

# Characterizing Robocalls with Multiple Vantage Points

Sathvik Prasad, Aleksandr Nahapetyan, Bradley Reaves

{snprasad, anahape, bgreaves}@ncsu.edu

North Carolina State University

**Abstract**—Telephone spam has been among the highest network security concerns for users for many years. In response, industry and government have deployed new technologies and regulations to curb the problem, and academic and industry researchers have provided methods and measurements to characterize robocalls. Have these efforts borne fruit? Are the research characterizations reliable, and have the prevention and deterrence mechanisms succeeded?

In this paper, we address these questions through analysis of data from several independently-operated vantage points, ranging from industry and academic voice honeypots to public enforcement and consumer complaints, some with over 5 years of historic data. We first describe how we address the non-trivial methodological challenges of comparing disparate data sources, including comparing audio and transcripts from about 3 Million voice calls. We also detail the substantial coherency of these diverse perspectives, which dramatically strengthens the evidence for the conclusions we draw about robocall characterization and mitigation while highlighting advantages of each approach. Among our many findings, we find that unsolicited calls are in slow decline, though complaints and call volumes remain high. We also find that robocallers have managed to adapt to STIR/SHAKEN, a mandatory call authentication scheme. In total, our findings highlight the most promising directions for future efforts to characterize and stop telephone spam.

## 1. Introduction

Robocalls have devastated consumer trust in the telephone network in the United States, and the problem is growing in other countries. Some estimates have claimed that as many as half of all calls are unwanted spam or fraud [1]. While many consumers simply do not answer calls from unknown numbers, those who do often hear misleading sales offers or impersonations of government agencies or well-known brands attempting to shake down their victims. These calls are effective and extract millions of dollars each year from victims.

The onslaught of fraud and abuse has made countering robocalls a major public policy issue in the United States. Presidential administrations, FCC and FTC Commissioners, and state attorneys general are among those who have made phone spam a key issue in their platforms. The attention has led to new laws like the TRACED Act [2] and a swath of changes to regulations by the FCC and FTC. These regula-

tions have expanded the definition of illegal calls and allowed or mandated that carriers implement technical measures to reduce or prevent illegal calls. Consumer providers have also responded by engaging and advertising their spam-call labeling services. Researchers, motivated by personal frustration and public need, have also expended significant energy to develop tools and techniques to characterize and counter spam calls.

With the substantial public and private investment in combatting the problem comes a need to assess our progress and re-evaluate what we believe we know about illegal calling. Industry sources publish regular reports on call trends, but using proprietary and opaque methodologies. For example, call labelling<sup>1</sup> must happen with limited context, and is therefore difficult or impossible to evaluate accuracy. Academic research uses transparent methods and evaluations, but comprise “one-shot” efforts describing a single moment in time. Collectively, the technical community and the public at large need a reassessment of the state of the network.

The goal of this paper is to aid researchers, regulators, industry operators, and the public in understanding and assessing robocalling and what effects, if any, the totality of anti-abuse efforts have had. To provide the most relevant view to-date, we fuse data from previously studied resources (FTC Consumer Complaints and an academic voice honeypot) and newly available vantage points (data from FTC’s Project Point of No Entry investigations, IRS fraud threat intelligence feeds, and a commercial voice honeypot). All of these individual sources have value and limitations. By comparing and contrasting these sources, we can get a more complete picture of robocalling. In the process, we can directly investigate how operational choices affect the findings, providing valuable insights for future researchers. Accordingly, this paper makes the following methodological contributions:

- We describe and evaluate call audio clustering for campaign identification using audio embeddings. Unlike prior work [3], embeddings can be computed deterministically in parallel, are vastly more efficient to query at the scale of data we study here, and remain comparable in accuracy.
- We show how to reliably and repeatably compare campaigns across vantage points.
- We develop techniques to determine if a campaign is interactive or static, permitting us and future work to track the growth of interactive illegal calling over time.

1. E.g., applying “Scam Likely” to caller ID information

Interactive calls hide information from honeypots and are analogous to cloaking detection in web phishing, and we anticipate generative AI will accelerate this trend.

We also contribute the following results (among others) and lessons for stakeholders:

- Longitudinal data indicates a secular decline in robocall volumes, though the magnitude of the decline may range from 25%-50%. This contradicts findings from a longitudinal robocall study from 2020 [3]. We cannot attribute the decline to any single factor.[Sec 5.1]
- We present the first analysis of the adoption of STIR/SHAKEN call authentication protocol by robocalls. Notwithstanding the decline mentioned previously, we show that remaining robocallers have succeeded in assimilating to the new regime. We further discuss reasons that explain why STIR/SHAKEN has failed to dramatically reduce robocall volumes and propose recommendations to address them.[Sec 5.1 and Sec 6]
- Implementing interactivity in a voice honeypot dramatically improves outcomes. We saw up to a 60% increase in collected call audio content, and nearly an order of magnitude increase in callback numbers collected in an interactive honeypot.[Sec 5.1.2]
- We show that estimates from a 2018 study on caller ID blocking effectiveness [4] still hold. However, they all but fail when one source of caller ID blacklist is applied to another vantage point.[Sec 5.2.7]
- We also show that manual call campaign labeling, whether by consumers or experts may be inconsistent. We recommend researchers and engineers to adopt our automated approach and release our source code to identify robocall campaigns from bulk robocall audio recordings using audio embeddings.[Sec 5.1.2]
- Using traceback data, we find that about 99% of campaigns subject to enforcement actions through the Project Point of No Entry (PPoNE) investigation ceased to operate after the enforcement actions. The small fraction of campaigns that did not cease to operate do so by establishing resilient operations that spread their calls over as many as 10 different service providers.[Sec 5.3.5]
- We estimate that PPoNE investigated campaigns associated with 5.5% of all robocalls observed in this study, though on average the investigation lagged the first observed call by **387 days**. [Sec 5.3.5]

## 2. Background and Related Work

In the 1990's, declining voice call costs due to deregulation led to a dramatic rise in telemarketing calls in the US. While the majority of these calls were for legitimate goods and services, people found them bothersome. In response, Congress empowered the Federal Trade Commission (FTC) to take measures to reduce unwanted calls.

The first measure was to set and maintain a policy called the Telemarketing Sales Rule (TSR), and the second was to

establish the National Do Not Call Registry (DNC). Among its many provisions, prerecorded telemarketing calls are prohibited entirely unless the caller has written consent from the consumer. All other telemarketing calls are also prohibited without written consent if a telephone number is listed in the DNC. The TSR also requires all telemarketing calls to transmit correct, accurate Caller ID (CLI) information. The FTC and other regulator are allowed to seek civil penalties of over \$50,000 for each and every violation of the TSR.

While the stakes are high, the TSR and DNC have done little to prevent an epidemic of automated sales or scam calls reaching virtually every American on a regular basis. The reason for this policy failure is partially technical and partially economic. The economic issue is that robocalling is so profitable that far more call center operators engage in unlawful activity than regulators and attorneys general have resources to bring cases against them.

Robocalling is profitable and easy because deregulation led to the North American Public Switched Telephone Network consisting of a federation of over 1,000 telephone service providers, most of whom use VoIP to transmit calls. All networks, whether using VoIP or legacy call technologies, deploy gateways at their network edge to protect their internal networks and handle media and signaling protocol conversions. Most calls transit many networks between their origination and termination networks. In the process, only the barest information — the caller's claimed ID — is transmitted from network to network. Each network maintains records of every call it originates, terminates, or transits, but these records only know the previous "hop" in the call path. The result is that when an illegal call is reported, the call source has to be identified through a laborious process called "traceback."

Network operators, at the mandate of regulators and legislators, have implemented two schemes to address these problems. The first is to mandate a clearinghouse for traceback requests called the Industry Traceback Group (ITG) to serve as a central point of contact between traceback seekers and network operators. The FCC has mandated that all providers respond to traceback requests within a 24 hour period. The ITG has drastically improved traceback responsiveness, conducting over 300 traceback requests per month.

The second scheme was a requirement for all VoIP operators to support a call authentication mechanism called STIR/SHAKEN. STIR/SHAKEN adds an additional authentication header to call setup messages that is intended to allow the provider at the receiving end to identify the originating network and authenticate that the call information (especially Caller ID) is accurate. STIR/SHAKEN specifies three attestation levels. The highest level of authentication is the 'A' level attestation, where the originating provider is (i) responsible for originating the call into the network, (ii) has a direct relationship with the customer (call originator) and the ability to identify them if required, and (iii) is directly associated with the phone number used to originate the call. 'B' level attestation indicates that the provider meets the first two criteria as the 'A' level, but the provider does not directly

own the phone number used to originate the call. ‘C’ level attestation indicates that the provider has no relationship with the call originator. In 2020, the FCC mandated [5] the adoption of STIR/SHAKEN as a call authentication framework across the North American Phone Network, pursuant to the TRACED Act [2]. The mandate imposed a deadline of June 30, 2021. However, the mandate included exception for smaller service providers and provisions for extensions when appropriate.

STIR/SHAKEN is analogous to DKIM in that it asks as a sort of “postmark” on a call request to indicate its source. It does not actually authenticate the source of the call, nor does it offer mutual authentication. It also does not provide mechanisms that are compatible with legacy infrastructure, so calls that originate, terminate, or transit non-VoIP networks irrecoverably lose this information. In addition to fundamental limits on what it guarantees or how compatible it is with the network, the short deployment timeframe led to a host of failures and incompatibilities so severe that many legitimate calls could not be authenticated [6].

