PinDr0p: Using Single-Ended Audio Features To Determine Call Provenance

Vijay A. Balasubramaniyan, Aamir Poonawalla, Mustaque Ahamad,
Michael T. Hunter and Patrick Traynor
Converging Infrastructure Security (CISEC) Laboratory
Georgia Tech. Information Security Center (GTISC)
Georgia Institute of Technology
266 Ferst Dr. NW, Atlanta, GA 30313, USA
{vijayab, aamirp, mustaq, mhunter, traynor}@cc.gatech.edu

ABSTRACT

The recent diversification of telephony infrastructure allows users to communicate through landlines, mobile phones and VoIP phones. However, call metadata such as Caller-ID is either not transferred or transferred without verification across these networks, allowing attackers to maliciously alter it. In this paper, we develop PinDr0p, a mechanism to assist users in determining call provenance — the source and the path taken by a call. Our techniques detect and measure single-ended audio features to identify all of the applied voice codecs, calculate packet loss and noise profiles, while remaining agnostic to characteristics of the speaker's voice (as this may legitimately change when interacting with a large organization). In the absence of verifiable call metadata, these features in combination with machine learning allow us to determine the traversal of a call through as many as three different providers (e.g., cellular, then VoIP, then PSTN and all combinations and subsets thereof) with 91.6% accuracy. Moreover, we show that once we identify and characterize the networks traversed, we can create detailed fingerprints for a call source. Using these fingerprints we show that we are able to distinguish between calls made using specific PSTN, cellular, Vonage, Skype and other hard and soft phones from locations across the world with over 90% accuracy. In so doing, we provide a first step in accurately determining the provenance of a call.

Categories and Subject Descriptors

H.4.3 [**Information Systems Applications**]: Communications Applications

General Terms

Security

Keywords

provenance, telephony, VoIP, fingerprinting

CCS'10, October 4–8, 2010, Chicago, Illinois, USA.

1. INTRODUCTION

The current telephony infrastructure allows users to communicate using a variety of technologies. Circuit switched landlines continue to provide telephony to the majority of homes and businesses. Mobile phones now offer service to more than four billion users [54]. Voice over IP (VoIP) allows users to inexpensively communicate with each other irrespective of the geographical distances, with systems such as Skype [10] currently serving over 400 million users [8]. Each of these telecommunication networks adopt their own set of standards, from the underlying transport protocols to the codecs used to transmit media. Yet, they seamlessly interact through a variety of conversion mechanisms. A call may traverse multiple such networks, taking advantage of the benefits offered by each before reaching its final destination.

The diversification of telephony infrastructure significantly reduces the integrity associated with call metadata, such as Caller-ID [1], as it is either not transferred across these networks or is transferred without verification. This allows easy manipulation of metadata by hardware and software including soft phones on desktop computers. For example, between January 21st and 26th of 2010, customers of banks in four states received calls asking them to reveal personal information including credit card and PIN details. Many of these attacks use VoIP phones to anonymously and inexpensively dial a large number of customers while forging the Caller-IDs of these banks [30].

In this paper, we develop PinDr0p[1], an infrastructure to assist users in determining the provenance of a call — the source and the path taken by a call. Through a combination of signal processing and machine learning, we show that regardless of the claimed source, the audio delivered to the receiver exhibits measurable features of the networks through which the call was delivered. For example, calls that traverse a VoIP network experience packet loss that results in perceivable effects in the final call audio. Such artifacts are noticeably absent in calls that have only traversed cellular or Public Switched Telephone Networks (PSTNs). In particular, the codec transformations applied by multiple intermediary PSTNs, VoIP and cellular networks, in combination with packet loss and noise characteristics, allow us to develop profiles for various call sources based solely on features extracted from the received audio. In the absence of any verifiable metadata, these features offer a means of developing source fingerprints that help compare and distinguish different incoming calls.

We make the following contributions:

[1]Our mechanisms take advantage of audio and path artifacts that, like the sound made by the drop of a pin, are largely inaudible to the human ear.

- **Identify robust source and network path artifacts extracted purely from the received call audio:** We show that the received call audio provides extractable features that are strong identifiers of the networks that the call has traversed, allowing us to determine the provenance of a call. These include degradations (packet loss in VoIP) and noise characteristics of codecs unique to each network.

- **Develop call provenance classifier architecture:** We develop a multi-label machine learning classifier based on the extracted features to correctly identify the provenance of an incoming call with 91.6% accuracy with as little as 15 seconds of audio. Because PinDr0p does not rely on metadata available in some networks (e.g., VoIP) or cryptography, it is more readily deployable across the diverse devices and networks that make up modern telephony systems.

- **Demonstrate our robustness in identifying call provenance for *live* calls:** We make calls using PSTN phones, cellular phones, Vonage, Skype and other soft phones from locations across the world and are able to distinguish between them with 90% accuracy with only a small sample being labeled. As we increase the number of such labels we are able to distinguish between these calls with 100% accuracy. This demonstrates that PinDr0p makes VoIP-based phishing attacks harder and provides an important first step towards a Caller-ID alternative.

We note that while our approach does not provide the same guarantees as the use of end-to-end cryptography, it is also not encumbered with the difficulties of key distribution, management and the requirement that both endpoints are capable of such operations. The guarantees provided by our approach are instead more akin to traceback techniques from IP networks [41]. However, PinDr0p does not mandate the modification of the core infrastructure to attach additional metadata in-transit as our provenance information is extracted directly from the received audio. While adversaries may attempt to modify their attack in order to circumvent PinDr0p (e.g., change codecs, replicate the noise profile *and* change the physical location from which an attack is launched to match packet loss characteristics), our approach significantly increases the difficulty of successfully launching such an attack and improves the chances of identifying an attacker.

The remainder of this paper is organized as follows: Section 2 provides background information on telephony networks and codecs; Section 3 discusses the details of our proposed call provenance mechanism; Section 4 details our experimental setup and results; Section 5 presents experimental results from a real-world attack scenario; Section 6 offers further insight into our scheme and discusses trade-offs and limitations; Section 7 presents related work; Section 8 provides concluding remarks and future work.

2. BACKGROUND

Telephony networks are exceedingly complex systems. While once designed, manufactured and run by a single company, today's networks are an elaborate combination of many different technologies. We offer a very high-level description of these systems, how voice is encoded in them and the transformations that occur as voice crosses between different classes of networks.

As shown in Figure 1, there are three general classes of telephony networks. PSTNs represent traditional circuit-switched telephony systems. These networks are generally characterized by lossless connections and high fidelity audio. While pieces of the core of some of these networks are being replaced by IP connections, these

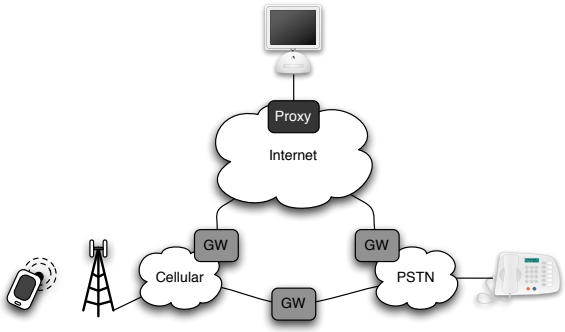

Figure 1: A high-level description of modern telephony systems. Note that a call between two endpoints may cross a variety of networks. At each gateway, calls are re-encoded using that network's codec.

Codec	Networks	Applications
G.711	PSTN, VoIP	Standard Telephony
GSM-FR	Cellular	Cellular Telephony
iLBC	VoIP	VoIP over Cable
Speex	VoIP	XBox Live
G.729	VoIP	SkypeOut/SkypeIn

Table 1: Audio Codecs and their typical deployment. G.711 is widely used in both PSTN and VoIP networks

.

private links are tightly controlled to ensure near zero packet loss. Like PSTN systems, cellular networks have a circuit switched core, with portions currently being replaced by IP links. While these networks can have considerably different technologies deployed in their wireless interfaces, their cores are extremely similar. Finally, VoIP networks by name run on top of IP links and generally share the same paths as all other Internet-based traffic. Accordingly, VoIP systems virtually always experience packet loss.

Voice is encoded and decoded in each of these networks using a variety of *codecs*. Specific codecs are selected for different networks based on competing goals including sound quality, robustness to noise and bandwidth requirements. While a large number of codecs exist, we describe and study the five most commonly used narrowband codecs in this work. We summarize these codecs and their typical environments in Table 1.

