Network Working Group
Internet-Draft
Intended status: Informational
Expires: January 16, 2014

J. Peterson
NeuStar, Inc.
H. Schulzrinne
Columbia University
H. Tschofenig
Nokia Siemens Networks

July 15, 2013

Secure Origin Identification: Problem Statement, Threat Model, Requirements, and Roadmap
draft-peterson-secure-origin-ps-01.txt

Abstract

Over the past decade, SIP has become a major signaling protocol for voice communications, one which has replaced many traditional telephony deployments. However, interworking SIP with the traditional telephone network has ultimately reduced the security of Caller ID systems. Given the widespread interworking of SIP with the telephone network, the lack of effective standards for identifying the calling party in a SIP session has granted attackers new powers as they impersonate or obscure calling party numbers when orchestrating bulk commercial calling schemes, hacking voicemail boxes or even circumventing multi-factor authentication systems trusted by banks. This document therefore examines the reasons why providing identity for telephone numbers on the Internet has proven so difficult, and shows how changes in the last decade may provide us with new strategies for attaching a secure identity to SIP sessions.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

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."

This Internet-Draft will expire on January 16, 2014.

Table of Contents

1. Introduction

In many communication architectures that allow users to communicate with other users the need for identifying the originating party that initiates a call or a messaging interaction arises. The desire for identifying the communication parties in the end-to-end communication attempt arises from the need to implement authorization policies (to grant or reject call attempts) but has also been utilized for charging. While there are a number of ways to enable identification this functionality has been provided by the Session Initiation Protocol (SIP) [2] by using two main types of approaches, namely using P-Asserted-Identity (PAI) [4] and SIP Identity [1], which are described in more detail in Section 4. The goal of these mechanisms is to validate that originator of a call is authorized to use the From identifier. Protocols, like XMPP, use mechanisms that are conceptional similar to those offered by SIP.

Although solutions have been standardized it turns out that the current deployment situation is unsatisfactory and, even worse, there is little indication that it will be improve in the future. In [8] we illustrate what challenges arise. In particular, the interworking with different communication architectures (e.g., SIP, PSTN, XMPP, RTCWeb) breaks the end-to-end semantic of the communication interaction and destroys the identification capabilities. Furthermore, the use of different identifiers (e.g., E.164 numbers vs. SIP URIs) creates challenges for determining who is able to claim "ownership" for a specific identifier.

After the publication of the PAI and SIP Identity specifications

various further attempts have been made to tackle the topic but unfortunately with little success. The complexity resides in the deployment situation and the long list of (often conflicting) requirements. A number of years have passed since the last attempts were made to improve the situation and we therefore believe it is time to give it another try. With this document we would like to start an attempt to develop a common understanding of the problem statement as well as requirements to develop a vision on how to advance the state of the art and to initiate technical work to enable secure call origin identification.

2. Problem Statement

In the classical public-switched telephone network, a limited number of carriers trusted each other, without any cryptographic validation, to provide accurate caller origination information. In some cases, national telecommunication regulation codified these obligations. This model worked as long as the number of entities was relatively small, easily identified (e.g., through the concept of certificated carriers) and subject to effective legal sanctions in case of misbehavior. However, for some time, these assumptions have no longer held true. For example, entities that are not traditional telecommunication carriers, possibly located outside the country whose country code they are using, can act as voice service providers. While in the past, there was a clear distinction between customers and service providers, VoIP service providers can now easily act as customers, originating and transit providers. For telephony, Caller ID spoofing has become common, with a small subset of entities either ignoring abuse of their services or willingly serving to enable fraud and other illegal behavior. For example, recently, enterprises and public safety organizations [14] have been subjected to telephony denial-of-service attacks. In this case, an individual claiming to represent a collections company for payday loans starts the extortion scheme with a phone call to an organization. Failing to get payment from an individual or organization, the criminal organization launches a barrage of phone calls, with spoofed numbers, preventing the targeted organization from receiving legitimate phone calls. Other boiler-room organizations use number spoofing to place illegal "robocalls" (automated telemarketing, see, for example, the FCC webpage [15] on

this topic). Robocalls is a problem that has been recognized already by various regulators, for example the Federal Communications Commission (FCC) recently organized a robocall competition to solicit ideas for creating solutions that will block illegal robocalls [16]. Criminals may also use number spoofing to impersonate banks or bank customers to gain access to information or financial accounts.

In general, number spoofing is used in two ways, impersonation and anonymization. For impersonation, the attacker pretends to be a specific individual. Impersonation can be used for pretexting, where the attacker obtains information about the individual impersonated, activates credit cards or for harassment, e.g., by causing utility services to be disconnected, take-out food to be delivered, or by causing police to respond to a non-existing hostage situation ("swatting", see [18]). Some voicemail systems can be set up so that they grant access to stored messages without a password, relying solely on the caller identity. As an example, the News International phone-hacking scandal [17] has also gained a lot of press attention where employees of the newspaper were accused of engaging in phone

hacking by utilizing Caller ID spoofing to get access to a voicemail. For numbers where the caller has suppressed textual caller identification, number spoofing can be used to retrieve this information, stored in the so-called Calling Name (CNAM) database. For anonymization, the caller does not necessarily care whether the number is in service, or who it is assigned to, and may switch rapidly and possibly randomly between numbers. Anonymization facilitates automated illegal telemarketing or telephony denial-of-service attacks, as described above, as it makes it difficult to blacklist numbers. It also makes tracing such calls much more labor-intensive, as each such call has to be identified in each transit carrier hop-by-hop, based on destination number and time of call.

Secure origin identification should prevent impersonation and, to a lesser extent, anonymization. However, if numbers are easy and cheap to obtain, and if the organizations assigning identifiers cannot or will not establish the true corporate or individual identity of the entity requesting such identifiers, robocallers will still be able to switch between many different identities.

It is insufficient to simply outlaw all spoofing of originating telephone numbers, because the entities spoofing numbers are already

committing other crimes and thus unlikely to be deterred by legal sanctions. Also, in some cases, third parties may need to temporarily use the identity of another individual or organization, with full consent of the "owner" of the identifier. For example:

The doctor's office: Physicians calling their patients using their cell phones would like to replace their mobile phone number with the number of their office to avoid being called back by patients on their personal phone.

Call centers: Call centers operate on behalf of companies and the called party expects to see the Caller ID of the company, not the call center.

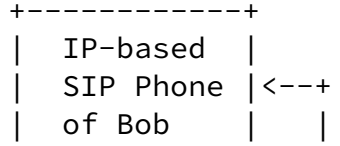
3. Use Cases

In order to explain the requirements and other design assumptions we will explain some of the scenarios that need to be supported by any solution. To reduce clutter, the figures do not show call routing elements, such as SIP proxies, of voice or text service providers. We generally assume that the PSTN component of any call path cannot be altered.

3.1. VoIP-to-VoIP Call

For the IP-to-IP communication case, a group of service providers that offer interconnected VoIP service exchange calls using SIP end-to-end, but may also deliver some calls via circuit-switched facilities, as described below. These service providers use telephone numbers as source and destination identifiers, either as the user component of a SIP URI (e.g., sip:12125551234@example.com) or as a tel URI [7].

