Preventing Spam For SIP-based Instant Messages and Sessions

Kumar Srivastava
Henning Schulzrinne
Department of Computer Science
Columbia University
{kumars, hgs}@cs.columbia.edu

October 28, 2004

Abstract

As IP telephony becomes more widely deployed and used, tele-marketers or other spammers are bound to start using SIP-based calls and instant messages as a medium for sending spam. As is evident from the fate of email, protection against spam has to be built into SIP systems otherwise they are bound to fall prey to spam. Traditional approaches used to prevent spam in email such as content-based filtering and access lists are not applicable to SIP calls and instant messages in their present form. We propose Domain-based Authentication and Policy-Enforced for SIP (DAPES): a system that can be easily implemented and deployed in existing SIP networks. Our system is capable of determining in real time, whether an incoming call or instant message is likely to be spam or not, while at the same time, supporting communication between both known and unknown parties. DAPES includes the deployment of reputation systems in SIP networks to enable real-time transfer of reputation information between parties to allow communication between entities unknown to each other.

1 Introduction

"Spam" can be defined as Unsolicited Bulk Communications (UBC) [9] [3] or calls and emails sent to groups of recipients who have not requested it. Spam is a major problem that plagues communication networks. Not only does it hog expensive network resources but reduces productivity by wasting precious man-hours. While the use of caller-id is the only way to block out spam in phone networks, there are few mechanisms to eliminate spam in email such as access lists for the sender (either black or white lists), content based filtering and message limits enforced through computations and Turing tests. However, the ease of creating and assuming new identities renders access lists too restrictive while new ways of hiding content in email and instant messages make content filtering either ineffective or prone to false positives. Content-based filtering techniques such as the Naive Bayes model can have false positive rates of up to 12% [19].

In the context of SIP-based sessions, UBC can be in the form of SIP-based calls or instant messages(IMs). Growth of IP telephony networks and SIP-based networks is bound to attract the attention of spammers who will find it lucrative to exploit such networks for the purposes of sending UBC. Mechanisms such as access lists (lists that explicitly allow or block communication from users on the list) are too restrictive and are not effective on a large scale. Content-based filtering could be used for SIP-based IMs but can not be used with SIP-based calls. To restrict spam in SIP-based sessions, there is need for sender filters that restrict a single person or application from easily assuming new identities to evade blacklists or restrict them from assuming the identity of a third party [25] [24].

We propose DAPES[1](Domain-based Authentication and Policy-Enforced for SIP), an architecture that aims to address the most likely source of SIP-based spam. DAPES relies on the common use of outbound proxies in the SIP architecture, reflected in the so-called "SIP trapezoid" consisting of the two user agents and the outbound and destination proxy. The outbound proxy verifies the identity of the caller within its local domain. However, identity verification is insufficient to remove spam, as some domains allow the easy creation of new identities. Thus, we add domain descriptions and reputation systems to allow the recipient to decide whether an identity is trustworthy and if it is likely to send spam.

[1]This work is supported through a grant from Nortel Networks

DAPES is not simply an extension to access lists. DAPES classifies incoming SIP-based calls or IMs according to the likelihood of them being spam. It deals with all such calls or IMs in a specific manner by using relevant information for making spam-related decisions. DAPES ensures communication between both known and unknown users while at the same time making informed decisions about spam in real time through the use of domain descriptions and reputation systems.

The remainder of the paper is structured as follows. Section 2 gives a more in-depth look at the goals of DAPES. Section 3 lists some of the assumptions and requirements of DAPES followed by Section 4 which categorizes all possible domains into five categories. Section 5 describes the two main stages that ensure authentication and verification in DAPES. Section 6 presents several different scenarios of communication and describes how DAPES deals with incoming calls in such scenarios. Section 7 describes DAPES's reputation system followed by acknowledgments in Section 8.

2 Goals

Traditional approaches to control spam such as spam filters and access lists restrict access to email or calls that might be potentially useful to users while at the same time requiring constant maintenance [15] [2]. They also prevent unknown but legitimate users from contacting other users. DAPES is aimed to support communication with both previously known and unknown entities. Previous approaches to controlling spam have concentrated on implementing versions of white-lists or black-lists and the use of content filters. Whitelists block all calls from unknown users including potentially non-spam calls or prevent previously blocked callers from making genuine "non-spam" calls. It suffers from the problem of scalability as creating and maintaining such lists can quickly become unmanageable. Content-filtering, though popular in email systems, is not applicable to SIP calls since there is no way to filter calls based on their signaling content. A mechanism is needed that can perform real-time authentication of the calling entity and prevent any SIP calls with spoofed *From* header fields or calls originating from spoofed domains or user identities.

DAPES can be extended to prevent spam in email. It can be used to devise a mechanism to ascertain the trustworthiness of a business partner, friend or investment adviser. However, initially, we are only concerned with determining if a user is likely to send spam via SIP.

A separate, but related problem is that of nuisance calls and messages where the caller does not necessarily place a large number of indiscriminate calls. For example, there is anecdotal evidence that users with female-sounding names on Skype are receiving unsolicited messages and calls from individuals. The domain-based authentication mechanisms described in this document are of limited use to prevent these cases. Section 6.5 describes some possible approaches.

3 Assumptions and Requirements

This system has a number of assumptions and requirements:

- We assume the two-hop model as shown in Figure 1. All calls leaving a SIP domain are routed through outbound SIP proxies. Incoming calls from foreign domains are accepted only if they are preceded by Transport Layer Security (TLS) or IPsec authentication between the SIP proxies of the SIP domains [6].
- The architecture is compatible with existing SIP standards and implementations to facilitate deployment.
- Outbound proxies have certificates signed by well-known Certificate Authorities (CA) [10].
- We assume that individual users do not have public key certificates signed by a well-known CA. (They may have self-signed certificates, but these are insufficient to ensure the authenticity of the caller unless such certificates are created through a bootstrapping architecture that allows for the certificates to be used within a specific environment. [12] An example of such an architecture is 3GPP [31]).
- Calls are routed through outbound proxies belonging to the domain of the caller.
- All signaling associations between proxies use TLS or IPsec.
- Outbound proxies verify the identity of the caller, as described in Section 5.