The codec used all over the world in PSTNs is G.711 [47], with North America and Japan using the mu-law compression algorithm and Europe and the rest of the world using A-law. Both the algorithms generate a 64 kbps (20 ms audio frames) Constant Bit Rate (CBR) stream for speech sampled at 8kHz, which is relatively bandwidth intensive when compared to other codecs. In cellular networks, the GSM full rate (GSM-FR) [22] codec was the first digital cellular coding standard and is still widely used in networks around the world. Unlike G.711, which is a waveform coder, GSM-FR uses predictive coding, which is more common among modern codecs and allows a large reduction in bandwidth requirements, with GSM-FR having an average bit rate of 13 kbps.

A plethora of codecs have been specifically designed for VoIP systems. The Internet Low Bit-rate codec (iLBC) [20] is extremely robust to packet losses and operates on a bit rate of 13.33 kbps (30 ms audio frames) and 15.20 kbps (20 ms audio frames). iLBC is a mandatory standard for VoIP over Cable and is also used by Google Voice and Skype [10]. Speex [6] is a Variable Bit Rate (VBR)

codec that supports a wide range of bit-rates from 2.15 kbps to 44 kbps and uses 20 ms audio frames. Speex, in addition to being supported on many VoIP soft phones, is commonly used in gaming teleconferencing systems such as Xbox Live [7]. A large number of VoIP systems also use G.729 (10 ms audio frames) [48], which requires very low bandwidth as it supports a CBR of 8kbps. Skype also uses G.729 when making and receiving calls to landlines and mobile phones (SkypeOut/SkypeIn service). It is also used by most Cisco hard IP phones [9]. Finally, a number of VoIP phones also support G.711, which is used in PSTN systems.

Audio must be reencoded when passing between two different telephony networks. For instance, whereas the audio in a call between two PSTN users is likely to only have been encoded in G.711, both G.711 and GSM-FR will be applied to the audio for a conversation between users on a PSTN and cellular network, respectively. Encoding changes occur in media gateways located at the edge of telephony networks, meaning that VoIP calls can traverse multiple Internet autonomous systems without necessarily being reencoded. Through this infrastructure, phone calls are delivered seamlessly between users. *To establish call provenance, we seek to measure these transformations as well as characteristics of the underlying networks.*

3. CALL PROVENANCE

The provenance of a call describes the characteristics of the source and traversed networks. This information can be used to create fingerprints that help distinguish and compare different calls in the absence of verifiable end to end metadata. For example, provenance can be used to identify if a call has passed through a VoIP network and, if it has not typically done so, alert the receiver of the change. At the very least, provenance must be able to distinguish between traffic that has traversed different telephony networks: PSTN, cellular and VoIP. We investigate whether this can be achieved with only the audio content available at the receiving end of a call. This approach is attractive as provenance can be determined without access or modification to intermediate network elements such as gateways or routers.

As a call traverses multiple networks, the audio is repeatedly reencoded with the network's choice of codec. To illustrate, a Skype call to a landline is initially encoded using G.729 and re-encoded using G.711 when it encounters the VoIP-PSTN gateway. If we can extract artifacts of each of the applied codecs from the received audio then simple codec to network translation (G.729 $\Longrightarrow$ VoIP) determines call provenance. In addition, identifying the codec used in a particular network helps characterize that network. However, codecs like G.711 are widely used in both PSTN and VoIP systems, implying codec detection alone is insufficient. Therefore, we seek additional differentiators.

Networks themselves introduce degradations into call audio. In VoIP, there are packet losses which are not seen in circuit switched PSTN networks. Similarly, mobile phones have bit errors due to fading effects on radio channels. The loss of an entire packet containing 20 ms of speech is measurably different from a small number of incorrect bits. These features are more robust than simply extracting codec information as packet loss and bit errors are hard for an adversary to control — *an adversary bounded by a lossy connection, many miles away, cannot spoof a lossless, dedicated PSTN line to a bank.*

Solution Overview: To identify and characterize the different networks a call has traversed we focus on degradations specific to each network. We first demonstrate how we can identify and characterize a VoIP network by detecting packet loss or concealed packet loss in the received audio. We then show how PSTN and cellular networks can be identified and characterized due to their vastly different noise characteristics. Finally, since the quality of the received audio significantly degrades with the number of networks traversed, we extract quality specific features. We create a feature vector that aggregates feature values obtained from the packet loss, noise and quality measurements and use it to train a multi-label classifier to identify the networks that a call originated and traversed. In addition, we demonstrate how the feature vector provides call provenance fingerprints that can be used to consistently identify a call source.

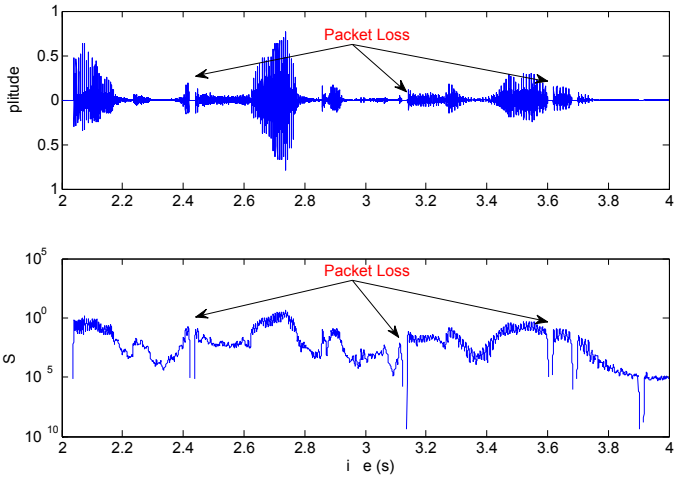

Figure 2: Packet Loss and Corresponding Energy Drop. The breaks in the signal (top) that occur due to packet loss are more accurately determined using the short time energy (bottom) of the signal.

3.1 Identifying VoIP Networks

3.1.1 Detecting Packet Loss

Within an IP network a lost packet can be easily identified using the sequence numbers present in each packet (metadata). However, these sequence numbers are lost once the call is retransmitted over another telephony network. Accordingly, we must identify artifacts of these lost packets from the received audio. The top graph in Figure 2 shows two seconds of speech encoded with G.711 and transmitted through a VoIP network with a packet loss rate of 5%. The effect of a lost packet is sometimes visibly identifiable by a break in the waveform (annotated by arrows). However, such loss can be detected more accurately by determining the short-time average energy of the signal, as shown in the bottom graph in Figure 2.

Short-time average energy (STE) is traditionally used in speech analysis to detect words surrounded by pauses as they cause abrupt drops in energy levels. This can be adapted to detect a packet loss, which also causes an abrupt decrease in energy. STE for a signal $y(n)$ is defined as:

$$E_n = \sum_{m=-\infty}^{\infty} y^2(m) \cdot w(n-m),$$

where E_n is the STE for a window of speech $w(n)$. Specifically, $w(n)$ is a sliding Hamming window of length N, where the speech samples closer to n are weighted more than those at the window edge. For the codecs we consider, a packet contains at least 10 ms of audio represented by 80 samples of speech. By making our

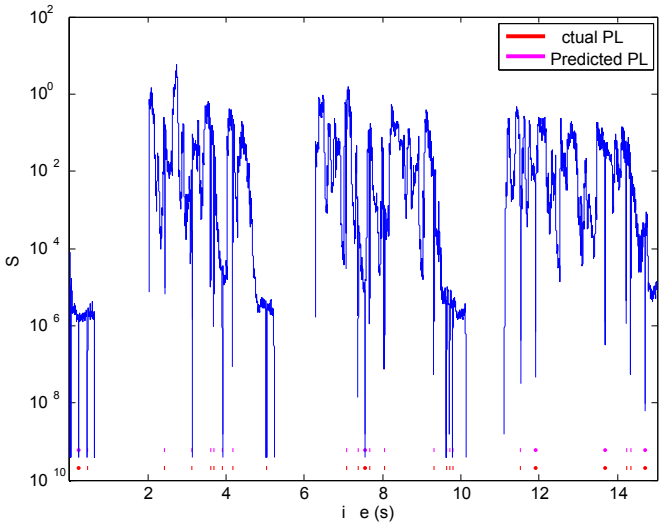

Figure 3: Packet Loss Prediction. The dots below show the actual losses and the ones above are identified by our algorithm. The close correspondence between the two indicates that we detect lost packets accurately.

window length less than 80, multiple values of E_n are completely influenced by a dropped packet. This results in the breaks in energy shown in Figure 2. We detect packet loss by looking for a significant drop in energy followed by an energy floor, accompanied by a significant energy rise.