Contemporary and subsequent research efforts provided alternatives that modified existing legacy protocols [7], [8] or used in-band [9] or out-of-band [10], [11], [12] protocols to provide cryptographic mutual call authentication. Other efforts have explored preventing spoofing [13], [14], fingerprinting call audio or data features [15], [16], automating tracebacks [17], intercepting calls with voice assistants [18], [19], or detecting or preemptively blocking calls based solely on metadata or calling patterns [20], [21], [22], [23], [24], [25], [26]. Finally, researchers have studied the vulnerability of humans to phone scams [27], [28] and how they could be helped [29], [30].

Most similar to our own work is data-driven characterization of robocalls, particularly using call complaint data or voice honeypots. While prior to 2012 there was a research emphasis on VoIP spam and other forms of abuse [31], the focus was on the impacts to VoIP infrastructure directly exposed to the public Internet. The first documented discussion of a voice honeypot was a M3AAWG report in 2014 [32], followed by work covering the deployment and initial results [33] of the described honeypot. This first honeypot paper explored geographic trends and overlap with consumer complaint data over 1.3M calls collected in a 7-week period, though audio analysis was limited to manual case studies. Soon after, a paper explored collecting voice and SMS using paid mobile SIM cards, finding that a small vantage point of 8 SIM cards gave a small vantage point of roughly 600 calls in 7 months. Pandit et al. studied the potential of honeypot and complaint data to serve as an estimate of Caller ID blacklisting effectiveness, claiming up to a 55% unwanted call reduction [4]. Finally, in 2020 Prasad and colleagues presented initial results from audio and metadata analysis of 11 months of honeypot data, including describing a method for clustering related calls together based on audio and showing a steady rate of unwanted calls [3]. Their followup work [34] applied natural language processing to call audio to analyze call content. Among other efforts, they demonstrated extraction of callback numbers that could be

used to identify and prosecute scammers.

Compared with prior work, the first principal contribution of this paper is to substantially improve and add to the growing body of robocall analysis techniques. Our new campaign identification approach scales to the volumes of historical and current robocalls, and we are the first to study STIR/SHAKEN and the impact of honeypot interaction. As for results, we confirm or refute prior findings with much greater confidence than before. We overturn previous findings of blacklist effectiveness and long-term static call volumes, while we confirm previous studies of low overlap in honeypot and consumer complaint data. We also show that robocalling is such a large problem that two large honeypots can have high call volumes but low overlap.

### 3. Methodology

Collecting, processing, and extracting insights from telephony abuse data is non-trivial. We develop specialized techniques to process robocall audio data at scale. In this section, we describe these methods, rationalize the design choices, evaluate their performance, and use these techniques to derive insights from a multitude of robocall-related data sources.

#### 3.1. Audio Aggregation

A fundamental unit of measurement in the robocalling ecosystem is a **robocall campaign**. We define a campaign as a set of phone calls that play the same audio message. In the context of data collecting in honeypots, a campaign is a collection of audio recordings where the audio in each recording is identical or nearly identical. Our technique enables us to efficiently extract campaigns from bulk robocall audio data that spans millions of calls in a reasonable time with high accuracy on a commodity desktop machine with reasonable memory and CPU resources. A strict definition of a campaign allows us to rigorously evaluate the performance of our aggregation pipeline through manual analysis, due to lack of ground truth. Next, we describe each step of the audio aggregation pipeline shown in Figure 1.

**3.1.1. Silence Detection.** We use Voice Activity Detection (VAD) to identify and discard silent calls. Silent calls are common when characterizing unsolicited phone calls collected in a honeypot. Since they do not contain any audio content, they offer little value in the context of robocall campaign analysis. Using WebRTC VAD, we measure the duration of audio and the duration of silence for each call. Calls containing at least five seconds of audio and at least 10% of audio content are retained for further processing. These thresholds are chosen based on estimates of how long a typical English speaker takes to deliver a meaningful short sentence in the context of a robocall campaign. In addition to receiving robocalls, honeypots receive unsolicited FAX calls and calls that merely play music-on-hold when answered. We identify and filter such calls at the transcription level by discarding calls with fewer than 30 characters in the

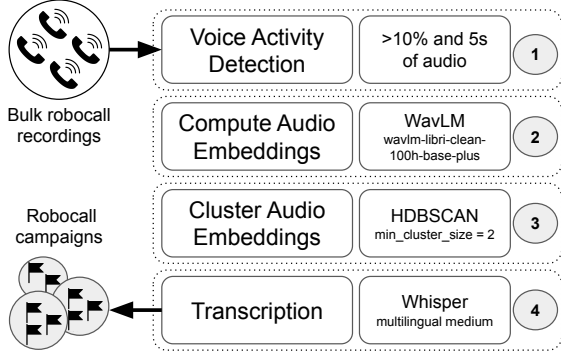


Figure 1: Audio processing pipeline

transcript. After this pre-processing step, we discard silent calls and retain calls with meaningful audio content for further processing.

**3.1.2. Computing Audio Embeddings.** We extract audio features from robocalls with substantial audio content by computing audio embeddings using pre-trained models. Recent advances in transformer-based speech language models enabled us to extract meaningful representations from audio data. Using *WavLM* [35], we compute 768-dimensional embeddings from the audio of each robocall. *wavlm-libri-clean-100h-base-plus* model is a fine-tuned version *WavLM-Base-Plus*. The base model was pre-trained using more than 90,000 hours of human speech data and can capture human speaker characteristics. The *wavlm-libri-clean-100h-base-plus* model is the fine-tuned version of the base model for Automatic Speech Recognition (ASR) on the LibriSpeech dataset, which has a WER of 0.0683 [36]. Since the model is pre-trained on 16kHz audio data, we resample robocall recordings to 16kHz before computing the embeddings. By performing a forward pass of the audio data through the model, we extract the output of the hidden states. The base model architecture has 12 Transformer encoder layers [37], a hidden state of 768 dimensions, and 8 attention heads. We aggregate the embeddings at the last hidden state using mean pooling to obtain a single 768-dimensional embedding for each audio recording. Since the *WavLM* model architecture captures human speaker characteristics, the computed embeddings represent the human speech within the robocall audio recordings. [38]

**3.1.3. Advantages of Embeddings.** The audio embedding-based approach offers several advantages over fingerprint-based audio analysis. Audio fingerprint-based techniques require a database of fingerprints to compare against. This leads to space and time complexity tradeoffs, as we need to compare hundreds of thousands of calls against each other. The computational complexity and the memory footprint grows rapidly, making it infeasible to scale to the volume of data presented in this work. In contrast, embeddings are compact vectors that can have a notion of similarity in the vector space. This allows us to compare embeddings using an

appropriate similarity metric, such as cosine similarity, which is not possible with fingerprints. As a consequence, relative comparison drastically reduces the computational complexity and the memory footprint, making it feasible to scale to the volume of data presented in this work. Furthermore, we can use distance-metric based cluster evaluation technique to measure the quality of robocall campaigns, whereas fingerprint-based techniques require manual evaluation.

**3.1.4. Clustering Audio Embeddings.** After computing embeddings for each audio recording, we cluster the embeddings to identify robocall campaigns. We use the HDBSCAN clustering algorithm to group similar embeddings into campaigns. HDBSCAN is a density-based clustering algorithm that can cluster high-dimensional data. Although most calls collected in honeypots are robocalls, there are mis-dials from real humans or calls from humans trying to reach the previous owner of the phone number. Since these are not robocalls, our clustering approach must discard them as outliers. Since HDBSCAN identifies outliers, it enables us to discard such misdialled calls and retain only robocalls for further analysis. Finally, HDBSCAN is computationally efficient and converges relatively quickly when compared to other clustering algorithms [39], making it a suitable choice for clustering the large volume of data presented here.

**3.1.5. Transcription and Language Identification.** After identifying robocall campaigns, we transcribe the audio recordings within each cluster using Whisper’s multi-lingual medium model. Although we explored other STT models, we found that Whisper’s speech-to-text models are one of the best-performing models [40] for transcribing noisy multi-lingual robocall recordings. Another advantage of using Whisper to transcribe robocall recordings is that it is capable of identifying the language spoken in the audio. The medium multi-lingual model [41] is trained on 100 languages and detects the language spoken in the audio by sampling the first 30 seconds of the audio. During downstream analysis, we use the detected language to filter out calls that are not in English since our analysis is focused on English robocalls.

## 3.2. Performance Evaluation of Campaign Detection

Prior work has used audio fingerprinting techniques to identify and cluster robocall campaigns. However, such techniques fail to scale to the volume of data presented in this work, where we process data that is at least an order of magnitude larger than previously reported datasets of robocall audio recordings.

