Measurement and Evaluation of a Real World Deployment of a Challenge-Response Spam Filter

Jelena Isacenkova, Davide Balzarotti

Eurecom, Sophia Antipolis, France
{isachenk,balzarotti}@eurecom.fr

ABSTRACT

Despite the number of existing solutions, spam still accounts for a large percentage of the email traffic on the Internet. Both the effectiveness and the impact of many common anti-spam techniques have already been largely studied and evaluated against multiple datasets. However, some of the less known solutions still lack a proper experimental validation. For example, Challenge-Response (CR) systems have been largely discussed, and often criticized, because they shift the effort to protect the user's mailbox from the recipient to the sender of the messages. In addition, these systems are believed to produce a lot of backscattered emails that further deteriorate the global Internet situation.

In this paper we present the first comprehensive measurement study of a real anti-spam system based on a challenge-response technique. In our work we analyze a large amount of data, collected for a period of six months from over forty companies protected by a commercial challenge-response product. We designed our experiments from three different point of views: the end user, the system administrator, and the entire Internet community. Our results cover many different aspects such as the amount of challenges sent, the delay on the message delivery, and the likelihood of getting the challenge server blacklisted.

Our aim is neither to attack nor to defend CR-based solutions. Instead, we hope that our findings will shed some light on some of the myths about these kind of systems, and will help both users and companies to take an informed decision on the topic.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: Measurement techniques, Performance attributes

General Terms

Experimentation, Measurement

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
IMC'11, November 2–4, 2011, Berlin, Germany.
Copyright 2011 ACM 978-1-4503-1013-0/11/11 ...$10.00.

Keywords

Spam, whitelist, challenge-response, blacklisting

1. INTRODUCTION

Despite the considerable effort and the large amount of proposed solutions to detect and filter unsolicited emails, according to the MessageLabs Intelligence Annual Security Report [35] published by Symantec, in 2010 spam still accounted for 89.1% of the emails on the Internet. Even worse, according to the same Symantec's report, in 2010 the spam volume increased by 1.4% compared with the previous year.

Traditional anti-spam solutions are based on two common techniques: filtering emails based on their content, or filtering them based on their senders. The first category includes content-based text classification techniques [15, 19, 32, 33] that aim at finding (often using supervised learning) the tokens commonly associated to spam messages. The second category includes instead detection methods based on some properties of the sender [24, 30, 31], of his reputation [11, 36], or of the domain from which the email is delivered [18, 22, 36].

Even though these two categories cover most of the widely adopted techniques, a number of other solutions have also been proposed to fight spam. One notable example is represented by Challenge-Response (CR) filters [21, 29], a solution based on the observation that the large majority of good emails are delivered from senders that are already known to, and trusted by, the recipient. The name of the approach comes from the fact that, whenever the sender of an email is unknown (i.e., not yet in the user's personal whitelist), the system temporarily quarantines the email and automatically sends back a message to the sender, asking him to solve a simple *challenge* to verify his legitimacy. This technique somehow changes the traditional approach of treating incoming emails, shifting the delivery responsibility from the recipient to the sender of the message.

Since their first introduction, CR-based techniques have been considered an extremely controversial solution [13, 2]. On the one hand, they seem to be able to completely block any unsolicited email, but, on the other hand, they also have a number of side-effects that can seriously hamper their adoption on a large scale.

In particular, it is possible to group the main criticisms against CR systems around three main points. First, the *social* and *usability* issues that, on one side, are related to the efforts required from the user to maintain a proper whitelist, and, on the other, to the annoyance for the sender that has to invest time to solve a challenge in order to have his

message delivered. Previous studies, in particular Erickson et al. [21], have already studied the usability of CR systems in controlled experiments. Their study concludes that such systems are very effective when accompanied with already existing anti-spoofing techniques. The authors also measure that CR solutions outperform traditional systems like SpamAssassin, generating on average 1% of false positives with zero false negatives.

The second point against CR systems concerns the fact that they can introduce a (possibly conspicuous) delay in the emails delivery due to the quarantine period applied to previously unknown senders. Finally, the last (and one of the main) critique against CR systems is due to the challenge emails sent in response to spam messages. Since unsolicited emails often contain spoofed sender addresses, the challenges are often delivered to non-existing recipients or to innocent users. These misdirected messages (often referred as "backscattered" spam) pollute the Internet with unnecessary traffic and damage other users that may receive challenges for emails they never sent. From this point of view, CR antispam filters seem to literally bounce the spam back towards other innocent users. However, supporters of the CR approach often rebut by saying that well-designed systems only send back a challenge to a few percents of the spam messages they receive. Therefore, considering the fact that real forged addresses are not too common, normal users are very unlikely to often receive misdirected challenges. Unfortunately, since both sides lack real data to support their own hypothesis, it is hard for users and companies to tell which is the truth and take a conscious decision.

To the best of our knowledge, this paper presents the first study on both the effectiveness and the impact of a real-world deployment of a challenge-based antispam solution. In our work we measure and analyze a large amount of data collected for a period of six months from 47 companies protected by a commercial CR-based antispam product.

In particular, we conduct our measurements to analyze the behavior of CR systems from three different perspectives:

1. From the *end user* point of view, to measure how this technique affects the delivery of both spam and normal messages to the end user's mailbox;

2. From the *server's administrator* point of view, focusing on some of the problems of maintaining a CR installation in a real company;

3. From the *Internet* point of view, to measure the amount and the impact of backscattered messages and misdirected challenges.

It is important to stress the fact that the purpose of this study is neither to attack nor to defend CR-based solutions. Instead, our goal is to provide real-world figures and statistics that can help both users and companies to take an informed decision based on our study. Our results can also help to shed some light on some of the myths related to CR antispam techniques.

The rest of the paper is organized as follows. In Section 2 we introduce our data collection methodology and the dataset we used in our measurements. Section 3 presents a study of the amount of challenges sent by a CR system. Section 4 describes the effectiveness of CR systems in distinguishing spam from legitimate messages. Section 5 introduces some of the problems related to maintaining this type of antispam filter. Then, we summarize our findings in Section 6, present a survey of related work in Section 7, and finally conclude the paper in Section 8.

2. DATA COLLECTION

In this section, we describe the dataset we used in our experiments and we provide a short overview of our data collection methodology.

System Overview

Our study has been carried out within a company providing an anti-spam solution based on a challenge-response technique. Figure 1 presents the overall system architecture and a "weighted" lifecycle of the incoming emails. The *CR filter* consists of two main components: a message dispatcher and a set of additional spam filters.

The dispatcher receives the incoming messages from the company's Incoming Mail Transfer Agent (MTA-IN) server. Some of the email servers were configured to work as *open relays*, serving emails also for a *restricted* number of domains that are different from the ones in which the systems are installed. This configuration allows the server to accept messages not targeting to, or originating from, known users in the system.

The MTA-IN server first checks if the email address is well formed (according to RFC822 [17]) and then if it is able to resolve the incoming email domain. In addition, if the server is not configured as an open relay, it also verifies that the recipient exists in the system.

Our study shows that this first layer of simple checks is responsible to drop more than 75% of the incoming messages (see Figure 2), while open-relay systems pass most of the messages to the next layer. These results are perfectly in line with similar values reported by the other analysis of spam delivery rate [34, 26]. The reasons behind the dropped messages are summarized in the following table:

Dropped Percentage	Reason
0.06%	Malformed email
4.19%	Unable to resolve the domain
2.27%	No relay
0.03%	Sender rejected
62.36%	Unknown Recipient

The second check point for the incoming emails is at the internal email dispatcher. This component is the core of the CR infrastructure and it is the one responsible for deciding to which category the email belongs to: white, black or gray.

The white and black spools are controlled by the user's *whitelist* and *blacklist*. Emails in the black category are dropped immediately, while emails from senders in the whitelist are delivered to the user's INBOX. Emails matching none of the previous lists fall in the gray category. These messages are then filtered with additional antispam techniques (e.g., virus scan, reverse DNS and IP blacklisting). If an email passes the filters, then dispatcher sends a challenge-response message to the original sender containing a request to solve a CAPTCHA. Otherwise, the email is considered spam and it is dropped.

Figure 1 also reports the average number of messages for each spool, assuming that 1,000 emails are received by the MTA-IN. The figures are computed by aggregating the data of all the monitored servers not configured as open relay.