We note that the presence of all three of these characteristics is necessary to detect packet loss as each appears individually even in speech that has not experienced any packet loss. For instance, in Figure 2, we see a significant rise in energy at approximately 2 seconds due to the start of a speech segment. This is a result of Voice Activity Detection (VAD) in VoIP systems where packets are only sent during active speech to reduce bandwidth. Similarly, when a speech segment ends there is a significant drop in energy. Figure 3 shows the STE of a 15 second speech sample, encoded with G.711 and transmitted through a network with 5% packet loss. The dots at the bottom are the actual packet losses and the ones above are the packet losses identified by our detection mechanism. The close correspondence between the two shows that our detection mechanism identifies packet loss accurately.

Figure 4 shows false positive and false negative cases for our detection mechanism. In the top graph, a packet loss occurs at the start of a speech segment (7 seconds). Since we classify packet losses based on an energy drop, floor and rise, such losses are not detected. Note that this conservative approach reduces our false negatives at the cost of potentially missing a small number of losses at the beginning and end of speech. False negatives are shown in the bottom graph in Figure 4 at 3.2 seconds and occur in the rare case when speech stops and starts in quick succession, with the stop duration corresponding to a multiple of 80. This pattern occurs only when there is a voiced "plosive," or a stop sound in speech, such as the *b* sound in the word "*about*."

Each time a packet loss is detected, the length of the energy floor also reveals the codec used in a particular VoIP network. Figure 5 shows the effect of packet loss on two VoIP networks using different codecs: iLBC which encodes 30ms and Speex which encodes 20 ms of speech per packet. The length of the energy floor is larger for iLBC than Speex. In addition, since G.729 encodes 10 ms and G.711 encoded 20 ms per packet by default, the length of the energy

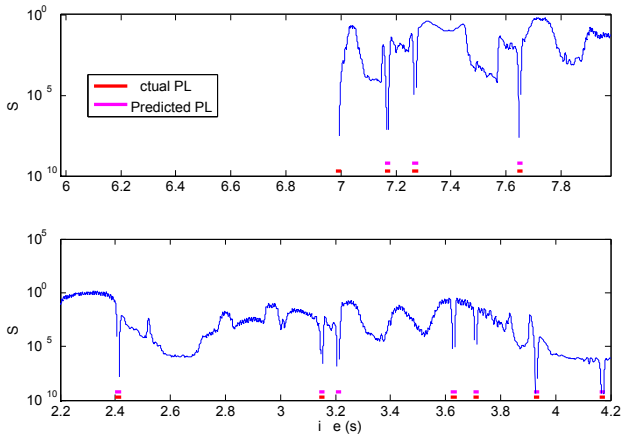

Figure 4: Scenarios showing a false negative (top at 7 seconds) and a false positive (bottom at 3.2 seconds).

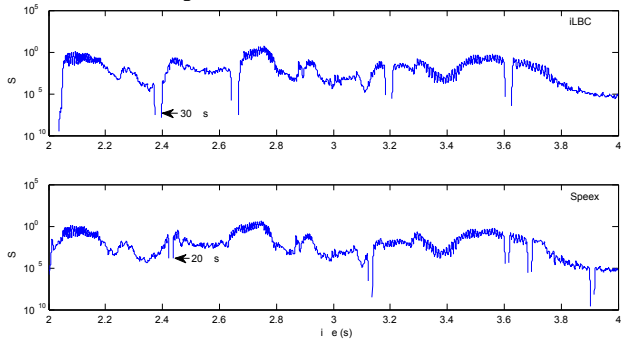

Figure 5: Packet loss affect codecs differently. iLBC encodes 30 ms of audio per packet and therefore a packet loss results in more audio lost in comparison to Speex which encodes 20 ms of audio.

floor is a good indication of the codec used. We might identify the wrong codec when consecutive packets are dropped as two consecutive packets dropped in a network using G.729 (10 ms audio) will be similar to a single packet dropped in a network using G.711 (20 ms audio). However, the probability of consecutive packets being dropped is lower that the probability of a single dropped packet and we can identify the codec based on the most commonly occurring energy floor length.

To summarize, short time energy provides a highly accurate mechanism to determine packet losses and the detection mechanism can also be used to identify the codec used. Therefore, when a call traverses a potentially lossy VoIP network, the packet loss rate and the codec used in that network can be extracted from the received audio.

3.1.2 Detecting Concealed Packet Loss

Some VoIP systems employ packet loss concealment (PLC) algorithms to prevent short speech gaps from affecting call quality. Such concealment can be carried out at the receiver (reactive) or with the assistance of the sender (proactive). In reactive recovery, the lost packet is concealed either with silence, noise or is regenerated by interpolating previously received packets. Proactive recovery algorithms include redundant information such as the previous packet's audio with each packet. This approach incurs a bandwidth overhead and is rarely used. We focus on identifying the effects

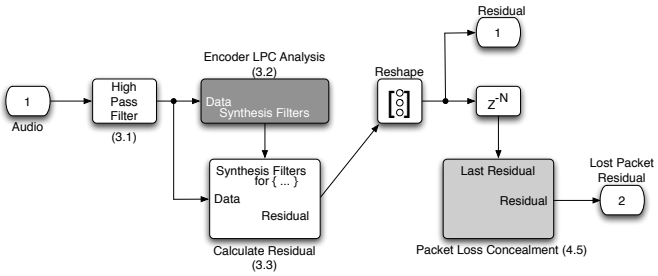

Figure 6: The iLBC packet loss concealment detection algorithm. Because lost packets are regenerated in a largely deterministic fashion from the residual and synthesis filters of the previous packet, such packets can be detected by measuring the correlation between the residuals of sequential packets.

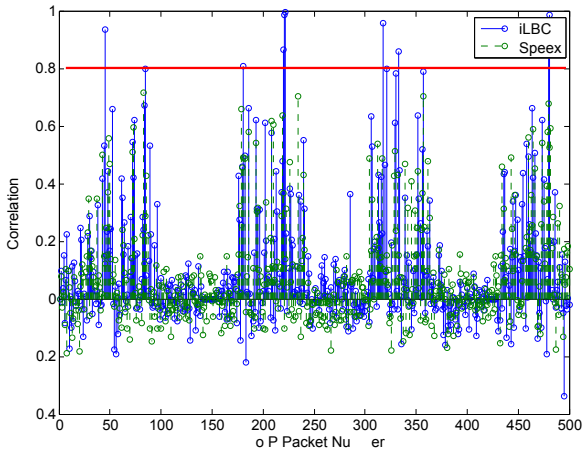

Figure 7: The result of testing for the presence of highly correlated in-sequence packets based on the iLBC packet loss concealment algorithm. The algorithm specifically detects iLBC (solid blue lines) while remaining agnostic to other codecs such as Speex (dotted green lines)

of receiver side recovery algorithms on the audio and leave sender side algorithms to future work.

When the concealment mechanism is silence or noise substitution, the STE-based algorithm from the previous section can be used to detect packet losses by suitably adjusting the energy floor to correspond to the noise floor. Most VoIP codecs, however, reconstruct lost packets from previous packets. G.711 uses waveform substitution to repeat a portion of the previous packet [49]. In codecs designed specifically for VoIP such as iLBC or Speex, the concealment algorithm is more elaborate in order to improve robustness to high packet loss rates. Fortunately, we observe that concealment techniques are predominantly deterministic and a detection mechanism can be created that exploits the correlation between reconstructed packets and previous packets. We discuss the details of the PLC algorithm in iLBC to provide further clarity.

iLBC uses a linear predictive coding (LPC) algorithm to represent speech in a significantly compressed form. LPC is based on the source filter model of speech production, where the larynx (source) produces sound energy, which when voiced consists of a fundamental frequency (pitch) and its harmonics. This sound energy is then shaped (synthesis filters) by the vocal tract (throat and mouth) into enhanced frequency bands known as formants, which provide speech its intonation. The LPC algorithm inverse-filters the formants from the speech signal to leave behind the original sound energy, known as the residual. A codec like iLBC uses the residual, the synthesis filters and dynamic codebook encoding to reduce the original speech into a set of parameters which can be transmitted. The decoder uses these parameters to reconstruct the residual and the synthesis filters which when combined re-synthesize the speech. When a packet is lost, the decoder uses the residual from the previous packet and creates a new pitch synchronous residual for the packet to be concealed. Additionally, a random excitation is added to the new residual (non-deterministic part). The new residual along with the synthesis filters from the previous packet are used to create speech that will be substituted for the lost packet. Therefore the new residual will be strongly correlated to the previous packet's residual. To detect PLC in iLBC we first split the received audio into packets containing 30 ms audio each (the default for iLBC's). We then create a pitch synchronous residual from each packet and compare it to the residual extracted from the next packet. As these quantities are generally not highly correlated, the detection of an association between sequential packets is a very strong indicator of iLBC's packet loss concealment algorithm. The packet loss concealment algorithms for the other codecs, though different, can be detected based on how sequential packets are correlated.

