Concerns around the Applicability of RFC 4474

draft-rosenberg-sip-rfc4474-concerns-00

Status of this Memo

By submitting this Internet-Draft, each author represents that any applicable patent or other IPR claims of which he or she is aware have been or will be disclosed, and any of which he or she becomes aware will be disclosed, in accordance with Section 6 of BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt.

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

This Internet-Draft will expire on August 20, 2008.

Copyright Notice

Copyright (C) The IETF Trust (2008).

Abstract

RFC 4474 defines a mechanism for secure identification of callers in the Session Initiation Protocol (SIP). This mechanism has been used as the foundation for some recent additional work, including connected party identification, anti-spam, and secure media. However, concerns have been raised about the applicability of RFC 4474 in real deployments and the actual level of security services it provides. This document describes those concerns.
Table of Contents

1. Introduction ................................................. 3
2. The Problem with Numbers ................................. 3
3. Attacks Introduced by Usage of Numbers ................. 5
   3.1. The Re-Sign Attack ................................. 5
   3.2. The False Number Attack ....................... 7
4. Consequences ............................................... 8
   4.1. Unsecure Caller ID for Phone Numbers ............ 8
   4.2. Comparison with RFC3325 .......................... 9
   4.3. Interactions with DTLS-SRTP ...................... 10
5. Conclusions ................................................. 11
6. Security Considerations ................................. 11
7. IANA Considerations ....................................... 12
8. Informative References ..................................... 12

Author’s Address .............................................. 13

Intellectual Property and Copyright Statements ........... 14
1. Introduction

The Session Initiation Protocol (SIP) [RFC3261] defined a simple mechanism for conveying the identity of the caller - a basic From field which was inserted by the calling UA. This mechanism has all of the security pitfalls of the From header field in email.

A short term solution was standardized in [RFC3325], which provides a network asserted identity. This mechanism is better than a basic From field, but works only in closed user groups and communities that have mutual trust. A longer term solution was developed, generally called SIP Identity, and published in [RFC4474]. SIP Identity makes use of domain-based signatures to provide security. An originating proxy authenticates the user, verifies that it matches their identity in the From header field, and if so, includes an Identity header field. This header field contains a cryptographic signature over the From header field along with other parts of the message, and is then signed by the originating domain. A recipient that receives the request can verify the signature, and compare the domain name in the signers certificate with the domain name in the From header field. If they match, the recipient knows that the originating domain has vouched for the identity of the caller.

The SIP Identity mechanism improves upon RFC 3325 by allowing it to work across several transit networks. There is no requirement of mutual trust along every hop; the terminating domain or user directly verifies the assertion made by the calling domain.

SIP Identity has been identified as a core SIP protocol [I-D.ietf-sip-hitchhikers-guide]. It has been used as the basis for connected identity, which delivers the identity of the called party [RFC4916]. More recently, SIP Identity is used to provide the integrity services needed to secure DTLS-SRTP [I-D.ietf-avt-dtls-srtp]. In essence, it is becoming a core part of the SIP security story.

However, recently concerns have been raised about the applicability of SIP Identity, especially in the presence of phone numbers. This document explores those concerns.

2. The Problem with Numbers

The mechanism in RFC 4474 is intimately intertwined with the domain part of the From URI. On the receiving side, the identity is considered verified when two conditions exist:
o The signature placed in the Identity header field matches the signature computed by the verifier, and

o the domain owner, as identified in the signing certificate, matches the domain part of the From header field

Both of these properties are needed to provide the security that SIP Identity provides. Unfortunately, in reality, most SIP deployments at the time of writing make use of phone numbers, and not traditional email-style user@domain identifiers. Of course, a phone number does not have a domain part. How, then, can RFC 4474 be used? The specification itself has this to say on the subject:

11. Identity and the TEL URI Scheme

Since many SIP applications provide a Voice over IP (VoIP) service, telephone numbers are commonly used as identities in SIP deployments. In the majority of cases, this is not problematic for the identity mechanism described in this document. Telephone numbers commonly appear in the username portion of a SIP URI (e.g., ‘sip:+17005551008@chicago.example.com;user=phone’). That username conforms to the syntax of the TEL URI scheme (RFC 3966 [13]). For this sort of SIP address-of-record, chicago.example.com is the appropriate signatory.

It is also possible for a TEL URI to appear in the SIP To or From header field outside the context of a SIP or SIPS URI (e.g., ‘tel:+17005551008’). In this case, it is much less clear which signatory is appropriate for the identity. Fortunately for the identity mechanism, this form of the TEL URI is more common for the To header field and Request-URI in SIP than in the From header field, since the UAC has no option but to provide a TEL URI alone when the remote domain to which a request is sent is unknown. The local domain, however, is usually known by the UAC, and accordingly it can form a proper From header field containing a SIP URI with a username in TEL URI form. Implementations that intend to send their requests through an authentication service SHOULD put telephone numbers in the From header field into SIP or SIPS URIs whenever possible.

