Internet Engineering Task Force (IETF)
Request for Comments: 7340
Category: Informational
ISSN: 2070-1721

J. Peterson
NeuStar, Inc.
H. Schulzrinne
Columbia University
H. Tschofenig

September 2014

Secure Telephone Identity Problem Statement and Requirements

Abstract

Over the past decade, Voice over IP (VoIP) systems based on SIP have replaced many traditional telephony deployments. Interworking VoIP systems with the traditional telephone network has reduced the overall level of calling party number and Caller ID assurances by granting attackers new and inexpensive tools to 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. Despite previous attempts to provide a secure assurance of the origin of SIP communications, we still lack effective standards for identifying the calling party in a VoIP session. This document 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. It also gives high-level requirements for a solution in this space.

Status of This Memo

Table of Contents

1. Introduction

In many communication architectures that allow users to communicate with other users, the need arises for identifying the originating party that initiates a call or a messaging interaction. The desire to identify communication parties in end-to-end communication derives 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) [RFC3261] by using two main types of approaches, namely, P-Asserted-Identity (PAI) [RFC3325] and SIP Identity [RFC4474], which are described in more detail in Section 5. The goal of these mechanisms is to validate that the originator of a call is authorized to claim an originating identifier. Protocols like the Extensible Messaging and Presence Protocol (XMPP) use mechanisms that are conceptually 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 improve in the future. In [SECURE-ORIGIN], we illustrate what challenges arise. In particular, interworking with different communication architectures (e.g., SIP, Public Switched Telephone Network (PSTN), XMPP, Real-Time Communications on the Web (RTCWeb)) or other forms of mediation breaks the end-to-end semantic of the communication interaction and destroys any identification capabilities. (In this document, we use the term "PSTN" colloquially rather than in a legal or policy sense, as a common shorthand for the circuit-switched analog and time-division multiplexing (TDM) digital telephone system, often using Signaling System #7 (SS7) to control call setup and teardown.) 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; although domain-based identifiers (sip:user@example.com) might use certificate or DNS-related approaches to determine who is able to claim "ownership" of the URI, telephone numbers do not yet have any similar mechanism defined.

After the publication of the PAI and SIP Identity specifications ([RFC3325] and [RFC4474], respectively), further attempts have been made to tackle the topic but, unfortunately, with little success, due to the complexity of deploying solutions 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 to develop a common understanding of the problem

statement as well as basic 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, there were a limited number of carriers, all of whom trusted each other to provide accurate caller origination information in an environment without any cryptographic validation. 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., in the manner carriers are certified in the United States), 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 there was a clear distinction between customers and service providers in the past, VoIP service providers can now easily act as customers or either originating or transit providers. Moreover, the problem is not limited to voice communications, as growth in text messaging has made it another vector for bulk unsolicited commercial messaging relying on impersonation of a source telephone number or, sometimes, an SMS short code. 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 have been subjected to telephony denial-of-service attacks [TDOS]. 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 US Federal Communications Commission webpage on this topic [ROBOCALL-FCC]). Robocalls are a problem that has been recognized already by various regulators; for example, the US Federal Trade Commission (FTC) recently organized a robocall competition to solicit ideas for creating solutions that will block illegal robocalls [ROBOCALL-CHALLENGE]. 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 and, for example, activates credit cards, or for harassment, e.g., causing utility services to be disconnected, take-out food to be delivered, or police to respond to a non-existing hostage situation ("swatting"; see [SWATTING]). 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, in the News International phone-hacking scandal [NEWS-HACK], employees of the newspaper were accused of engaging in phone hacking by utilizing Caller ID spoofing to get access to 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 identify perpetrators and craft policies to block them. It also makes tracing such calls much more labor-intensive, as each call has to be identified in each transit carrier hop-by-hop, based on destination number and time of call.

It is insufficient to simply outlaw all spoofing of originating telephone numbers because the entities spoofing numbers are already committing other crimes and are thus unlikely to be deterred by legal sanctions. 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.