Figure 6 shows a detailed block diagram for the iLBC PLC detection algorithm. Since the encoding procedure in iLBC already extracts the residual from the audio, we first split the audio into 30 ms chunks and apply the encoding steps defined in Section 3.1 to 3.3 of iLBC RFC 2951 [20]. This includes running a high pass filter to remove noise in the audio, performing LPC analysis to extract the synthesis filters and then using the synthesis filters along with the data to extract the residual, r. We use r to generate a pitch synchronous residual r' as defined in Section 4.5 of iLBC RFC 2951. r' will be strongly correlated to the residual from the next chunk of 30 ms of audio if that packet had been lost. We calculate r and r' for each chunk and report high correlations between as indications of PLC.

Figure 7 shows the correlation between residuals of a 15 second speech sample encoded with the iLBC codec (solid blue lines) and transmitted through a VoIP network with a loss rate of 10%. At each high correlation point (above 0.8) we confirm from our logs that the particular packet was lost. To show that the PLC detection algorithm is specific to iLBC, we run it on the same 15 second speech sample encoded with Speex instead and transmitted through the 10% loss rate VoIP network. The results are again shown in Figure 7 as the dashed green lines. Though packets were lost in this case too, the detection algorithm does not show any high correlation between residuals, confirming that we can create PLC detection algorithms specific to the way each codec conceals packets. Since all the codecs use different concealment strategies, in addition to detecting concealed packet losses our algorithms also provide a strong indication of the codec used in a particular VoIP network.

Finally, in Figure 7 we observe that for the 15 second sample encoded with iLBC, 54 out of the 501 packets (loss rate $= \frac{52}{501} = 10.38\%$) were lost and we are only able to identify 9 correlations. This is largely due to the fact that the PLC algorithm is not completely deterministic (random excitation). However, the number of concealed packets detected is still indicative of the loss rate. To show this, we ran our detection algorithm over 15 seconds of 20 male and female American English speech samples from the

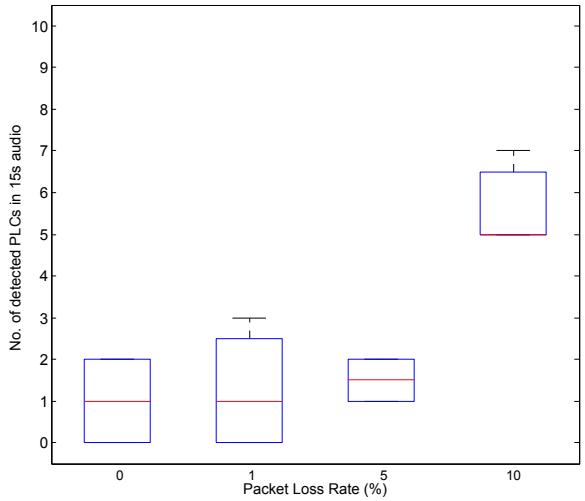

Figure 8: Number of concealed packets detected with increasing loss rate in a 15s speech sample. The median number of concealed packets detected by our algorithm increases with increasing loss rate.

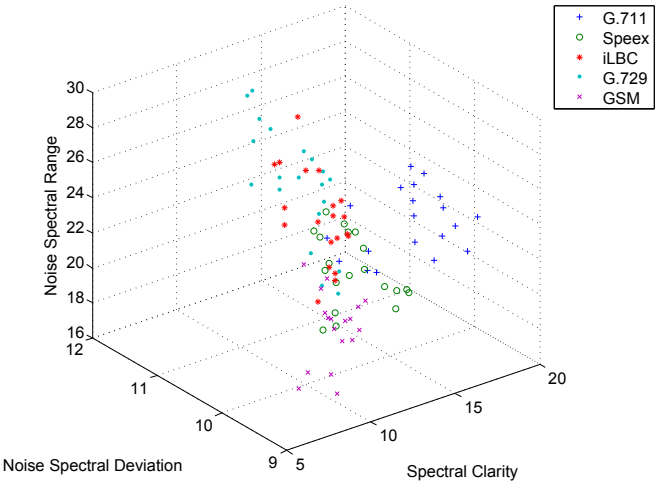

Figure 9: The noise profile of G.711 is significantly different from other codecs, allowing us to identify it when it is used in a network.

Open Speech Repository [55] encoded with iLBC and transmitted through VoIP networks with 0, 1, 5 and 10% loss rates. The association between the number of concealed packets detected and the packet loss rate are shown in Figure 8. Figure 8 shows the median and the 25th and 75th percentiles with whiskers specified as .5 times the interquartile range. We see that the median number of concealed packets increases significantly as the loss rates increase. Therefore, the PLC detection algorithm can make approximations of the loss rate but is not as accurate as the detection algorithm for unconcealed packet losses.

Our packet loss and packet loss concealment detection algorithms identify three aspects about the provenance of a call: (1) Whether the call traversed a VoIP network, (2) the packet loss rate in that network and (3) the codec used in that network. (1) identifies if there are VoIP networks in the path of a call and (2) and (3) characterize the VoIP network.

3.2 Identifying PSTN and Cellular Networks Through Noise Profiling

Now that we are able to identify and characterize VoIP networks, we can look for codec specific artifacts in the received audio to identify PSTN and cellular networks.

Waveform codecs like G.711 are used mostly in PSTN networks as they capture speech without any compression and require much higher bandwidth (64 kbps) than most other codecs. They tend to introduce noise only during speech activity resulting in a strong correlation between the noise and the signal. This is known as multiplicative noise and its presence can be determined based on spectral statistic metrics: spectral level range and the spectral level deviation. Furthermore, the spectral clarity for such a codec, or the measured crispness of the audio, is very high. In contrast, since cellular networks require efficient use of bandwidth they use high compression codecs like GSM-FR (13 kbps). The spectral clarity of such codecs suffer due to the significant compression. Spectral clarity quantifies the perceptible difference in call quality that we experience when talking on a landline versus a mobile phone. Figure 9 shows the spectral clarity, the spectral level range and deviation for 20 male and female American English speech samples from the Open Speech Repository [55] encoded and decoded using the different codecs. We see that G.711 and GSM-FR can be clearly identified. Once we identify the codec using these metrics we can do a simple codec to network translation to determine if a call has traversed a PSTN network or has originated from a cellular network. Furthermore these three metrics provide a noise profile of the network thereby characterizing it.

3.3 Extracting Provenance Data

We have seen how packet loss and packet loss concealment detection identifies and characterizes any traversed VoIP network. Similarly, the noise profiles identify and characterize any PSTN and cellular network. Together, we can create fingerprints that detail the provenance of a call.

Call provenance fingerprints consist of two parts: (1) the path traversal signature and (2) detailed characterization of each network in the path traversal signature. The path traversal signature identifies the networks that a call traversed and the codec used. The characterization provides more details of each network. The features we extract can be used towards both these parts as shown in Figure 10. To obtain the path traversal signature we first train a multi-label classifier as shown in Figure 10 using a repository of speech samples. Each sample is subjected to codec transformations and network degradations depending on the networks it traverses (details in Section 4). For each of the resulting audio samples, we first look for packet losses. If present, we calculate the packet loss rate which forms the packet loss profile and then add the extracted codec information and the rate (as G.711 with some loss rate indicates a VoIP network) to the feature vector. Next, we apply the correlation algorithm to detect packet loss concealment. If the correlation algorithm finds concealed losses, the corresponding codec is again added to the feature vector along with the number of concealed packets (PLC profile). We then extract the noise profile for the call audio and add the spectral metrics to the feature vector. Since the quality of speech degrades with the number of networks traversed we also obtain call quality metrics from a single ended quality tool, P.563 [50] and add this to the feature vector. The multi-label classifier is then trained on each sample's feature vector and label. A sample has five labels, each indicating the presence or absence of a codec. For example a speech sample in our reposi-

Configuration	Scenario	# Simulated Samples
Single Network Traversal		
PSTN - PSTN	Plain old telephone call	20
Mobile - Mobile	Short distance call b/w cell phones	20
VoIP - VoIP	Unfederated call b/w VoIP clients e.g., Google Talk	60
Two Network Traversal		
PSTN - Mobile	Call b/w PSTN landline and cell phone	320
PSTN - VoIP	Call b/w PSTN landline and VoIP client e.g., SkypeOut	360
Mobile - VoIP	Call b/w cell phone and VoIP client	560
Three Network Traversal		
PSTN - VoIP - Mobile	International call using calling cards	1200
PSTN - VoIP - PSTN	Same as above	240
Mobile - VoIP - Mobile	VoIP call bridging b/w two mobile phones e.g., Google Talk	960
Mobile - PSTN - VoIP	Call b/w mobile using a PSTN core network and a VoIP client	400
Mobile - PSTN - Mobile	Similar as above	80
VoIP - PSTN - VoIP	Call b/w two commercial VoIP clients e.g., typical Vonage call	720
		Total = 4940

Table 2: Call Traversal Scenarios. Each of the traversal configurations is representative of a real world scenario.