As illustrated in Figure 1, if Alice calls Bob, the call will use SIP end-to-end. (The call may or may not traverse the Internet.)



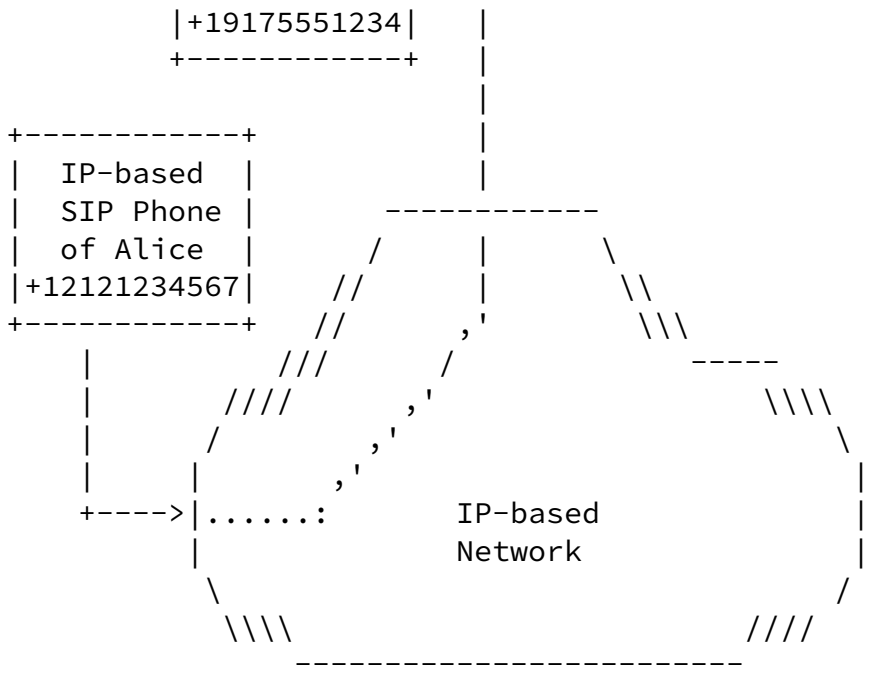
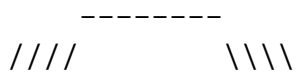
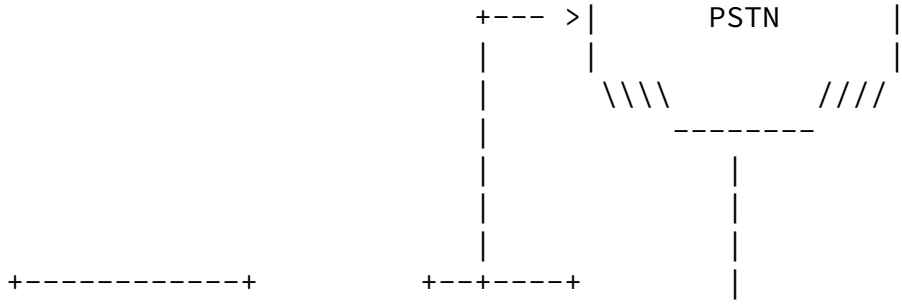


Figure 1: VoIP-to-VoIP Call.

3.2. IP-PSTN-IP Call

Frequently, two VoIP-based service providers are not directly connected by VoIP and use TDM circuits to exchange calls, leading to the IP-PSTN-IP use case. In this use case, Dan's VSP is not a member of the interconnect federation Alice's and Bob's VSP belongs to. As far as Alice is concerned Dan is not accessible via IP and the PSTN is used as an interconnection network. Figure 2 shows the resulting exchange.





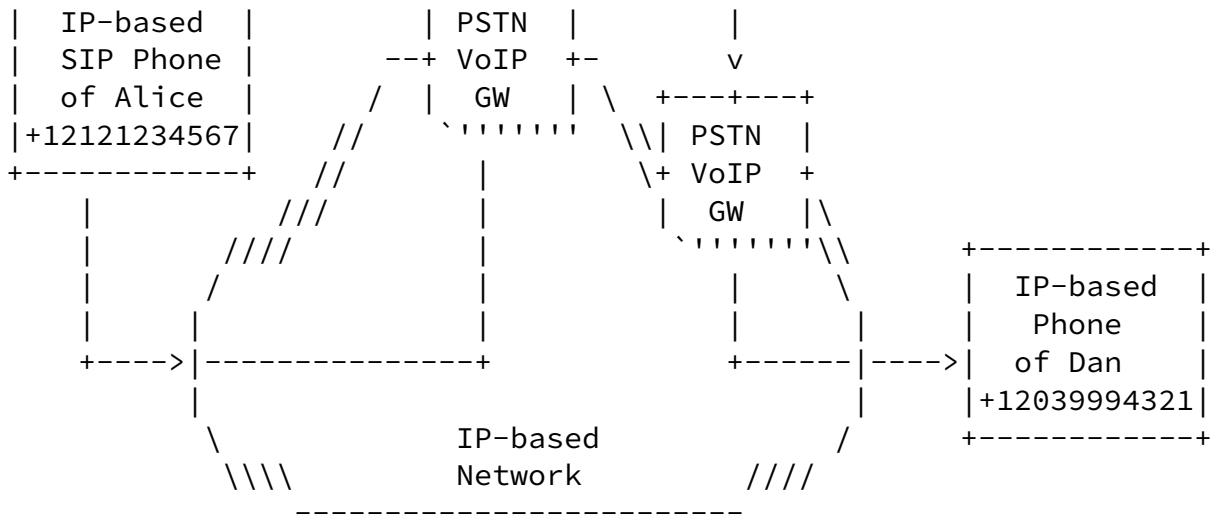
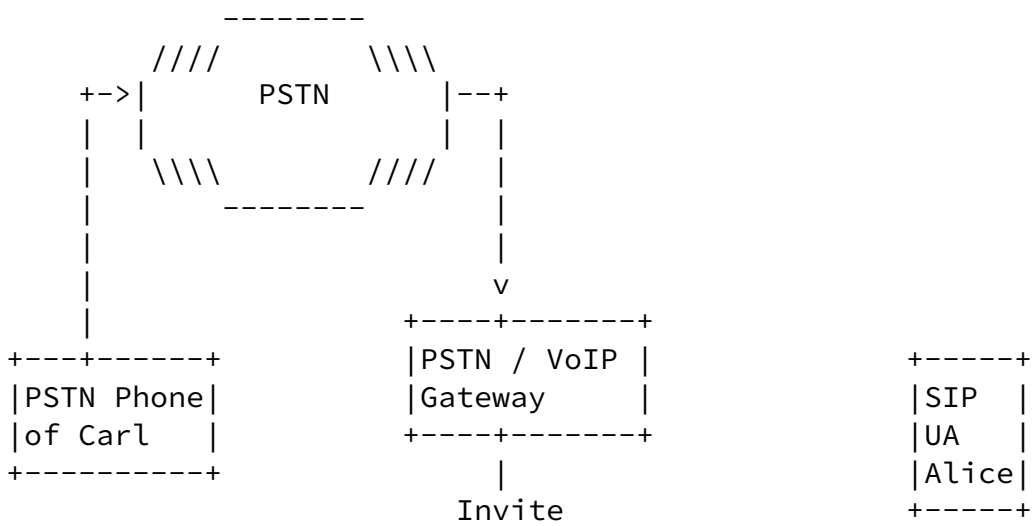


Figure 2: IP-PSTN-IP Call.