The problem space is further complicated by a number of use cases where entities in the telephone network legitimately send calls on behalf of others, including "Find-Me/Follow-Me" services. Ultimately, any SIP entity can receive an INVITE request and forward it to any other entity, and the recipient of a forwarded message has little means to ascertain which recipient a call should legitimately target (see [SIP-SECURITY]). 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:

Doctors' offices: 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. Terminology

The following terms are defined in this document:

In-band Identity Conveyance: In-band conveyance is the presence of call origin identification information conveyed within the control plane protocol(s) setting up a call. Any in-band solution must accommodate in-band intermediaries such as Back-to-Back User Agents (B2BUAs).

Out-of-Band Identity Verification: Out-of-band verification determines whether the telephone number used by the calling party actually exists, whether 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 because the in-band technique does not work in all cases, as when certain intermediaries are involved or due to interworking with circuit-switched networks.

Authority Delegation Infrastructure: The delegation authority infrastructure determines how the authority over telephone numbers is used when numbers are ported and delegated. It also describes how the existing numbering infrastructure is reused to maintain the lifecycle of number assignments.

Canonical Telephone Number: In order for either in-band conveyance or out-of-band verification to work, entities must be able to canonicalize telephone numbers to arrive at a common syntactical form.

4. 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.

4.1. VoIP-to-VoIP Call

For the VoIP-to-VoIP 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 in separate use cases 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 [RFC3966].

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.)

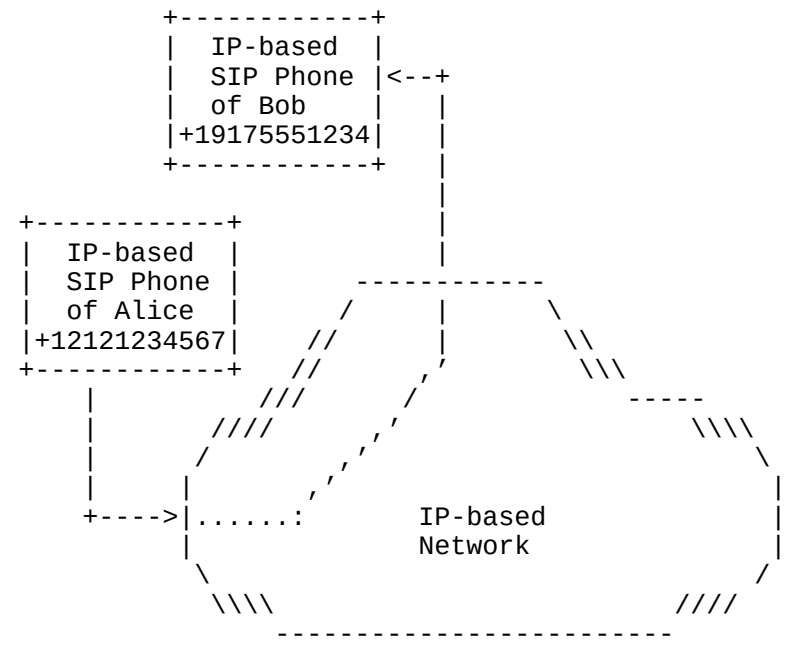

Figure 1: VoIP-to-VoIP Call

4.2. VoIP-PSTN-VoIP Call

Frequently, two VoIP-based service providers are not directly connected by VoIP and use Time Division Multiplexer (TDM) circuits to exchange calls, leading to the IP-PSTN-IP use case. In this use case, Dan’s Voice Service Provider (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.