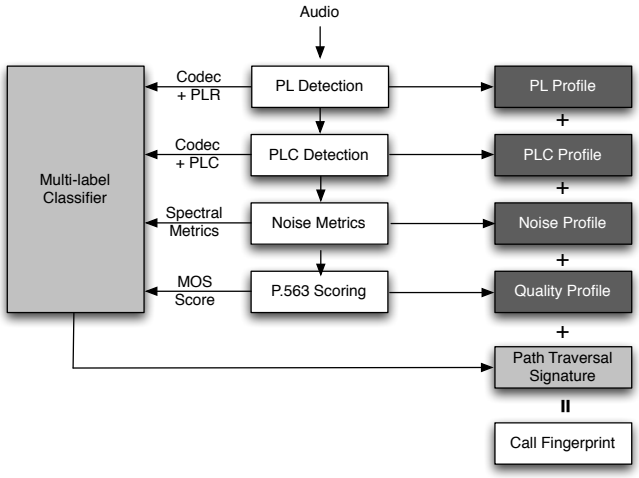

Figure 10: The PinDr0p call provenance extraction algorithm. After the applied codecs have been detected, packet loss rates are compared against individual source profiles. The resulting signature can be used to judge the provenance of an incoming call.

tory that was encoded using GSM-FR (originated at a cellphone), then re-encoded using iLBC (traversed a VoIP network) and finally re-encoded using G.711 (receiving end point is a landline) would have a '1' for three labels (GSM-FR, iLBC and G.711) and a '0' for Speex and G.729. Multi-label classifiers have been used significantly in text categorization [52, 33, 60] and we use a set of standard reduction techniques to convert the multi-label data into a single label model. The classifier then learns which features best predict the presence or absence of a label.

For any new call audio we perform the same procedure, but do not add any label as the classifier will predict a set of labels based on the learned model. The prediction of the classifier for the path traversal signature, along with packet loss, noise and quality profiles, represents the call provenance fingerprint for a particular source in PinDr0p.

3.4 Security Implications

The path traversal signature and the complete provenance fingerprint provide a useful security framework in the absence of any verifiable metadata. The traversal signature alone can be used against adversaries who are bound by operating constraints. For example, adversaries trying to spoof a dedicated line to the bank might use VoIP due to the fact that they can remain largely anonymous and can make a large number of inexpensive calls. However, the path traversal signatures for these two calls will be different. To address this, the adversary can switch to a landline, in which case he has lost the ability to easily make a large number of calls and potentially compromised his anonymity.

We can also use the complete provenance fingerprint against adversaries as it also characterizes individual networks. Since this involves capturing detailed profiles of these networks traversed, an adversary trying to spoof a call needs to be able to match all these profiles. We show in Section 5 that our fingerprints are able to discriminate between sources that are in the same city using the same provider, demonstrating that matching an entire fingerprint is extremely difficult. Accordingly, we believe our approach is a significant first step in creating suitable defenses against a host of attacks possible in today's diverse telephony infrastructure.

4. EVALUATION

We evaluate our approach based on two metrics: (1) the accuracy of our multi-label classifier in predicting the correct network traversal signature of a call and (2) the ability of our provenance fingerprint to consistently identify a call source. We discuss the evaluation of the first in this section and analyze the second in the following section.

4.1 Experimental Setup

We train and test the multi-label classifier against a repository of speech samples that are subjected to a representative set of real-world call traversal scenarios and network degradations. We assume calls can traverse one, two or three networks as most call scenarios fall into one of these cases; however, our methodology can be extended to deal with additional transcoding. Table 2 shows the considered call traversal configurations. Single network traversals represent calls that are contained within one system. For example, the VoIP-VoIP scenario occurs when two Skype users call each other. Since both clients are connected to the Internet, they communicate through a set of relays (supernodes) and the call stays completely within the IP network. Two network traversals are calls from users on one telephony technology to users on another. There are six possible combinations and for brevity we only list three of them, in each case subsuming the symmetric traversal scenario (i.e., PSTN-Mobile and Mobile-PSTN are categorized as a single

scenario). Finally, three network traversals occur when providers attempt to take advantage of the benefits offered by each telephony technology. For instance, while calls between two Vonage clients within the US can be completely VoIP-VoIP, Vonage specifically transmits the call over the PSTN backbone due to its QoS guarantees. Similarly, most international calling card services use VoIP across the Internet as this provides an inexpensive calling alternative.

Our experiments use speech samples from the Open Speech Repository [55], which contains samples of 20 different American English speakers, 10 male and 10 female, speaking phrases from the Harvard sentence list [3]. These samples are used for standardized testing of PSTN, VoIP and cellular systems as recommended by the IEEE Recommended Practices for Speech Quality Measurements [4]. Each sample is 40 seconds long, but we consider only the first 15 seconds, as call quality algorithms such as P.563 typically use this length to determine call quality metrics.

We consider the most popular narrowband codecs for encoding calls in our experiments. Specifically, we use G.711 for PSTN systems, G.711, G.729, iLBC and Speex for VoIP systems, and GSM for cellular systems. Calls traversing two telephony networks (e.g., VoIP to cellular) are transcoded to the new codec.[2] Since transcoding is not always defined for a pair of codecs, we follow the common practice of converting to and from an intermediate G.711 form. We use the PJSIP [38] suite of applications to encode and perform the necessary conversions between codecs. PJSIP contains open source SIP and media stacks and is part of the European Broadcasting Union Audio over IP standard [17]. It supports G.711, iLBC, Speex and GSM. For G.729, we integrate the Intel Integrated Performance Primitives Library [25] into PJSIP.

In addition to the codecs, each traversed network is characterized by its signal degradation characteristics. VoIP networks experience packet losses which typically increase in correlation with factors such as routing distances, "last-mile" unreliability, network congestion and over-subscription. For VoIP networks, we simulate packet loss rates of 1, 5 and 10%. For bit errors occurring from multi-path fading radio channels in mobile networks, we use a GSM traffic channel simulator developed for Simulink [32].

Experiments are conducted by taking one speech sample from the Open Speech Repository and encoding it with the appropriate codec using PJSIP. Samples corresponding to packet losses or signal degradations found in the traversed telephony network are also generated and tested (e.g., packet loss in iLBC, multi-path fading in GSM). We also append the codec multi-label for each generated sample. We aggregate all possible resulting speech samples into a corpus. The number of samples for each of the traversal scenarios is shown in Table 2.

We run the feature extraction algorithms described in Section 3.3 on each of the speech samples and then train and test a multi-label classifier on the resulting feature vector and label. We use Mulan [51], an open source Java library for multi-label learning, to create our machine learning classifier.

4.2 Classification Results

Multi-label classifiers can use a variety of reduction techniques including Binary Relevance (BR), Label Power (LP) set and Random k-Labelsets (RAkEL) [53] to convert the multi-label into a single label. The resulting labels can then be classified by any of the traditional single-label classifiers. We use C4.5 decision trees as the underlying single-label classifier as it outperforms other classifiers that we considered including Naive Bayes and Neural Networks.

[2]Recall that VoIP calls can cross multiple autonomous systems throughout the Internet without being transcoded.

Metric	Definition	BR	LP	RAkEL
Hamming Loss	$\frac{1}{\|D\|} \cdot \sum_{i=1}^{\|D\|} \frac{\|Y_i \triangle P_i\|}{\|L\|}$	.09	.1	.05
Accuracy	$\frac{1}{\|D\|} \cdot \sum_{i=1}^{\|D\|} \frac{\|Y_i \cap P_i\|}{\|Y_i \cup P_i\|}$	83.7%	83.7%	91.6%
Precision	$\frac{1}{\|D\|} \cdot \sum_{i=1}^{\|D\|} \frac{\|Y_i \cap P_i\|}{\|P_i\|}$	91.5%	89.3%	93.7%
Recall	$\frac{1}{\|D\|} \cdot \sum_{i=1}^{\|D\|} \frac{\|Y_i \cap P_i\|}{\|Y_i\|}$	90.3%	89.3%	97%

Table 3: Accuracy of multi-label classifier using C 4.5 decision trees. RAkEL outperforms the simpler binary relevance and label power set reduction techniques.