3.3. PSTN-to-VoIP Call

Consider Figure 3 where Carl is using a PSTN phone and initiates a call to Alice. Alice is using a VoIP-based phone. The call of Carl traverses the PSTN and enters the Internet via a PSTN/VoIP gateway. This gateway attaches some identity information to the call, for example based on the information it had received through the PSTN, if available.



| ^

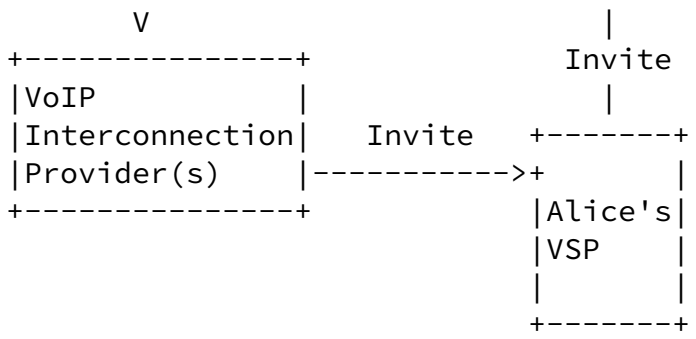
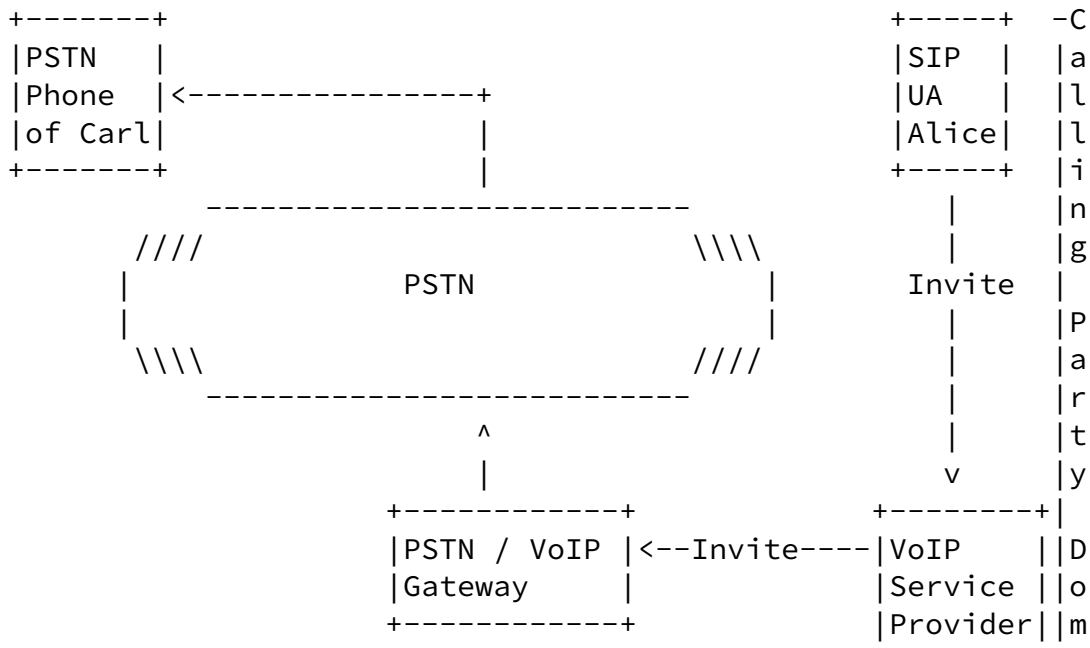


Figure 3: PSTN-to-VoIP Call.

Note: A B2BUA/Session Border Controller (SBC) exhibits behavior that looks similar to this scenario since the original call content would, in the worst case, be re-created on the call origination side.

3.4. VoIP-to-PSTN Call Call

Consider Figure 4 where Alice calls Carl. Carl uses a PSTN phone and Alice an IP-based phone. When Alice initiates the call the E.164 number needs to get translated to a SIP URI and subsequently to an IP address. The call of Alice traverses her VoIP provider where the call origin identification information is added. It then hits the PSTN/VoIP gateway. Ideally, Alice would like to know whether she, for example, talks to someone at her bank rather than to someone intercepting the call. If Alice wants to be assured that she's being connected to the right party, it is a slightly different aspect to what [4][1]. Problem statements and solutions are offered with [9] and [6].



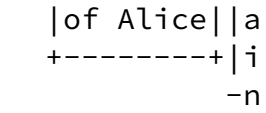


Figure 4: IP-to-PSTN Call.

3.5. PSTN-VoIP-PSTN Call

Consider Figure 5 where Carl calls Alice. Both users have PSTN phones but interconnection between the two PSTN networks is accomplished via an IP network. Consequenly, Carl's operator uses a PSTN-to-VoIP gateway to route the call via an IP network to a gateway to break out into the PSTN again.