- We do not deal with the case that a user is known to the callee and previously having a good reputation sends an unsolicited bulk instant message, such as the annoying urban legends forwarded to friends. Similarly, the case of a compromised host placing unauthorized calls using its own identity can not be detected, but blacklists can temporarily prevent such calls until the compromised end system has been repaired.
- For most of the paper, we assume the standard SIP two-hop model. If a call is forwarded across more than two hops, the identity can't be confirmed by the third and subsequent proxy hops. However, the first forwarding proxy can determine if the domain of the caller corresponds to the domain of the outbound proxy. Thus, we essentially rule out the equivalent of email "open relays". An open relay outbound proxy may be necessary for opening pin holes in a firewall in a visited domain, however, we describe the problem in more detail in Section 6.3.

4 Domain Classification

DAPES performs domain-based authentication and verification of incoming calls to determine whether they are spam or not. Most UBC originates from domains created by spammers for sole purposes of spamming. DAPES tries to determine the trustworthiness of the incoming call's source domain in order to determine the likelihood of the call being spam. Before we introduce the domain classification used in DAPES, we present some of the common "trustworthy" communication domain types that exist presently.

4.1 Common Communication Domain Types

In practice, sources of none-UBC domains can be divided into a small number of families:

Employer: Some domains, such as microsoft.com, are used exclusively by employees (or students and alumni) of a (typically larger) organization.

ISP: Internet service providers typically provide email accounts to their customers and are likely to offer SIP-based VoIP services as well.

Associations: Associations such as IEEE or ACM offer their members permanent email identifiers whose incoming mail is usually redirected to an employer or ISP account.

Personal: Some individuals, families and small organizations obtain their own domains and use them for sending and receiving emails, as web hosting providers often provide email services with their hosting services.

Mailbox providers: A number of organizations, ranging from portals to email-only service providers, offer identities under one or more of their domains, for free (advertising-supported) or for some fee-for-service arrangement. Examples include yahoo.com, gmail.com, 21.com, hotmail.com, pobox.com and dozens of others.

All of these types of domains offer email services today and may well offer SIP IM and VoIP services in the near future. In the SIP space, services like Free World Dialup (FWD) and iptel.org offer free memberships, allowing consumers to reach others within that same domain by some short-hand identifier. In the wider VoIP space, services such as Skype also offer freemail-like rendezvous services. Currently, VoIP service providers (VSPs), such as Vonage and Primus, typically identify their users by E.164 telephone numbers and thus are likely to fall under the same telemarketing restrictions as traditional, circuit-switched telephone companies.

Rather than this classification, however, we base the domain classification used by DAPES according to identity management procedures employed by domains. Since authentication in DAPES is domain-based, we present a model that classifies domains into categories based on their trustworthiness and the probability of affiliated users sending out spam. Domain classification is important as it defines the likelihood of a domain being used to send out spam. Classifying domains according to their trustworthiness enables us to implement dynamic access lists that force incoming calls from such domains to be dealt with appropriately. For example, if a domain's procedures to detect spam activity and revoke mail privileges for user accounts found to be sending out spam is verifiable and indicate that the domain is unlikely to be a source of spam, the destination SIP domain can be assured that the incoming call has a very low likelihood of being spam. On

the other hand, a call originating from a domain like Hotmail, which has no restrictions on creation of new accounts and identities and provides no guarantee on outgoing calls, has a higher likelihood of being spam and thus can be dealt with accordingly by the destination SIP domain. DAPES classifies all possible domains into five categories depending on their identity management policies and authentication, authorization and accounting procedures [17]. DAPES uses the following criteria to classify domains:

- Does the domain verify user identity on calls?
- How easy is it to get a user-name within the domain and what mechanisms (personal verification, financial payment, anti-bot measures) are used for identity creation?
- What happens if the user is found to be spamming, for example, loss of money or identity?
- Are the domain users limited in the number of messages that can be sent and are there procedures in place that prevent account creation by bots?

According to the above criteria, DAPES classifies domains into one of the five domain types shown below, in decreasing order of likelihood of being spam sources.

Admission-controlled domains: Admission controlled domains have long-term and personal relationships with their members. These domains have extensive identity instantiation procedures which are linked to the admission or membership to the domain. For example, employees of a company or students in a university are members in admission-controlled domains.

Bonded domains: Members of such domains do not need to have a personal relationship with the domain administrator and neither does the administrator need to keep track of each member personally. However, membership to the domain is contingent on the posting of financial bonds that are tied to observing certain rules of behavior and conforming to the policies of the domain. Types of such behavior expectations can be a guarantee of not sending unsolicited messages or conforming to a daily outgoing message limit. The members of such domains agree to forfeit the bond if they are found to be violating these rules or policies. In such domains, anonymity and the use of pseudonyms is not a problem as long as each identity has a unique bond associated with it.

Membership domains: Members of such domains are not known to the domain administrator but are required to provide verifiable identification such as credit card information (assuming the credit card company has throughly verified the identity of the user) or proof of membership to other membership, admission or bonded domains to the domain. The credit card is linked to the identity and makes it difficult for the user to create multiple identities within the same domain.

Open, rate-limited domains: We can also envision a domain that is open but limits the number of messages per time unit. In order for rate limits to be meaningful, they have to prevent account creation by bots. Such domains could enforce a rate-limit on the creation of new accounts. Applicants can also be required to solve riddles that are meant to consume CPU cycles. Since legitimate users would not need to create multiple identities, this would impose a restriction only on potential spammers. Yahoo, for example, uses screen reading to prevent automatic account creation by bots.

Open domains: Membership to these domains require no user authentication. Multiple identities can be created with no limit or check on usage. Webmail systems such Hotmail or ISP trial accounts fall under this category of domains.