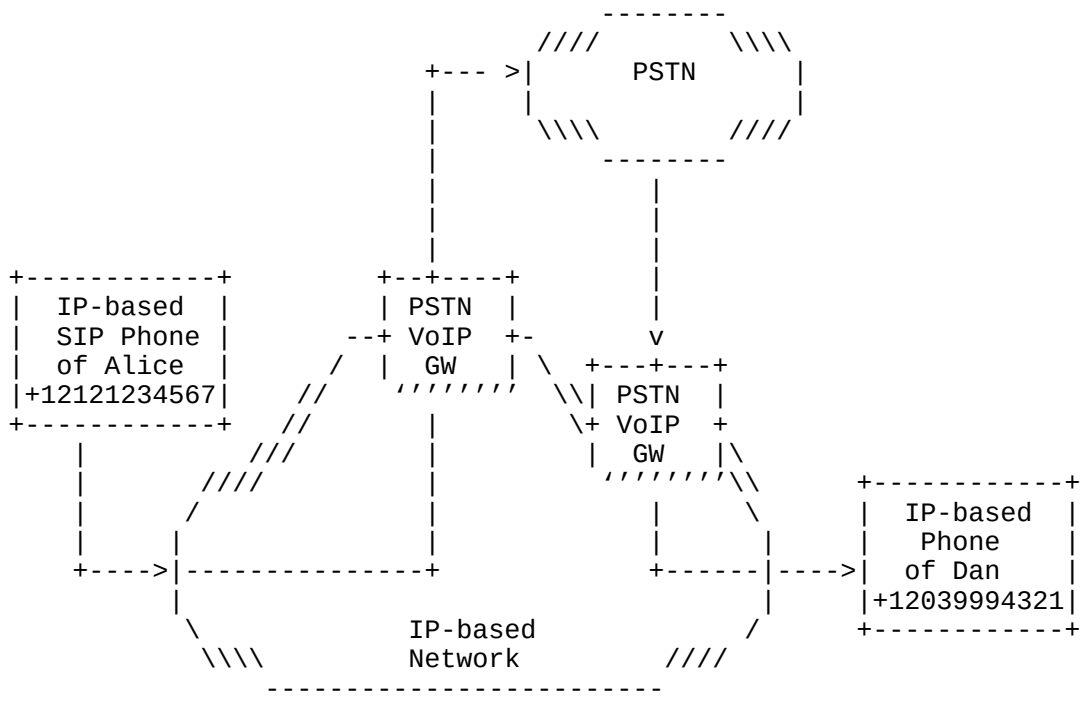

Figure 2: IP-PSTN-IP Call

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.

4.3. 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 from 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 caller identification information it had received through the PSTN, if available.