Method	Total Clusters	Max Cluster Size	Total Calls Clustered	Data Clustered (%)	Cluster Perfection (%)	Intra Cluster Precision (%)
Wav2Vec2	119	44	740	67.27	94.11	97.68
WavLM	145	51	842	76.55	93.79	97.84
Echoprint	165	19	586	53.27	96.97	98.83

TABLE 1: Campaign Identification Performance Evaluation

We evaluate the performance of our audio aggregation and clustering techniques using a public real-world dataset of

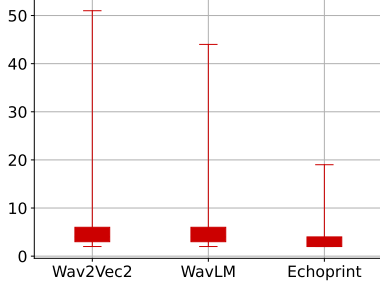


Figure 2: WavLM and Wav2Vec2 uncover large (outlier) campaigns while maintaining a high median cluster size

robocalls collected from various sources [42]. We benchmark them against the fingerprint-based technique used in prior work [3], [34], and compare the performance of *WavLM* and *Wav2Vec2* based embeddings. We find that *WavLM* audio embedding-based approach outperforms the others while scaling to the volume of data presented in this work. *WavLM* embeddings cluster the highest fraction of calls into campaigns (76.6%), while maintaining a high cluster perfection score (93.8%) and intra-cluster precision (97.8%). It is also able to uncover large campaigns while discarding individual examples of robocalls that are not part of a campaign.

Prior fingerprint approaches [3] used ad-hoc metrics, described in Figure 15 and Figure 16. By manually listening to the audio recordings and by reading the transcripts, we label each cluster. While computing cluster perfection score, we use a strict criteria where we mark a campaign as a bad campaign even if a single misplaced call is present in the audio cluster. The Intra-cluster precision scores for *Wav2Vec2*, *WavLM* and fingerprint-based techniques are 97.7%, 97.8%, and 98.8% respectively. The Cluster Perfection scores for *Wav2Vec2*, *WavLM* and fingerprint-based techniques are 94.11%, 93.79%, 96.97%, as listed in Table 1.

**Measuring campaign size:** We analyze the distribution of cluster sizes and the size of outlier clusters to measure the ability of our techniques to uncover large campaigns, while maintaining good cluster quality. Figure 2 shows the distribution of cluster sizes for each method, where *WavLM* and *Wav2Vec2* maintain a higher mean cluster size compared to fingerprinting-based approach, while successfully identifying larger campaigns.

**Measuring data coverage:** We evaluate the fraction of data clustered and the fraction of data discarded as outliers to measure the ability of our techniques to cluster robocalls into campaigns, while discarding individual examples of robocalls that are not part of a campaign and any misdials from human callers. *WavLM* based approach clusters the highest fraction of calls into campaigns (76.55%). Since it is able to uncover large campaigns while maintaining high cluster quality and performs at-par or better than two other techniques, we use *WavLM* for the rest of the analysis in this work.

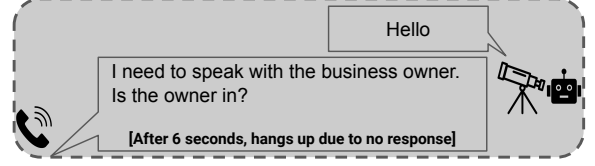


Figure 3: Interactive campaign disconnecting the call due to lack of engagement (no response) from Robocall Observatory



Figure 4: Interactive campaign engaging with RRAPTOR (an interactive honeypot) which yields additional details

### 3.3. Comparing Campaigns Across Honeypots

After identifying robocall campaigns in English, we compare representative robocall transcripts of campaigns in each data source to identify common campaigns across multiple data sources. We randomly sampled one representative transcript from each campaign and compared them against a randomly sampled representative transcript from campaigns uncovered in a different data source. For each comparison, we compute a distance metric between the two transcripts to measure the similarity between the campaigns. We build a similarity matrix by identifying the closest matching pairs.

**Comparing Transcripts vs Comparing Audio Embeddings:** Using transcription to compute campaign overlap between two data sources is more reliable than audio embeddings for a few reasons.

- *Audio collection environment can differ:* Honeypots are often deployed with different configurations, lacking consistency in data collection. For example, different honeypots may set different maximum call durations, or use different codes (G711A, G729, etc) in their setup.
- *Interactive vs Non-Interactive:* Comparing a two-channel audio stream from an interactive honeypot with a single-channel audio stream from a non-interactive honeypot is non-trivial. However, comparing the tran-

scripts of audio coming from the calling party (robocaller) allows us to compare the content of the calls.

- *Varying Network Conditions*: The network conditions under which the data is collected can substantially influence the audio quality (packet loss, jitter, etc.). A cloud-based vs. an on-prem honeypot also influences the network parameters of the raw audio. Transcripts are less sensitive to these variations.

**Transcript Similarity Metric**: Transcripts of phone call recordings are often noisy because of the underlying network characteristics. The call audio is affected by packet loss, jitter, and other network conditions. As a consequence, the transcripts from audio recordings that sound similar to a human ear may generate slightly different transcripts (different punctuations, capitalization, etc.). Therefore, we choose a well-established string similarity metric. We pre-process the transcripts into tokens<sup>3</sup> by splitting them into words, removing numbers, punctuations, and special characters, and normalizing the case.

**Jaccard Similarity over Tokens**: Jaccard similarity is a set-based similarity metric that measures the similarity between two sets. Comparing tokenized transcripts with Jaccard similarity identifies campaigns that are identical or nearly identical in content.

**Longest Common Subsequence (LCS) over Tokens**: LCS is a string similarity metric that measures the similarity between two strings by finding the longest subsequence that is common to both strings. Using a similar tokenization and pre-processing approach, we use LCS over tokenized transcripts to capture campaigns that have the same initial pitch but behave differently after the initial greeting. Such campaigns are designed to wait for user engagement and continue their pitch only after they hear a response from the callee, as shown in the example conversation in Figure 4. Token-based LCS preserves the order of the tokens, and captures such interactive campaigns while also capturing non-interactive campaigns that have identical or nearly identical transcripts across different data feeds.

**Selecting a similarity threshold**: For all comparisons using LCS and Jaccard similarity, we use a similarity threshold of 90% to determine common campaigns across multiple data sources. A 90% threshold allows us to capture campaigns accurately, while also accounting for minor variations in the transcripts due to network conditions, audio quality, and other factors. The 10% margin approximately translates to half a sentence (average transcript length is 70 words), which accounts for robocalls that terminate the call prematurely or may have minor transcription variations. With these thresholds, the expert author manually sampled and verified more than 100 example campaigns to ensure that these thresholds result in accurate matches across honeypots. By setting a high threshold, we prioritize precision over recall.

### 3.4. Ringless Voicemail Detection

Ringless voicemail or voicemail injection campaigns are a type of robocall campaign where the audio message is delivered directly to the user's voicemail without ringing the user's phone. These campaigns are noteworthy since they override the traditional call presentation behavior, where the user can see the incoming call along with crucial information like the "Spam/Scam Likely" label, the caller ID, call signer information, etc.<sup>2</sup> Voicemail injection campaigns bypass this call presentation and directly deliver the message to the user's voicemail. A common approach to deliver ringless voicemail is to place an initial call so that the user's line is busy, and then place a well-timed robocall so that the call goes directly to voicemail. We analyze the timing information in the call signaling information (SIP signaling) with matching calling and called numbers and measure the number of campaigns that adopt ringless voicemail delivery systems.

### 3.5. Callback Number Extraction

Callback numbers are phone numbers embedded within the call audio. Robocall originator provides these numbers instructing the call recipient to engage with the call originator. For legal telemarketing calls, callback numbers are toll-free numbers that allow the call recipient to opt-out of future calls. Interestingly, some campaigns use low-quality text-to-speech (TTS) to read out the callback number, which makes it challenging to extract the callback number from the transcript. For example, the callback number "844-924-XXXX" is transcribed as "eight four four nine two four...". We refer to such callback numbers as *vocalized callback numbers*. We extract callback numbers from the transcripts of robocall campaigns using a curated set of regular expressions and use *phonenumber* library to process them further.

### 3.6. Caller ID Normalization and Comparison

We normalize caller IDs to E.164 format to discard invalid phone numbers and user entries in crowd-sourced data. Since the size of our data spans tens of millions of entries, we implement an efficient tree-based caller ID search framework to compare caller IDs across data sources. This framework ingests valid caller IDs from each feed, normalizes them to E.164 format, and stores them in a tree-based data structure, providing a logarithmic time complexity lookup API for caller ID comparison.

### 3.7. Hardware Resources for Computation

All the analysis presented in this work were performed using on-prem computing and storage resources. Our workloads were distributed across two Intel i9 machines with four NVIDIA GeForce RTX 4090 in total in combination

2. In 2022 the FCC passed a rule which treats ringless voicemail as robocalls, and requires them to adhere to the TCPA: <https://www.fcc.gov/document/fcc-finds-ringless-voicemails-are-subject-robocalling-rules>

3. SpaCy tokenizer using `en_core_web_sm` model

with a machine equipped with NVIDIA GeForce RTX 3070. It took multiple months of compute time and multiple TBs of storage to pre-process audio, compute audio embeddings and transcribe millions of robocall recordings.

## 4. Data Collection and Characterization

Vantage Point Name	Voice Honeypot	Dataset Size	Call Audio	STIR/SHAKEN
Robocall Observatory	Yes	10.05M	Yes	Yes
RRAPTOR	Yes	948.76k	Yes	Yes
PPoNE Enforcement Actions	No	1.35k	Yes	No
FTC Do Not Call Complaints	No	13.71M	No	No
IRS APWG Data Feed	No	668.99k	No	No

TABLE 2: Vantage points, dataset sizes and characteristics

In this section, we describe each data source used in this paper, explain the data collection process, and discuss how we apply our data processing pipelines to each data source. We also discuss the ethical considerations of our work.