It is important that all domains be classified uniquely as one of the above mentioned domains. To ensure that spammers do not simply buy cheap domains and claim that all spammer accounts are personally known to that domain, an independent third party or a set of third parties are required to maintain records about domains. These records contain relevant information about the domains such as account policies, restrictions and the type of domain. In addition, such records can also contain information about the reputation of these domains, information that is publicly available to all interested parties.

5 Stages of Caller Verification in DAPES

Any SIP communication in DAPES goes through two stages of verification. As mentioned above, DAPES uses the "SIP-trapezoid" consisting of the two user agents and the outbound and destination proxy as shown in Figure 1. The first stage of verification deals with verifying the identity of the caller within the local domain. This ensures that every call leaving the domain has been properly authenticated and the identity behind the caller is properly verified. This is a significant step because for domains that can be verified as being "trustworthy", a destination proxy simply has to verify that the call is indeed originating from the registered outbound proxy for that domain. If this verification is successful, the destination proxy can let the call through. However, for open domains, the destination proxy cannot trust the identity verification in the local domain as new identities in open domains can be easily created. We can still prevent the spoofing of legitimate addresses through a second step in verification. DAPES uses TLS authentication between the outbound proxy and the destination proxy and verification of the outbound SIP proxy through Domain Name System (DNS) Service Records (SRV) [27]. SIP proxies of the source and destination SIP domain perform mutual TLS authentication using CA signed certificates. Following successful TLS authentication, the destination SIP proxy queries a DNS server for the SRV records of the source SIP domain. The destination SIP proxy verifies that the outbound SIP proxy of the source domain(listed as the outbound proxy for the call in the INVITE, and also the outbound proxy that was authenticated through TLS authentication) is listed as a legitimate outbound SIP proxy for the source domain. This allows the destination SIP proxy to verify that the outbound SIP proxy of the incoming call is authorized to forward calls from that domain. The stages of verification in DAPES are discussed in more detail below.

5.1 Verifying Local User Identities

As mentioned above, it is very important that local user identities are thoroughly authenticated within the local SIP domain. This guarantees inbound proxies of other SIP domains that incoming calls have been properly authenticated before being forwarded to them. This authentication can be done in two different ways, namely:

5.2 Digest authentication on INVITE

The outbound proxy of a SIP domain performs digest authentication on INVITE messages by challenging all requests coming from users that belong to that domain. This forces all users that have this domain name in their *From* header field, to get authenticated from the outbound proxy before their calls are allowed to leave that domain. This enforces authentication at the outbound proxy thereby guaranteeing that all calls originating from this outbound proxy are authenticated and belong to callers registered with this domain [26].

5.3 Digest authentication on REGISTER and INVITE with address verification on INVITE

The authentication of both REGISTER and INVITE messages is always required. Digest authentication of REGISTER is required to prevent hijacking of calls for users and also to inform the server of the current location of the user. Any subsequent INVITE messages from a registered user have to contain the same contact information as in the REGISTER message thereby confirming that the call is being placed by a previously authenticated user [26]. Digest authentication on INVITE ensure that unregistered users are not able to place call without being authenticated by the proxy server.

An outbound proxy that has access to the registrar database may be able to verify that the INVITE originates from one of the IP addresses registered earlier in the *Contact* header of a REGISTER. This will generally work for users behind NATs only if the registrar notes the external address of the registration request.(we assume the common scenario that all outbound connections from a single end system have the same external IP address) Address-based verification is weak, but may suffice for the purposes described here. In order to prevent IP address spoofing, the outbound proxy has to at least perform null authentication as described in Section 22 of RFC 3261.

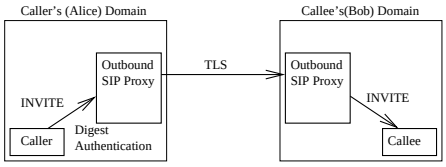

Figure 1: Verification process

5.4 Verifying Previously Known Outbound Proxies

DAPES ensures the authenticity of an incoming call by authenticating the outbound proxy of the SIP domain from which the call is originating. This verification is done using TLS and makes use of TLS certificates [6]. Once the SIP proxy for the callee's domain receives a call from an outbound SIP proxy of another domain, the two proxies engage in a TLS handshake to authenticate themselves to each other. Following this, the SIP proxy of the callee's domain proceeds to carry out domain verification of the caller's outbound SIP proxy. It tries to verify that the proxy is indeed registered as the SIP proxy for the incoming call's origin domain. It sends a DNS SRV or Naming Authority Pointer (NAPTR) query to the local DNS server and tries to verify that the address of the caller's proxy listed in the contact header indeed refers to the registered SIP proxy for that domain [27]. This approach counters SIP calls with spoofed *From* addresses as it allows the destination proxy to make sure that the IP address from which the incoming SIP call was received refers to a legitimate outbound SIP proxy of the source domain. The TLS handshake does not pose too much of an overhead as it does not have to be done on a per-message basis.

The above approach can also be used in the case when TLS support is not available, "inverse MX" records that have recently been proposed for SMTP [18] [16] could be approximated for SIP by having all outbound proxies register as proxies for a domain, even if they do not accept inbound SIP sessions. If they do not accept inbound SIP requests, they are simply registered with the lowest priority, ensuring that they are only contacted if all other SIP proxies for the domain are out of order. In most cases, outbound-only proxies will require digest authentication and thus fail these inbound requests.

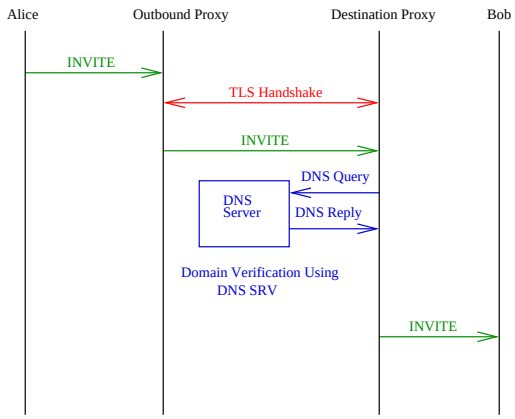

Figure 2: Message flow for domain verification

5.5 Verifying roaming users

To authenticate and verify roaming users or users that are visiting domains different from their home domains, DAPES provides a mechanism where such users can register with their home domains and route their messages through the home domain. Thus, even though the incoming request might be originating from an unknown or untrusted domain, if it is being routed through the user's home domain, it will be accepted. TLS authentication at each step is still assumed.