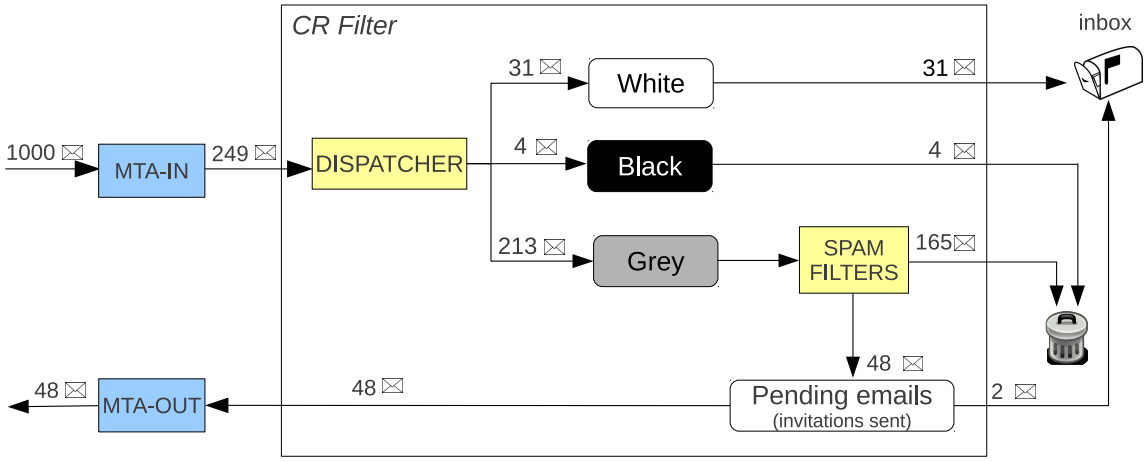


Figure 1: Lifecycle and distribution of incoming emails

Figure 3 shows that the other spam filters included in the CR engine drop on average 54% of the gray emails. Challenge messages are instead generated for 28% of emails. In the open relay cases, the engine filters have a lower performance rate, and the number of challenges sent increases by an extra 9%. This shows that, in an open relay configuration, the CR system receives more junk messages and it is more likely to reply with a challenge to illegitimate emails.

Whitelisting process

The process of email whitelisting involves both parties: the sender and the recipient. There exist several alternative ways for the email address to get added to a user's whitelist. In particular, the system we tested in our experiments supported the following mechanisms:

- The sender solves a challenge sent by the CR system as a response to one of his messages;
- The user authorizes the sender from the daily message digest;
- The address is manually added to the whitelist by the user;
- The user previously sent an email to that address.

In the general scenario, suppose that Alice sends an email to Bob, a user protected by a challenge-response system. If this is the first communication between Alice and Bob, the system temporarily stores the email in a "gray" spool and sends back a message to Alice. The message includes a link to a webpage that contains a CAPTCHA (the challenge) that Alice has to solve to get her email delivered and her address added to Bob's whitelist. After this simple authentication step, Alice's address is considered trustworthy, and the CR system will not interfere in any future communication between the two users, promptly delivering to Bob any further message coming from Alice.

If Alice does not solve the challenge, the email stays in the gray spool for a period of 30 days, after which it is dropped by the system. Bob also receives a daily digest that summarizes the quarantined messages, so that he can manually authorize them or delete them from the list.

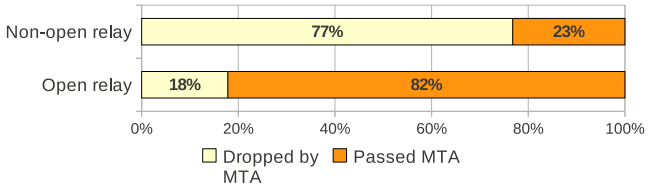


Figure 2: MTA-IN email treatment

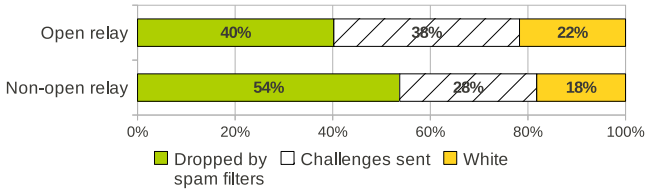


Figure 3: Message category at the internal email processing engine

General Statistics

In our experiment we collected statistical data about a commercial system deployed in 47 companies of different sizes. The monitoring period lasted for 6 months, between July and December 2010. For some of the servers we had access to the data for the entire time frame, while for other companies our collection was limited to a shorter period of time (with a minimum of 2 months).

In total we collected statics for 90 millions of incoming emails. All the results were sanitized to protect both the end users and the companies privacy. In particular, we never got access to the message bodies and we stored only aggregated figures obtained from the automated analysis of the email headers.

The data collection was performed on a daily basis by analyzing the logs of the MTAs and of the challenge-response engines. In addition, information about the solved CAPTCHAs was collected by analyzing the access logs of the web-server serving the challenges. The extracted information was stored

General Statistics

Number of Companies	47	Challenge Sent	4,299,610
Open Relays	13	Emails Whitelisted from digest	55,850
Users protected by CR	19,426	Solved CAPTCHAs	150,809
Total incoming emails	90,368,573	Messages Dropped because of:	
Messages in the Gray spool	11,590,532	reverse DNS filter	3,526,506
Messages in the Black spool	349,697	RBL filter	4,973,755
Messages in the White Spool	2,737,978	Antivirus filter	267,630
Total Messages Dropped at MTA	75,690,366	Total Messages Dropped by filters	7,290,922

Daily Statistics

Emails (per day)	797,679	Challenges sent (per day)	53,764
Messages in the White spool (per day)	31,920	Total number of days	5,249

Table 1: Statistics of the collected data

in a Postgres database and later analyzed and correlated by a number of Python scripts.

Table 1 shows some general statistics about the dataset we collected. Each company's server was configured to protect certain users with the challenge-response system, while protecting other accounts by traditional anti-spam techniques. In this paper we limit our analysis to the 19,426 users protected by the CR solution (this number includes normal users as well as administrative accounts and other rarely used email addresses). The table also shows the total number of the messages that we analyzed, the breakdown in the different spools (white, black, and gray), and some statistics about the effectiveness of the other spam filters included in the system (discussed in more details in Section 5).

Finally, since the number of days in which we were able to collect data varies between companies (for a total of 5,249 analyzed days), the table also report some *daily* statistics.

3. PART I: THE INTERNET POINT OF VIEW

In this section we focus on the consequences of adopting CR spam filters from a global point of view. In particular, we present an evaluation of the amount of challenge emails sent out by a challenge-response system during normal operation.

These *backscattered messages* are often criticized for two main reasons: the fact that misdirected challenges can be delivered to innocent users, and the fact that a large amount of useless messages are poured into the Internet, thus increasing the global traffic and overloading third parties email servers.

In the rest of the section we provide real-world measurements to estimate the impact of these two phenomena.

3.1 Email Backscattering

From an external point of view, a challenge response system can be approximated by a black box that receives emails from the Internet and separates them in three categories: some (the *white* set) are delivered to the users Inbox, while others (the *black* set) are immediately flagged as spam and discarded. The remaining messages (the *gray* set) are the ones for which the system is unable to take a decision. Therefore, for each email in this set, the system sends back to the sender another email containing a challenge to be solved. In this simplified model, a challenge-response system can be seen as a software that receives a certain amount of emails, and "reflects" a fraction of them back to the senders. This fraction, that we call *Reflection Ratio* $\mathcal{R}$, is an important parameter of a CR system.

By using the numbers in Figure 1, it is easy to compute the average reflection ratio: $\mathcal{R} = 48/249 = 19.3\%$ for the emails reaching the CR filter (or, $\mathcal{R} = 48/1000 = 4.8\%$ if we consider all the emails reaching companies' MTA-INs).

Understanding the Reflection Ratio

Is 19.3% a good value for $\mathcal{R}$? If not, what would be a reasonable value?
Unfortunately, it is very hard to answer these questions since it is not clear how to estimate which is an acceptable range for the reflection ratio.

To explain why, let us consider two extreme cases. In the first case, the CR system does not contain any other spam detector or blacklist mechanism. Therefore, the amount of challenges it sends is roughly the same as the amount of spam it receives, currently estimated between 80 and 90% [35] of the total email traffic. Values of $\mathcal{R}$ close to this range are obviously unacceptable, since, from a global point of view, the system would just act as a spam multiplier.