This text makes it sound like there really isn’t much of a problem; that the originator merely needs to use the SIP version of his SIP URI, and then there will be no problem. However, the usage of phone numbers in this way is problematic. The presence of the domain part in the URI is artificial; as an identifier for a user, the domain part is not relevant.
When phone numbers are used, users only know the association between the phone number itself and a user. That is, Bob knows that "+1 (973) 865-4321" corresponds to Alice. The domain part is not relevant to him. Bob doesn’t use the domain part to reach Alice; he just dials the number. Bob might look up Alice in his phone book, but he’ll only see the number there. Bob will use that number when dialing Alice from his cell phone or non-VoIP equipment, and not even have the opportunity to provide a domain name. Alice, in turn, gives out only her number, without the trailing domain part. Almost all user equipment today that provides a form of caller ID, will NOT render the domain part of the URI if the user part is a phone number. Indeed, its very likely that Bob doesn’t even know that Alice gets service from a.com. Most users don’t know the providers of the other people they interact with. Local number portability has made that association even more ephemeral. Users move between providers relatively easily now, keeping the same number, even though the domain part effectively changes.

Consequently, the domain part of the SIP URI, when used in conjunction with phone numbers, is not relevant to users in establishing the identity associated with that number. For email-style identifiers, this is not true – the domain part is highly relevant.

Unfortunately, this problem is a FUNDAMENTAL PROPERTY OF PHONE NUMBERS. No specifications or efforts on the part of IETF can fix this problem. Phone numbers are fundamentally NOT scoped to a domain, and attempts to represent them in any other form are ultimately futile from an identification perspective.

3. Attacks Introduced by Usage of Numbers

The fact that the domain part is irrelevant for SIP URI containing phone numbers introduces two attacks into RFC 4474.

3.1. The Re-Sign Attack

Consider the network topology of Figure 2.

```
+---------+   +---------+   +---------+   +---------+
|         |   |         |   |         |  +---------+
|       |   |         |   |         |   |         |  |       |
+-------+   | a.com    |---| b.com    |---| c.com    |---| Bob     |
| Alice |   +---------+   |         |   |         |   |         |  |       |
+-------+                 +---------+   |         |   |         |  +---------+
```
Alice, who gets VoIP service from a.com, has a phone number of +1 (973) 865-4321, and sends an INVITE to Bob. Alice has no user@domain form of AoR; a.com is providing strictly voice services. Per the instructions in RFC 4474, Alice populates the From header field of her INVITE as sip:+19738654321@a.com;user=phone. The a.com proxy as as the authentication service as defined in RFC 4474, and adds an Identity header field. It uses a certificate, signed by a root CA, asserting that it is a.com. This INVITE passes to b.com. The proxy at b.com, for purposes of malice or otherwise, does the following:

- Removes the Identity and Identity-Info header fields
- Modifies the From header field to sip:+19738654321@b.com;user=phone. In other words, it replaces the domain part with its own domain name.
- Adds a new Identity and Identity-Info header field, containing its own signature, performed using a certificate it holds for the b.com domain.
- Forwards the request to the target in c.com.

This request arrives at c.com, and the request is passed to Bob. Bob’s agent runs the verification process described in RFC 4474. The signature on the request is valid, and the domain in the From header field matches the domain in the certificate of the signer. Bob’s UA shows the identity of the caller as "+1 (973) 865-4321" and indicates that the identity has been verified. This number does in fact match the number of the caller, Alice. And so, as far as Bob is concerned, everything seems fine. However, b.com has been able to insert itself, and do whatever it wants.

What happened here? It is illustrative to look at this attack in the case where Alice had used a user@domain identifier, for example, alice@a.com. In this case, if the b.com transit domain had done the same processing described above, the request would have arrived to Bob’s user agent. The signature would be valid, and the domain of the signer would match the domain in the From field. However, the identity shown to Bob would be "alice@b.com", which doesn’t match Alice’s AOR as far as Bob knows. Consequently, Bob will be alerted to the fact that something is going on.

It’s important to note that, even with email-style identifiers, the
re-sign attack might not be noticed by Bob. If Bob doesn’t know Alice; he has no way to know that the caller isn’t actually alice@b.com. Consequently, the attack would succeed in that case as well. Indeed, this weakness is outlined in RFC 4474 itself. From Section 13.1:

In the end analysis, the Identity and Identity-Info headers cannot protect themselves. Any attacker could remove these headers from a SIP request, and modify the request arbitrarily afterwards. However, this mechanism is not intended to protect requests from men-in-the-middle who interfere with SIP messages; it is intended only to provide a way that SIP users can prove definitively that they are who they claim to be.