### 4.1. Robocall Observatory

The Robocall Observatory is a non-interactive automated call answering system designed to collect and store information about unsolicited phone calls. It is a collection of VoIP phone numbers mapped to a PBX, where the PBX is configured to either answer or decline a phone call. When answering a call, Robocall Observatory plays a simple greeting — “Hello”. It records any answered calls, and stored the audio as lossless WAV files. It also stores the raw SIP signaling information (include STIR/SHAKEN headers) along with the call detailed records (CDRs).

**Data Characterization:** The data collected from the Robocall Observatory spans from Apr 1st, 2019 to May 25th, 2024. During this period, the observatory received more than 10 million (10,056,650) calls. On average, there were approximately 5,349 calls per day, with a high rate of calls per day during the initial years followed by a slow decline in the call rate over the years.

STIR/SHAKEN was enabled on the observatory on 1st Nov 2021. Since then, Robocall Observatory has received 3,234,362 calls. Throughout the duration of Robocall Observatory’s operation, about 30.40% (983,519) of calls were signed with an A-level attestation, followed by 21.59% (698,278) with B-level attestation, 9.52% (307,986) with C-level attestation, and 38.48% (1,244,579) were unsigned calls.

The Robocall Observatory answered and recorded over 2.3 Million (2,302,606) unsolicited phone calls. Out of all calls recorded, only 25.5% (588,097) calls met our filtering criteria of containing enough audio information for further downstream audio processing. This is a decrease from the 37.25% seen in prior work by Prasad *et al.* [3].

The preprocessed calls containing substantial audio information were further processed to compute their respective audio embeddings using WavLM. These embeddings were

clustered together using the campaign detection technique described previously in Section 3.1.4. A total of 360,723 (61.33%) calls were clustered into 43,822 campaigns, with 227,374 discarded as outlier calls. These outlier calls are sometimes mis-dials or individual calls not part of a larger volume of traffic. The average campaign size was 8.23 calls per campaign, with some campaigns containing thousands of calls each. The largest campaign contained over 3,000 calls. The total time to cluster all calls was about 55 hours on a single machine with 32 cores and 128GB of RAM.

**Cluster Evaluation:** To evaluate the quality of the clusters, we compute the silhouette score with cosine distance and the Calinski-Harabasz score, using sklearn’s implementation. The clusters resulted in a silhouette score of 0.56 and a Calinski-Harabasz score of 1,166. Since the silhouette score ranges from -1 to +1, a score of 0.56 indicates that the clusters are well separated [43]. Additionally, a higher value of Calinski-Harabasz indicates dense and well separated clusters [44]. Crucially, we were able to automatically compute these cluster evaluation metrics without any ground truth labels (unlike audio fingerprint based clustering), since each audio embedding can be mapped to a spatial domain.

**Language Distribution:** Using OpenAI’s Whisper, we found that 94.8% of all calls clustered into campaigns were in English. The next prominent language was Spanish (2.5%) and Chinese (1.5%). The top three distribution aligns with the spoken languages in the United State, where English, Spanish and Mandarin are the top three spoken languages [45]. The relative volume of Mandarin and Spanish robocalls in the Robocall Observatory were substantially higher in 2020-2021, and declined in 2022. Despite this decline, we uncovered numerous Spanish and Mandarin campaigns active in 2023-2024.

**Callback Numbers:** Throughout Robocall Observatory’s operation, we observed 6,010 callback numbers from 17,447 campaigns. About 3.7% (222) callback numbers were vocalized callback numbers, which were not captured by the techniques described by [34]. Callback numbers continue to be an effective approach used by robocallers and telemarketers to engage with the target. In 2024 alone, we uncovered 247 callback numbers from 356 campaigns.

Out of the 6,010 callback numbers uncovered, 3,114 had a lifetime of less than 1 day, indicating that these callback numbers were used for short-lived campaigns. Often, fraudulent or illegal campaigns use such short-lived callback numbers to evade detection.

The Robocall Observatory encountered 481 callback numbers with a lifetime of more than 1 year, with the oldest callback number that has been in use for over 5 years. Often such callback numbers are used for useful robocalls and for legal telemarketing campaigns. For example, callback numbers with lifetimes over many years are used by campaigns for public announcements (school closures, weather alerts, etc.), pharmacy notifications, and other legal telemarketing campaigns. Many such legal campaigns provide a callback number, and instruct the target to call back to and ask the call originator to opt out of such calls from the campaign, similar to an unsubscribe mechanism in email marketing.



## 4.2. RRAPTOR

RRAPTOR [46] is an interactive commercial honeypot that automatically answers calls made to the numbers assigned to it. Unlike the Robocall Observatory, RRAPTOR plays a set of pre-recorded prompts at various stages of a call to engage the caller in conversation. It collects and stores audio data as two-channel mp3 files, and keeps track of the calling number, the date of the call and the STIR/SHAKEN attestation level of the call.

**Data Characterization:** The data collected from RRAPTOR, spans from 2024-01-01 to 2024-05-23. During this period, the honeypot received 948,775 calls at a rate of about 6,682 calls per day, much higher rate than Robocall Observatory.

STIR/SHAKEN was enabled throughout the duration of RRAPTOR’s operation. About 44.3% (420,639) of unsolicited calls received by RRAPTOR were signed with an A-level attestation, followed by 24.5% (232,344) with C-level attestation, 12.8% (121,026) with B-level attestation. About 18.4% (174,766) calls were unsigned calls.

Unlike the robocall observatory, each call received by RRAPTOR was answered and recorded. Furthermore, the maximum audio duration of each recorded call was 90 seconds. In total, 36.28 % (344,254) calls met the criteria of containing more than 10% of audio and more than 5 seconds of audio and were processed further to extract robocall campaigns. This is a slight decrease from prior work of 37.25% and slightly higher than Robocall Observatory’s 25.5%. [3]

The robocall campaign aggregation process clustered 31.33% (107,531) calls with substantial audio into 15,475 campaigns. The mean campaign size was 6.94, with the largest campaign containing 3714 calls. The total time to cluster the calls was about 20 hours on a machine with 32 cores and 128GB of RAM.

**Cluster Evaluation:** The RRAPTOR clusters resulted in a cosine-distance based silhouette score of 0.499 and a Calinski-Harabasz score of 540.154. Both scores indicate that the clusters are well-separated and dense, similar to the clusters obtained from the robocall observatory data. [43], [44]

**Language Distribution:** Similar to the data collected in the robocall observatory, the majority of calls clustered into a robocall campaign in RRAPTOR were in English (96.9% or 104,094 calls), with 1.4% (1,454) in Spanish and 0.9% (1,004) in Mandarin.

**Callback Numbers:** Within a span of less than five months, RRAPTOR uncovered 2,325 callback numbers, of which 6 were vocalized. These callback numbers were used across 2,519 campaigns.

## 4.3. FTC Complaint Data

The Federal Trade Commission’s Do Not Call Complaints feed consists of consumer reported complains of unwanted phone calls. This data feed is available to the public via api.data.gov. Each report consists of a mandatory field where the consumer reports the caller ID of the unwanted call and the date of the call. It also contains a “subject” field where



Figure 5: Distribution of the top 15 categories of tracebacks of illegal robocalls used in PPoNE enforcement actions

the consumer can select from 16 categories to describe the nature of the call. The FTC’s DNC feed spans from 2020-02-14 to 2024-05-23, during which 13,707,524 reports were collected at a rate of about 8,787 reports per day.

## 4.4. PPoNE

As part of the Project Point of No Entry (PPoNE) initiative, the Federal Trade Commission (FTC) released a set of 31 enforcement actions against gateway service providers. Each of these enforcement action contains a collection of traceback requests and their outcome, and is presented as a table at the end of the report. We developed an automated system to download, process and extract information from these enforcement letters. This automated pipeline parses the PDF enforcement letters and extract the audio URLs. We also manually verified the results and used semi-automated PDF table parsing tools to extract tabular data embedded within the PDFs.

The PPoNE dataset consists of a collection of 1,353 audio examples ranging from 2020-12-30 to 2023-10-11. Each audio example was used to complete a traceback request to identify the source of the illegal traffic. This is a small fraction of the total tracebacks conducted during this period, and is not representative of all tracebacks.

This dataset crucially contains a collection of 118 different label, where a label is assigned to each audio example, (eg: Amazon-AuthorizeOrder, SocialSecurityAdministrationImposter, AutoWarranty, etc). Figure 5 shows the distribution of the top 15 labels in the dataset. These labels are broad categorization of the type of the call based on the audio content. These labels were likely assigned by human investigators, and are often quite broad — including more than one campaign under the same label.



## 4.5. IRS APWG

The Anti-Phishing Working Group (APWG) IRS Impersonation Scam feed consists of a caller ID field that is supposedly associated with originating illegal IRS impersonation calls, or other fraudulent robocalls that are related to taxes or other tax related services. The datafeed contains the caller ID and the timestamp of when the call was reported to the APWG community. We use APWG’s APIs to collect this data. This feed was available to the members of the APWG community, but has since been discontinued. The APWG [47] data feed spans from 2012-11-07 to 2023-02-14, during which 668,991 reports were collected at a rate of about 178 reports per day.