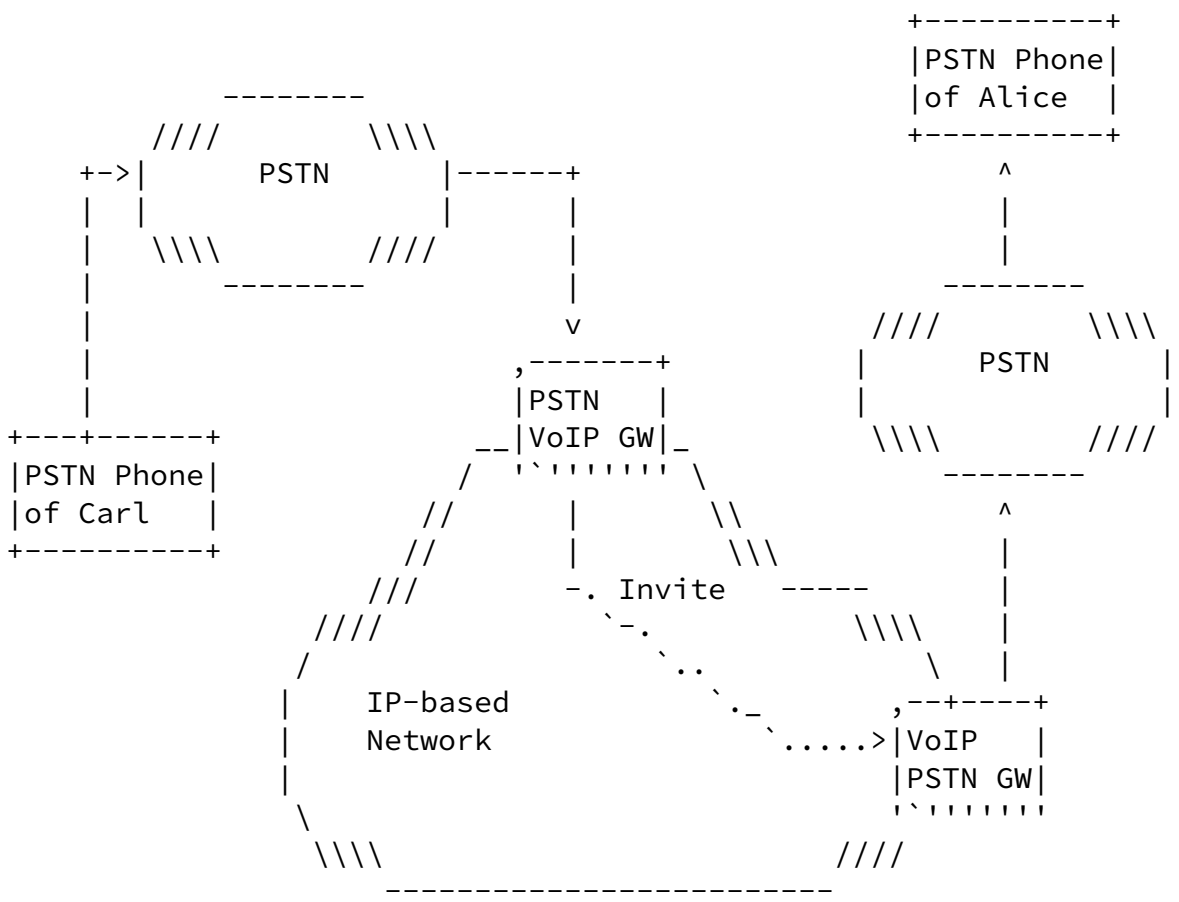


Figure 5: PSTN-VoIP-PSTN Call.

3.6. PSTN-to-PSTN Call

For the "legacy" case of a PSTN-to-PSTN call, otherwise beyond improvement, we may be able to use out-of-band IP connectivity at both the originating and terminating carrier to validate the call

information.

4. Limitations of Current Solutions

From the inception of SIP, the From header field value has held an arbitrary user-supplied identity, much like the From header field value of an SMTP email message. During work on [2], efforts began to provide a secure origin for SIP requests as an extension to SIP. The so-called "short term" solution, the P-Asserted-Identity header described in [4], is deployed fairly widely, even though it is limited to closed trusted networks where end-user devices cannot alter or inspect SIP messages and offers no cryptographic validation. As P-Asserted-Identity is used increasingly across multiple networks, it cannot offer any protection against identity spoofing by intermediaries or entities that allow end users to set the P -Asserted-Identity information.

Subsequent efforts to prevent calling origin identity spoofing in SIP include the SIP Identity effort (the "long term" identity solution) [1] and Verification Involving PSTN Reachability (VIPR) [12]. SIP Identity attaches a new header field to SIP requests containing a signature over the From header field value combined with other message components to prevent replay attacks. SIP Identity is meant both to prevent originating calls with spoofed From headers and intermediaries, such as SIP proxies, from launching man-in-the-middle attacks to alter calls passing through. The VIPR architecture attacked a broader range of problems relating to spam, routing and identity with a new infrastructure for managing rendezvous and security, which operated alongside of SIP deployments.

As we will describe in more detail below, both SIP Identity and VIPR suffer from serious limitations that have prevented their deployment at significant scale, but they may still offer ideas and protocol building blocks for a solution.

4.1. SIP Identity

The SIP Identity mechanism [1] provided two header fields for securing identity information in SIP requests: the Identity and Identity-Info header fields. Architecturally, the SIP Identity mechanism assumes a classic "SIP trapezoid" deployment in which an

authentication service, acting on behalf of the originator of a SIP request, attaches identity information to the request which provides partial integrity protection; a verification service acting on behalf of the recipient validates the integrity of the request when it is received.

The Identity header field value contains a signature over a hash of selected elements of a SIP request, including several header field values (most significantly, the From header field value) and the

entirety of the body of the request. The set of header field values was chosen specifically to prevent cut-and-paste attacks; it requires the verification service to retain some state to guard against replays. The signature over the body of a request has different properties for different SIP methods, but all prevent tampering by man-in-the-middle attacks. For a SIP MESSAGE request, for example, the signature over the body covers the actual message conveyed by the request: it is pointless to guarantee the source of a request if a man-in-the-middle can change the content of the message, as in that case the message content is created by an attacker. Similar threats exist against the SIP NOTIFY method. For a SIP INVITE request, a signature over the SDP body is intended to prevent a man-in-the-middle from changing properties of the media stream, including the IP address and port to which media should be sent, as this provides a means for the man-in-the-middle to direct session media to resource that the originator did not specify, and thus to impersonate an intended listener.

The Identity-Info header field value contains a URI designating the location of the certificate corresponding to the private key that signed the hash in the Identity header. That certificate could be passed by-value along with the SIP request, in which case a "cid" URI appears in Identity-Info, or by-reference, for example when the Identity-Info header field value has the URL of a service that delivers the certificate. [1] imposes further constraints governing the subject of that certificate: namely, that it must cover the domain name indicated in the domain component of the URI in the From header field value of the request.

The SIP Identity mechanism, however, has two fundamental limitations that have precluded its deployment: first, that it provides Identity only for domain names rather than other identifiers; second, that it

does not tolerate intermediaries that alter the bodies, or certain header fields, of SIP requests.

As deployed, SIP predominantly mimics the structures of the telephone network, and thus uses telephone numbers as identifiers. Telephone numbers in the From header field value of a SIP request may appear as the user part of a SIP URI, or alternatively in an independent tel URI. The certificate designated by the Identity-Info header field as specified, however, corresponds only to the domain portion of a SIP URI in the From header field. As such, [1] does not have any provision to identify the assignee of a telephone number. While it could be the case that the domain name portion of a SIP URI signifies a carrier (like "att.com") to whom numbers are assigned, the SIP Identity mechanism provides no assurance that a number is assigned to any carrier. For a tel URI, moreover, it is unclear in [1] what entity should hold a corresponding certificate. A caller may not

want to reveal the identity of its service provider to the callee, and may thus prefer tel URIs in the From header field.

This lack of authority gives rise to a whole class of SIP identity problems when dealing with telephone numbers, as is explored in [10]. That document shows how the Identity header of a SIP request targeting a telephone number (embedded in a SIP URI) could be dropped by an intermediate domain, which then modifies and resigns the request, all without alerting the verification service: the verification service has no way of knowing which original domain signed the request. Provided that the local authentication service is complicit, an originator can claim virtually any telephone number, impersonating any chosen Caller ID from the perspective of the verifier. Both of these attacks are rooted in the inability of the verification service to ascertain a specific certificate that is authoritative for a telephone number.

As deployed, SIP is moreover highly mediated, and mediated in ways that [2] did not anticipate. As request routing commonly depends on policies dissimilar to [13], requests transit multiple intermediate domains to reach a destination; some forms of intermediaries in those domains may effectively re-initiate the session.