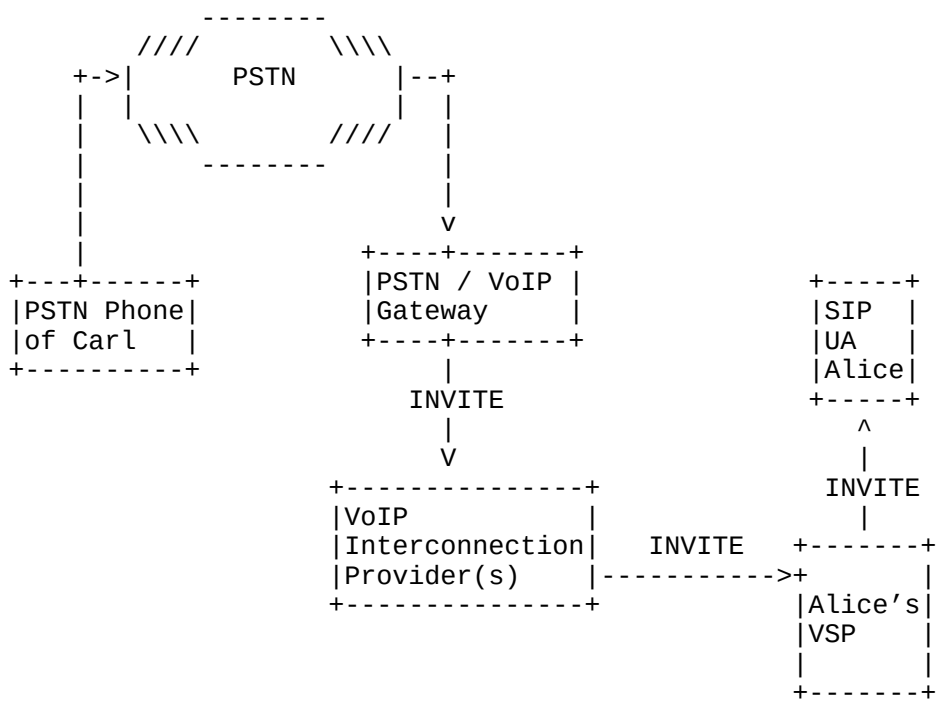

Figure 3: PSTN-to-VoIP Call

4.4. VoIP-to-PSTN Call

Consider Figure 4, where Alice calls Carl. Carl uses a PSTN phone, and Alice uses an IP-based phone. When Alice initiates the call, the E.164 number is 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. It is desirable that the gateway verify that Alice can claim the E.164 number she is using before it populates the corresponding calling party number field in telephone network signaling. Carl's phone must be able to verify that it is receiving a legitimate call from the calling party number it will render to Carl.

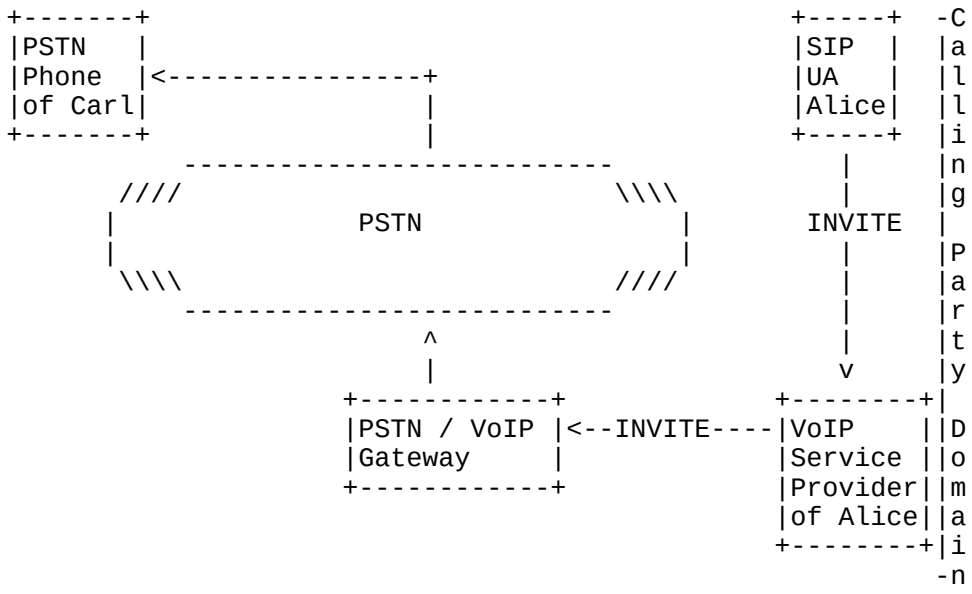

Figure 4: VoIP-to-PSTN Call

4.5. PSTN-VoIP-PSTN Call

Consider Figure 5, where Carl calls Alice. Both users have PSTN phones, but interconnection between the two circuit-switched parts of the PSTN is accomplished via an IP network. Consequently, 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.