## 4.6. Ethical Consideration

The Institutional Review Board (IRB) and our university’s legal team has reviewed and approved all of the data collection methods used in this paper. The maintainers of the commercial honeypot donated their data to the project. The Robocall Observatory is set up in a first-party consent state, where only one party needs to consent to the recording of the call. When a call originates from outside the state, Federal law applies, which is also a first-party consent scenario. While we attempt to analyze automated mass call campaigns, we are conscious that the honeypots may receive calls from individuals. Through HDBSCAN discarding outliers and VAD-based pre-processing discarding short-duration calls, this work focusses on robocalling campaigns (collection of calls with substantial audio that are identical or nearly identical) and not individual call recordings. The Robocall Observatory data collection technique ensures that it gracefully terminates calls at 60 seconds (RRAPTOR at 90 seconds) to minimize data collection from non-robocalls. Furthermore, we have confidentiality agreements in place that prevent us from sharing raw data. For our other sources of threat intelligence, we have data sharing agreements with the APWG. Finally, the PPoNE and the FTC DNC data were collected from public documents and APIs.

## 5. Results

This section characterizes the robocalling ecosystem from each independent vantage point, quantifies the relative overlap of caller IDs across multiple data feeds, measures the overlap of campaigns between Robocall Observatory and RRAPTOR, quantifies the impact of PPoNE enforcement actions, and distills the key takeaways from these measurements.

### 5.1. Metadata analysis

#### 5.1.1. Robocall Volumes and STIR/SHAKEN Adoption.

One of the most visible changes to the telecom ecosystem in United States over the past few years is the mandatory deployment of STIR/SHAKEN. We present the current state of the longitudinal robocall volumes and delve into how

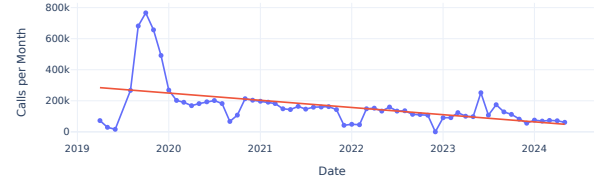


Figure 6: Robocall Observatory’s monthly call volume

STIR/SHAKEN deployment has influenced the robocalling ecosystem from multiple vantage points.

**Finding 1:** *The volume of robocalls is declining in the long term.* The monthly volume of robocalls received by Robocall Observatory has been declining over the span of 5 years, as seen in Figure 6 in comparison to the flat rate seen in prior work [3]. A linear regression for the monthly volume of calls received by Robocall Observatory had a slope of -127 calls per month ( $R^2 = 0.227$ ). Although Robocall Observatory has been receiving fewer calls over time, the reported sum of money lost by victims has increased from 2022 to 2023, as per the FTC’s annual report [48]. A linear regression across the FTC-DNC daily complaints on the other hand, we see a slope of -1.5 complaints per day ( $R^2 = 0.005$ ) over the span the 5 years. Illegal robocalls continue to pose a significant threat to consumers in the United States.

**Finding 2:** *The volume of robocalls with STIR/SHAKEN attestation is increasing over the long term.* Robocall Observatory started receiving STIR/SHAKEN attested calls from 1st Nov 2021. Over a span of about 2.5 years, Robocall Observatory observes an increasing trend of the fraction of calls signed using STIR/SHAKEN. In Nov 2021, about 60% of all unsolicited phone calls were signed. Whereas, in May 2024, about 80% of all calls were signed. Further, the volume of unsigned robocalls has been decreasing over time. In terms of overall traffic, using a linear regression we see the slope of unattested traffic to be -0.00031 normalized signed calls per day ( $R^2 = 0.089$ ).

**Finding 3:** *Attestation level of robocalls is increasing over time.* STIR/SHAKEN offers three levels of attestation — A, B, and C, as described in Section 2. As seen in Figure 8 and Figure 7, majority of robocalls are signed with A-level attestation. About 44.3% (420,639) of unsolicited calls received by RRAPTOR were signed with an A-level attestation, followed by 24.5% (232,344) with C-level attestation, 12.8% (121,026) with B-level attestation. About 18.4% (174,766) calls were unsigned calls. This trend is consistent with the data from Robocall Observatory, where we see 51.2%(179,544) of the calls signed with A-level attestation, followed by 13.8%(49,937) with B-level attestation, and 14.2%(48,442) with C-level attestation since 2024. The fraction of unsigned calls decreased from 38.5% in 2021 to 20.7% in 2024, demonstrating a steady increase of STIR/SHAKEN signed robocalls. This trend is consistent not only across our data feeds, but also with recent industry reports on robocalls signed with STIR/SHAKEN in 2024, according to which 80% of robocalls are signed with A-level attestation [49]

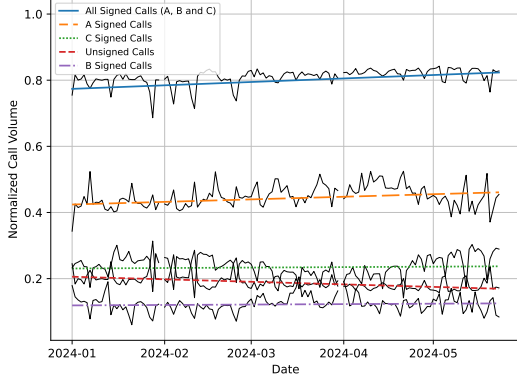


Figure 7: Calls received by RRAPTOR signed with STIR/SHAKEN in 2024 (normalized by total calls per day)

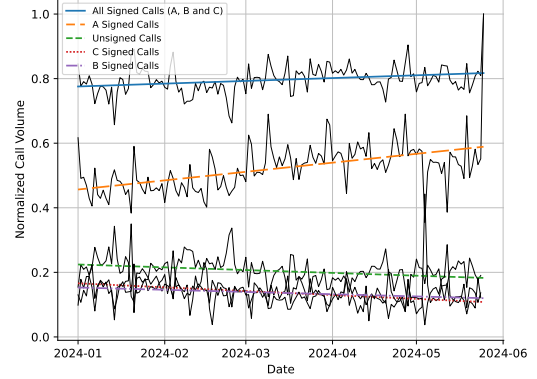


Figure 8: Calls received by Robocall Observatory signed with STIR/SHAKEN in 2024 (normalized by total daily calls)

### 5.1.2. Role of Honeypots in Characterizing Robocalls.

Honeypots serve as valuable vantage points to study robocalls. We discuss and quantify the impact of adding interactivity to honeypots and how that can improve our visibility into the robocalling ecosystem.

**Finding 4:** *Large-scale and long-running honeypots are reliable vantage points to measure broader robocalling phenomena.* To study the extent to which honeypots can be used to measure broader phenomena within the robocalling landscape, we compare two phenomena inherent to the robocalling ecosystem. Namely, fraction of unsolicited phone calls signed with STIR/SHAKEN and distribution of languages used by robocalling campaigns. The fraction of unsolicited phone calls signed with STIR/SHAKEN is consistent across three distinct vantage points — RRAPTOR, Robocall Observatory and independent industry reports. As seen in Figure 7 and Figure 8, the fraction of signed calls is about 80% across both honeypots as of May 2024, while following a similar trend throughout 2024. Furthermore, A-attested calls are the most common, with C-attested calls being the second most while B-attested calls were the third most common category. Finally, in both honeypots, the largest fraction of robocall campaigns use English. Spanish and Mandarin are the second and third most common languages, respectively. By comparing the signing rates and language distribution, we find that broader phenomena measured using Robocall Observatory and RRAPTOR as two different vantage points are consistent. This demonstrates that honeypots operating at the scale of Robocall Observatory and RRAPTOR are reliable sources of data for measuring broader robocalling phenomena. They can independently report and provide valuable insights into the robocalling ecosystem.

Honeypots can differ in their operational characteristics by being more or less interactive, as described in Section 2 [32]. However, developing, deploying and operating an interactive honeypot (versus a non-interactive honeypot) is substantially more challenging. It requires well-timed interactive prompts, more computational resources, and non-trivial design overhaul to detect and play audio in near-real time. Such fundamental changes to a honeypot’s design is not justified if it yields only marginal benefits.

**Finding 5:** *An interactive honeypot captures more analysable audio from a robocall than a non-interactive honeypot.* We find that interactive honeypots are crucial to study the interactive campaigns that are active in the current robocalling ecosystem. This is a notable shift from a previous study from 2020 by Prasad et. al. [3] where a non-interactive honeypot was sufficient to capture substantial amount of audio information. RRAPTOR captured 66.8% calls with more than 10% audio while Robocall Observatory captured 42% calls with more than 10% audio. Operationally, the maximum duration of recordings for RRAPTOR (interactive) was set to 90 seconds, while Robocall Observatory (non-interactive) recorded calls were limited to 60 seconds. Robocall originators are employing interactivity to limit the amount of information a honeypot can extract from a robocall. This behavior is analogous to cloaking techniques used by phishing websites to evade detection.

**Finding 6:** *An interactive honeypot captures nearly 10 times more callback numbers from a robocall when compared to a non-interactive honeypot.* Information captured from a robocall extends beyond the amount of audio, such as callback numbers extracted from the audio. In 2024, RRAPTOR captured 2,325 unique callback numbers from 2,519 campaigns, while Robocall Observatory captured 247 callback numbers from 356 campaigns. This may indicate that independent of the call volume the honeypot receives, employing an interactive honeypot to engage with the robocaller allows for more callback numbers to be extracted. We discuss this phenomena in more detail in Section 5.3.

**Finding 7:** *The lifetime of a callback number can vary significantly, with some being used for a short duration and other being used for extended periods of time.* To better understand the life cycle of callback numbers, we computed the lifetime of callback numbers by measuring the first and the last instance when a callback number was used by a campaign found in Robocall Observatory. We see that the average lifetime of a callback number was 88 days, with a standard deviation of 219 days.