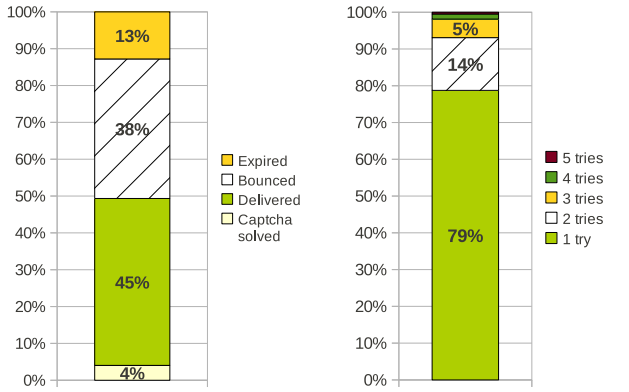
In the second scenario, the CR system has been carefully configured and it has been associated with another perfect spam detector. In this case, the system never replies to spam and only sends back challenges to legitimate messages whose senders are not already in the recipients whitelist. In this case (represented by very low values of $\mathcal{R}$) the system does not generate any backscattered emails. Therefore, it may seem to be the final goal to reach in a perfect CR system.

Unfortunately, a very low value of $\mathcal{R}$ also corresponds to a completely useless system. In fact, if the internal spam filter can already distinguish good messages from spam, there is no need to add a challenge response system on top of it. In other words, in order to be useful a CR system has to be able to "substantially" reduce the amount of spam received by the users. However, this can only happen if the system sends back an equivalent "substantial" number of backscattered messages.

To conclude, the reflection ratio is a good indicator of the amount of challenges generated by a CR system. At the same time, it is important to be extremely careful to use this value alone to draw conclusions about the quality of such systems.

3.2 Misdirected Challenges

So far, we focused on the amount of challenges generated by a CR system. However, this value only measures the *amount* and not the real *impact* of the generated emails. In fact, not all the challenges are the same. Some of them reach the real senders and, despite being a little nuisance, could be tolerated as an acceptable price to pay for fight-



(a) Challenge delivery status distribution

(b) Tries required to solve CAPTCHA

Figure 4: Challenge statistics

ing spam. We refer to them as *legitimate challenges*. A second class of them is directed to non-existing addresses, and, thus, constitutes garbage traffic on the network. Finally, some misdirected challenges are delivered to existing spoofed email addresses, therefore reaching other innocent users. This category is much more harmful, and it is often referred to as *backscatter spam* (note that not all the backscattered *messages* are spam).

In order to distinguish the three categories of challenges, we analyzed the status of the challenge delivery in the servers' logs. In the systems under analysis, we found that only 49% of the challenges were successfully delivered to the destination servers. The remaining 51% were either bounced, or expired after many unsuccessful attempts (see Figure 4(a)). In the bounced set, a small portion has been stopped because the server that sent the challenges has been temporarily blacklisted (the problem will be discussed in more details in Section 5), while the large majority (71.7%) has been bounced due to the fact that the recipient address did not exist. This value provide a reasonable estimation of the amount of challenges that belong to the second category.

Another piece of the puzzle can be found by measuring the number of challenges that were actually solved. Previous work [21], conducted in a controlled environment, estimated that about 50% of the challenges were never solved. Unfortunately, our study shows a completely different picture. According to the web servers' logs of the companies we analyzed, on average 94% of the CAPTCHA URLs included in the delivered challenges were never even opened. The remaining were either solved (4%) or were visited by the user but not solved (0.25%). Figure 4(b) also shows the average number of attempts required to solve the CAPTCHAs. The fact that we never observed more than five attempts support the fact that probably there are still no real cases of attack against CR systems based on trying to automatically solve the challenges.

So far, we estimated the legitimate challenges to be at least 4% and the ones sent to non-existing recipients to be around 36.6% (71.7% of the 51% of undelivered messages). The third category, i.e., the backscattered spam, can instead be approximated with the number of challenges correctly delivered but never solved, i.e. somewhere between 0 and 45 %.

By combining the percentage of backscattered spam with the reflection ratio we presented before, we obtain the *Backscattered Ratio* β, i.e., the ratio of incoming emails for which the CR system sends back a misdirected challenge to the wrong user. In our experiments, we obtain, in the worst case, $\beta = 8.7\%$ (at the CR filter) or 2.1% (at the MTA-IN).

However useful, these figures must be considered only approximate upper bounds. For example, it is possible that challenge messages get dropped by some spam filter after being successfully delivered, or that real users ignore or intentionally decide to not solve a particular challenge. Finally, there are automatically generated emails (notifications from websites, mailing lists, receipts of purchase, ...) to take into account. When a user expects to receive such messages, he should either use an email address not protected by the CR system (functionality provided by the commercial product we have evaluated), or manually add the address to the whitelist.

Unfortunately, this is not always the case. In fact, we measured that around 2% of the message addresses in the gray spool have been whitelisted manually by the users from the daily digest. In other words, the challenge was not delivered or it was not solved, but the user recognized that the message was not spam and he manually added the sender to his whitelist to allow future communications.

3.3 Traffic Pollution

The reflection ratio only measures the number of messages, without taking into account their size. Therefore, it is not a very accurate indicator to estimate the amount of traffic generated by a challenge response system. For that purpose, we need to extend the previous definition by introducing the *ReflecteD Traffic ratio* $\mathcal{R}_T$, that represents the ratio between the amount of traffic received by the system and the amount of email traffic generated for the challenges.

To measure this new value, we deployed to all the servers a new sensor that extracts from the email headers the total size of the incoming messages and the total size of the outgoing challenges. Over a month period, the average ratio we measured at the CR filter was $\mathcal{R}_T = 2.5\%$. Unfortunately, we could not get a similar measure at the entrance of the MTA-IN servers. However, since the number of messages at MTA-IN is in average four times bigger than at the CR filter (see Figure 1), we can estimate that a large scale deployment of challenge-response spam filters would increase the email traffic on the Internet of around 0.62%.

3.4 Data Variability

In previous sections we reported aggregated figures for a number of different factors that can be used to estimate the "external" impact of a CR system.

In order to preserve the companies' privacy, each value was obtained by combining together the data collected from all the monitored installations. However, it is interesting to investigate what the variance of those values is, and if the size of the company affects in some way the presented indicators. For instance, it could be the case that CR filters work better for small companies, but fail to scale well to larger installations.

Figure 5 shows a scatter plot of five variables: the number of protected accounts (`users`), the amount of emails re-

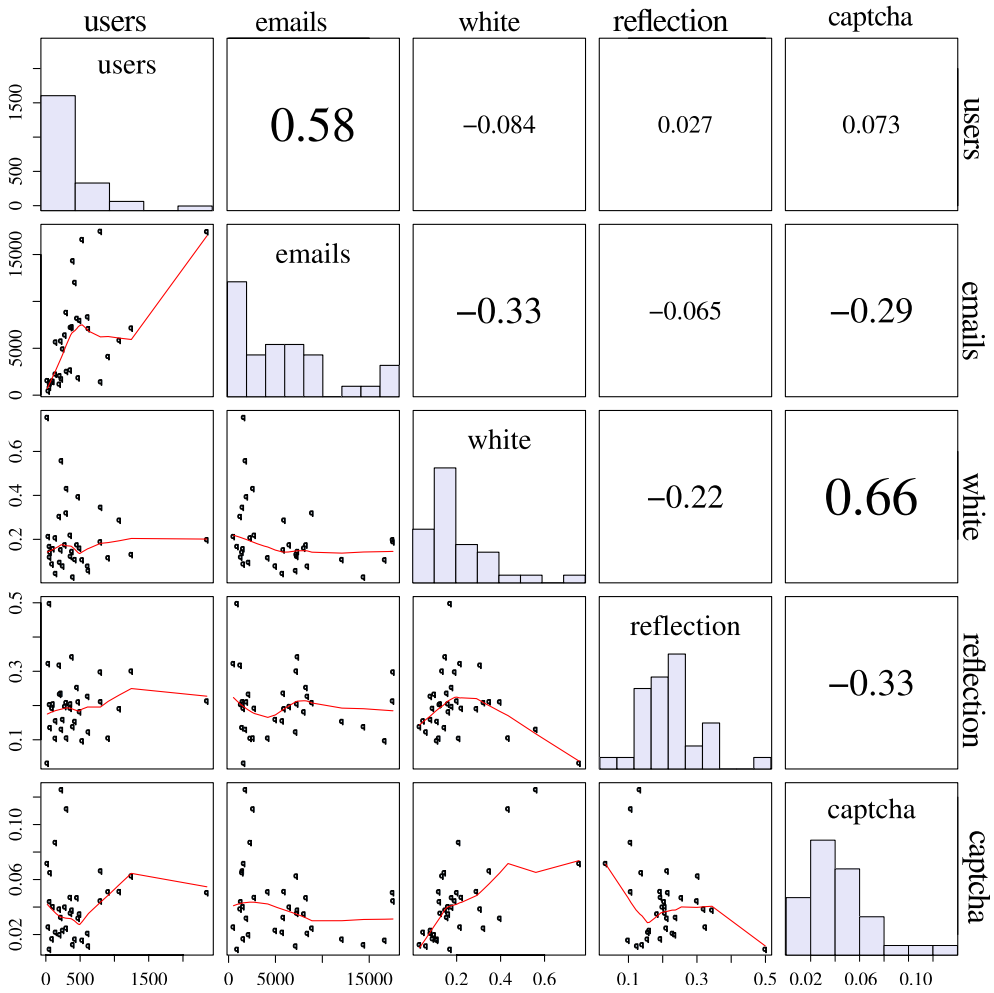


Figure 5: Histograms and correlations between different dimensions. Graphs on the diagonal represent the data histogram. Below the diagonal are the plots of the correlation between every pair of variables, summarized by the correlation values reported above the diagonals.

ceived daily (`emails`), the percentage of emails delivered in the white spool (`white`), the reflection ratio at the CR filter (`reflection`), and the percentage of challenges solved (`captcha`).