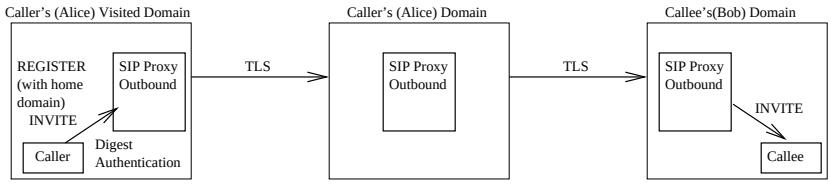

Figure 3: Message flow for verification of roaming users

6 Models for Classification of Callers and Domains

DAPES uses domain classification to distinguish between SIP-based sessions and the likelihood of them being spam. This is important as we do not want regular known callers to go through extensive authentication and verification procedures while at the same time, unknown callers should be properly authenticated before their call is processed. This requires us to model all calls as one of a finite set of scenarios which can then be dealt with accordingly. These scenarios are described below.

6.1 Known User

Users that are known to the caller through personal relationships or previous communication (incoming call log) can be placed on whitelists. For calls made by such users, the only requirement for further processing of the call is the authentication of the user. To achieve this, we propose two methods.

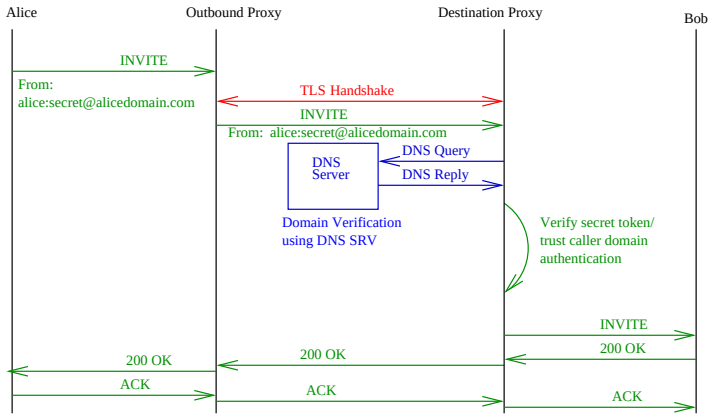

Figure 4: Message flow for the authentication of a known user in a trusted domain. Local user authentication between Alice and the outbound SIP proxy is assumed

- **Trusting caller's local authentication**: The callee can simply trust the caller domain authentication of the caller. This builds on the previous stated requirement of the system that all INVITE or REGISTER messages have to be authenticated through digest authentication. Since all callers conforming to this protocol are guaranteed to have been authenticated by their outbound SIP proxies as a precondition to further processing of their call, the callee's SIP proxy can trust that the caller is indeed who he is claiming to be.
- **Authentication through secret tokens**: Users that already have an existing relationship with the caller and have been authenticated using the authentication procedures defined for unknown callers and unknown domains or unknown callers and known domains (explained below) can be awarded secret tokens. These tokens can then be presented at all successive calls and can be used by the callee's SIP proxy to authenticate the caller (see figure 4). The usefulness of this approach is demonstrated in the use of mailing lists that can have unlimited recipient size. To prevent calls that are part of a mailing list from being trapped and dropped by SIP proxy, secret tokens can be included with such calls with multiple recipients essentially implementing a whitelist of users authorized to place such

calls. Currently, there is no mechanism to carry such secret tokens in the INVITE message, however using such tokens involves two parts. First, the callee awards the secret token to the caller in encrypted form. For any subsequent communication, the secret token is appended to the SIP message also in encrypted form. Encryption could use a symmetric key or any of the PKI methods discussed in this section.

- **Authenticated identity and the saml based approach** If the incoming SIP message conforms to the Authenticated Identity Body (AIB) format according to the specifications in [22] i.e. certain headers of the SIP message have been included in a signed MIME body, the identity of the caller can be easily verified.
- **End-to-end security using sipping-cert** DAPES also supports the use of Address of Records (AOR) to locate the certificate of the caller. If such a public key can be retrieved, S/MIME based PKI authentication can be easily accomplished. According to the procedure defined in [11], such certificates can be created, stored and retrieved easily.

6.2 Unknown User, Trusted Domain

Users that have been authenticated by their domain's outbound proxy but are not known to the callee go through the following authentication and verification procedure. Once the callee's outbound proxy has authenticated the incoming call's outbound proxy and has determined it to be a trustworthy domain, it can then request the SIP proxy of the other domain or independent providers of such services to provide information about the caller and specific domain policies. This information is then used to determine the trustworthiness of the caller.

The information that is requested by the callee's SIP proxy is given below.

- The authentication method indirectly used to verify the identity of the user: This information can be about digest authentication or mutual PKI-based authentication through TLS [26].
- The domain's policy regarding the generation of unsolicited messages: These can range from no policy, to daily outgoing message limits, penalties for sending unsolicited messages through financial penalties and identity revocation. This information can be requested from policy services that maintain and store up-to-date information about a domain's policies. For example, a admission domain will have strict procedures for checking spam activity and penalizing its members. This information could be used by the caller's domain for making a decision on the likelihood of the call being spam.