However, these man-in-the-middle attacks (when an email-style From URI is being used) will be detectable for cases where the called party knows the caller. Even when the called party doesn’t know the caller, attempts to return their calls would fail, and future discussions and communications with the caller would probably quickly reveal that the wrong identifier had been delivered.

Our conclusion is that, when email-style identifiers are used, RFC 4474 provides a reasonable degree of protection against re-signing and, in fact, provides a reasonable degree of message integrity over the parts of the message covered by the Identity signature. However, when phone numbers are used, RFC 4474 provides no protection against re-signing, and its message integrity is easily subject to MITM attacks.

3.2. The False Number Attack

Consider once more the topology of Figure 2. However, in this discussion, Alice is a malicious user, and their provider a.com is also malicious. Alice wishes to make a call to Bob, but wishes to lie about her phone number in an attempt to mislead Bob as to the identity of the caller. Consequently, even though Alice’s actual phone number is +1 (973) 865-4321, she would like to initiate a call and claim to have the number +1 (202) 456-1414, which is the United States White House switchboard number.

Alice sends her INVITE with a From header field of sip:+12024561414@a.com;user=phone. Her proxy at a.com, which is an accomplice in this fabrication, signs the request anyway and includes a pointer to its perfectly valid certificate into the Identity-Info header field. The b.com domain is a large provider and doesn’t have the resources to check up on the behavior of its transit partners.
So, it passes the INVITE on to Bob’s domain, c.com.

This call is then passed to Bob. Since this is a perfectly valid From header field value, and the Identity signatures and certificates are all valid, Bob accepts the request. His user agent, noticing that this is a phone number, renders just the phone number, which Bob recognizes as the White House switchboard number. He then answers the call.

The reason this attack is possible is that, for phone numbers, only the user part, and NOT the domain part, are relevant for identification. Furthermore, there is nothing within the scope of RFC 4474 which allows a recipient of a request to determine that a particular domain is, in fact, a legitimate owner of a particular phone number. Indeed, the very notion of ownership is a complex one. Thus, it is possible for a domain to claim a particular phone number in its user part, even if that phone number is not in fact ‘owned’ by that domain. The only protection offered against this attack is the trustworthiness of the domain itself. A domain cannot lie about who they are, but they can lie about what numbers they own. Thus, if a user trusts that a particular domain won’t lie, they can determine that a call is from that domain, and therefore trust the phone number only due to their belief in the truthfulness of that domain.

4. Consequences

There are several important consequences of these attacks.

4.1. Unsecure Caller ID for Phone Numbers

The false number attack described in Section 3.2 means that RFC 4474 does not readily provide ‘secure caller ID’ for phone numbers. The definition of secure in this context is that the phone number present in the From header field URI does in fact represent the number of the party that originated the call.

The term ‘readily’ above is important. As noted in Section 3.2, the asserted calling number can be considered valid if the signing domain is trusted by the called UA. As a general rule, a UA will not have an easy way to ascertain the trustworthiness of any particular domain. A UA might, perhaps, be configured with the domains of providers it does trust. For example, a UA might trust the large service providers in its country (generally a small and enumerable set), and so it could have ‘secure’ caller ID only for calls from those providers.

If a UA is only willing to trust its own provider, this would likely
result in a model whereby each provider determines the trustworthiness of the previous provider, and for those it trusts, it modifies the From header field URI and resigns the request. Such an approach would introduce a transitive trust model to possibly alleviate this problem.

4.2. Comparison with RFC3325

A critical question to be answered, then, is whether RFC 4474 provides any additional security properties above RFC 3325.

Firstly, it is clearly the case that for email-style identifiers, RFC 4474 is far superior to RFC 3325. RFC 3325 allows the false number attack even for those identifiers. A malicious originating domain could assert a caller identity of sip:george@whitehouse.gov, and this would be the identity rendered to a called party. RFC 3325 provides an overall level of secure caller ID equal to the trustworthiness of the LEAST trustworthy domain in the network. As the size of a network grows, this level of trustworthiness approaches zero. This is not true in RFC 4474; an attacker could not assert identity within whitehouse.gov.

Considering phone numbers, two cases must be considered. In one model, each domain in a chain of domains resigns the request and changes the domain name to point to itself. This is possible because the re-sign attack. In this case, it is not because a domain is being malicious; but rather, because it will only choose to re-sign if it trusts its upstream domain. This allows a UA to be configured with a single domain to trust - its own provider - and obtain transitive trust overall. This model is, for all intents and purposes, identical to the trust model outlined for RFC 3325. Consequently, RFC 4474 is no better than RFC 3325 here.