This graph represents a very efficient way to convey a large amount of information about the five variables. On the diagonal, it shows the histograms of the values of each variable. For example, the first element on the diagonal shows that most of the companies have less than 500 users, with few exceptions that have more than 2,000 users. Some values have a very high variability, such as the percentage of white emails that varies from less than 10% to over 70%. However, the two main coefficients we have measured in this Section, i.e. the reflection ratio and the percentage of solved challenges, seem to stay constant between different installations. The percentage of solved challenges only varies between 2% and 12%, and the reflection ratio stays in the range of 10% to 25%.

In Figure 5, the plots below the diagonals show the correlation between every pair of variables, while the upper part of the graph reports the corresponding correlation values (the font size allows to immediately focus on the higher values). Notably, the percentage of challenges sent by a CR system (`reflection`) is not correlated to the size of the companies (`users`) or to the amount of emails received. Not surprisingly, a small inverse correlation exists instead with the percentage of white emails. In other words, servers that receive a large amount of white emails (and therefore a lower amount of spam), tend to send less challenges and vice versa.

The rate of solved challenges (`captcha`) shows more correlations with other values, and in particular it is also strongly correlated with the white percentage. However, as the histogram shows, the variability of the `captcha` variable is so small that it can be considered almost a constant between the different installations.

4. PART II: THE USER POINT OF VIEW

Despite the backscattering phenomenon described in the previous section, CR systems are often considered one of the most effective ways to protect users from spam. In theory, if the challenge-response system is properly configured, these systems should be able to achieve a 100% detection rate, thus blocking all unsolicited emails. However, previ-

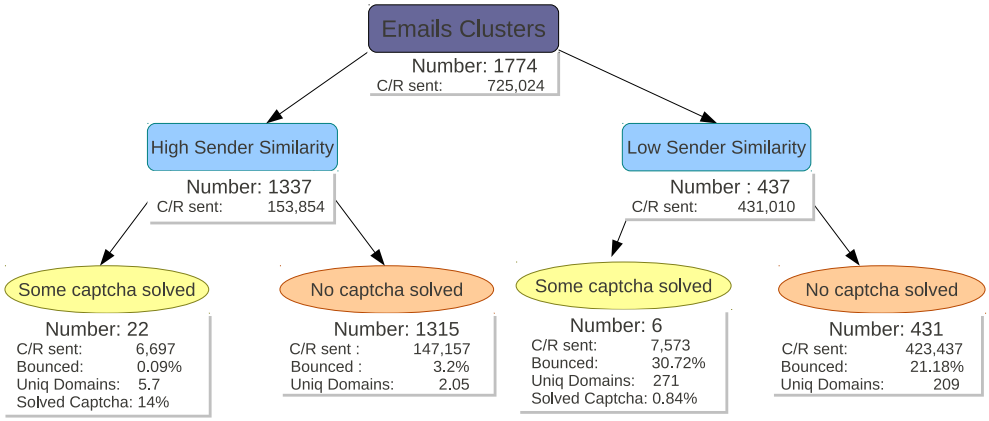


Figure 6: Spam clustering statistics

ous studies [21] that confirmed this value were conducted on prototype systems evaluated in a controlled environment.

In this section we measure if this is actually the case in a real-world installation, and we evaluate the real impact for the end users protected by a CR system. In particular, we measure the delay introduced in the legitimate emails delivery, and the amount of spam that is able to reach the final users despite the CR filter. In addition, we also measure the change rate of the users' whitelists, one of the foundations of this kind of antispam solution.

4.1 Spam Protection

The main purpose of a CR system is to block all automatically generated emails coming from addresses previously unknown to the recipient. The first obvious consequence of this approach is that CR solutions are ineffective by design against targeted attacks, i.e., attacks in which the attacker manually composes a malicious message to target a particular individual. In fact, if the attacker receives back the challenge message, he can easily solve it and have his message delivered to the recipient. However, a recent Symantec report [35] estimated that only one out of 5,000 spam messages contains a targeted attack. In addition, all the existing anti-spam mechanisms can be easily evaded by targeted attacks, and, therefore, we can consider this threat beyond reach of all existing anti-spam solutions.

Unfortunately, targeted attacks are not the only ones that can pass through a CR filter. By studying a dataset of bounced challenges, we noticed that a large number of messages had the same subject and the same size. Per se, this is not surprising. However, a closer look revealed that while most of the messages were bounced or dropped by the filter, in some cases one of those emails was successfully delivered to the final user's mailbox.

To better investigate the reason behind this sporadic phenomenon, we decided to analyze the behavior, in terms of challenges and delivered messages, of a number of large spam campaigns.

For our experiment we applied standard clustering algorithms to the subject of the messages in the gray spool (i.e., the ones for which the system generated a challenge message). In particular, we put in the same cluster the messages with the same subject, limiting the analysis to the ones at least 10 words long. Finally, we discarded the clusters containing less than 50 emails. These very conservative thresholds were adopted to greatly reduce the risk of misclassification. In reality, the large majority of emails (including spam) have much shorter subjects, or they have enough variability to elude our simple comparison algorithm. However, our target was not to be able to cluster and identify all the incoming emails or all the spam campaigns, but just to identify a number of them with a low percentage of false positives.

The results obtained over a three month monitoring period are summarized in Figure 6. Our system identified 1,775 clusters, containing between 50 and 3696 messages each. In the next step, we divided the clusters in two categories, based on the sender email similarity. In the first group we put all the clusters where emails are sent by a very limited number of senders, or in which the sender addresses are very similar to each other (for example, `dept-x.p@scn-1.com`, `dept-x.q@scn-1.com`, and `dept-x.p@scn-2.com`). These clusters are likely associated to newsletters or marketing campaigns. The second group contains instead the clusters with a very low sender similarity, i.e., the ones in which most of the emails originate from different domains and different senders' addresses. This behavior is very common in spam campaigns sent by malware infected machines.

Figure 6 shows that only 28 out of 1774 clusters contain at least one solved challenge. Moreover, these few clusters have very different characteristics, depending on whether they belong to the first or the second category. The ones with high sender similarity have a higher rate of solved challenges (some clusters as high as 97%) and almost no bounced emails. The clusters with low sender similarity have instead on average 31% of emails bounced because of non-existing recipient, and only one or two captchas solved each.