Using the corpus described above, we use 10-fold cross validation to measure the accuracy of the multi-label classifier under the three reduction techniques. Our results are described in Table 3. We define the metrics as specified in the multi-label classification literature [52]. Let the multi-label dataset consist of $|L|$ labels (five in our case) and $|D|$ instances in the test set, with each instance i represented by feature vector f_i and label Y_i. The classifier C makes label predictions $P_i = C(f_i)$ for each instance f_i. For a test instance with known path traversal signature, the classifier predicts a label using only the feature vector. The metrics defined help quantify the difference between the predicted and actual labels.

We find that RAkEL has the lowest Hamming loss and the highest accuracy of 91.6%. The results show that we are able to predict which networks a call traversed with high accuracy. We also find that the majority of misclassifications occur for samples that traversed a VoIP network with 0% packet loss rate. We plan to study how to distinguish VoIP, PSTN and cellular codecs when there are virtually no degradations as part of future work.

5. REAL-WORLD TESTING

The complete provenance fingerprint of a call consists of the path traversal signature, and profiles for packet loss, concealment, noise and quality. If this fingerprint remains consistent for a call source, it provides valuable metadata that can be used to identify and distinguish different calls purely from the received audio. We asked different users to make a set of 10 live calls to our testbed in Atlanta, GA from 16 different locations around the world, including Australia, India, United Arab Emirates, United Kingdom and France. The complete list of locations is shown in Figure 11.

Each call lasts approximately 20 seconds. We extract features and profiles from the received audio and then label all calls from a call source with the same unique label. We then train a neural network classifier for N sets of the 10 call sets (set = one call from each source). We vary N from one to five and then test with five new call sets. This represents the scenario that a user labels a set of calls and expects subsequent calls coming from the same source to be labeled correctly by our algorithm. Our experiment evaluates the tradeoff between labeling effort and accuracy.

The results show that even if a single set of 16 calls is labeled, the remaining five sets of calls from the 16 different locations are identified with the correct call source label with 90% accuracy. The accuracy increases quickly to 96.25% for two, 97.5% for three, 97.5% for four and 100% for five labeled sets. Figure 12 shows the confusion matrix for 1, 3 and 5 training sets.

Even with a singly labeled training set (Fig 12(a)) we find that all VoIP calls are correctly identified as they are easily distinguishable from the other networks. They are also distinguishable among themselves as they are geographically spread across Atlanta, Maryland, Dubai, Pune and Australia and each has a different packet loss concealment profile. In some cases we were pleasantly surprised by

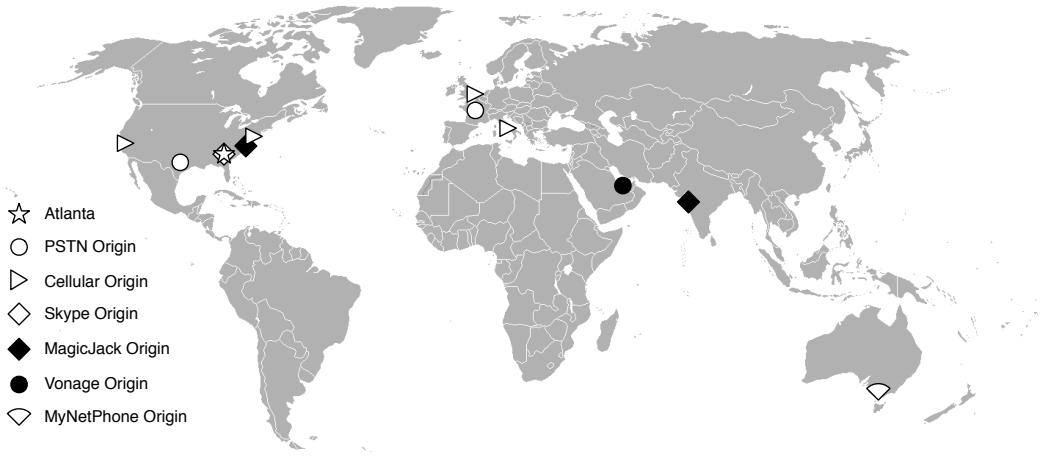

Figure 11: We tested our system using multiple sources from four continents: North America, Europe, Asia and Australia. Specifically, we recorded incoming calls from five different PSTN phones in Atlanta, GA, Dallas, TX, and France; four different mobile phones in Atlanta, GA, New York City, NY, San Jose, CA and London, UK; six VoIP phones in Atlanta, GA (Skype and Vonage), Baltimore, MD(MajicJack), Pune, India(MagicJack), Dubai, UAE(Vonage) and Melbourne, Australia (MyNetPhone).

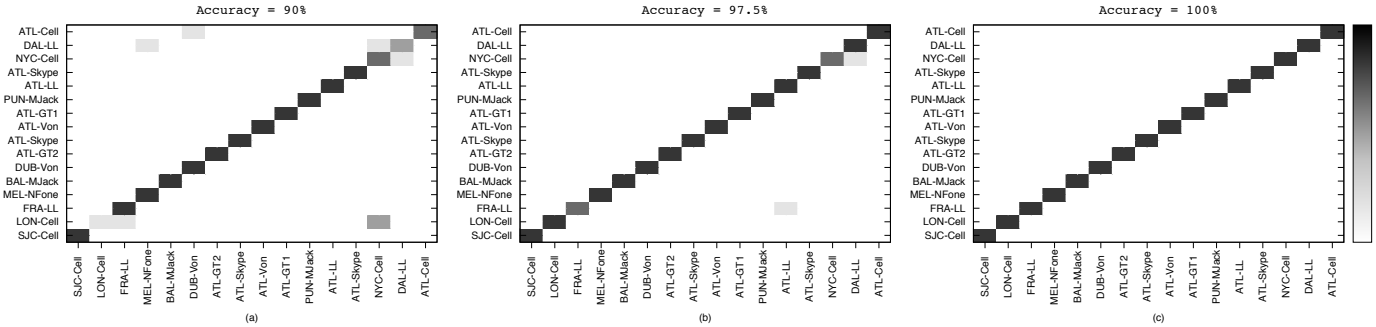

Figure 12: The confusion matrix for the live-captured call data trained with labels for (a) one set of calls, (b) three sets of calls and (c) five sets of calls from all call sources. The accuracy on even a singly labeled training set is 90% and quickly jumps to 100% with 5 labeled training sets.

the actual differentiator. We found that Vonage calls from Atlanta were distinguishable from all the other VoIP calls based on its high spectral level range (noise profile) rather than the packet loss profile. We suspect that this is due to the fact that Vonage calls almost immediately transfer to the PSTN backbone for quality of service, while other services predominantly use VoIP. However, we did not observe this in the international Vonage call from Dubai where the call path would be predominantly VoIP, instead, to make the call affordable.

Figure 12(a) shows that even with a singly labeled training set we are able to distinguish between the three landlines from Atlanta, including the two from within the Georgia Tech campus, demonstrating that even for similar call sources the characteristics can be significantly different. We also see that three of the five calls from the London mobile phone are misclassified as a mobile phone call from New York and one call was misclassified as a landline call from France. The provenance of the call from London seems to be misclassified based on either the distance similarity (both coming from Europe) or the same origin network (cell). The number of misclassifications for the test set containing 80 calls ($16 locations \times 5$) drops significantly from 10 to 3 and then to 2 with increasing the number of training sets. With five labeled call sets being trained we have no misclassification showing that with each extra label the classifier becomes increasingly accurate.

The profiles that we capture for each source are consistent for the same call source but have enough variability to allow us to distinguish different call sources. Although we still require 15 seconds of call audio before being able to identify the provenance, we believe that an attempt to steal sensitive information (e.g., bank account numbers) from a potential victim requires significantly more time. Accordingly, users should be sure to wait at least this amount of time before disclosing such information. We plan to investigate the uniqueness of a larger number of call sources as part of future work.

6. DISCUSSION

In this section, we investigate some of the limitations of our current infrastructure and discuss a number of future extensions that will both improve the accuracy of our detection and its resistance to more active adversaries.

6.1 Limitations

Our call provenance infrastructure is designed to detect codecs and path characteristics associated with a given source. In spite of its relative strength, there exist a number of limitations associated

with our current system. For instance, unlike Caller-ID systems, our call provenance infrastructure requires that the receiver answer the call before its source can be verified. This may not be useful to those using Caller-ID as a means of deciding whether or not to take a call. This shortcoming could potentially be addressed by pushing our mechanism into the cloud. Incoming calls could potentially be forced to first interact with a recording, which could collect sufficient audio for analysis, before reaching the intended target. We leave the details of such a system to future work.

We currently rely heavily on packet loss characteristics of the path between sources and our testbed to differentiate VoIP fingerprints. While instantaneous packet loss rates certainly fluctuate, paths and their corresponding loss patterns are relatively stable in the Internet [37]. However, we recognize that our packet loss profiles may need to be more accepting of diurnal cycles and temporary anomalies and plan to study such issues as we continue to extend the PinDr0p infrastructure.

