary Outlook spam filter operating in series with the Forefront filter, the reliability in terms of legitimate and spam precision and recall rates is not satisfactory. Both a large number of false positives and false negatives were observed requiring the user to frequently inspect messages caught by this spam filter rendering it virtually useless. This is not because this filter is necessarily poorly designed, but rather due to the fact that this filter has a much more challenging job than the other filters since it is working only with messages that have already been filtered. All of the "obvious" spam messages have already been removed now allowing this filter only the opportunity to offer a "second opinion" on messages allowed by the primary filter to pass.

As mentioned earlier, of the three types of spam messages, the Type I messages are generally the ones of the least concern. The main concern with these messages is that they fill up the inbox often causing legitimate messages to be overlooked, and they consume unnecessary bandwidth and storage space from the network and server. If the messages of Type I all have unsubscribe functions provided can these messages be eliminated completely? The answer to this question, based on the messages observed during this test period is both yes and no. Indeed each mailing subscription can and usually will be terminated following the submission of an unsubscribe request. Figures 3 and 4 show time series plots of the number of daily Type I spam messages received for each of the two e-mail accounts. These plots suggest a significant decrease in the number of Type I spam messages as result of using the unsubscribe function.

However, it is anticipated that the relief in Type I spam is only temporary. Using the "unsubscribe" function may be seen as analogous to spraying the weeds in our yards with weed-killer. The present weeds are killed at the root, but eventually new weeds will come back. A user must carefully weigh the effort required to send "subscribe" requests for all Type I spam messages received against the effort needed to simply delete them as they arrive

Conclusion

The results of the case study presented in this article suggest that current state-of-the-art filters are successful in reducing spam to a manageable problem. However, while filters have clearly become much more sophisticated, at the same time spammers are becoming significantly more sophisticated and creative. In recent years, image spam (Uemura 2008, Soranamageswari and Meena 2010) has been widely adopted by spammers along with website spam including use of social networking sites for distribution of spam (Araujo and Martinez-Romo 2007, Ridzuan et al 2010). It is suspected that in the near future spammers will be targeting mobile communication, in particular text messaging (Hoanca 2006). As result of these mechanisms, future spam filters will need to be frequently updated, most likely to the same extent as virus and other security protection software is subjected to daily live updates.

References

Araujo, L. and Martinez-Romo, J. (2010) Web Spam Detection: New Classification Features Based on Qualified Link Analysis and Languiage Models. *IEEE Transactions on Information Forensics and Language Models*, 6. Early Preview.

Hoanca, B. (2006). How Good Are Our Weapons in the Spam Wars? *IEEE Technology and Society Magazine*, Spring 2006, 22-30

Issac, B., Jap, W. and Sutanto, J. (2009) Improved Bayesian Anti-Spam Filter – Implementation and Analysis on Independent Spam Corpuses. *Proceedings of the 2009 International Conference on Computer Engineering and Technology*, 326-330

Ridzuan, F., Potdar, V., Teleski, A. and Smyth, W. (2010). Key Parameters in Identifying Cost of Spam 2.0. *Proceedings of the 24th IEEE International Conference of Advanced Information Networking Applications*, 789-796

Sahami, M., Dumais, S., Heckerman, D. and Horvitz, E. (1998) A Bayesian Approach to Filtering Junk E-Mail. *Proceedings of the AAAI ’98 Workshop on Learning for Text Categorization*, 55-62

Sanchez, F., Duan, Z. and Dong, Y. (2010) Understanding Forgery Properties of Spam Delivery Paths, *Proceedings of the Seventh Annual Collaboration, Electronic Messaging, Anti-Abus and Spam Conference*.

Soranamageswari, M. and Meena, C. (2010) Statistical Feature Extraction for Classification of Image Spam Using Artificial Neural Networks. *Proceedings of the Second International Conference on Machine Learning and Computing*, 101-105.