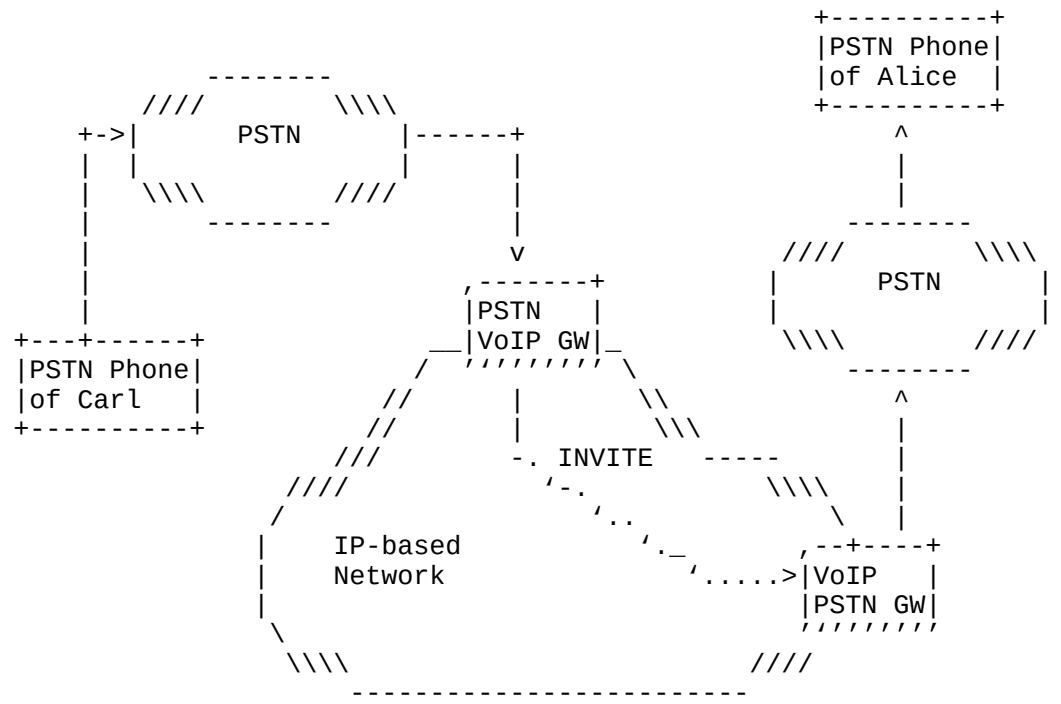

Figure 5: PSTN-VoIP-PSTN Call

4.6. 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.

5. 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 [RFC3261], 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 [RFC3325], 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 untrusted entities

to set the P-Asserted-Identity information. An overview of addressing spam in SIP and an explanation of how it differs from similar problems with email appeared in [RFC5039].

Subsequent efforts to prevent calling-origin identity spoofing in SIP include the SIP Identity effort (the "long-term" identity solution) [RFC4474] and Verification Involving PSTN Reachability (VIPR) [VIPR-OVERVIEW]. 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 to prevent both (a) SIP UAs from originating calls with spoofed From headers and (b) intermediaries, such as SIP proxies, from launching man-in-the-middle attacks by altering calls as they pass through the intermediaries. 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 on a significant scale, but they may still offer ideas and protocol building blocks for a solution.

5.1. P-Asserted-Identity

The P-Asserted-Identity header field of SIP [RFC3325] provides a way for trusted network entities to share with one another an authoritative identifier for the originator of a call. The value of P-Asserted-Identity cannot be populated by a user, though if a user wants to suggest an identity to the trusted network, a separate header (P-Preferred-Identity) enables them to do so. The features of the P-Asserted-Identity header evolved as part of a broader effort to reach parity with traditional telephone network signaling mechanisms for selectively sharing and restricting presentation of the calling party number at the user level while still allowing core network elements to know the identity of the user for abuse prevention and accounting.

In order for P-Asserted-Identity to have these properties, it requires the existence of a trust domain as described in [RFC3324]. Any entity in the trust domain may add a P-Asserted-Identity header to a SIP message, and any entity in the trust domain may forward a message with a P-Asserted-Identity header to any other entity in the trust domain. If a trusted entity forwards a SIP request to an untrusted entity, however, the P-Asserted-Identity header must first be removed; most end-user devices are outside trust domains. Sending a P-Asserted-Identity request to an untrusted entity could leak potentially private information, such as the network-asserted calling

party number in a case where a caller has requested presentation restriction. This concept of a trust domain is modeled on the trusted network of devices that operate the traditional telephone network.