**Finding 8:** *We found that 362 campaigns were using voicemail injection techniques to deliver the message.* Since

Robocall Observatory collects both signaling and call audio, this allows us to detect campaigns that leave a voicemail instead of a traditional phone call. We uncovered 3,784 campaigns with at least one call containing more than one call attempt within a 15 second window. We further filtered these campaigns to determine if more than 90% of calls within the campaign had more than one call request message (SIP INVITE) within 15 seconds. A total of 362 campaigns were using voicemail injection techniques to deliver robocalls. Such calls directly reach the user’s voicemail and bypasses the standard call presentation to the user [29], where labels like “Scam Likely” and “Spam Risk” are displayed when the phone is ringing.

**Finding 9:** *Audio-based campaign identification is more specific than human-assigned labels*. The PPONE dataset described in Section 4.4 contains campaign names that are human interpretable, and were likely assigned by human analysts. A total of 118 such labels exist in the PPONE dataset. However, on clustering these raw audio recordings using the audio aggregation pipeline described in Section 3.1, 174 campaigns were identified. These clusters had a cosine distance based Silhouette score of 0.67 and Calinski Harabasz Score of 211.71. Meanwhile, when clustered using the PPONE human labels, the Silhouette score was -0.38 and Calinski Harabasz Score was 17.1. The lack of consensus on the definition of a robocall campaign makes it challenging to reliably draw conclusions from human-assigned labels. Therefore, using reliable and automated campaign identification techniques like the one presented in Figure 1 is crucial to track campaigns across data sources and study their evolution.

## 5.2. Caller ID Overlap Analysis

Each vantage point listed in Table 2 contains caller ID information. Comparing the relative overlap in caller IDs contained within these feeds offer insights into effective strategies of harnessing originating phone number based threat intelligence about robocalling campaigns. To quantify the relative overlap between these data feeds, (i) we identify the temporal overlap between two feeds, (ii) measure the total number of reported “events” (calls, complaints, etc), (iii) identify unique originating phone numbers that are valid NANP numbers, and (iv) report the overlap between the two feeds. Table 3 describes these fine-grained results.

**5.2.1. Honeypots and Crowd-sourced Consumer Complaints.** We compared the caller IDs uncovered by the two honeypots (RRAPTOR and Robocall Observatory) with the FTC Do Not Call consumer complaints. Over a span of about 4 months, 21,225 unique phone numbers found in RRAPTOR were also reported by frustrated consumers to the FTC via the Do Not Call complaints reporting mechanism. This overlap translates to about 4% of the unique phone numbers in RRAPTOR and 8% of unique phone numbers in the FTC DNC feed. The top three categories consumers selected when reporting these calls were “Others”, “Dropped Call or No Message” and “No Subject Provided”.

Unlike RRAPTOR, Robocall Observatory data collection overlapped with the FTC DNC for over 4 years. During this time, Robocall Observatory collected 6.7 million calls, and the FTC DNC received 12.9 million complaints, which is almost twice the number of calls received by Robocall Observatory. 306,224 unique and valid NANP phone numbers uncovered by Robocall Observatory were also reported by consumers to the FTC via the Do Not Call complaints. This is the largest overlap in magnitude between any two feeds which translates to about 10% of the unique phone numbers in Robocall Observatory and 4.4% in the FTC DNC feed. The top three categories reported by consumers for the calls intersecting Robocall Observatory and FTC DNC were “Other”, “Dropped Call or No Message”, and “No Subject Provided”. This reinforces the observation that a large proportion of unsolicited phone calls are silent calls.

### 5.2.2. PPONE Enforcement Actions and Caller ID Feeds.

The PPONE is a small dataset of selected examples used in enforcement actions by the FTC. Therefore it does not represent a large-scale data collection effort like the other feeds. By comparing historical PPONE data with newer feeds, we can detect if campaigns subject to prior enforcement are still active.

A total of 75 numbers referenced in PPONE enforcement actions were also reported by consumers via the FTC DNC reporting mechanism. Due to the nature and size of the PPONE dataset, drawing conclusions on the scale of overlap is not significant. It is noteworthy that about 6% of phone numbers in the PPONE were reported to the FTC DNC registry, and in 56% of cases the FTC DNC feed reported the phone number first. Among these overlapping calls, the top category selected by consumers in the FTC DNC complaints was “Calls pretending to be government, businesses, or family and friends”. The PPONE dataset and IRS APWG feed overlap for about 2 years, however only 11 unique phone numbers were present in both.

**5.2.3. PPONE Enforcement Actions and Honeypots.** The PPONE dataset and Robocall Observatory overlap for over 2 years during which Robocall Observatory received about 4.4 million calls (both answered and unanswered) from 1.9 million valid NANP phone numbers. There were 35 unique caller IDs in Robocall Observatory that had a matching caller ID in the PPONE feed, totaling 103 calls made to Robocall Observatory.

The PPONE enforcement actions and RRAPTOR do not overlap since the enforcement actions pre-date RRAPTOR’s data collection time period. However, we still uncovered 9 phone numbers listed in the PPONE enforcement actions generating robocalls to RRAPTOR. While caller ID should be treated as a weak signal of campaign activity, we examine campaign audio content in the next section to understand if robocall campaigns that were subject to prior enforcement action remain active.

**5.2.4. Robocall Observatory and RRAPTOR (independent honeypots).** There were 15,692 unique phone numbers that

	Total Call Attempts	Unique Calling Numbers	Overlap (Unique Numbers)	Overlap Period (days)
Robocall Observatory	344,344	151,937		
RRAPTOR	948,775	544,129	15,692	143
Robocall Observatory	8,428,978	3,143,974		
IRS APWG	668,991	262,931	9,190	1,415
Robocall Observatory	6,712,525	3,001,945		
FTC DNC	12,982,167	6,885,768	306,224	1,560
Robocall Observatory	4,356,546	1,942,778		
PPoNE	1,350	1,350	35	1,015
RRAPTOR	948,775	544,129		
FTC DNC	1,349,281	281,582	21,225	143
RRAPTOR	948,775	544,129		
PPoNE	1,350	1,350	9	0
FTC DNC	9,772,625	5,764,999		
IRS APWG	111,053	92,281	33,749	1,096
FTC DNC	8,365,227	4,695,484		
PPoNE	1,350	1,350	75	1,015
IRS APWG	30,734	26,039		
PPoNE	1,350	1,350	11	1,015

TABLE 3: Caller ID overlap across data feeds

were observed in both honeypots (Robocall Observatory and RRAPTOR) over a span of about 5 months. About 3% of caller IDs seen in RRAPTOR were also seen in Robocall Observatory. However, over 10% of the caller IDs seen in Robocall Observatory were also seen in RRAPTOR. This indicates RRAPTOR has a broader view of the robocalling ecosystem than Robocall Observatory.

**5.2.5. Robocall Observatory and IRS APWG.** A total of 9,190 distinct phone numbers were found in both Robocall Observatory and IRS APWG, which made up 3% of the unique phone numbers in the IRS APWG feed, but 0.3% of the unique phone numbers in Robocall Observatory. Although the overlap period spanned more than three years, there was low overlap between these two feeds.

**5.2.6. FTC DNC and IRS APWG.** A total of 33,749 unique phone numbers were shared between the feeds. This translates to 37% of the unique NANP phone numbers in the IRS APWG, but less than 1% of the unique phone numbers in the FTC DNC. About 90% of the common phone numbers across these two feeds were first found in the FTC DNC. Since the IRS APWG feed is curated towards tax scams, IRS impersonation scams, and other tax-related scams, we explored the overlap between the category selected by consumers when reporting to the FTC DNC portal. The subject for over 40% of the reports that matched in the FTC DNC were labeled as “Calls pretending to be government, businesses, or family and friends”.

### 5.2.7. Key findings and takeaways.

In a 2018 study by Pandit et. al. [4], researchers applied a sophisticated methodology to design and characterize a caller ID blocklist. While we did not fully replicate this methodology, we can present an upper bound on effectiveness of caller ID blocking using various data feeds. The intuition

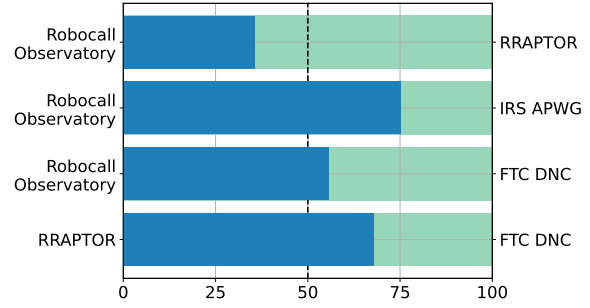


Figure 9: Honeypots consistently uncover the reported NANP phone before crowd-sourced feeds.

is that the best a blocklist source can do is to receive a call from a unique caller ID, which can then be used to block all future calls originating from the same caller ID. Therefore, the upper bound for a single vantage point call blocking approach is the number of calls that matched a caller ID within the blocklist.