One of the main reasons that SIP deployments mimic the PSTN architecture is because the requirement for interconnection with the

PSTN remains paramount: a call may originate in SIP and terminate on the PSTN, or vice versa; and worse still, a PSTN-to-PSTN call may transit a SIP network in the middle, or vice versa. This necessarily reduces SIP's feature set to the least common dominator of the telephone network, and mandates support for telephone numbers as a primary calling identifier.

Interworking with non-SIP networks makes end-to-end identity problematic. When a PSTN gateway sends a call to a SIP network, it creates the INVITE request anew, regardless of whether a previous leg of the call originated in a SIP network that later dropped the call to the PSTN. As these gateways are not necessarily operated by entities that have any relationship to the number assignee, it is unclear how they could provide an identity signature that a verifier should trust. Moreover, how could the gateway know that the calling party number it receives from the PSTN is actually authentic? And when a gateway receives a call via SIP and terminates a call to the PSTN, how can that gateway verify that a telephone number in the From header field value is authentic, before it presents that number as the calling party number in the PSTN?

Similarly, some SIP networks deploy intermediaries that act as back-to-back user agents (B2BUAs), typically in order to enforce policy at

network boundaries (hence the nickname "Session Border Controller"). As a common practice, these entities modify SIP INVITE requests in transit in such a way that they no longer satisfy the transaction-mapping semantics of [2], commonly changing the From, Contact and Call-ID header field values, as well as aspects of the SDP, including especially the IP addresses and ports associated with media. The policies that motivate these changes may be associated with topology hiding, or may alter messages to interoperate successfully with particular SIP implementations, or may simply involve network address translation from private address space. But effectively, a SIP request exiting a B2BUA has no necessary relationship to the original request received by the B2BUA, much like a request exiting a PSTN gateway has no necessary relationship to any SIP request in a pre-PSTN leg of the call. An Identity signature provided for the original INVITE has no bearing on the post-B2BUA INVITE, and, were the B2BUA to preserve the original Identity header, any verification service would detect a violation of the integrity protection.

The SIP community has long been aware of these problems with [1] in practical deployments. Some have therefore proposed weakening the security constraints of [1] so that at least some deployments of B2BUAs will not violate (or remove) the integrity protection of SIP requests. However, such solutions do not address one key problem identified above: the lack of any clear authority for telephone numbers, and the fact that some INVITE requests are generated by intermediaries rather than endpoints. Removing the signature over the SDP from the Identity header will not, for example, make it any clearer how a PSTN gateway should assert identity in an INVITE request.

4.2. VIPR

Verification Involving PSTN Reachability (VIPR) directly attacks the twin problems of identifying number assignees on the Internet and coping with intermediaries that may modify signaling. To address the first problem, VIPR relies on the PSTN itself: it discovers which endpoints on the Internet are reachable via a particular PSTN number by calling the number on the PSTN to determine whom a call to that number will reach. As VIPR-enabled Internet endpoints associated with PSTN numbers are discovered, VIPR provides a rendez-vous service that allows the endpoints of a call to form an out-of-band connection over the Internet; this connection allows the endpoints to exchange information that secures future communications and permits direct, unmediated SIP connections.

VIPR provides these services within a fairly narrow scope of applicability. Its seminal use case is the enterprise IP PBX, a device that has both PSTN connectivity and Internet connectivity,

which serves a set of local users with telephone numbers; after a PSTN call has connected successfully and then ended, the PBX searches a distributed hash-table to see if any VIPR-compatible devices have advertised themselves as a route for the unfamiliar number on the Internet. If advertisements exist, the originating PBX then initiates a verification process to determine whether the entity claiming to be the assignee of the unfamiliar number in fact received the successful call: this involves verifying details such as the start and stop times of the call. If the destination verifies successfully, the originating PBX provisions a local database with a route for that telephone number to the URI provided by the proven

destination. The destination moreover gives a token to the originator that can be inserted in future call setup messages to authenticate the source of future communications.

Through this mechanism, the VIPR system provides a suite of properties, ones that go well beyond merely securing the origins of communications. It also provides a routing system which dynamically discovers mappings between telephone numbers and URIs, effectively building an ad hoc ENUM database in every VIPR implementation. The tokens exchanged over the out-of-band connection established by VIPR moreover provide an authorization mechanism for accepting calls over the Internet that significantly reduces the potential for spam. Because the token can act as a nonce due to the presence of this out-of-band connectivity, the VIPR token is less susceptible to cut-and-paste attacks and thus needs to cover with its signature far less of a SIP request.

Due to its narrow scope of applicability, and the details of its implementation, VIPR has some significant limitations. The most salient for the purposes of this document is that it only has bearing on repeated communications between entities: it has no bearing on the classic "robocall" problem, where the target receives a call from a number that has never called before. All of VIPRs strengths in establishing identity and spam prevention kick in only after an initial PSTN call has been completed, and subsequent attempts at communication begin. Every VIPR-compliant entity moreover maintains its own stateful database of previous contacts and authorizations, which lends itself to more aggregators like IP PBXs that may front for thousands of users than to individual phones. That database must be refreshed by periodic PSTN calls to determine that control over the number has not shifted to some other entity; figuring out when data has grown stale is one the challenges of the architecture. As VIPR requires compliant implementations to operate both a PSTN interface and an IP interface, it has little apparent applicability to ordinary desktop PCs or similar devices with no ability to place direct PSTN calls.

The distributed hash table also creates a new attack surface for impersonation. Attackers who want to pose as the owners of telephone numbers can advertise themselves as routes to a number in the hash table. VIPR has no inherent restriction on the number of entities

that may advertise themselves as routes for a number, and thus an originator may find multiple advertisements for a number on the DHT even when an attack is not in progress. As for attackers, even if they cannot successfully verify themselves to the originators of calls (because they lack the call detail information), they may learn from those verification attempts which VIPR entities recently placed calls to the target number: it may be that this information is all the attacker hopes to glean. The fact that advertisements and verifications are public results from the public nature of the DHT that VIPR creates. The public DHT prevents any centralized control, or attempts to impede communications, but those come at the cost of apparently unavoidable privacy losses.

Because of these limitations, VIPR, much like SIP Identity, has had little impact in the marketplace. Ultimately, VIPR's utility as an identity mechanism is limited by its reliance on the PSTN, especially its need for an initial PSTN call to complete before any of VIPR's benefits can be realized, and by the drawbacks of the highly-public exchanges requires to create the out-of-band connection between VIPR entities. As such, there is no obvious solution to providing secure origin services for SIP on the Internet today.

5. Environmental Changes

5.1. Shift to Mobile Communication

In the years since [1] was conceived, there have been a number of fundamental shifts in the communications marketplace. The most transformative has been the precipitous rise of mobile smart phones, which are now arguably the dominant communications device in the developed world. Smart phones have both a PSTN and an IP interface, as well as an SMS and MMS capabilities. This suite of tools suggests that some of the techniques proposed by VIPR could be adapted to the smart phone environment. The installed base of smart phones is moreover highly upgradable, and permits rapid adoption out-of-band rendezvous services for smart phones that circumvent the PSTN: for example, the Apple iMessage service, which allows iPhone users to send SMS messages to one another over the Internet rather than over the PSTN. Like VIPR, iMessage creates an out-of-band connection over the Internet between iPhones; unlike VIPR, the rendezvous service is provided by a trusted centralized database of iPhones rather than by a DHT. While Apple's service is specific to customers of its smart phones, it seems clear that similar databases could be provided by neutral third parties in a position to coordinate between endpoints.