P-Asserted-Identity has been very successful in telephone replacement deployments of SIP. It is an extremely simple in-band mechanism, requiring no cryptographic operations. Since it is so reminiscent of legacy mechanisms in the traditional telephone network and interworks so seamlessly with those protocols, it has naturally been favored by providers comfortable with these operating principles.

In practice, a trust domain exhibits many of the same merits and flaws as the traditional telephone network when it comes to securing a calling party number. Any trusted entity may provide P-Asserted-Identity, and a recipient of a SIP message has no direct assurance of who generated the P-Asserted-Identity header field value: all trust is transitive. Trust domains are dictated by business arrangements more than by security standards; thus, the level of assurance of P-Asserted-Identity is only as good as the least trustworthy member of a trust domain. Since the contents of P-Asserted-Identity are not intended for consumption by end users, end users must trust that their service provider participates in an appropriate trust domain, as there will be no direct evidence of the trust domain in the SIP signaling that end-user devices receive. Since the mechanism is so closely modeled on the traditional telephone network, it is unlikely to provide a higher level of security than that.

Since [RFC3325] was written, the whole notion of "P-" headers intended for use in private SIP domains has also been deprecated (see [RFC5727]) largely because of overwhelming evidence that these headers were being used outside of private contexts and leaking into the public Internet. It is unclear how many deployments that make use of P-Asserted-Identity in fact conform to the Spec(T) requirements of [RFC3324].

P-Asserted-Identity also complicates the question of which URI should be presented to a user when a call is received. Per [RFC3261], SIP user agents would render the contents of the From header field to a user when receiving an INVITE request, but what if the P-Asserted-Identity contains a more trustworthy URI, and presentation is not restricted? Subsequent proposals have suggested additional header fields to carry different forms of identity related to the caller, including billing identities. As the calling identities in a SIP request proliferate, the question of how to select one to render to the end user becomes more difficult to answer.

5.2. SIP Identity

The SIP Identity mechanism [RFC4474] provides 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 that 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 Session Description Protocol (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 a resource that the originator did not specify and thus 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. [RFC4474] 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, it provides identity only for domain names rather than other identifiers, and second, 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, [RFC4474] 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 particular number has been assigned to any specific carrier. For a tel URI, moreover, it is unclear in [RFC4474] 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 [CONCERNS]. 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 re-signs 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.

Moreover, as deployed, SIP is highly mediated and is mediated in ways that [RFC3261] did not anticipate. As request routing commonly depends on policies dissimilar to [RFC3263], requests transit multiple intermediate domains to reach a destination; some forms of intermediaries in those domains may effectively reinitiate 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. 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 denominator 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 delivered 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 provide policy or interworking functions at network boundaries (hence, the nickname "Session Border Controller"). These functions range from topology hiding, to alterations necessary to interoperate successfully with particular SIP implementations, to simple network address translation from private address space. To implement these functions, these entities modify SIP INVITE requests in transit, potentially 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. Consequently, a SIP request exiting a B2BUA does not necessarily bear much resemblance to the original request received by the B2BUA, just as an SS7 request exiting a PSTN gateway may transform all aspects of the SIP request in the VoIP 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 [RFC4474] in practical deployments. Some have therefore proposed weakening the security constraints of [RFC4474] so that at least some deployments of B2BUAs will be compatible with integrity protection of SIP requests. However, such solutions do not address the key problems 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.

5.3. 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 rendezvous 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 Private Branch Exchange (IPBX), 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. Moreover, the destination 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 that 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 also provide an authorization mechanism for accepting calls over the Internet, which significantly reduces the potential for spam. Because the token can act as a cookie 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 far less of a SIP request with its signature.

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 solution to the classic "robocall" problem, where the target typically receives a call from a number that has never called before. All of VIPR's 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 more to 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 of 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 (DHT) 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; thus, an originator may find multiple advertisements for a number on the DHT even when an attack is not in progress. Attackers may learn from these validation attempts which VIPR entities recently placed calls to the target number, even if they cannot impersonate the target since they lack the PSTN call detail information. 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 required 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.