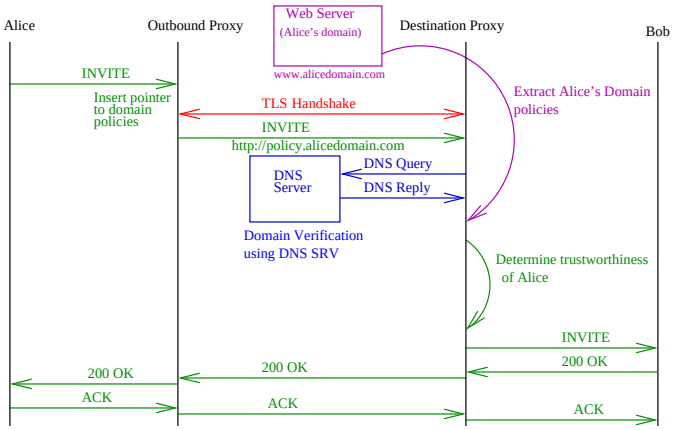

Figure 5: Message flow for the authentication of a unknown user in a trusted domain

There are several methods that the domain can use to provide such information to the callee's proxy server. This information can be included in a SIP message, embedded in the SIP proxy certificates or it can be published in a machine-readable form such as XML on the domain's public server and the SIP proxy for that domain can return a pointer to this file [28]. The callee's proxy server can retrieve the policies and

subsequently make its decision on whether to terminate the session or forward the call to the callee. It is important that these policies have limited lifetimes and that they are updated regularly and verified by an independent registrar to prevent false advertising and misuse.

A second approach to verify unknown users from trusted domains is to allow such users to include in their initial INVITE message, the SIP contact information of common, trusted friends between the caller and the callee. The callee can then contact and check the roster of this common friend and confirm the status of the caller. The underlying basis of this approach is the use of social networks that permit the flow of transitive trust from one user to another. Such functionality is demonstrated through social networks such as Orkut [36] or Friendster [32] and can be easily implemented using SIP [4] [20] [8] [7] [14] [1].

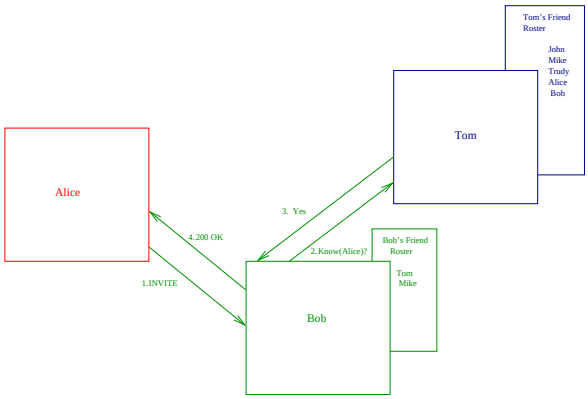

Figure 6: Flow of transitive trust between friends of friends

6.3 Roaming Users

As noted in Section 3, the DAPES model does not allow open relays as the destination proxy needs to verify that the domain in the TLS certificate and the domain in the SIP *From* header agree. For the case of the roaming user, e.g., using the services of a hotel or airport network, the outbound proxy cannot verify the user's identity, but can limit the message rate for each IP address to make spamming unlikely. Such a message limit effectively transforms the domain into a rate-limited domain. To provide stronger assurances, a user forced to use an outbound proxy in a visited domain should route its messages through its home proxy, authenticating itself to the proxy as described in Section 5.1.

6.4 Call Redirection

Call forwarding via proxying or redirection responses does not pose a problem as long as the recipient can trust the previous hop to have verified the caller identity. If the destination proxy forwards a call to another domain, the destination proxy server will detect a difference between the proxy domain and the domain in the *From* header field and reject the call. However, calls are generally only forwarded from domains that the callee is familiar with, as forwarding calls to random domains is unlikely to be useful. Thus, the final destination should ascertain that the domain of the previous hop is known to the callee. It may also obtain the verification policy of that domain to make its decision.

6.5 Nuisance Calls

In the PSTN, nuisance calls, including prank calls and stalkers, have been a problem for a long time, with a variety of mechanisms to limit their impact. We distinguish nuisance calls from UBC by their individualized and possibly non-commercial nature.

Scarcity of identifiers: Since telephone numbers are hard to get, requiring subscription to telephone service, blacklisting of numbers is effective at limiting repeated nuisance calls from the same person. Determined stalkers can still use pay phones, but this incurs significant inconvenience.

Traceback and legal remedies: After repeated nuisance calls, the telephone company will trace the identity of the caller and possibly initiate legal proceedings. The authors are not aware of data on how often this mechanism is used successfully.

Cost of distance: Since long-distance and international calls still have non-zero costs, the set of nuisance callers is somewhat limited.

VoIP and IM do not suffer from any of these limitations. The removal of these limitations is a core advantage of these services, but also requires new remedies to deal with nuisance callers. Indeed, any system that allows communication between strangers is likely to suffer from nuisance call problems. While nuisance communication does not pose a significant problem for email, as the cost of ignoring such messages is low for most recipients, the more intrusive nature of IM and VoIP calls makes even a small number a major problem, particularly as their geographic reach may well increase.

The identity and domain classification services introduced earlier help to some extent. For example, they allow setting up blacklists. More effective is likely the use of collaborative rating systems for individual callers so that repeat offenders, cranks and other "problem" callers can be blocked by a callee before they reach that particular callee. This approach will be helped if public devices are not identified by the device but rather by the caller using it, as otherwise most pay-phones would quickly make any such blacklist. Also, due to the unpleasant personal consequences of making nuisance calls from one's employment-related address, there are likely to be domains that are going to be rarely sources of such nuisance calls. Open and open rate-limited domains appear to be the most likely source of nuisance calls, as well as public devices such as payphones.

One of the oft-cited cases why blacklisting of unknown callers is not desirable are variations of the "spouse calling from pay-phone on highway" problem. However, there may well be a relatively simple solution for such calls that fail all other vetting attempts (good reputation of caller; not in address book or outgoing call log; admission-controlled, membership or bonded domain, etc.). For example, an automated answering system could ask the caller or IM sender to supply some personal information about the callee that any friend or relative is likely to know.