5.2. Failure of Public ENUM

At the time [1] was written, the hopes for establishing a certificate authority for telephone numbers on the Internet largely rested on public ENUM deployment. The e164.arpa DNS tree established for ENUM could have grown to include certificates for telephone numbers or at least for number ranges. It is now clear however that public ENUM as originally envisioned has little prospect for adoption. That said, national authorities for telephone numbers are increasingly migrating their provisioning services to the Internet, and issuing credentials that express authority for telephone numbers to secure those services. This new class of certificate authority for numbers could be opened to the public Internet to provide the necessary signatory authority for securing calling partys' numbers. While these systems are far from universal, the authors of this draft believe a certificate authority can be constructed for the North American Numbering Plan in a way that numbering authorities for other country codes could follow.

5.3. Public Key Infrastructure Developments

Also, there have been a number of recent high-profile compromises of web certificate authorities. The presence of numerous (in some cases, of hundreds) of trusted certificate authorities in modern web browsers has become a significant security liability. As [1] relied on web certificate authorities, this too provides new lessons for any work on revising [1]: namely, that innovations like DANE [5] that designate a specific certificate preferred by the owner of a DNS name could greatly improve the security of a SIP identity mechanism; and moreover, that when architecting new certificate authorities for telephone numbers, we should be wary of excessive pluralism. While a chain of delegation with a progressively narrowing scope of authority (e.g., from a regulatory entity to a carrier to a reseller to an end user) is needed to reflect operational practices, there is no need to have multiple roots, or peer entities that both claim authority for the same telephone number or number range.

5.4. Pervasive Nature of B2BUA Deployments

Given the prevalence of established B2BUA deployments, we may have a further opportunity to review the elements signed by [1] and to decide on the value of alternative signature mechanisms. Separating the elements necessary for (a) securing the From header field value and preventing replays, from (b) the elements necessary to prevent men-in-the-middle from tampering with messages, may also yield a strategy for identity that will be practicable in some highly mediated networks. It could be possible, for example, to provide two

signatures: one over the elements required for (b), and then a

separate signature over the elements necessary for (a) and the signature over (b); this would allow verification services in mediated networks to ignore the failure of a (b) signature while still verifying (a). Any solution along these lines must however always secure any cryptographic material necessary to support DTLS-SRTP or future security mechanisms.

5.5. Stickiness of Deployed Infrastructure

One thing that has not changed, and is not likely to change in the future, is the transitive nature of trust in the PSTN. When a call from the PSTN arrives at a SIP gateway with a calling party number, the gateway will have little chance of determining whether the originator of the call was authorized to claim that calling party number. Due to roaming and countless other factors, calls on the PSTN may emerge from administrative domains that have no relationship with the number assignee. This use case will remain the most difficult to tackle for an identity system, and may prove beyond repair. It does however seem that with the changes in the solution space, and a better understanding of the limits of [1] and VIPR, we are today in a position to reexamine the problem space and find solutions that can have a significant impact on the secure origins problem.

5.6. Relationship with Number Assignment and Management

Currently, telephone numbers are typically managed in a loose delegation hierarchy. For example, a national regulatory agency may task a private, neutral entity with administering numbering resources, such as area codes, and a similar entity with assigning number blocks to carriers and other authorized entities, who in turn then assign numbers to customers. In many countries, individual numbers are portable between carriers, at least within the same technology (e.g., wireline-to-wireline). Separate databases manage the mapping of numbers to switch identifiers, companies and textual caller ID information.

As the PSTN transitions to using VoIP technologies, new assignment policies and management mechanisms are likely to emerge. For example, it has been proposed that geography could play a smaller

role in number assignments, and that individual numbers are assigned to end users directly rather than only to service providers, or that the assignment of numbers does not depend on providing actual call delivery services.

Databases today already map telephone numbers to entities that have been assigned the number, e.g., through the LERG (originally, Local Exchange Routing Guide) in the United States. Thus, the transition

to IP-based networks may offer an opportunity to integrate cryptographic bindings between numbers or number ranges and service providers into databases.

5.7. Threat Model

The primary enabler of robocalling, vishing and related attacks is the capability to impersonate a calling party number. The most stark example of these attacks are cases where automated callees on the PSTN rely on the calling number as a security measure, for example to access a voicemail system. Robocallers use impersonation as a means of obscuring identity; while robocallers can, in the ordinary PSTN, block (that is, withhold) their caller identity, callees are less likely to pick up calls from blocked identities, and therefore calling from some number, any number, is prefereable. Robocallers however prefer not to call from a number that can trace back to the robocaller, and therefore they impersonate numbers that are not assigned to them.

The scope of impersonation in this threat model pertains solely to the rendering of a calling telephone number to an end user or automaton at the time of call set-up. The primary attack vector is therefore one where the attacker contrives for the calling telephone number in signaling to be a particular chosen number, one that the attacker does not have the authority to call from, in order for that number to be rendered on the terminating side. The threat model assumes that this attack simply cannot be prevented: there is no way to stop the attacker from creating calls that contain attacker-chosen calling telephone numbers in their signaling. The solution space therefore focuses on ways that terminating or intermediary elements might differentiate authorized from unauthorized calling party numbers, in order that policies, human or automatic, might act on that information.

Rendering an authenticated calling party number during call set-up time does not entail anything about the entity or entities that will send and receive media during the call itself. In call paths with intermediaries and gateways as described below, there may be no way to provide any assurance in the signaling about participants in the media. In those end-to-end IP environments where such an assurance is possible, it is highly desirable, but in the threat model considered in this document, the threat of impersonation does not extend to impersonating an authorized listener after a call has been completed. Attackers that could impersonate an authorized listener require powers that robocallers and voicemail hackers are unlikely to possess, and historically such attacks have not played a role in enabling robocalling or related problems.

In protocols like SIP, call signaling can be renegotiated after the call has been completed, and through various transfer mechanisms common in telephone systems, callees can easily be connected to, or conferenced in with, telephone numbers other than the original calling number once a call has been set up. These post-setup changes to the call are outside the scope of impersonation considered in this model. Furthermore, impersonating a reached number to the originator of a call is outside the scope of this threat model.

In much of the PSTN, there exists a supplemental service that translates calling party numbers into regular names, including the proper names of people and businesses, for rendering to the called user. These services (frequently termed 'Caller ID') provide a further attack surface for impersonation. The threat model explored in this document focuses only on the calling party number, though presenting a forged calling party number can let the attacker cause a forged 'Caller ID' name to be rendered to the user as well. Providing a verifiable calling party number therefore does improve the security of Caller ID systems, but this threat model does not consider attacks specific to Caller ID, such as attacks on the databases consulted by the terminating side of a call to provide Caller ID, or impersonators choosing to forge a particular calling party number in order to present a misleading Caller ID to the user.

Finally, the scope of impersonation in this threat model does not consider simple anonymity as a threat. The ability to place

anonymous calls has always been a feature of the PSTN, and users of the PSTN today have the capability to reject anonymous calls should they wish to.