6. Environmental Changes

6.1. Shift to Mobile Communication

In the years since [RFC4474] was conceived, there have been a number of fundamental shifts in the communications marketplace. The most transformative has been the precipitous rise of mobile smartphones, 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 SMS and Multimedia Messaging Service (MMS) capabilities. This suite of tools suggests that some of the techniques proposed by VIPR could be adapted to the smartphone environment. The installed base of smartphones is, moreover, highly upgradable and permits rapid adoption of out-of-band rendezvous services for smartphones that bypass the PSTN. Mobile messaging services that use telephone numbers as identities allow smartphone users to send text messages to one another over the Internet rather than over the PSTN. Like VIPR, such services create an out-of-band connection over the Internet between smartphones; unlike VIPR, the rendezvous service is provided by a trusted centralized database rather than by a DHT, and it is the centralized database that effectively verifies and asserts the telephone number of the sender of a message. While such messaging services are specific to the users of the specific service, it seems clear that similar databases could be provided by neutral third parties in a position to coordinate between endpoints.

6.2. Failure of Public ENUM

At the time [RFC4474] 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, some national authorities for telephone numbers are migrating their provisioning services to the Internet and issuing credentials that express authority for telephone numbers to secure those services. These new authorities for numbers could provide to the public Internet the necessary signatory authority for securing calling party numbers. While these systems are far from universal, the authors of this document believe that a solution devised for the North American Numbering Plan could have applicability to other country codes.

6.3. Public Key Infrastructure Developments

There have been a number of recent high-profile compromises of web certificate authorities. The presence of numerous (in some cases, hundreds) trusted certificate authorities in modern web browsers has become a significant security liability. As [RFC4474] relied on web certificate authorities, this too provides new lessons for any work on revising [RFC4474], namely, that innovations like DNS-Based Authentication of Named Entities (DANE) [RFC6698], which 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 considering 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.

6.4. Prevalence of B2BUA Deployments

Given the prevalence of established B2BUA deployments, we may have a further opportunity to review the elements signed using the SIP Identity mechanism [RFC4474] 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. Solutions in this space must, however, remain mindful of the requirements for securing cryptographic material necessary to support Datagram Transport Layer Security for Secure RTP (DTLS-SRTP) or future security mechanisms.

6.5. 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 were not assigned the originating number. 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 [RFC4474] 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.

6.6. Concerns about Pervasive Monitoring

While spoofing the origins of communication is a source of numerous security concerns, solutions for identifying communications must also be mindful of the security risks of pervasive monitoring (see [RFC7258]). Identifying information, once it is attached to communications, can potentially be inspected by parties other than the intended recipient and collected for any number of reasons. As stated above, the purpose of this work is not to eliminate anonymity; furthermore, to be viable and in the public interest, solutions should not facilitate the unauthorized collection of calling data.

6.7. 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. Resellers with looser regulatory obligations can complicate the picture, and in many cases, it is difficult to distinguish the roles of enterprises from carriers. 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, that individual numbers could be assigned to end users directly rather than only to service providers, and that the assignment of numbers does not have to 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 (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.

7. Basic Requirements

This section describes only the high-level requirements of the STIR effort, which we expect will be further articulated as work continues:

Generation: Intermediaries as well as end systems must be able to generate the source identity information.

Validation: Intermediaries as well as end systems must be able to validate the source identity information.

Usability: Any validation mechanism must work without human intervention, for example, without mechanisms like CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart).

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

Reflecting existing authority: Must stage credentials on existing national-level number delegations, without assuming the need for an international golden root on the Internet.

Accommodating current practices: Must allow number portability among carriers and must support legitimate usage of number spoofing (e.g., doctors' offices and call centers).

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

Efficiency: Must minimize RTTs for any network lookups and minimize any necessary cryptographic operations.

Privacy: A solution must minimize the amount of information that an unauthorized party can learn about what numbers have been called by a specific caller and what numbers have called a specific called party.