Uemura, M and Tabata, T. (2008) Design and Evaluation of a Bayesian-filter-based Image Spam Filtering Method. *Proceedings of the 2008 International Conference on Infromation Security and Assurance*, 46-51

Yang, Y. and Elfayoumy, S. (2007) Anti-Spam Filtering Using Neural Networks and Bayesian Classifiers. *Proceedings of the 2007 IEEE International Symposium on Computational Intelligence in Robotics and Automation*, 272-278

Yeh, C. and Chiang, S. (2008) Revisit Bayesian Approaches for Spam Detection. *Proceedings of the 9th International Conference for Young Computer Scientists*, 659-664

	Primary Spam Filter	Secondary Spam Filter	Inbox
Days Recorded	30	30	30
Legitimate E-mail	0	10	442
Type 1 Spam	1	16	33
Type 2 Spam	77	15	6
Type 3 Spam	34	0	0
All Spam	112	31	39
Total Messages (by type)	112	41	481
Total Messages (by e-mail address)			634
Daily Traffic			
Legitimate E-mail	0.00	0.33	14.73
Type 1 Spam	0.03	0.53	1.10
Type 2 Spam	2.57	0.50	0.20
Type 3 Spam	1.13	0.00	0.00
All spam	3.73	1.03	1.30
Precision Rates			
All spam types combined	100.00%	75.61%	91.89%
Type 2 and 3 spam only	100.00%	60.00%	98.66%
Type 3 spam only	100.00%	0.00%	100.00%
Recall Rates			
All spam types combined	61.54%	44.29%	97.79%
Type 2 and 3 spam only	84.09%	71.43%	
Type 3 spam only	100.00%		
Distribution of Messages by Type			
Legitimate E-mail	452		
Type 1 Spam	50		
Type 2 Spam	98		
Type 3 Spam	34		
All Spam	182		

Table 1: Summary of Messages Received on the University Hosted E-mail Account

	Spam Filter	Inbox
Days Recorded	30	30
Legitimate E-mail	0	39
Type 1 Spam	11	8
Type 2 Spam	41	6
Type 3 Spam	5	0
All Spam	57	14
Total Messages (by type)	57	53
Total Messages (by e-mail address)		110
Daily Traffic		
Legitimate E-mail	0.00	1.30
Type 1 Spam	0.37	0.27
Type 2 Spam	1.37	0.20
Type 3 Spam	0.17	0.00
All spam	1.90	0.47
Precision Rates		
All spam types combined	100.00%	73.58%
Type 2 and 3 spam only	100.00%	86.67%
Type 3 spam only	100.00%	100.00%
Recall Rates		
All spam types combined	80.28%	100.00%
Type 2 and 3 spam only	88.46%	
Type 3 spam only	100.00%	
Distribution of Messages by Type		
Legitimate E-mail	39	
Type 1 Spam	19	
Type 2 Spam	47	
Type 3 Spam	5	
All Spam	71	