**Finding 10:** *Caller ID blocking is only effective when a blocklist is derived from a specific feed and applied to block calls made to the same feed.* A blocklist curated from FTC DNC feed is 21.1% effective when applied to block calls made to Robocall Observatory, and is 12.1% effective when applied to block calls made to RRAPTOR. This is an upper bound because it assumes that the FTC DNC observed the call before the call was made to either of the honeypots. However, we found FTC DNC lags honeypots and often reported a caller ID after they were seen in honeypots. A blocklist curated from RRAPTOR when applied to block calls made to Robocall Observatory is only 16.3% effective. Similarly, a blocklist curated from Robocall Observatory when applied to block calls made to RRAPTOR is 9.8% effective.

For comparison, blocking calls from caller IDs seen within the same feed would result in a blockrate of 42.6% for Robocall Observatory, and 55.3% for RRAPTOR. These findings are consistent with findings from the 2018 study [4], where a block list was created and used to block future unwanted calls made to the same set of phone numbers. This means that cross-source blocking of calls based on the caller ID is only marginally effective.

**Finding 11:** *Honeypots are able to capture active campaigns earlier than crowd-sourced consumer complaints.* Anecdotally, it is known that robocalling campaigns are observed first in honeypots before frustrating consumers to a point where they report these calls to the FTC’s DNC or their respective service providers. These reports often lead to tracebacks, and enforcement actions are taken by regulator agencies based on traceback results. From Figure 9, we see that in most of the cases honeypots observe the reported caller IDs even before they are reported by consumers. Therefore, honeypots serve as early-visibility vantage points to track emerging robocall campaigns.

### 5.3. Campaign Overlap Analysis

In this subsection, we analyze the overlap between the campaigns seen across Robocall Observatory, RRAPTOR, and PPONE — the three data sources that include call audio, as described in Table 2.

**5.3.1. Non-Interactive Campaigns seen in both Robocall Observatory and RRAPTOR.** We compare representative transcripts of the campaigns seen in both Robocall Observatory and RRAPTOR to identify the overlap between them, as described in Section 3.3. Robocall Observatory and RRAPTOR overlapped for 143 days; for that timewindow, we process audio data to identify audio campaigns in Robocall Observatory and RRAPTOR, independently. We compare representative transcripts from each campaign in Robocall Observatory and RRAPTOR, tokenize and pre-process the transcript, discard short transcript calls, compute the token-level Jaccard similarity and identify matching campaigns. We find that 18.2% of RRAPTOR clustered calls (6% of campaigns) were also found in Robocall Observatory; while from the vantage point of Robocall Observatory, 16.6% of clustered calls (14.25% of campaigns) were seen in RRAPTOR.

By using the token-level Jaccard similarity as the distance metric, we ensure that calls and campaigns matched are identical or near-identical to each other. Therefore, the overlap represents the campaigns that were seen by both the interactive and non-interactive honeypot, and behaved similarly in both honeypots. This overlap represents the fraction of calls that do not adapt their behavior based on the honeypot’s engagement level.

**5.3.2. Interactive and Non-Interactive campaigns seen in RRAPTOR and Robocall Observatory.** We quantify the volume of interactive and non-interactive campaigns seen in RRAPTOR and Robocall Observatory by following a similar approach, except that we use the token-level Longest Common Subsequence (LCS) similarity metric to compare the transcripts. By using token-level LCS, we meet two requirements: (1) we ensure that identical calls are still flagged as identical, (2) we can now match calls that begin with the same script, but wait for a response from the honeypot in the middle of the call before engaging further (example in Figure 4).

We find that 33.2% of RRAPTOR calls (10.5% of campaigns) were also found in Robocall Observatory; while from the vantage point of Robocall Observatory, 20.7% of calls (19.24% of campaigns) were also found in RRAPTOR. These overlap numbers represents the collection of interactive and non-interactive robocall campaigns seen in both RRAPTOR and Robocall Observatory. In both methods of identifying campaigns across feeds, the overlap between RRAPTOR and Robocall Observatory is higher when looking at calls as part of campaigns, rather than caller-IDs.

**5.3.3. Campaigns seen in Robocall Observatory that were subject to enforcement actions.** To understand the extent to

which the campaigns in Robocall Observatory were subject to enforcement actions, we compare the campaigns seen in Robocall Observatory with the calls in the PPONE data. Although the PPONE data doesn’t represent all enforcement actions, it does provides the first lower bound estimates of campaigns seen in Robocall Observatory that were subject to enforcement action. We process the calls in the PPONE data to identify campaigns using our audio campaign detection pipeline, described in Section 3.1. Each call in the PPONE data represents a traceback. Further, each call is a robocall (and not a misdial or a human call). Therefore we do not discard the outliers and treat the outliers as separate campaigns. Within a time period of over 2 years (1,015 days), we find that 5.5% calls in Robocall Observatory (19,735/360,723) were subject to enforcement action through the PPONE initiative. This represents 4.4% (1,925/43,822) of campaigns seen in Robocall Observatory.

**5.3.4. Campaigns that continue to operate despite enforcement actions.** The last reported call that was subject to enforcement action via the PPONE initiative was on Oct 11th 2023. However, a small fraction of illegal campaigns that were previously subject to enforcement actions continue to operate in 2024 and can be seen in commercial honeypots. In Robocall Observatory, we find that 0.5% calls (53/9,548 clustered calls) and 0.49% (8/1,621) campaigns active in 2024 were subject to enforcement actions via the PPONE initiative. In RRAPTOR, we find that 1.4% calls (1,448/103,947 clustered calls) and 0.62% campaigns (94/15,084) active in 2024 were subject to enforcement actions via the PPONE initiative.

**5.3.5. Takeaways from Campaign Overlap Analysis.**

**Finding 12:** *Relying entirely on caller ID is an ineffective strategy to characterize illegal robocalling campaigns.* The relative overlap of caller ID across different data feeds are low, as seen in Figure 12 and Figure 11. However, the relative overlap of calls based on campaign overlap analysis is much higher, as seen in Figure 14 and Figure 11. This indicates that the fraction of activity captured by merely relying on caller ID is a much smaller fraction of total activity captured using campaign overlap analysis. Therefore, caller ID is a useful feature to track illegal robocalls only when used in conjunction with other features or in tracebacks. Importantly, caller IDs need to be understood as ephemeral and not as a persistent feature of robocalling campaigns.

**Finding 13:** *On average, Robocall Observatory observed a call from a campaign that was subject to enforcement action about 387 days earlier.* We compute the time difference between the first call in a PPONE campaign that was subject to enforcement action and the first call in the same campaign that was seen in Robocall Observatory. Out of a total of 101 PPONE campaigns seen in Robocall Observatory, 96 were seen first in Robocall Observatory. For these 96 campaigns, we find that the average time difference between the first call in a PPONE campaign and the first call in the same campaign seen in Robocall Observatory is 387 days.



**Finding 14:** *Illegal robocalling operations establish resilient operations by using multiple carriers to route their calls into the phone network.* In Section 5.3.3, we uncovered campaigns in Robocall Observatory that were subject to enforcement actions. The traceback data within the enforcement actions provides the details about the originating carrier that was responsible for routing the calls into the phone network. Of the 101 PPoNE campaigns seen in Robocall Observatory, we find that these campaigns used an average of 1.62 different carriers to route calls into the phone network. Such operational strategies by illegal robocall originators enables them to continue operating despite the best efforts from enforcement agencies to block such calls by sending cease and desist orders to specific carriers using traceback data.

The most resilient campaign seen in Robocall Observatory was distributing their traffic across 10 different carriers. This campaign claimed to be from the Disconnection Department of an Electricity Company, warning the recipient about a disconnection of service within 30 minutes due to non-payment on the account. The recipient was urged to press 1 to speak with a representative to avoid the disconnection.

The second most resilient campaign seen in Robocall Observatory was using 5 different carriers to route calls into the phone network. This campaign was impersonating Amazon and warning the recipient about a fictional order of Apple iPhone placed using their Amazon account, urging them to engage with the call to authorize or cancel the order.

**Finding 15:** *Illegal Robocalling Campaigns are persistent and continue to operate despite enforcement actions.* In Section 5.3.4, we found that most campaigns that were subject to enforcement actions via the PPoNE initiative stopped operating after the enforcement actions. However, a small fraction of campaigns continue to operate despite enforcement actions. By manually inspecting the active campaigns, we find that most of these active campaigns impersonate Amazon, and therefore are obvious scams. The fact that these campaigns continue to operate despite enforcement actions suggest that these operations expect to be shut down by their respective carriers, and use alternate carriers to continue their operations. Therefore, enforcement actions are crucial to shut down illegal operations. However, we also need technical solutions that are swift and efficient in protecting consumers from illegal robocalls.

## 6. Discussion and Recommendations

*Stakeholders need systematic and uniform campaign and topic classification:* Outside of academic literature, the construct of a “call campaign” is undefined. In academic usage, only virtually identical audio comprises a campaign. We found that human-labeled campaign identifiers in the PPoNE dataset mixed constructs. In some cases labelers mixed call recordings with thematically-similar but otherwise-different audio. We suspect that most fraud teams work with a similar mixed-construct model. Data sharing and collaboration is virtually impossible without a common model of robocalling. By starting with a rigorous definition of campaigns, we

were able to effectively compare campaigns across multiple data sources and modalities.

Similarly, content classification is also ad-hoc. While the PPoNE labelers had free-text fields to write what seemed appropriate, the FTC DNC complaint database is limited to those that were derived at the beginning of data collection. As the scam landscape has changed, so have our needs and what makes sense as an organizing criteria. The top category consumers selected when reporting calls that we observed in RRAPTOR and Robocall Observatory was “Others”. More robust and updated categories can enable tracking of the types of scams active in the ecosystem from consumer complaints.