This second category is particularly interesting for our study. Each cluster in this group contains hundreds of emails, coming from many different domains, and often from non-existing sender addresses. However, out of these messages, sometimes one of the challenges was solved and, therefore, the email got whitelisted and delivered to the recipient's mailbox. These spam messages that are able to pass through the CR defense are likely a side effect of backscattered challenges that are sometimes erroneously delivered to innocent users. As a result, it is possible that one of these users solves a challenge for a mail he never sent. This phenomenon is,

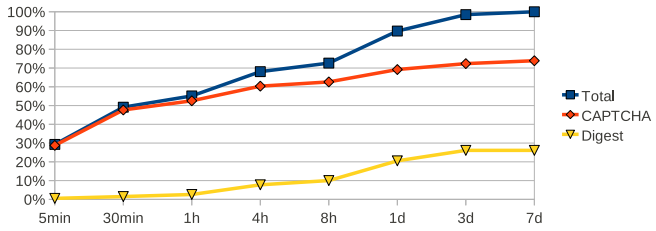


Figure 7: Cumulative effect of Captcha and Digest whitelisting

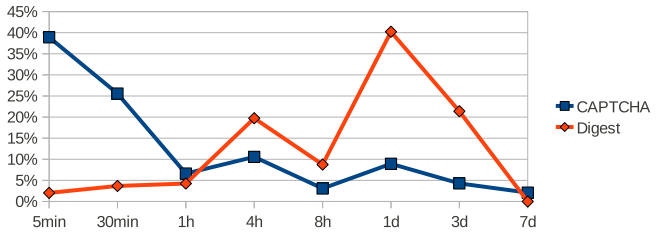


Figure 8: Time distribution of whitelisted messages

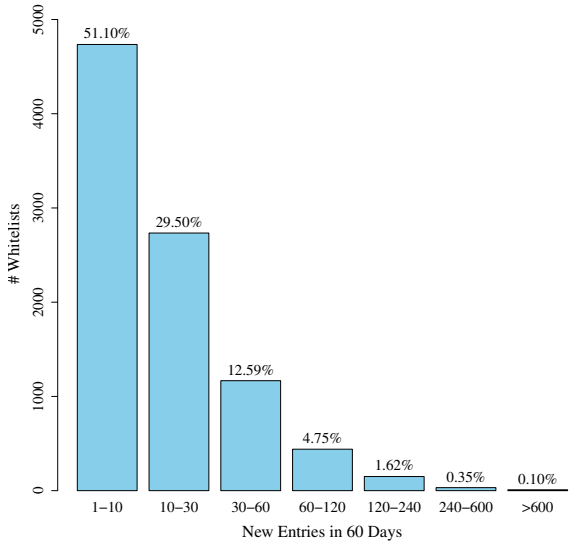


Figure 9: Distribution of the number of changes in users' whitelist

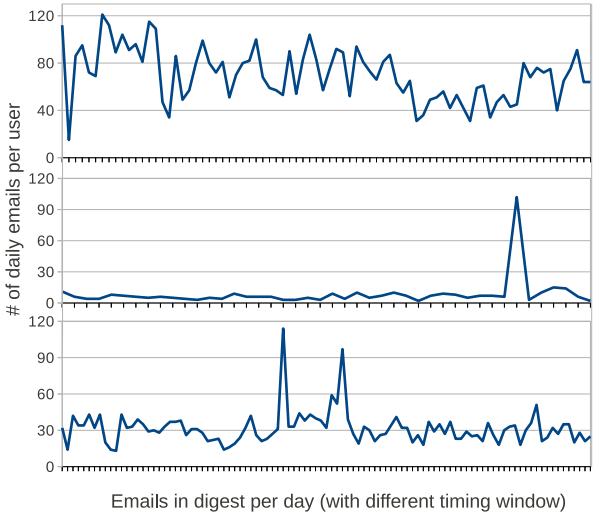


Figure 10: Daily pending email distribution of 3 different users

however, extremely rare. According to our measurements, we estimate that this kind of spurious spam delivery occurs ∼1 every 10,000 challenges sent. According to Table 1, this rate translates to an average of five spam delivery a day, over the 47 companies in our dataset. Excluding these isolated cases, CR systems are actually able to block all incoming spam messages.

4.2 Impact on Messages Delivery

Another consequence of blocking automatically generated emails is the fact that also normal emails can get blocked and remain in the user's graylist waiting for the corresponding challenges to be solved. This can happen for two reasons: because the sender still has to solve the challenge, or because the email is sent by an automatic system and the challenge is, therefore, dropped or never delivered. In both cases, the user fails to receive a (maybe important) email.

Figure 7 shows the CDF of the messages that were moved from the graylist to the whitelist in the monitored servers. The two curves report the percentage of messages that were whitelisted because the sender solved the challenge, and the ones that were whitelisted manually by the user from the daily digest. According to the graph, 30% of the messages are delayed less than 5 minutes, and half of them are delivered in less than 30 minutes. However, if the challenge was not solved after 4 hours, then it is likely that it will not be solved at all (Figure 8). In those cases, the user has to manually select the messages from the digest, with a delivery delay that is on average between 4 hours and 3 days.

Combining the values from these figures with the number of white and whitelisted emails (31 and 2 respectively) in Figure 1, we can conclude that:

- 31/33 = 94% of the emails in the user's INBOX are sent from addresses already in the whitelist, and, therefore, are delivered instantly.
- Out of the remaining 6% (2/33) of the messages that are quarantined in the gray spool, half of them are delivered in less than 30 minutes because the sender solved the challenge.
- Only 0.6% (10% of the 6%) of the messages were delivered with more than one day of delay.

4.3 Whitelists' Change Rate

We conclude this section on the user point of view with an analysis of the rate at which the users' whitelists change over time. For this experiment we monitored the number of changes in the users' whitelists for a period of two months. Email addresses can be added to a whitelist in four different ways, two manual and two automated. A user can manually import a new entry or he can whitelist a certain address from the digest. Automatically, new entries are included

when the user sends a mail to those addresses or when the senders solve the challenges generated by the CR system.

During the monitored period, 9267 whitelists were modified at least once. Out of them, only 6.8% had at least 1 new entry per day (on average), and the percentage drops even further when we look at higher changing rates (2.1% of the whitelists had at least 2 new entries per day, and 0.2% at least 5). Figure 9 presents a more detailed histogram of the frequency of changes. The graph shows how the large majority of the whitelists are, in fact, constantly in a steady state.

Finally, we monitored the amount of new messages present in the daily digest. This value varies greatly between users and also between different days. Figure 10 shows examples extracted from three different users. Some of them have constantly a large number of messages in the gray spool, while others have normally very small daily digests with anomalous peaks in conjunction to particular user behavior or unusually large amount of spam messages.

Again, a large size of the digest is at the same time a good and a bad factor for a CR system. In fact, a high number of messages means that the system is blocking a substantial amount of spam that would have been otherwise delivered to the user (remember that these are messages that successfully pass through the antivirus, reverse DNS, and the SpamHouse blacklist). On the other side, a large digest is also a negative factor as it increases the amount of work for the user that has to manually verify its content to double-check that he did not miss any important message. Finally, this also demonstrates that the degree to which CR system works depends a lot on the interplay of users' involvement. Some recipients may diligently weed out their digest, while others may let it grow hoping for the senders to respond to the challenges.

5. PART III: THE ADMINISTRATOR POINT OF VIEW

In this section we analyze some of the issues related to maintaining challenge-response systems from the system administrator point of view. In particular, we focus on the effort required to maintain the challenge-response infrastructure, and on the additional antispam techniques that can be integrated in the system to reduce the backscattering effect.

5.1 Server Blacklisting

As we already described in Section 4, when a CR system sends a challenge in response to a message with a spoofed sender's address, the challenge is delivered to a recipient that may not exist. As a result, these challenge-response messages can hit a spam trap [31], i.e., a set of email addresses that are maintained and distributed with the sole purpose to lure spam.

The emails collected by those traps are often adopted by popular services (e.g., SpamHaus [11], SORBS [7], SpamCop [10]) to update their blacklists. Hence, the IP used to send the challenges can itself get blacklisted as a result of the backscattered spam it sends. In order to reduce the impact of being blacklisted, one third of the systems we tested in our experiments were configured to rely on two MTA-OUT servers (with different IP addresses): one to send the challenges and another to send the outgoing user emails.

Our initial hypothesis was that the probability that a server has to get blacklisted should have been somehow proportional to the size of the email server, represented either by the number of users, or by the number of the received emails. In other words, we expected that systems sending more challenges were blacklisted more often, thus making CR solutions more difficult to maintain for large companies.

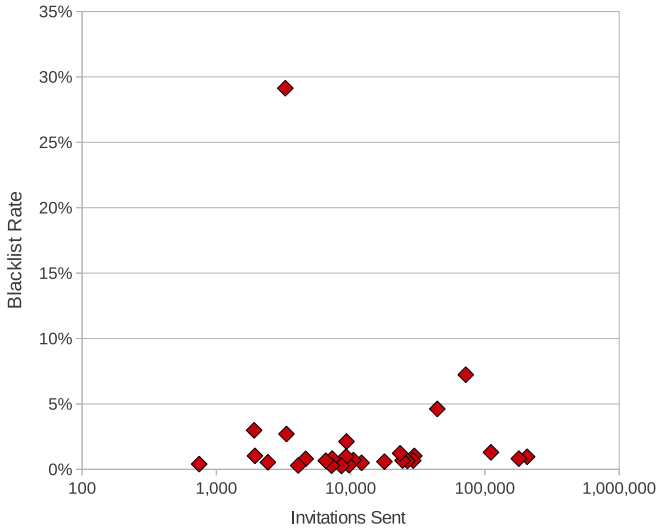


Figure 11: Server blacklisting rate

Surprisingly, our experiments proved us wrong. Using the data we collected we were able to estimate the rate at which various challenge server IPs get blacklisted. In particular, we followed two parallel approaches. In the first, we analyzed one month of data for 32 companies, measuring the ratio between the number of challenges sent and the number of blacklist-related error messages received from the challenge-response recipients. The result, summarized on a logarithmic scale in Figure 11, shows that while most of the servers had no problem at all with blacklisting, some of them were often blacklisted, even for a few days in a row. However, there seems to be no relationship between the server blacklisting ratio and the number of challenges it sends.