Table 2: Summary of Messages Received on the Yahoo Hosted E-mail Account

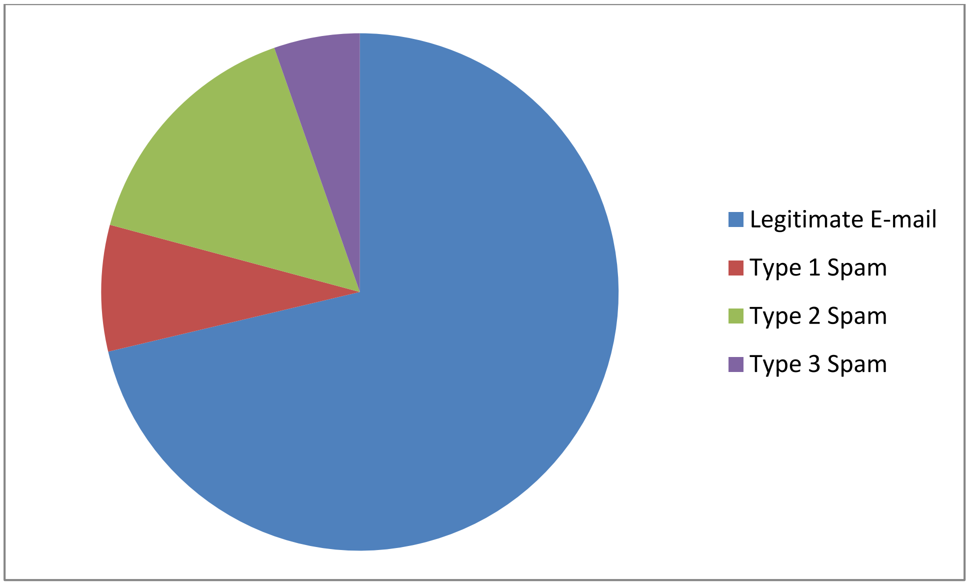

Figure 1: University Hosted E-mail Address distribution of incoming e-mail by type.

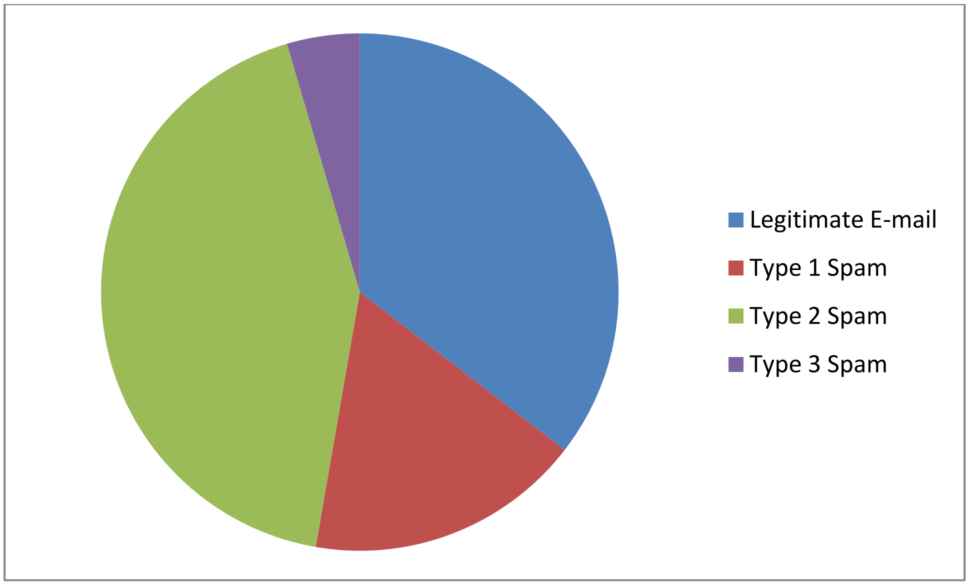

Figure 2: Yahoo Hosted E-mail Address distribution of incoming e-mail by type.