5.7.1. Actors

5.7.1.1. Endpoints

There are two main categories of end-user terminals, a dumb device (such as a 'black phone') or a smart device:

> Dumb devices comprise a simple dial pad, handset and ringer, optionally accompanied by a display that can show only a limited number of characters (typically, enough for a telephone number and an accompanying name, sometimes less). These devices are controlled by service providers in the network.
>
> Smart devices are general purpose computers with some degree of programmability and the capacity to access the Internet, along with a rich display. This includes smart phones, telephone applications on desktop and laptop computers, IP private branch exchanges, and so on.

There are also various hybrid devices, such as terminal adapters which attach dumb devices to a VoIP service, but which may in turn use auxiliary screens as displays for rich information (for example, some cable deployments use the television screen to render caller ID). These devices expose little programmability to end users.

There is a further category of automated terminals without an end user. These include systems like voicemail services that consume the calling party number without rendering it to a human. Though the capability of voicemail services varies widely, many today have Internet access and advanced application interfaces (to render 'visual voicemail,' to automatically transcribe voicemail to email, and so on).

5.7.1.2. Intermediaries

We assume that a call between two endpoints traverses a call path. The length of the call path can vary considerably: it is possible in VoIP deployments for two endpoint entities to send traffic to one

another directly, but more commonly several intermediaries exist in a VoIP call path. One or more gateways may also appear on a call path.

Intermediaries forward call signaling to the next entity in the path. These intermediaries may also modify the signaling in order to improve interoperability, to enable proper network-layer media connections, or to enforce operator policy. This threat model assumes there are no restrictions on the modifications to signaling that an intermediary can introduce.

Gateways translate call signaling from one protocol into another. In the process, they tend to consume any signaling specific to the original protocol (elements like transaction-matching identifiers) and may need to transcode or otherwise alter identifiers as they are rendered in the destination protocol.

The threat model assumes that intermediaries and gateways can forward and retarget calls as necessary, which can result in a call terminating at a place the originator did not expect, and that this is an ordinary condition in call routing. This is significant to the solution space, however, because it limits the ability of the originator to anticipate what the telephone number of the respondent will be.

Furthermore, we assume that some intermediaries or gateways may, due to their capabilities or policies, discard calling party number information, as a whole or in part. Today, many IP-PSTN gateways simply ignore any information available about the caller in the IP leg of the call, and allow the telephone number of the PRI line that

the gateway happens to use to be sent as the calling party number for the PSTN leg of the call. A call might also gateway to a multifrequency network where only a limit number of digits of automatic numbering identification (ANI) data are signaled, for example. Some protocols may render telephone numbers in a way that makes it impossible for a terminating side to parse or canonicalize a number. In these cases, providing authenticated identity may be impossible. This is not however indicative of an attack or other security failure.

5.7.1.3. Attackers

We assume that an attacker has the following powers:

The attacker can create telephone calls at will, originating them on either the PSTN or over IP, and can supply an arbitrary calling party number.

The attacker can capture and replay signaling previously received. [TBD: should this include a passive attacker that can capture signaling that isn't directly sent to it? Not a factor for robocalling, but perhaps for voicemail hacking, say.]

The attacker has access to the Internet, and thus the ability to inject arbitrary traffic over the Internet, to access public directories, and so on.

There are many potential threats in which an attacker compromises intermediaries in the call path, or captures credentials that allow the attacker to impersonate a target. Those system-level threats are not considered in this threat model, though secure design of systems to prevent these sorts of attacks is necessary for any of these countermeasures to work.

This threat model also does not consider a case in which the operators of intermediaries or gateways are themselves adversaries who intentionally suppress identity or send falsified identity with their own credentials.

5.7.2. Attacks

5.7.2.1. Voicemail Hacking via Impersonation

A voicemail service allows users calling from their mobile phones access to their voicemail boxes on the basis of the calling party number. An attacker wants to access the voicemail of a particular target. The attacker therefore impersonates the calling party number using one of the scenarios described below.

In all cases, the countermeasure to this threat is for the voicemail service to have an expectation that calls to its service will supply an authenticated identity, and in the absence of that identity, for it to adopt a different policy (perhaps requiring a shared secret to be dialed as a PIN). Authenticated identity alone provides a

positive confirmation only when an identity is claimed legitimately; the absence of authenticated identity here is not evidence of malice, just of uncertainty.

If the voicemail service could know ahead of time that it should always expect authenticated identity from a particular number, that would enable the voicemail service to adopt different policies for handling a request without authenticated identity. Since users contact a voicemail service repeatedly, this is something that a voicemail server could learn, for example, the first time that a user contacts it. Alternatively, it could access a directory of some kind that informs verifiers that they should expect identity from particular numbers.

5.7.2.2. Unsolicited Commercial Calling from Impersonated Numbers

The unsolicited commercial calling, or for short robocalling, threat is similar to the voicemail threat, except in so far as the robocaller does not need to impersonate any specific number, merely a plausible number. A robocaller may impersonate a number that is not a valid number (for example, in the United States, a number beginning with 0), or an unassigned number. The robocaller may change numbers every time a new call is placed, even selecting numbers randomly.

The countermeasures to robocalling are similar to the voicemail example, but there are significant differences. One important potential countermeasure is simply to verify that the calling party number is in fact valid and assigned. Unlike voicemail services, end users typically have never been contacted by the number used by a robocaller before, so they can't rely on past association to know whether or not the calling party number should always supply authenticated identity. If there were a directory that could inform the terminating side of that fact, however, that would help in the robocalling case.

When alerting a human is involved, the time frame for executing these countermeasures is necessarily limited. Ideally, a user would not be alerted that a call has been received until any necessary identity checks have been performed. This could however result in inordinate post-dial delay from the perspective of legitimate callers. Cryptographic operations and network operations must be minimized for these countermeasures to be practical.

The eventual effect of these countermeasures would be to force robocallers to either block their caller identity, in which case end users could opt not to receive their calls, or to use authenticated identity for numbers traceable to them, which would then allow for other forms of redress.

5.7.2.3. Attack Scenarios

Impersonation, IP-PSTN

An attacker on the Internet uses a commercial WebRTC service to send a call to the PSTN with a chosen calling party number. The service contacts an Internet-to-PSTN gateway, which inserts the attacker's chosen calling party number into the CPN field of an IAM. When the IAM reaches the endpoint terminal, the terminal renders the attacker's chosen calling party number as the calling identity.

Countermeasure: out-of-band authenticated identity

Impersonation, PSTN-PSTN

An attacker with a traditional PBX (connected to the PSTN through an ISDN PRI) sends a Q.931 SETUP request with a chosen calling party number which a service provider inserts into the corresponding SS7 CPN field of an IAM. When the IAM reaches the endpoint terminal, the terminal renders the attacker's chosen calling party number as the calling identity.

Countermeasure: out-of-band authenticated identity

Impersonation, IP-IP

An attacker with an IP phone sends a SIP request to an IP-enabled voicemail service. The attacker puts a chosen calling party number into the From header field value of the INVITE. When the INVITE reaches the endpoint terminal, the terminal renders the attacker's chosen calling party number as the calling identity.