The main problem with this approach is that the error messages received when delivering the challenges were not always very accurate, providing results that may not be completely reliable. Therefore, we decided to complement our analysis with a second technique, based on an automated script that periodically checked for the IP addresses of the CR servers in a number of services that provide an IP blacklist for spam filtering. In particular, our tool queried the blacklists provided by Barracuda [1], SpamCop [10], SpamHause [11], Cannibal [9], Orbit [4], SORBS [7], CBL [3], and Surriel [5]. The queries were performed every 4 hours for a period of 132 days (between September 2010 and January 2011).

The results of this second analysis confirm our previous finding. In more than four months, 75% of the servers never appeared in any blacklists. Few servers were blacklisted for less than one day, while the remaining four servers experienced some serious problems, appearing in at least one of the blacklists for many consecutive days (17, 33, 113, and 129 respectively). Again, between the top 3 server (according to the traffic and the number of challenges sent) none appeared

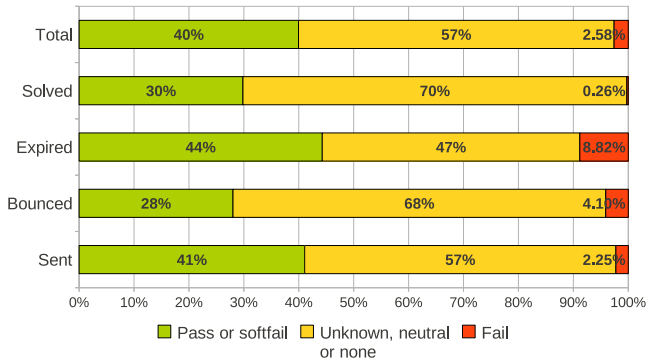


Figure 12: SPF validation test

in any of the blacklists during our experiment. Thus, proving again that there is no direct link between the number of times a server gets blacklisted and the server size.

5.2 Combining CR Systems with Other Spam Filters

Our final evaluation focuses on the combination of CR systems with other antispam solutions. As we already mentioned in Section 2, the product we analyzed in our experiments includes three other spam filters in order to reduce the number of useless challenges sent in response to spam messages. It employs a traditional antivirus to scan the emails, an IP blacklist provided by SpamHause [11] to filter out known spammers, and a reverse DNS lookup to exclude suspicious origins.

According to Table 1 and Figure 1, the combination of these filters was responsible for dropping 77.5% of the messages in the gray spool. One may argue if this is good enough, or if a much better result could be obtained by introducing other antispam techniques. This is a difficult question to answer, since the main drawback of adding new filters is that they also introduce false positives, to avoid which CR systems were introduced in the first place.

However, we decided to experiment with one additional spam filter based on the verification of the Sender Policy Frawework [36] (SPF). SPF was introduced to detect source address spoofing, that is one of the main problems of CR systems. Since SPF checks were not included in the product we evaluated in this paper, we decided to evaluate the potential impact of this filter by using an offline tool to automatically test all the emails in the gray spool. Figure 12 shows the results of our experiment, grouped by different message categories. For instance, by dropping the emails for which the SPF test fails, it would be possible to reduce by almost 9% the challenges that cannot be delivered (`expired`), and 4.10% of the bounced ones. The overall result shows that SPF can further reduce the number of "bad" challenges by 2.5%, at the cost of loosing 0.25% of the challenges that are actually solved by the sender.

6. DISCUSSION

Even though the aim of this work is neither to attack nor to defend challenge-response systems, it may be natural to ask what conclusions about this class of antispam solutions could be drawn from our measurements.

In the rest of this section we summarize the main findings we presented in the previous three sections.

Whitelist Assumptions

All approaches based on white-lists share two main assumptions: first, that the large majority of the "good" emails come from senders already in the recipient's whitelist, and, second, that these lists eventually reach a steady state where changes do not occur very often.

Both claims are supported by our experiments. In fact, over 43 companies, only $2/33 = 6.1\%$ of the incoming emails delivered to the users' INBOX require a challenge-response phase (see Figure 4) and 2% require the user to manually pick the message from the digest.

The stability of the whitelists was already evaluated by Erickson et al. [21], showing that the major burden on the user is concentrated in the first three weeks, after which the number of changes drops on average to one per day. Our experiments show that, in a real deployment, there are on average 0.3 new entry per user per day (excluding new users). Only 6.8% of the users had at least one daily change in their whitelists.

Delivery Delay

Another common critique of CR systems is due to the fact that the challenge-response step introduces a delay in the incoming email delivery. This is obviously an unavoidable side-effect, but our measurements show that it also has a limited impact. In fact, according to our study, it concerns only 4.3% of incoming emails and in half of the cases the delay is below 30 minutes. Even though the remaining 2.15% may still be an unacceptable inconvenient for certain users, we believe that for most of the users it would be a reasonable price to pay to protect against spam.

Challenge Traffic

Most of the criticisms against CR systems, and most of the hassles for the system administrators, come from the challenges that are sent to the wrong recipients. If they correspond to existing email accounts, the misdirected challenges become a spam for other users. On the other hand, if the addresses do not exist, the challange may hit a spamtrap. And on top of that, they constitute useless traffic over the Internet.

Our study shows that, on average, a CR system sends back one challenge for every 21 emails it receives (see Section 3), accounting for a traffic increase of less than 1%. These figures depend on the amount of spam received by the server, and seems to be more or less constant between small and large servers.

Unfortunately, the large part of the challenges sent are indeed useless (only about 5% of them are solved). But, as we already explained in the paper, these challenges are "required" to justify the system. In other words, without useless challenges, it would be the CR system to be useless. Therefore, this can be considered an intrinsic and unavoidable limitation of systems based on a challenge-response approach.

Our findings confirm that the backscattered phenomenon is the main problem of solutions based on challenge-response technique. Each installation must be carefully configured in order to minimize the impact of misdirected challenges on other real users. The administrator also has to decide

which other additional antispam techniques should be combined with the CR filter to maximize the benefits and, at the same time, to reduce the side effects and the risk of having the servers' IP blacklisted. However, the backscattered phenomenon is intrinsic in the behavior of a CR system and cannot be completely eliminated. From a company, the single most negative argument against the adoption of CR system is the fact that the challenge server can occasionally get blacklisted. Even worse, an attacker could intentionally forge malicious messages with the goal of forcing the server to send back the challenge to spam trap addresses, thus increasing the likelihood of getting the server IP added to one or more blacklist.

Other Limitations

This paper does not cover all aspects related to the adoption of a challenge-response system. We focused on performing a set of measurements based on real installations that were not under our direct control. Therefore, we intentionally excluded from our studies any evaluation of potential attacks against CR systems (like trying to spoof the sender address using a likely-whitelisted address).

In addition, in order to protect the users and the companies' privacy, we limited our study to the statistical information that can be automatically extracted from the headers of the messages. This ruled out other potentially interesting experiments that would have required access to the email bodies.

7. RELATED WORK

Most of the spam blocking techniques proposed by previous research fall into two categories: content-based and sender-based filtering. Content-based spam filtering techniques rely on signatures and classification algorithms applied to the emails content [32, 14, 19, 15, 33]. Although content-based filters were initially very effective and provided an acceptable defense against spam [27], with the evolution of the spam sophistication they became less effective over time.

Sender-based spam filtering techniques aim instead at blocking spams by analyzing information about the sender. To date, a wide range of sender-based filtering solutions has been proposed, including sender authentication (e.g., SPF [36], DMIK [18], and Occam [22]), sender IP reputation [11], network-level feature categorization, sender whitelisting, and detecting anomalies on the sender's behavior.

Sender authentication techniques authenticate the sender either by verifying it's domain [36, 18] or by providing a protocol to authenticate the server at each message delivery [22]. These kind of solutions are quite effective to prevent email spoofing, a phenomenon that is very common among spammers.