Variations of such a solution have been employed for email, requiring recipient liveness verification. However, these approaches, while effective against computer-generated spam, also tend to fail for legitimate business-related and mailing list messages.

6.6 Unknown User, Unknown Domain

Users that are not known to the caller and are affiliated with domains that are also not known to the callee's proxy server are verified through the following procedure. Since no information is known about the domain or the user, the proxy server needs to retrieve information about the trustworthiness of this domain and caller from a third party. To achieve this purpose, we propose the implementation of reputation systems for SIP. Unknown users originating from unknown or open domains can include in their session setup, information about a reputation server which stores reputation records for the user. The callee's SIP proxy can then inquire about the reputation of the caller and use the reputation information to further process the call.

7 DAPES's Reputation System

DAPES's reputation system is similar to the concept of "karma" in community web sites such as Slashdot.org or seller/buyer ratings on auction sites such as eBay. In DAPES, the destination proxy can make reputation queries to a centralized registrar assuming the role of a reputation server containing reputation records for users and domains.

The reputation system is implemented as a set of distributed servers that maintain lists of all registered users, domains and their reputation ratings as a database. It is very important that this reputation system be protected from modifications by untrusted users for either improving or denigrating the reputation of other users on the system. This can be implemented through proper authentication of users that attempt to access or modify the database. Access to this reputation system can be modeled as a request to the reputation registrar and that call can be processed in a similar fashion as shown in Section 6.1.

It is also important to make sure that access to the reputation system is not just contingent on the entity (trying to make the modification) being recognized as a member of a trusted domain by the registrar maintaining the reputation system. The callee has to prove that it indeed received a call from the caller

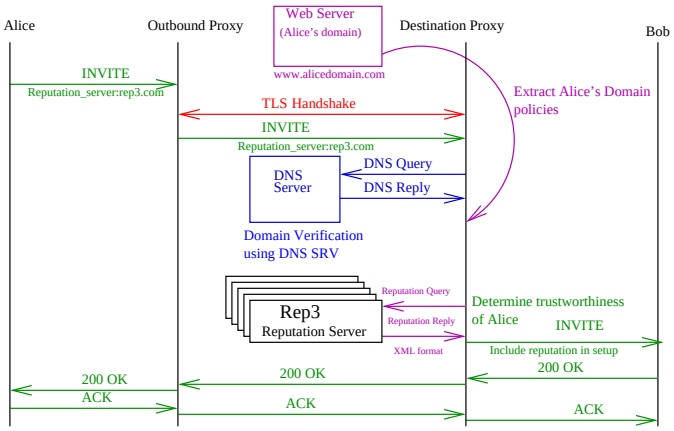

Figure 7: Message flow for the authentication of a unknown user in an unknown domain

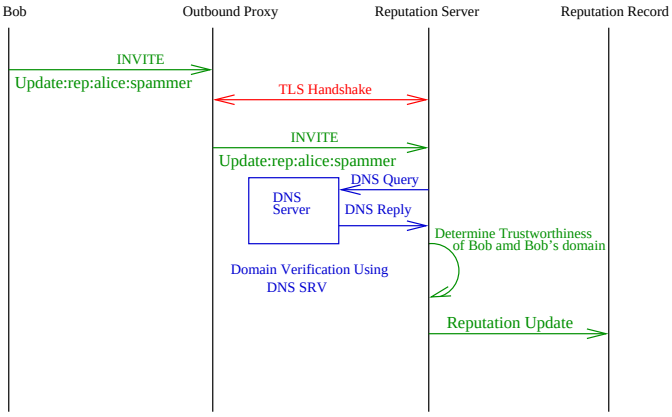

Figure 8: Message flow for a reputation query from a callee

whose ratings they are trying to modify. This might not be possible for some scenarios mentioned earlier however, for most cases, such information could be retrieved and matched quite easily. Evaluations of callers have to be limited to one. Multiple evaluations for a single call would allow malicious users posing as callees to repeatedly modify reputations for the same caller. In addition, the system needs to be able to enforce the number of allowed evaluations by a single caller. A callee should be able to modify the rating of a caller only once even if they have received several calls from the caller in a short period of time. In order to give the callee's or their outbound proxies making use of this reputation system the ability to verify the credibility of these evaluations, the system can assign domain-based weights to evaluations depending on the reputations and trustworthiness of the callee's domain and consequently of its members. If a callee from a highly reputed domain evaluates a call, that evaluation should hold more weight than an evaluation from a callee of an open domain such as Hotmail.

A potential problem that unknown users might face while trying to include information about a reputation server in their session setup request, is of anticipating which reputation server will be acceptable to the callee's SIP proxy. This problem could be solved by implementing the distributed reputation servers as a network with the ability to search for and propagate requests to the correct reputation server.

We summarize three popular reputation systems below.

7.1 eBay's Reputation System

eBay [34] uses a system where every user has an associated "Feedback Profile". This profile consists of comments from other eBay users who were involved in a transaction with that user. The feedback rating is essentially a sum of points. A positive comment from another user adds one point to the total and a negative comment reduces the total by one. Zero points are given for neutral comments. An additional "star icon" is given to the user's rating for 10 or more feedback points. eBay maintains user profiles consisting of the feedback score, positive feedback, numbers of members leaving positive and negative feedback, all positive feedback, recent ratings, and the number of bid retractions.

eBay's reputation system incorporates many of the features that we propose as part of our reputation system. eBay users have only 90 days to rate the other users in a transaction and the ratings made are permanent and can not be edited or removed by the user. The rating is added to the permanent record of that evaluated user. According to eBay, these ratings are also subject to defamation suits under the Communications Decency Act. eBay does allow mutual removal of the feedback rating but the user comments remain as part of the rating. It also has a list of rules that define whether a rating can be considered as "feedback abuse" and such ratings can be removed from the user's profile.