Some requirements specifically outside the scope of the effort include:

Display name: This effort does not consider how the display name of the caller might be validated.

Response authentication: This effort only considers the problem of providing secure telephone identity for requests, not for responses to requests; no solution is proposed for the problem of determining to which number a call has connected [RFC4916].

8. Acknowledgments

We would like to thank Sanjay Mishra, Fernando Mousinho, David Frankel, Penn Pfautz, Mike Hammer, Dan York, Andrew Allen, Philippe Fouquart, Hadriel Kaplan, Richard Shockey, Russ Housley, Alissa Cooper, Bernard Aboba, Sean Turner, Brian Rosen, Eric Burger, and Eric Rescorla for the discussion and input that contributed to this document.

9. Security Considerations

This document is about improving the security of call origin identification; security considerations for specific solutions will be discussed in solutions documents.

10. Informative References

[CONCERNS] Rosenberg, J., "Concerns around the Applicability of RFC 4474", Work in Progress, February 2008.

[NEWS-HACK] Wikipedia, "News International phone hacking scandal", June 2014, <http://en.wikipedia.org/w/index.php?title=News_International_phone_hacking_scandal&oldid=614607591>.

[RFC3261] 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.

[RFC3263] Rosenberg, J. and H. Schulzrinne, "Session Initiation Protocol (SIP): Locating SIP Servers", RFC 3263, June 2002.

[RFC3324] Watson, M., "Short Term Requirements for Network Asserted Identity", RFC 3324, November 2002.

[RFC3325] 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.

[RFC3966] Schulzrinne, H., "The tel URI for Telephone Numbers", RFC 3966, December 2004.

[RFC4474] Peterson, J. and C. Jennings, "Enhancements for Authenticated Identity Management in the Session Initiation Protocol (SIP)", RFC 4474, August 2006.

[RFC4916] Elwell, J., "Connected Identity in the Session Initiation Protocol (SIP)", RFC 4916, June 2007.

[RFC5039] Rosenberg, J. and C. Jennings, "The Session Initiation Protocol (SIP) and Spam", RFC 5039, January 2008.

[RFC5727] Peterson, J., Jennings, C., and R. Sparks, "Change Process for the Session Initiation Protocol (SIP) and the Real- time Applications and Infrastructure Area", BCP 67, RFC 5727, March 2010.

[RFC6698] Hoffman, P. and J. Schlyter, "The DNS-Based Authentication of Named Entities (DANE) Transport Layer Security (TLS) Protocol: TLSA", RFC 6698, August 2012.

[RFC7258] Farrell, S. and H. Tschofenig, "Pervasive Monitoring Is an Attack", BCP 188, RFC 7258, May 2014.

[ROBOCALL-CHALLENGE] Federal Trade Commission (FTC), "FTC Robocall Challenge", <http://robocall.challenge.gov/>.

[ROBOCALL-FCC] Federal Communications Commission (FCC), "Robocalls", April 2013, <http://www.fcc.gov/guides/robocalls>.

[SECURE-ORIGIN] Cooper, A., Tschofenig, H., Peterson, J., and B. Aboba, "Secure Call Origin Identification", Work in Progress, November 2012.

[SIP-SECURITY] Peterson, J., "Retargeting and Security in SIP: A Framework and Requirements", Work in Progress, February 2005.

[SWATTING] The Federal Bureau of Investigation (FBI), "Don’t Make the Call: The New Phenomenon of ’Swatting’", February 2008, <http://www.fbi.gov/news/stories/2008/february/swatting020408>.

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

[VIPR-OVERVIEW] Barnes, M., Jennings, C., Rosenberg, J., and M. Petit-Huguenin, "Verification Involving PSTN Reachability: Requirements and Architecture Overview", Work in Progress, December 2013.

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@cs.columbia.edu
URI: http://www.cs.columbia.edu

Hannes Tschofenig
Hall, Tirol 6060
Austria

EMail: Hannes.Tschofenig@gmx.net
URI: http://www.tschofenig.priv.at