However, in the second model, intermediate domains do not resign requests. Furthermore, UA’s utilize white lists and black lists of domains that are known to be trustworthy (or not). Today, such lists do exist and are provided for email spam. One can imagine a UA contacting such a service periodically, or upon an incoming call, to verify the signing domain against the list.

In this model, RFC 4474 is still superior to RFC3325. In RFC3325, the trustworthiness of the caller ID is as trustable as the least trustworthy domain. As noted above, this approaches zero - for all calls - once the network reaches a reasonable size. However, in RFC 4474, the trustworthiness of the caller ID is as trustable as the domain from which the call came. Of course, calls may still arrive from untrusted domains, but in the case of RFC 4474, a UA will know that. As such, it is possible for a UA to separate out caller ID
from domains it does trust from those it doesn’t, and if it can obtain sufficient coverage from its whitelists so that many incoming caller IDs are trusted, the system works better. That, however, is the key. If the whitelists accessible by a UA cover only 3% of the total number of allocated numbers, most incoming calls will not have a trustable caller ID, and the system will be just as bad as RFC 3325.

If history is any guide, it is my belief that domains are likely to follow the first model, and not the second.

Consequently, the conclusion is that RFC 4474 may provide somewhat more trustable caller ID for phone numbers than RFC 3325, but in practice they are likely to be identical. However, for email-style identifiers, RFC 4474 is superior.

4.3. Interactions with DTLS-SRTP

When used with email identifiers, RFC 4474 provides a limited form of message integrity protection against MITM attacks. As discussed above, this is because modified domains in From fields are likely to be readily detected by end users, either right away or in the future. However, with phone numbers, RFC 4474 provides no message integrity against MITM attacks.

This weakness interacts with DTLS-SRTP [I-D.ietf-avt-dtls-srtp]. In particular, [I-D.ietf-sip-dtls-srtp-framework] describes how the DTLS handshakes are correlated with the signaling exchange by means of the message integrity mechanisms provided by RFC 4474. However, when the calling party has a phone number, that message integrity is lost.

A consequence of this is that any intermediate signaling entity could modify the DTLS fingerprints, insert itself as a media intermediary, and then decrypt and re-encrypt media on each side. Such an attack would not be detectable if the caller has a phone number. Indeed, it won’t be detectable even if they have an email-style identifier, if the called party doesn’t recognize that the caller’s identity doesn’t match what their UA is showing to them.

This, in turn, raises an important question - does, in fact, DTLS-SRTP provide "more secure" media than Sdescriptions [RFC4568]? One of the key weaknesses of Sdescriptions is that, assuming each hop is TLS, the keying material is still exposed to each and every intermediate proxy. This means that any intermediary could intercept and process the media. When used with phone numbers, DTLS-SRTP has this same weakness.

That said, DTLS-SRTP does provide some advantages even when used with
phone numbers:

- To protect against eavesdropping attacks from off-path attackers, Sdescriptions requires that each and every hop has properly encrypted the link with TLS (this is ignoring the possibility of end-to-end SMIME encryption). Thus, simple misconfiguration of an intermediary can expose Sdescriptions to attacks by offpath attackers. DTLS-SRTP does not rely on confidentiality services from intermediaries; indeed it relies on nothing special from them, except that they aren’t being malicious. Thus, DTLS-SRTP provides better protection against offpath attacks.

- With Sdescriptions, because the keying material is in the signaling, intermediaries will see it directly even when each signaling hop is encrypted. Consequently, it is possible that this material could ‘leak’ out unsecured channels, such as logs or application interfaces. This would allow other entities access to the keying material and thus allow them to manipulate the media stream. With DTLS-SRTP, this is possible only by active attack by such applications. Thus, DTLS-SRTP provides mildly better protection here.

Thus, DTLS-SRTP still provides better security than Sdescriptions. However, when used with phone numbers, it is by no means ideal. Most importantly, it does NOT provide guarantees that intermediaries have not been able to intercept and decrypt the media.

5. Conclusions

Unfortunately, there is no simple remedy to this problem. The problems with RFC4474 cannot just be fixed by an alternate security technique. They are deeply rooted in the domain independence of phone numbers.

Consequently, our recommendation at this point is to more clearly document the limitations of RFC4474 when used with phone numbers. We recommend that RFC4474 be revised, and that the updated document include a more detailed discussion of the weaknesses it has when used with phone numbers. Similarly, we recommend that [I-D.ietf-sip-dtls-srtp-framework] be revised to include a more thorough description of the limitations of DTLS-SRTP when used with phone numbers.

6. Security Considerations

This document is concerned entirely with security.
7. IANA Considerations

None.

8. Informative References


Author’s Address

Jonathan Rosenberg
Cisco
Edison, NJ
US

Email: jdrosen@cisco.com
URI: http://www.jdrosen.net