7.2 Amazon's Reputation System

Amazon [33] also has a rating system where costumers can rate sellers. This rating, like eBay, is a combination of user comments and a one-to-five star rating submitted by previous buyers. The member profile of each user consists of the seller's average rating for the past year, the lifetime average and comments that have been left on their account by other members. Like eBay, users have 90 days to submit such a rating. Amazon does not alter feedback ratings or comments once they have been submitted.

7.3 Slashdot's Reputation System

Slashdot [35] is a message board that is primarily used for posting articles and comments to those articles. To counter spam on the message board, Slashdot employs a moderation system that has a built in reputation system. When readers view the frontpage, they are randomly granted moderator or meta-moderator privileges, with a certain number of moderation (mod) points. Each comment that is moderated costs one point. Meta-moderators evaluate the previous moderation of comments.

Slashdot.org associates a "karma" with all of its users. Karma represents how a user's comments have been moderated in the past. The moderation rating can be any of the following: "Terrible, Bad, Neutral, Positive, Good and Excellent". Essentially, Karma is a "sum of [a user's] activity on Slashdot represented by integers". Karma also has a limit to which it can be increased to prevent misuse. A high value of karma allows users to moderate other users and a bad Karma rating of an account usually means that the account is being used to spam the bulletin board.

7.4 Comparison of DAPES's Reputation System with Traditional Reputation Systems

Reputation systems range from systems that have reputation ratings for sellers and buyers according to reviews of their past transactions (auctions), to expert sites where experts answer questions from users in a bit to gain higher ratings. Almost all reputation systems provide a reputation rating for users based on evaluations made by other users involved in some past transaction. This rating is an indication for the behavior of that user in future transactions. However, determining whether a user is likely to send spam in real time is a much harder problem. Our system has to classify an incoming session request as spam or not. Thus, traditional methods of aggregating information over several transactions are not entirely applicable here. The proposed system needs to be able to make real time queries to a reputation server and determine from the records stored for that caller whether the call is likely to be spam or not. The callee does not have the luxury of going through the reputation record of the caller offline and then decide whether or not to accept the call. However, similar to traditional reputation systems, DAPES's reputation system allows the evaluation of transactions that have already taken place and uses such evaluations to predict whether a user is likely to send spam.

Reputations for determining whether somebody is spamming are likely to be easier than comment-based evaluation or the evaluation of commercial transactions since users are likely to place more calls or send instant messages than sell or buy items. Also, there is no attempt made to evaluate the value of the content, simply whether a message is unsolicited bulk communication. It is likely that some types of messages, such as unsolicited conference announcements outside of mailing lists, may be considered spam by some, but not all, recipients. However, in such cases, a recipient can decide whether to accept calls from senders that have a mixed rating.

Apart from this, our system has the same basic purpose as other traditional reputation systems. They provide a "solution to the ubiquitous problem of trust in new short term relationships" [23]. All such systems give users the ability to predict the nature of their interaction with other users. However, traditional reputation systems suffer from several shortcomings which our system tries to overcome. These are [23]:

- **Forced feedback**: Unless feedback is made essential for completing a transaction, users have no incentive to provide any useful feedback. Our system forces all users to provide feedback for a call, even if the feedback is done automatically and leads to no change in the reputation of the caller. No feedback from the user can be used as an indication that the user did not consider this call as spam. Users can also be prompted to indicate whether they considered the call as spam or not. The response can then be used to modify reputation ratings at the reputation server. For calls that are determined not to be spam and the callee is in agreement with the distinction, the caller can be added to the list of callers known to the callee. In the other case, if a caller is determined to be a spammer, it is in the callee's interest that they do a reputation update on the domain and the caller in question. By building such incentives for reputation updates in the system, the reputation system can be kept accurate and up-to-date. Several mechanisms have been proposed to get good feedback from users such as those proposed by Miller et. all [21] and Jurca and Faltings [13] The proposed approaches range from financial rewards based on users providing ratings that match with the rating computed by the central reputation system to users buying reputation information from rating brokers before a transaction and then selling their rating to the brokers for money. Financial incentives for providing correct reputation ratings have been shown to work theoretically and thus can be utilized in the DAPES reputation system.

- **Negative feedback**: It is quite difficult to elicit negative feedback from users in traditional reputation systems [23]. eBay allows all users involved in a transaction to negotiate a settlement over a disagreement and thereby not let the disagreement affect the reputation ratings. Our system does not support modifications of existing reputations. Once a reputation is modified, it is final and no one is allowed to reverse the changes. Other proposals to counter this problem are presented by Chen and Singh [5] and Yu and Singh [29] [30]. Some of the solutions propose the use of ratings given to the same object to group raters, and giving different weights to different groups of raters. Other solutions offer mechanisms to distinguish reliable witnesses from deceptive witnesses by making use of referrals from agents personally known to the witness [30]. Such solutions try to weigh the reputation reports submitted by raters and weighs the reports accordingly while making a reputation decision. These are directly applicable to DAPES's reputation system as ratings submitted by users can be weighted as described earlier in this section.

- **Correct ratings**: Most systems have no way of determining whether a reputation rating is correct or if it has been motivated by personal discontent or any other unrelated reason. DAPES's reputation system can implement accountability if all SIP domains provide support in form of rules and regulations that impose penalties on malicious users.

- **Reputation queries**: Many systems suffer from the problem of multiple user identities, the majority of them being "pseudonyms". For example, eBay allows users to register with any name they want. Several traditional systems allow users to re-register with a new account. This enables them to lose their old reputation ratings and start afresh. Systems also suffer from the "lack of portability" [23] whereby reputations from one domain can not be used in a different domain.

- **Non-repudiation of calls or messages**: The reputation system must be able to verify that a user indeed placed a call. This is necessary to ensure that malicious users are not able to degrade another user's ratings by claiming that they received a spam call from the user. Non-repudiation of calls also enables the reputation system or the home domain of the caller to impose accountability on the caller and hold them responsible for any spam activity. This could be implemented by logging calls at multiple points of the system such as at the outbound SIP proxy, at the destination SIP proxy and at the callee and these records could be matched to check whether the call was placed or not. Calls could also be signed by the caller however, DAPES assumes that callers do not own certificates. A challenge-response mechanism could also be employed to ensure in real-time that the caller wants to place a call to the callee.