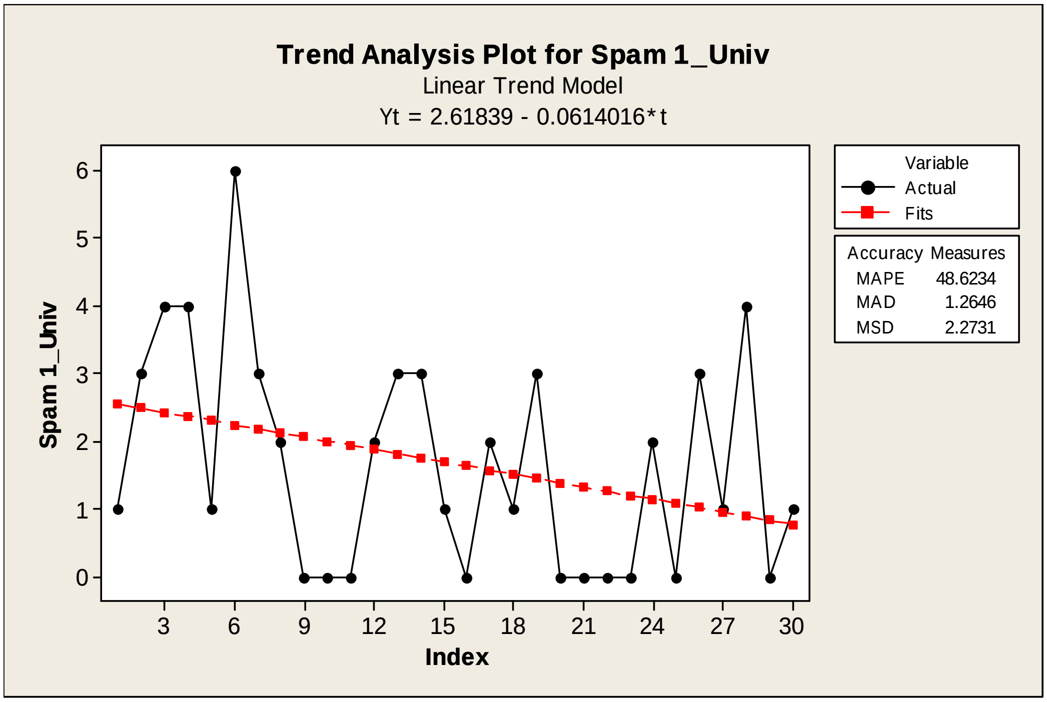

Figure 3: Time series analysis of the number of Type I spam messages received daily for the University Yahoo Hosted E-mail Address.

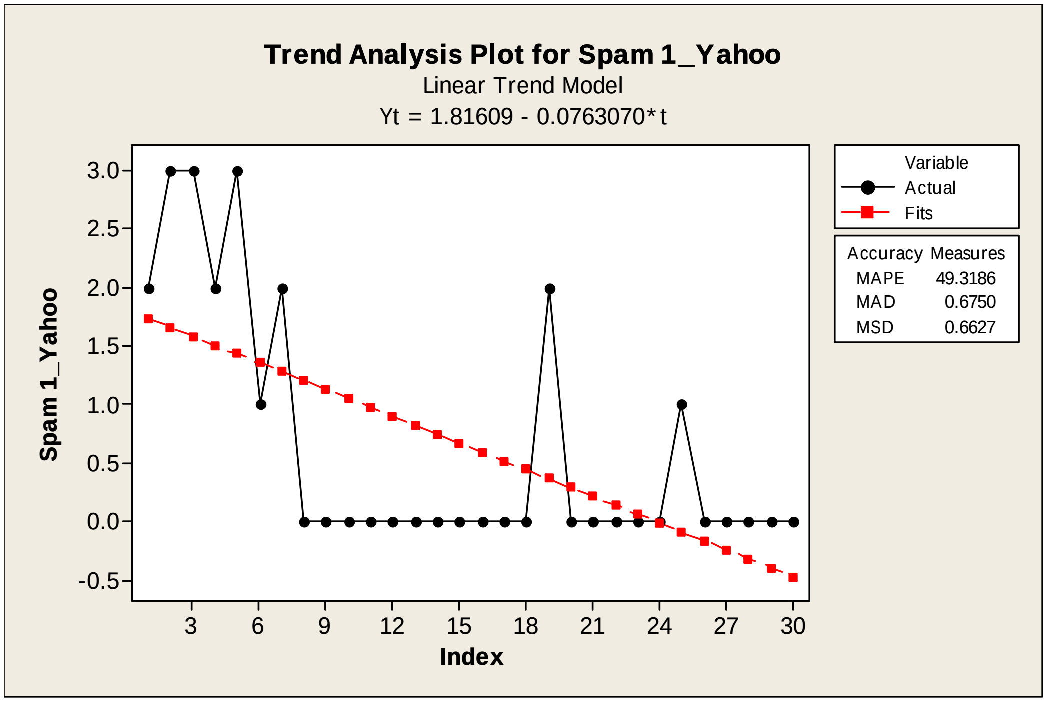

Figure 4: Time series analysis of the number of Type I spam messages received daily for the Yahoo hosted E-mail Address.