Approaches based on IP reputation [11, 25] rely on whitelists or/and blacklists of IP addresses that are known either to send spam or to be a trusted source. Therefore, these approaches are effective against static spammers and open-relay servers used for spam distribution. On the other hand, they are not able to provide the same degree of protection against spam sent through botnets, since botnets can change the sender's IP address by using a large number of different infected machines to deliver the messages.

Behavior-based solutions, like the ones proposed by Pathak et al. [30] or Ramachandran et al. [31], and network-level detection techniques, like the one proposed by Hao et al. [24], tend to react faster to spam campaigns and have a lower number of false-positive. However, these kind of solutions block only part of illegitimate emails, and, therefore, they have to be used in combination with other filters.

Beside the more common techniques presented so far, a number of other solutions have been proposed to protect users against spam. For example, fine-grained approaches have been proposed based on personal white and blacklists. The assumption behind such techniques is that users communicate mainly with a list of contacts that does not change much over time [16, 20, 21].

In this case, the main problem is to provide an automated way to populate and maintain the users whitelist. Garris et. al. [23] proposed a solution to this problem based on the idea of sharing the whitelist content with the user's friends on social networks. Their cryptographic solution addresses also the sender spoofing problem, and the protection of the privacy of the users during the friend list sharing process. The main limitation of their system is the fact that it requires a large-scale adoption by many social networking users.

One of the most wide-spread approaches for building and maintaining a list of trusted senders is based on the adoption of a challenge-response technique [28, 29, 8, 12, 6], already largely described in the rest of the paper.

Although challenge-response schemes are extremely successful in blocking spam, they also have a number of limitations that makes them disadvantageous over other solutions [21]. Additionally, CR solutions received a great amount of critiques from the anti-spam community [13, 2], often because of the amount of challenge emails they generate.

To the best of our knowledge, the only empirical study that analyzes challenge-response based whitelisting systems is presented by Erickson et.al. [21]. The authors focus on the deployment and the usability of such systems. The results of their evaluation support the usability of CR systems, but also show their limitations in coping with automatically generated legitimate emails, such as newsletters and notifications. On the other hand, the authors concluded that challenge-response systems are very successful to prevent spam and have lower false positives and false negatives rates compared to traditional content filtering techniques like SpamAssassin.

Our work aims instead to present a comprehensive study of a real-world whitelisting challenge-response antispam system, evaluating it's effectiveness and it's impact on the end-users, Internet, and server administration.

8. CONCLUSIONS

In this paper we present the first measurement study of the behavior of a real world deployment of a challenge-response antispam system. The experiments lasted for a period of six months, covering 47 different companies protected by a commercial CR solution.

In particular, we measure the amount of challenges generated by these systems and their impact in terms of traffic pollution and possible backscattered messages delivered to innocent users. We then measure the amount of emails that are delayed due to the quarantine phase, and the amount of spam that is able to pass through the filter and reach the users mailboxes. Finally, we focus on a problem that is less known, i.e., the fact that the invitations sent by these sys-

tems can accidentally hit a spamtrap and cause the email server to be blacklisted.

Our findings can be used to evaluate both the effectiveness and the impact of adopting this class of techniques. We hope that the figures provided in this paper may help to settle the long debate between advocates and opponents of CR systems.

9. REFERENCES

[1] Barracuda. http://www.barracudacentral.org/.

[2] Challenge-Response Anti-Spam Systems Considered Harmful. http://linuxmafia.com/faq/Mail/challenge-response.htm.

[3] Composite Blocking List. http://cbl.abuseat.org/.

[4] ORBITrbl. http://www.orbitrbl.com/.

[5] Passive Spam Block List. http://psbl.surriel.com/.

[6] Sendio. http://www.sendio.com/.

[7] Spam and Open-Relay Blocking System. http://www.sorbs.net/.

[8] Spam Arrest. http://www.spamarrest.com/.

[9] Spam Cannibal. http://www.spamcannibal.org/.

[10] SpamCop. http://www.spamcop.net/.

[11] The spamhaus project. http://www.spamhaus.org/.

[12] Total Block. http://www.totalblock.net/.

[13] Why are auto responders bad? http://spamcop.net/fom-serve/cache/329.html.

[14] I. Androutsopoulos, J. Koutsias, K. Chandrinos, G. Paliouras, and C. Spyropoulos. An evaluation of naive bayesian anti-spam filtering. 2000.

[15] A. Bergholz, J. Chang, G. Paaß, F. Reichartz, and S. Strobel. Improved phishing detection using model-based features. In *Proc. of the Conference on Email and Anti-Spam (CEAS)*, 2008.

[16] D. Bernstein. Internet mail 2000 (IM2000).

[17] D. Crocker. RFC822: Standard for ARPA Internet Text Messages. *Retrieved April*, 7:2008, 1982.

[18] M. Delany. Domain-Based Email Authentication Using Public Keys Advertised. In *the DNS (DomainKeys)", RFC 4870*, 2007.

[19] H. Drucker, D. Wu, and V. Vapnik. Support vector machines for spam categorization. *IEEE Transactions on Neural Networks*, 10(5):1048–1054, 1999.

[20] Z. Duan, Y. Dong, and K. Gopalan. Diffmail: A Differentiated Message Delivery Architecture to Control Spam. Technical report, 2004.

[21] D. Erickson, M. Casado, and N. McKeown. The Effectiveness of Whitelisting: a User-Study. In *Proc. of Conference on Email and Anti-Spam*, 2008.

[22] C. Fleizach, G. Voelker, and S. Savage. Slicing spam with occamâĂŹs razor. *CEASâĂŹ07*, 2007.

[23] S. Garriss, M. Kaminsky, M. Freedman, B. Karp, D. Mazieres, and H. Re. Reliable email. In *Proc. of NSDI*, 2006.

[24] S. Hao, N. Syed, N. Feamster, A. Gray, and S. Krasser. Detecting spammers with SNARE: Spatio-temporal network-level automatic reputation engine. In *Proc. of the 18th conference on USENIX security symposium*, pages 101–118. USENIX Association, 2009.

[25] J. Jung and E. Sit. An empirical study of spam traffic and the use of DNS black lists. In *Proc. of the 4th ACM SIGCOMM Conference on Internet measurement*, pages 370–375. ACM, 2004.

[26] C. Kanich, C. Kreibich, K. Levchenko, B. Enright, G. Voelker, V. Paxson, and S. Savage. Spamalytics: An empirical analysis of spam marketing conversion. In *Proceedings of the 15th ACM conference on Computer and communications security*, pages 3–14. ACM, 2008.

[27] D. Lowd and C. Meek. Good word attacks on statistical spam filters. In *Proc. of the second conference on email and anti-spam (CEAS)*, pages 125–132, 2005.

[28] R. Mastaler. Tagged message delivery agent. http://www.tmda.net/.

[29] M. Paganini. ASK: active spam killer. In *Proc. 2003 Usenix Annual Technical Conference*, 2003.

[30] A. Pathak, Y. Hu, and Z. Mao. Peeking into spammer behavior from a unique vantage point. In *Proc. of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats*, pages 1–9. USENIX Association, 2008.

[31] A. Ramachandran, N. Feamster, and S. Vempala. Filtering spam with behavioral blacklisting. In *Proc. of the 14th ACM conference on computer and communications security*, pages 342–351. ACM, 2007.

[32] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. A Bayesian approach to filtering junk e-mail. In *Learning for Text Categorization: Papers from the 1998 workshop*, volume 62, 1998.

[33] G. Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, C. Spyropoulos, and P. Stamatopoulos. A memory-based approach to anti-spam filtering for mailing lists. *Information Retrieval*, 6(1):49–73, 2003.

[34] B. Stone-Gross, T. Holz, G. Stringhini, and G. Vigna. The underground economy of spam: A botmaster2019s perspective of coordinating large-scale spam campaigns. In *USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET)*, 2011.

[35] Symantec's MessageLabs Intelligence. Messagelabs intelligence annual security report, 2010.

[36] M. Wong and W. Schlitt. Sender policy framework (SPF) for authorizing use of domains in e-mail, version 1. Technical report, RFC 4408, April, 2006.

Summary Review Documentation for

"Measurement and Evaluation of a Real World Deployment of a Challenge-Response Spam Filter"

Authors: J. Isacenkova, D. Balzarotti

Reviewer #1

Strengths: Real world data of a deployed system; spam is still a hot topic.

Weaknesses: In an effort to be objective, the authors (willfully?) avoid some fairly obvious conclusions about challenge-response systems. Intuitively, the throughput numbers per company and per day seem low, so I wonder about the generality of the study.