Our system allows only confirmed members of trustworthy domains to modify reputation ratings. This is achieved by using SIP to perform updates to the reputation system with all such requests being verified through DAPES. It is also a mandatory feature of the system that users have unique and permanent identities. In addition, the issue of portability can be easily solved by having a network of reputation servers that can facilitate the easy transfer of reputation ratings.

In a scenario where a call is proxied from the original destination domain to the final callee, the final destination will be unable to verify the caller. However, we assume that trustworthy domains will also filter calls before they are proxied, thus limiting the impact of such cases.

8 Acknowledgments

The authors would like to thank Hannes Tschofenig for his help in reviewing the paper.

9 Conclusion

We have presented DAPES, a system that can be deployed in SIP networks to prevent spam for SIP-based messages and sessions. DAPES makes use of domain descriptions to classify domains according to the likelihood of them facilitating spam. Using these classifications, incoming session requests are processed according to the protocol defined for those types of calls. DAPES is able to automatically recognize and block incoming spam for SIP-sessions while at the same time, allowing valid communications to proceed without any delays.

References

[1] Lada Adamic, Orkut Buyukkokten, and Eytan Adar. A social network caught in the web. *First Monday*, 8(6), 2003.

[2] Ron Anderson. Antispam techniques. *Security Pipeline*, May 2004.

[3] Chris Bonatti. A generalized mechanism for control of unwanted application communications. Internet draft, Internet Engineering Task Force, May 2004.

[4] Oscar Boykin and Vwanig Roychowdhury. Personal email networks: An effective anti-spam tool. *Preprint*, 2004.

[5] M. Chen and J. Singh. Computing and using reputations for internet ratings. *Proceedings of the Third ACM Conference on Electronic Commerce (EC01)*, 2001.

[6] T. Dierks and C. Allen. The TLS protocol version 1.0. RFC 2246, Internet Engineering Task Force, January 1999.

[7] Laura Garton, Caroline Haythornthwaite, and Barry Wellman. Studying online social networks. Technical report, JCMC, June 1997.

[8] Eric Gradman. Distributed social software. Technical report, LayerOne Conference, December 2003.

[9] Paul Hoffman. Unsolicited bulk email: Definitions and problems. Technical report, Internet Mail Consortium, 1997.

[10] R. Housley, W. Ford, W. Polk, and D. Solo. Internet X.509 Public Key Infrastructure Certificate and CRL profile. RFC 2459, Internet Engineering Task Force, January 1999.

[11] C. Jennings. Certificate discover for sip. Internet draft, Internet Engineering Task Force, August 2004.

[12] C. Jennings. Certificate discovery for SIP. Internet Draft draft-jennings-sipping-certs-02, Internet Engineering Task Force, February 2004.

[13] R. Jurca and B. Faltings. An incentive compatible reputation mechanism. *Proceedings of the 6th Int. Workshop on Deception Fraud and Trust in Agent Societies (at AAMAS03)*, 2003.

[14] Henry Kautz, Bart Selman, and Mehul Shah. Referral Web: Combining social networks and collaborative filtering. *Communications of the ACM*, 40(3):63–65, 1997.

[15] Neal Krawetz. Anti-spam solutions and security. *Security Focus*, February 2004.

[16] Christopher Laforet and Geoffrey Deasey. Enhancing SMTP Mail Services to Minimize SPAM. Internet draft, Internet Engineering Task Force, January 2004.

[17] J. Loughney and G. Camarillo. Authentication, Authorization, and Accounting, Requirements for the Session Initiation Protocol SIP. RFC 3702, Internet Engineering Task Force, February 2004.

[18] J. Lyon and M. Wong. MTA authentication records in DNS. Internet draft, Internet Engineering Task Force, June 2004.

[19] David Madigan. Statistics and the war on spam. *Rutgers University*, 2004.

[20] Christopher McCarty. Structure in personal networks. Technical report, Journal of Social Structure, July 2004.

[21] N. Miller, P. Resnick, and R. Zeckhauser. Eliciting honest feedback in electronic markets. *Working paper originally prepared for the SITE02 workshop*, 2003.

[22] J. Peterson. Sip authenticated identity body (aib) format. Internet draft, Internet Engineering Task Force, May 2004.

[23] Paul Resnick, Ko Kuwabara, Richard Zeckhauser, and Eric Friedman. Reputation systems. *Commun. ACM*, 43(12):45–48, 2000.

[24] J. Rosenberg, G. Camarillo, and D. Willis. A framework for consent-based communications in the session initiation protocol (SIP). Internet draft, Internet Engineering Task Force, July 2004.

[25] J. Rosenberg and C. Jennings. The Session Initiation Protocol (SIP) and Spam. Internet draft, Internet Engineering Task Force, July 2004.

[26] J. Rosenberg, Henning Schulzrinne, G. Camarillo, A. R. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler. SIP: Session Initiation Protocol. RFC 3261, Internet Engineering Task Force, June 2002.

[27] Henning Schulzrinne and J. Rosenberg. SIP: locating SIP servers. RFC 3263, Internet Engineering Task Force, June 2002.

[28] H. Tschofenig, J. Peterson, J. Polk, D. Sicker, and M. Tegnander. Using saml for sip. internet-draft, Internet Engineering Task Force, July 2004.

[29] B. Yu and M. Singh. An incentive compatible reputation mechanism. *Proceedings of the 4th International Workshop on Cooperative Information Agents*, pages 154–165, 2000.

[30] B. Yu and M. Singh. Detecting deception in reputation management. *Proceedings of the Second Int. Joint Conference on Autonomous Agents and Multiagent Systems*, pages 73–80, 2003.

[31] www.3gpp.com.

[32] www.friendster.com.

[33] www.amazon.com.

[34] www.ebay.com.

[35] www.slashdot.org.

[36] www.orkut.com.