As an implementation decision, we currently associate a source with a single fingerprint. This assumption is appropriate when dealing with an immobile source such as a corporate calling center. However, individual users may take advantage of the mobility allowed by VoIP software such as Skype to legitimately place calls from a number of different locations. The advantage in such a scenario is that the receiver is likely to recognize the caller's voice and can therefore manually associate new fingerprints to a particular source.

Lastly, we have attempted to analyze the most widely used codecs in our study. However, other less widely use codecs were not considered in this initial study. For instance, the Adaptive Multi-Rate (AMR) codec, which provides higher audio quality and is beginning to compete with GSM on mobile devices, and a handful of others such as the Enhanced Variable Rate Codec (EVRC) for CDMA networks will be considered as part of our future work.

6.2 Proposed Extensions and Future Work

We intend to continue improving the robustness of PinDr0p through a number of extensions. While admittedly difficult, an adversary capable of replicating all of the codecs and path characteristics associated with the path between a legitimate source and target receiver would potentially be able to be identified as the profiled source. This process not only implies that the adversary has correctly guessed all of the codecs applied by intermediary hops, but that they can ensure that their traffic exhibits similar packet loss, bit error and noise characteristics as a legitimate connection. *This is exceptionally difficult as an adversary, for example, can not decrease the packet loss characteristics of an intermediary network that they do not control.* Our approach therefore represents a significant improvement over the current state of the art.

While we currently detect the presence of as many as three different codecs applied to audio, our mechanisms do not uncover the order in which the codecs were applied. Determining codec order is an extremely difficult problem on the surface. Knowledge of this ordering will make spoofing attempts by an adversary located off the path more difficult.

Finally, we are interested in extending our analysis to include a larger number of intermediary networks. While highly uncommon, it is possible that some international calls may be transcoded by as many as five different codecs while in flight between their source and destination. The repeated decoding and encoding of audio information drastically reduces its quality at the receiver end of the call and may also obscure the presence of the intermediary networks given the elevated noise levels present in the sample.

6.3 Additional Applications

We have focused the work in this paper on using call provenance to address Caller-ID spoofing attacks. However, the utility of PinDr0p is not limited to this task. While stories of VoIP-based phishing (vishing) have become popular in the media [57, 45], the extent to which such calls are occurring compared to traditional telephony fraud is unknown. The deployment of our infrastructure in a distributed fashion may help to answer this question. In particular, the use of call provenance in this space can assist in determining the prevalence and potentially the identity of individual vishing campaigns. While we leave the details of such an infrastructure to future work, we hope to be able to provide the security community with a tool for better understanding such attacks.

PinDr0p may also be useful as a means of authenticating channels. For instance, credit card and home security companies often use Caller-ID information as a second factor of authentication when customers call with account questions. Such organizations could increase the number and difficulty of questions asked of the caller based on the measured provenance of the incoming call. In multi-factor authentication analysis, PinDr0p can be used to determine if information exchange through a website and a phone call are truly independent. Finally, PinDr0p could also be used by law enforcement agencies for call forensics.

7. RELATED WORK

The concept of data provenance in computing was first studied in database systems. These techniques seek to identify the source of a piece of data and the process by which it arrived at the database [13, 21, 11]. Such information can be proactively added at the source and transformation points as metadata [18, 15] or reactively obtained through techniques such as query inversion [58, 14]. Such techniques have been adapted and extended to other platforms including web servers with trusted hardware [35]. The presence of such mechanisms provides a significantly improved infrastructure for performing audits and determining data quality [34].

More recently, a number of researchers have attempted to provide provenance information for networks. Traceback techniques [41, **?**, 23] attempt to determine the true path of packets in the presence of potentially spoofed source information. Such information can either be added directly to the packets as metadata [41, 44, 59, 39], or by state stored and queried from within the routers themselves [61]. A range of watermarking tools also exist to identify the provenance of flows in IP networks [56, 28, 24]. The diversity of telephony networks (i.e., circuit switched PSTN, cellular and VoIP) makes such watermarks extremely difficult. Specifically, metadata introduced in one network (e.g., watermarks, path information) is generally lost when the call is transmitted over another network.

We are not aware of previous work that attempts to identify the provenance of a phone call in a diverse telephony environment. However, techniques in a purely Internet-based environment have been considered [46]. Perhaps the closest to our work are caller identification (Caller-ID) services that provide the caller's number or name in PSTN and mobile networks. Calls originating from IP networks traditionally have no unique associated number or name and therefore cannot be used to identify the caller [43]. Moreover, a variety of techniques already exist to spoof phone numbers [2]. Artifacts of calls themselves may provide significant provenance information. Specifically, because call quality relies greatly on a combination of the codec [31, 5], the range of end devices [12] and network degradations [16, 40, 29], the detection of these characteristics using tools designed to measure single-ended call quality [16,

31, 40] can potentially be used to further improve the provenance of a call.

8. CONCLUSION

Caller-ID has long been viewed as a reliable means of identifying the source of a call. However, this mechanism is now easily spoofable through a variety of free and low-cost techniques. In this paper, we take a first step towards a mechanism capable of determining call provenance — the source and the path taken by a call. We leverage attributes of the audio delivered to the receiver, including characteristics of the applied codes, packet loss profiles and bit error rates. We use these measurable elements to identify the codecs applied to incoming calls passing through as many as three intermediary types of telephony networks with a 91.6% accuracy. Moreover, fingerprints for specific sources were identified with between 90% and 100% accuracy with one and five training sets, respectively. Through additional device-specific fingerprinting mechanisms and distance estimation techniques, we believe that our mechanisms can be further improved and made more robust to attack.

Acknowledgments

We would like to thank Machigar Ongtang, William Froning, Daniel Sylvester, Patrick Greives, Karishma Babu, Maryam Poonawalla, Jeevan Poonawalla, Behlul Poonawalla, Insiya Poonawalla, Qusai Poonawalla, Arjun Maheshwaran, Viswanathan Mahalingam, Devdutt Patnaik, Jonathan Li, Pooja Karia, Frank Park, Karthik Balasubramaniyan, Nivedhya Ramaswamy, Davide Ariu, Krishnan Shankar Narayan, Shirpaa Manoharan, LVS Gopiraman, Mandar Harshe, Kevin Stumph and Naveen Tamilmani for assisting us in placing phone calls to our testbed. We would also like to thank Kevin Butler for his comments. This work was supported in part by the US National Science Foundation (CNS-0916047). Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation.

9. REFERENCES

[1] Caller ID FAQ v2.32 1st April 2004. http://www.ainslie.org.uk/callerid/cli_faq.htm.

[2] The Definitive Resource on Caller ID Spoofing. http://www.calleridspoofing.info/.

[3] The Harvard Sentences. http://www.cs.columbia.edu/~hgs/audio/harvard.html.

[4] IEEE Recommnded Pratice for Speech Quality Measurements. In *IEEE Transactions on Audio and Electroacoustics*, volume 17, 1969.

[5] The E-model, a Computational Model for Use in Transmission Planning. Technical Report ITU-T G. 107, ITU-T, February 2003.

[6] The Speex Codec. http://www.speex.org/, 2003.

[7] Xbox LIVE. http://www.xbox.com/en-US/LIVE/, 2005.

[8] Presentation on Q1 2009 Earning Report of Ebay Inc.k. http://www.slideshare.net/earningreport/presentation-on-q1-2009-earning-report-of-ebay-inc, 2009.

[9] IP Phone – Cisco. http://www.cisco.com/en/US/products/hw/phones/ps379/index.html, 2010.

[10] Skype. http://www.skype.com/, 2010.

[11] O. Benjelloun, A. Das, S. Alon, and H. J. Widom. Uldbs: Databases with Uncertainty and Lineage. In *In VLDB*, pages 953–964, 2006.

[12] S. R. Broom. VoIP Quality Assessment: Taking Account of the Edge-Device. *IEEE Transactions on Audio, Speech & Language Processing*, 14(6):1977–1983, 2006.

[13] P. Buneman, S. Khanna, and W. C. Tan. Why and Where: A Characterization of Data Provenance. In *Proceedings of the International Conference on Database Theory (ICDT)*, 2001.

[14] Y. Cui and J. Widom. Practical Lineage Tracing in Data Warehouses. In *Proceedings of the 16th International Conference on Data Engineering (ICDE)*, Washington, DC, USA, 2000.

[15] S. B. Davidson and J. Freire. Provenance and Scientific Workflows: Challenges and Opportunities. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD)*, 2008.