Comments to Authors: High-level: interesting paper that goes in-depth to a less often studies spam fighting technique. Despite the author's claims, I believe that this data actually speaks quite negatively about CR-style systems, and as a result, will be of interest to the community looking to provide numbers to what it (likely) already intuitively believes.

Specifically, in S5.1, I agree with your conclusion that there is not a good correlation of challenges sent to number of servers blacklisted, but I feel like this is missing the point. Any system that could result in your outgoing mail server being black listed seems inherently like a bad idea. You say that 75% of servers never appeared in a blacklist, but that means that 25% did get blacklisted at some point! This typically requires some amount of proactive work on the part of the admin to unblacklist the mail server, or worse, debug why the outgoing mail is getting filtered by the destination. It is nice to quantify what that probability is, but my guess is that these blacklisting rates will be enough to scare off most administrators. Further, the reality is that someone could maliciously trick a CR system into getting black listed with 100% certainty, which is even worse.

My belief is the back-scatter *rate* is a little bit of a red herring. Emails are ultimately very small volume in terms of bits on the wire relative to other Internet services. What's more of interest is the number of messages that actually annoy non-CR participating senders. Have you considered doing a user-survey to try to estimate this?

Judging from back-of-the-envelope numbers from talking to local campus mail admins, 90M mails in 6 months for 20K people seems on the low end (~25 emails per person day, including spam, so at 90% spam, ~3 non-spam messages/person/day?). How does your study compare in terms of scale (emails per person per unit time) of non-CR previous work?

The data in figure 9 would have been nice to see also as a time series to understand how bursty white list changes are.

Reviewer #2

Strengths: Real world data; detailed analysis of a less understood spam fighting

Weaknesses: Strong conclusions from insufficient.

Comments to Authors: I was most concerned about the server blacklisting, 25% seems high. Isn't that the most negative argument against CR systems? 25% of mail servers means a lot of email delays,and lots of offline work to get them off the blacklist.

Even if there is a large amount of backscatter challenges being sent to innocent users, is it really such a big deal from a network traffic point of view? I would be more worried about other costs to the innocent users, rather than unwanted traffic volumes.

Reviewer #3

Strengths: The results presented in the paper help to illustrate the impacts of using Challenge-Response system with real-world data. It provides useful information for making decision on adopting the system.

Weaknesses: While analyzing real-world data, authors shy away from drawing any conclusion and stop at providing numbers. Some of analysis is incomplete and lacks depth. See below for detail.

Comments to Authors: Overall, this paper is written well. It is interesting to see the results presented from the viewpoints of a user, a system administrator, and the Internet.

In section 3.1, on understanding backscattered ratio, clearly it depends on the effectiveness of spam filter. It is not clear the result of 19.3% backscattered ratio reflects the spam filter or the effectiveness of the CR system. On the other hand, CR system and spam filter collectively fight spams. High backscattered ratio will bother many "innocent senders". The tradeoff between accuracy of spam filter and backscatter spam will be interesting.

As part the message or conclusion of the measurement result, I find that only 4% of challenges to be legitimate bothersome (section 3.2). Furthermore, 45% of challenges seem to belong to the case that the challenges are sent to "innocent senders". This effectively introduces spams to the system. Furthermore, how would spam filter treat these newly introduced spams. Could it be that the spam filter has already filtered those challenge questions?

Figure 5 is a bit confusing to read. The plots below diagonals should be explained further.

Reviewer #4

Strengths: The main strength of this paper is that it is the first study of real world deployment of challenge response spam filters. The evaluation covers a lot of data accumulated from different sources (47 companies of different sizes).

Weaknesses: The main weakness of the paper is that it presents the results while reserving judgment on the results. The paper would

have been stronger if the authors explained how some results were good or bad when using CR systems. Also, in addition to evaluation of an existing CR system, the authors should have presented how certain design choices affect the results, and how CR systems could be improved.

Comments to Authors: This paper is a good first step in understanding the impact of CR systems in real world. The paper can be made stronger by providing guidance/insights regarding choices to make in design/deployment that will alleviate the problems of CR systems.

Reviewer #5

Strengths: Cool dataset, and analysis that provides some helpful numbers where people have often speculated. The paper looks at most aspects that came to mind while I read the paper.

Weaknesses: The studied systems are a niche solution and unlikely ever to make a real difference (which, to their credit, the authors acknowledge right from the start). The paper's execution is pretty rough at times and the presentation confusing.

Comments to Authors: In Section 2, it would help if you could describe your datasets earlier. I found it confusing to read about results of open/non-open relays before you describe the actual dataset. Similarly, it'd be nice to see the CR's architecture before presenting any results.

What size are the companies that have deployed the system? In which country/countries are they located?

90 million messages isn't exactly "almost 100 million", as you're stating in the "General Statistics" subsection. In Table 1, percentages would help; same for Figure 3.

In 3.1, separate notation for the two notions of R (accepted deliveries vs attempted deliveries) would help.

You need to define clearly what you mean by backscattered emails. In Section 3, you first describe them as "challenge emails sent out by a challenge-response system". Later it becomes clear that you mean only *misdirected* challenge messages. Then you conflate reflection ratio and backscatter ratio (which you mention in 3.1 but don't define until 3.2), etc. I found this quite frustrating.
I'm not sure whether looking at Figure 5 I should be in awe or send the information visualization police and have you all arrested on the spot. For sure, it's far too complicated and busy, because I keep needing to go back and forth between text and plot to understand what it shows. For example, I remained confused about how negative numbers could possibly represent a ratio until you explain that the numbers in fact are correlations. All I learn about the sizes of the companies in your dataset is that most (I can't tell the percentage, given the useless y-axis label) have less than 500 employees -- that's really rather coarse information. The alternating axis label locations are the icing on the cake. You have room in the paper, why not break the chart apart into individual, semantically coherent ones?

The description of Figure 7 in 4.2 doesn't match what Figure 7 shows in any straightforward way. I don't see how your observations regarding Figures 7 and 8 affect the numbers you report in the three bullet points ending Section 4.2.

I like 4.3. It reflects that the degree to which CR works depends a lot on the interplay of users' involvement, on *both* sides. Some recipients may diligently weed out their digest, while others may let it grow and hope for the senders to solve the problem by solving the CAPTCHAs. The flip-side applies to the senders -- why would they solve the CAPTCHA right away when there's a chance that the recipient is diligent and will see it in the digest anyway?

In Section 5 you start speaking about invitations. Are you referring to challenge messages?

The writing needs a lot of work. Fix: "the same Symantec's report" -- do you mean "the same year's Symantec report"? "1% of false positive", "possible conspicuous", "users that may [...] receives", "we collected statics", "let's", "challenge messages were never open", "between every pairs", "one of the most effective way", "withelists", "messages where bounced". Broken reference (or cut'n'paste bug?) in 4.2. :("SpamHause", "do not exists", "a challenge-response systems").

Response from the Authors

First of all, we want to thank all the reviewers for their insightful comments on our work. According to most of them, the main weakness of the paper was the lack of conclusions based on the performed experiments. Even though our main goal was to present and discuss the data, and not to draw personal conclusions, we decided to address the reviewers' comments by adding some overall conclusions to the discussion section.

Another point raised by two reviewers was that Figure 5 was hard to understand. We took this comment into account by better presenting those results and by adding a detailed description of each part of the figure.

We also found extremely useful all the suggestions about the ways in which we could improve the paper presentation and the description of the data. In particular, according to the authors comments, we completely reorganized section 2, corrected the confusing definitions of the backscattering and reflection ratios, and clarified the results on the impact on emails delivery (Section 4.2). Furthermore, we stressed some of the points (e.g., the fact that misdirected challenges could be filtered out by spam filters) and we added more precise comments on possible inaccuracy of some of the presented figures.

The first reviewer mentioned the fact that 90M emails for 20K accounts seems on the very low end. This misunderstanding is probably due to the fact that the 20K accounts also include a large number of internal mailboxes used for administrative purposes, accounts of previous employees, and mailboxes that were inactive during the period of our experiments. Additionally, as we already mentioned in the paper, not all of the servers were monitored for 6 months (the period during which we had access to the data varied between 2 and 6 months). And lastly, some of the accounts received hundreds of emails per day, while others only few per week, thus accounting for the low average mentioned by the reviewer.

Finally, we could not address few comments (e.g., about the size and physical location of the companies) in order to preserve the privacy of the data providers. In some cases we were not able to evaluate some of the reviewers proposals (e.g., by trying different spam filters) that, even though extremely interesting, would have required us to change the production servers.