Countermeasure: in-band authenticated identity

Impersonation, IP-PSTN-IP

An attacker with an IP phone sends a SIP request to the telephone number of a voicemail service, perhaps without even knowing that the voicemail service is IP-based. The attacker puts a chosen calling party number into the From header field value of the INVITE. The attacker's INVITE reaches an Internet-to-PSTN gateway, which inserts the attacker's chosen calling party number into the CPN field of an

IAM. That IAM then traverses the PSTN until (perhaps after a call forwarding) it reaches another gateway, this time back to the IP realm, to an H.323 network. The PSTN-IP gateway puts takes the calling party number in the IAM CPN field and puts it into the SETUP request. When the SETUP reaches the endpoint terminal, the terminal renders the attacker's chosen calling party number as the calling identity.

Countermeasure: out-of-band authenticated identity

5.7.2.4. Solution-Specific Attacks

[TBD: This is just forward-looking notes]

Threats Against In-band

Token replay

Removal of in-band signaling features

Threats Against Out-of-Band

Provisioning Gargbage CPRs

Data Mining

Threats Against Either Approach

Attack on directories/services that say whether you should expect authenticated identity or not

Canonicalization attack

6. Requirements

This section describes the high level requirements:

Usability Any validation mechanism must work without human intervention, e.g., CAPTCHA-like mechanisms.

Deployability Must survive transition of the call to the PSTN and the presence of B2BUAs.

Validation by intermediaries Intermediaries as well as end system must be able to validate the source identity information.

Display name The display name of the caller must also be validated or the callee must be able to determine that only the calling number has been validated.

Consider existing structures must allow number portability among carriers and must support legitimate usage of number spoofing (doctor's office and call centers)

Minimal payload overhead Must lead to minimal expansion of SIP headers fields to avoid fragmentation in deployments that use UDP.

Privacy Any out-of-band validation protocol must not allows third parties to learn what numbers have been called by a specific caller.

7. Roadmap

The authors of this document believe that the entire solution scope consists of a couple of separable aspects:

In-band caller ID Conveyance: This functionality allows call origin identification information to be conveyed within SIP, and takes the nature of E.164 numbers and the prevalence of B2BUAs into account. This may consist of a revised version of the SIP Identity specification that takes E.164 numbers into account and allows for separate validation of the SIP request headers and the SIP request body. This approach addresses the case where intermediaries do not remove header fields.

Out-of-Band Caller-ID Verification: This functionality determines whether the E.164 number used by the calling party actually exists, the calling entity is entitled to use the number and whether a call has recently been made from this phone number. This approach is needed when the in-band technique does not work due to intermediaries or due to interworking with PSTN networks.

Certificate Delegation Infrastructure: This functionality defines how certificates with E.164 numbers are used in number portability, and delegation cases. It also describes how the existing numbering infrastructure is re-used to maintain the lifecycle of number assignments.

Extended Validation: This functionality describes how to describes attributes of the calling party beyond the caller-id and these attributes (e.g., the calling party is a bank) need to be verified upfront.

8. Acknowledgments

We would like to thank Alissa Cooper, Bernard Aboba, Sean Turner, Eric Burger, and Eric Rescorla for their discussion input that lead to this document.

9. IANA Considerations

This memo includes no request to IANA.

10. Security Considerations

This document is about improving the security of call origin identification.

11. Informative References

[1] Peterson, J. and C. Jennings, "Enhancements for Authenticated Identity Management in the Session Initiation Protocol (SIP)", RFC 4474, August 2006.

[2] Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J., Sparks, R., Handley, M., and E. Schooler, "SIP: Session Initiation Protocol", RFC 3261, June 2002.

[3] Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J., Sparks, R., Handley, M., and E.

Schooler, "SIP: Session Initiation Protocol", RFC 3261, June 2002.

[4] Jennings, C., Peterson, J., and M. Watson, "Private Extensions to the Session Initiation Protocol (SIP) for Asserted Identity within Trusted Networks", RFC 3325, November 2002.

[5] Hoffman, P. and J. Schlyter, "The DNS-Based Authentication of Named Entities (DANE) Transport Layer Security (TLS) Protocol: TLSA", RFC 6698, August 2012.

[6] Elwell, J., "Connected Identity in the Session Initiation Protocol (SIP)", RFC 4916, June 2007.

[7] Schulzrinne, H., "The tel URI for Telephone Numbers", RFC 3966, December 2004.

[8] Cooper, A., Tschofenig, H., Peterson, J., and B. Aboba, "Secure Call Origin Identification", draft-cooper-iab-secure-origin-00 (work in progress), November 2012.

[9] Peterson, J., "Retargeting and Security in SIP: A Framework and Requirements", draft-peterson-sipping-retarget-00 (work in progress), February 2005.

[10] Rosenberg, J., "Concerns around the Applicability of RFC 4474", draft-rosenberg-sip-rfc4474-concerns-00 (work in progress), February 2008.

[11] Kaplan, H. and V. Pascual, "Loop Detection Mechanisms for Session Initiation Protocol (SIP) Back-to- Back User Agents (B2BUAs)", draft-ietf-straw-b2bua-loop-detection-00 (work in progress), April 2013.

[12] Barnes, M., Jennings, C., Rosenberg, J., and M. Petit-Huguenin, "Verification Involving PSTN Reachability: Requirements and Architecture Overview", draft-jennings-vipr-overview-04 (work in progress), February 2013.

[13] Rosenberg, J. and H. Schulzrinne, "Session Initiation Protocol (SIP): Locating SIP Servers", RFC 3263, June 2002.

[14] Krebs, B., "DHS Warns of 'TDoS' Extortion Attacks on Public Emergency Networks", URL: http://krebsonsecurity.com/2013/04/dhs-warns-of-tdos-extortion-attacks-on-public-emergency-networks/, Apr 2013.

[15] FCC, ., "Robocalls", URL: http://www.fcc.gov/guides/robocalls, Apr 2013.

[16] FCC, ., "FCC Robocall Challenge", URL: http://robocall.challenge.gov/, Apr 2013.

[17] Wikipedia, ., "News International phone hacking scandal", URL: http://en.wikipedia.org/wiki/News_International_phone_hacking_scandal, Apr 2013.

[18] Wikipedia, ., "Don't Make the Call: The New Phenomenon of 'Swatting'", URL: http://www.fbi.gov/news/stories/2008/february/swatting020408, Feb 2008.

Authors' Addresses

Jon Peterson
NeuStar, Inc.
1800 Sutter St Suite 570
Concord, CA 94520
US

Email: jon.peterson@neustar.biz

Henning Schulzrinne
Columbia University
Department of Computer Science
450 Computer Science Building
New York, NY 10027

US

Phone: +1 212 939 7004
Email: hgs+ecrit@cs.columbia.edu
URI: http://www.cs.columbia.edu

Hannes Tschofenig
Nokia Siemens Networks
Linnoitustie 6
Espoo 02600
Finland

Phone: +358 (50) 4871445
Email: Hannes.Tschofenig@gmx.net
URI: http://www.tschofenig.priv.at