[16] L. Ding, Z. Lin, A. Radwan, M. S. El-Hennawey, and R. A. Goubran. Non-Intrusive Single-Ended Speech Quality Assessment in VoIP. *Speech Commun.*, 49(6):477–489, 2007.

[17] European Broadcasting Union. Audio Contribution over IP. http://www.ebu-acip.org/.

[18] I. T. Foster, J.-S. Vöckler, M. Wilde, and Y. Zhao. Chimera: AVirtual Data System for Representing, Querying, and Automating Data Derivation. In *Proceedings of the International Conference on Scientific and Statistical Database Management (SSDBM)*, 2002.

[19] J. Franklin, D. McCoy, P. Tabriz, V. Neagoe, J. V. Randwyk, and D. Sicker. Passive Data Link Layer 802.11 Wireless Device Driver Fingerprinting. In *Proceedings of the USENIX Security Symposium (SECURITY)*, 2006.

[20] Global IP Solutions. The Internet Low Bitrate Codec (ILBC). http://tools.ietf.org/html/rfc3951, 2004.

[21] P. Groth, L. Moreau, and M. Luck. Formalising a Protocol for Recording Provenance in Grids. In *Proceedings of the UK OST e-Science Third All Hands Meeting 2004 (AHM'04)*, 2004.

[22] GSM. GSM-FR: GSM Full Rate (GSM 06.10). http://www.3gpp.org/FTP/Specs/html-info/0610.htm, 1995.

[23] I. Hamadeh and G. Kesidis. A Taxonomy of Internet Traceback. *International Journal of Secururity and Networks*, 1(1/2):54–61, 2006.

[24] A. Houmansadr, N. Kiyavash, and N. Borisov. RAINBOW: A Robust And Invisible Non-Blind Watermark for Network Flows. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2009.

[25] Intel. Intel Integrated Performance Primitives Library. http://software.intel.com/en-us/intel-ipp/.

[26] JK Audio - Telephone Audio Interface Products. THAT-1: Telephone Handset Audio Tap. http://www.jkaudio.com/that-1.htm, 2009.

[27] J. Kekäläinen. Binary and graded relevance in ir evaluations: comparison of the effects on ranking of ir systems. *Inf. Process. Manage.*, 41(5):1019–1033, 2005.

[28] N. Kiyavash, A. Houmansadr, and N. Borisov. Multi-flow Attacks Against Network Flow Watermarking Schemes. In *Proceedings of the USENIX Security Symposium (SECURITY)*, 2008.

[29] M. Lee and J. W. McGowan. Method and Apparatus for the Detection of Previous Packet Loss in Non-Packetized Speech. http://www.patentstorm.us/patents/7379864.html, May 2008.

[30] Linda McGlasson. Vishing Scam: Four More States Struck. http://www.bankinfosecurity.com/articles.php?art_id=2138, 2010.

[31] L. Malfait, J. Berger, and M. Kastner. P.563 - The ITU-T Standard for Single-Ended Speech Quality Assessment. *IEEE Transactions on Audio, Speech & Language Processing*, 14(6):1924–1934, 2006.

[32] Mathworks. Simulink - Simulation and Model-Based Design. http://www.mathworks.com/products/simulink/.

[33] A. K. McCallum. Multi-label text classification with a mixture model trained by em. In *AAAI 99 Workshop on Text Learning*, 1999.

[34] S. Miles, P. Groth, M. Branco, and L. Moreau. The Requirements of Recording and Using Provenance in e-Science Experiments. *Journal of Grid Computing*, 5(1), 2007.

[35] T. Moyer, K. Butler, J. Schiffman, P. McDaniel, and T. Jaeger. Scalable Web Content Attestation. In *Proceedings of the Annual Computer Security Applications Conference (ACSAC)*, 2009.

[36] J. Pang, B. Greenstein, R. Gummadi, S. Seshan, and D. Wetherall. 802.11 User Fingerprinting. In *Proceedings of the ACM International Conference on Mobile Computing and Networking (MOBICOM)*, 2006.

[37] V. Paxson. End-to-end routing behavior in the Internet. *ACM SIGCOMM Computer Communication Review*, 36(5):56, 2006.

[38] B. Prijono. PJSIP. http://www.pjsip.org/.

[39] A. Ramachandran, K. Bhandankar, M. B. Tariq, and N. Feamster. Packets with Provenance. http://www.cc.gatech.edu/research/reports/GT-CS-08-02.pdf, May 2008.

[40] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra. Perceptual Evaluation of Speech Quality (PESQ)-A New Method for Speech Quality Assessment of Telephone Networks and Codecs. In *Proceedings of the IEEE Internation Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Washington, DC, USA, 2001.

[41] S. Savage, D. Wetherall, A. Karlin, and T. Anderson. Practical Network Support for IP Traceback. *ACM SIGCOMM Computer Communication Review*, 30(4):295–306, 2000.

[42] M. Sherr, E. Cronin, S. Clark, and M. Blaze. Signaling Vulnerabilities in Wiretapping Systems. *IEEE Security & Privacy Magazine*, 3(6):13–25, November 2005.

[43] J. Slay and M. Simon. Voice over IP Forensics. In *Proceedings of the*

International Conference on Forensic Applications and Techniques in Telecommunications, Information, and Multimedia (e-Forensics), 2008.

[44] A. C. Snoeren, C. Partridge, L. A. Sanchez, C. E. Jones, F. Tchakountio, B. Schwartz, S. T. Kent, and W. T. Strayer. Single-Packet IP Traceback. *IEEE/ACM Transactions on Networking (TON)*, 10(6):721–734, December 2002.

[45] W. Stenehjem. Too Good To Be True: A Column on Consumer Trust Issues by Attorney General Wayne Stenehjem's Consumer Protection and Antitrust Division. `www.ag.state.nd.us/tgtbt/2008/03-05-08.pdf`, 2008.

[46] H. Tae, H. L. Kim, Y. M. Seo, G. Choe, S. L. Min, and C. S. Kim. Caller Identification System in the Internet Environment. In *In Proceedings of the USENIX Security Symposium (SECURITY)*, 1993.

[47] The International Telecommunication Union. G.711: Pulse Code Modulation (PCM) of Voice Frequencies. `http://www.itu.int/rec/T-REC-G.711/e`, 1972.

[48] The International Telecommunication Union. G.729: Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear Prediction. `http://www.itu.int/rec/T-REC-G.729/e`, 1996.

[49] The International Telecommunication Union. G.711 Appendix I. `http://www.itu.int/rec/T-REC-G.711/recommendation.asp?lang=en&parent=T-REC-G.711-199909-I!AppI`, 1999.

[50] The International Telecommunication Union. Recommendation P.563 - Single Ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony Applications. `http://www.itu.int/itudoc/itu-t/aap/sg12aap/history/p563/index.html`, 2004.

[51] The Machine Learning and Knowledge Discovery Group at Aristotle University of Thessaloniki. Mulan: An Open Source Library for Multi-Label Learning. `http://mlkd.csd.auth.gr/multilabel.html`, 2010.

[52] G. Tsoumakas and I. Katakis. Multi-label Classification: An Overview. *International Journal of Data Warehousing and Mining*, 2007:1–13, 2007.

[53] G. Tsoumakas and I. Vlahavas. Random k-labelsets: An ensemble method for multilabel classification. In *ECML '07: Proceedings of the 18th European conference on Machine Learning*, pages 406–417, Berlin, Heidelberg, 2007. Springer-Verlag.

[54] I. T. Union. Measuring the information society - the ict development index 2009. 2009.

[55] VoIP Troubleshooter.com. The Open Speech Repository. `http://www.voiptroubleshooter.com/open_speech/index.html`, 2010.

[56] X. Wang, S. Chen, and S. Jajodia. Tracking anonymous peer-to-peer VoIP calls on the Internet. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*, 2005.

[57] H. Weisbaum. Don't Get Hooked by Latest Phishing Scam. `http://www.msnbc.msn.com/id/18553590/`, 2007.

[58] A. Woodruff and M. Stonebraker. Supporting Fine-grained Data Lineage in a Database Visualization Environment. In *Proceedings of the International Conference on Data Engineering (ICDE)*, Washington, DC, USA, 1997.

[59] Y. Xiang, W. Zhou, Z. Li, and Q. Zeng. On the Effectiveness of Flexible Deterministic Packet Marking for DDoS Defense. In *Network and Parallel Computing (NPC) Workshops*, 2007.

[60] B. Yang, J. T. Sun, T. Wang, and Z. Chen. Effective multi-label active learning for text classification. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 917–926, New York, NY, USA, 2009. ACM.

[61] W. Zhou, E. Cronin, and B. T. Loo. Provenance-aware secure networks. In *Proceedings of the International Conference on Data Engineering Workshops (ICDE)*, 2008.