Tracking illegal robocalls using caller ID alone has limitations, as discussed in Section 5.3 and Section 5.2, but audio-based campaign comparison is more effective. The FTC DNC feed should enable consumers to upload audio recording examples or voicemail examples. Leveraging automated audio analysis techniques similar to the ones presented in this work can minimize manual analysis time and enable swift action against bad actors.

*Stakeholders need tools to accelerate enforcement against telephone abuse:* Enforcement actions are key to combat illegal robocalls [50]. As seen in Section 5.3.4, enforcement actions ended virtually all relevant campaign activity. The challenge is that the technical and legal frameworks for doing so have high latency and low throughput. Investigators seem to be doing a good job prioritizing the top offenders, such as PPoNE targeting 5.5% of all honeypot traffic with only 31 actions and 1,353 tracebacks. Despite incredible improvements to traceback, only a few hundred per month are feasible with current resources, and they still can take multiple days to trace a single call to its source. Moreover, the average campaign operated for *over a year* before the investigation took place. Stakeholders should look for ways to enable rapid enforcement based on early indicators available from honeypots. The recent 24-hour traceback requirements are a small yet promising step forward [51].

*Robocall mitigation strategies should be driven by data-driven characterization.* The findings in Section 5 demonstrate how the robocalling ecosystem adapts to changes intended to combat illegal robocalls. The deployment of STIR/SHAKEN, a call authentication framework intended to curb caller ID spoofing, has resulted in a growing trend of legitimate phone numbers being used to generate bulk robocalls that are authenticated by the originating carriers [52]. Carriers need to limit robocall originator’s access to legitimate phone numbers and restrict their ability to generate illegal robocalls that are authenticated. To this end, developing and adhering to robust Know Your Customer and Know Your Upstream Provider policies and practices are crucial.

*Deployment of interactive honeypots to study “smarter” robocall campaigns.* Interactive robocalling campaigns continue to become more prevalent. Easy-to-use generative-AI based audio services will likely accelerate this change. Widely deployed honeypots must adapt to this changing landscape. Researchers, anti-robocall product vendors, carriers and other stakeholders should develop better tools and techniques to study such interactive campaigns.

## 7. Acknowledgements

The authors would like to thank the shepherd and anonymous reviewers for their helpful comments. We thank Bandwidth Inc. for providing VoIP service and phone numbers for the Robocall Observatory. We also thank the APWG, David Frankel and ZipDX for sharing the data that made this study possible. This material is based upon work supported by the National Science Foundation under grant number CNS-2142930. This paper was partially supported by funds from the 2020 Facebook Internet Defense Prize. We would like to thank Abhinaya for her helpful suggestions. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, other funding agencies or financial supporters.

## References

- [1] “Over half of the calls people receive now are spam — zdnet.com,” <https://www.zdnet.com/article/over-half-of-the-calls-people-receive-now-are-spam/>, [Accessed 07-06-2024].
- [2] FCC, “TRACED Act,” <https://www.fcc.gov/TRACEDAct>, 2023, [Accessed 05-06-2024].
- [3] S. Prasad, E. Bouma-Sims, A. K. Mylappan, and B. Reaves, “Who’s Calling? Characterizing Robocalls through Audio and Metadata Analysis,” *USENIX Security Symposium*, 2020.
- [4] S. Pandit, R. Perdisci, M. Ahamad, and P. Gupta, “Towards Measuring the Effectiveness of Telephony Blacklists,” *NDSS*, 2018.
- [5] FCC, “Combating Spoofed Robocalls with Caller ID Authentication,” <https://www.fcc.gov/call-authentication>, 2023, [Accessed 05-06-2024].
- [6] “7. Industry Traceback 2023 and Beyond - SIP Forum — sipforum.org,” <https://www.sipforum.org/download/7-industry-traceback-2023-and-beyond/>, [Accessed 07-06-2024].
- [7] H. Tu, A. Doupe, Z. Zhao, and G.-J. Ahn, “Toward authenticated caller ID transmission: The need for a standardized authentication scheme in Q.731.3 calling line identification presentation,” *2016 ITU Kaleidoscope: ICTs for a Sustainable World (ITU WT)*, 2016.
- [8] —, “Toward Standardization of Authenticated Caller ID Transmission,” *IEEE Communications Standards Magazine*, 2017.
- [9] B. Reaves, L. Blue, and P. Traynor, “AuthLoop: End-to-End Cryptographic Authentication for Telephony over Voice Channels,” *USENIX Security Symposium*, 2016.
- [10] B. Reaves, L. Blue, H. Abdullah, L. Vargas, P. Traynor, and T. Shrimpton, “Authenticall: Efficient identity and content authentication for phone calls,” 2017.
- [11] Y. Chen, Y. Wang, Y. Wang, M. Li, G. Dong, and C. Liu, “CallChain: Identity Authentication Based on Blockchain for Telephony Networks,” *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 2021.
- [12] C. Du, “UCBlocker: Unwanted Call Blocking Using Anonymous Authentication,” 2023.
- [13] H. Mustafa, W. Xu, A. Sadeghi, and S. Schulz, “You Can Call but You Can’t Hide: Detecting Caller ID Spoofing Attacks,” *2014 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, 2014.
- [14] —, “End-to-End Detection of Caller ID Spoofing Attacks,” *IEEE Transactions on Dependable and Secure Computing*, 2016.
- [15] V. Balasubramaniyan, A. Poonawalla, M. Ahamad, M. T. Hunter, and P. Traynor, “PinDrOp: Using single-ended audio features to determine call provenance,” *CCS ’10*, 2010.
- [16] H. Sengar, “VoIP Fraud: Identifying a Wolf in Sheep’s Clothing,” *CCS*, 2014.
- [17] D. Adei, V. Madathil, S. Prasad, B. Reaves, A. Scafuro, “Jäger: Automated Telephone Call Traceback,” *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (ACM CCS)*, 2024.
- [18] M. Sahin, M. Relieu, and A. Francillon, “Using chatbots against voice spam: Analyzing Lenny’s effectiveness,” *SOUPS*, 2017.
- [19] S. Pandit, K. Sarker, R. Perdisci, M. Ahamad, and D. Yang, “Combating Robocalls with Phone Virtual Assistant Mediated Interaction,” *USENIX Security Symposium*, 2023.
- [20] V. A. Balasubramaniyan, M. Ahamad, and H. Park, “CallRank: Combating SPIT Using Call Duration, Social Networks and Global Reputation,” 2007.
- [21] H. K. Bokharaei, A. Sahraei, Y. Ganjali, R. Keralapura, and A. Nucci, “You can SPIT, but you can’t hide: Spammer identification in telephony networks,” *2011 Proceedings IEEE INFOCOM*, 2011.
- [22] N. Jiang, Y. Jin, A. Skudlark, W.-L. Hsu, G. Jacobson, S. Prakasam, and Z.-L. Zhang, “Isolating and analyzing fraud activities in a large cellular network via voice call graph analysis,” *MobiSys ’12*, 2012.
- [23] H. Li, X. Xu, C. Liu, T. Ren, K. Wu, X. Cao, W. Zhang, Y. Yu, and D. Song, “A Machine Learning Approach to Prevent Malicious Calls over Telephony Networks,” *2018 IEEE Symposium on Security and Privacy (SP)*, 2018.
- [24] J. Liu, B. Rahbarinia, R. Perdisci, H. Du, and L. Su, “Augmenting Telephone Spam Blacklists by Mining Large CDR Datasets,” *ACM Asia Conference on Computer and Communications Security*, 2018.
- [25] D. Ucci, R. Perdisci, J. Lee, and M. Ahamad, “Towards a Practical Differentially Private Collaborative Phone Blacklisting System,” *ACSAC*, 2020.
- [26] Y.-S. Wu, S. Bagchi, N. Singh, and R. Wita, “Spam detection in voice-over-IP calls through semi-supervised clustering,” *2009 IEEE/IFIP International Conference on Dependable Systems & Networks*, 2009.
- [27] H. Tu, A. Doupe, Z. Zhao, and G.-J. Ahn, “Users Really Do Answer Telephone Scams,” *USENIX Security Symposium*, 2019.
- [28] I. G. Harris, A. Derakhshan, and M. Carlsson, “A Study of Targeted Telephone Scams Involving Live Attackers,” *Lecture Notes in Computer Science*, 2020.
- [29] I. N. Sherman, J. Bowers, K. McNamara, J. Gilbert, J. Ruiz, and P. Traynor, “Are You Going to Answer That? Measuring User Responses to Anti-Robocall Application Indicators,” *Network and Distributed System Security Symposium*, 2020.
- [30] S. I. Hashmi, N. George, E. Saqib, F. Ali, N. Siddique, S. Kashif, S. Ali, N. Bajwa, and M. Javed, “Training Users to Recognize Persuasion Techniques in Vishing Calls,” *CHI Extended Abstracts*, 2023.
- [31] A. D. Keromytis, “A survey of voice over ip security research,” in *International Conference on Information Systems Security*. Springer, 2009, pp. 1–17.
- [32] “m3aawg.org,” [https://www.m3aawg.org/sites/default/files/document/M3AAWG\\_Telephony\\_Honeypots\\_BP-2014-08.pdf](https://www.m3aawg.org/sites/default/files/document/M3AAWG_Telephony_Honeypots_BP-2014-08.pdf), [Accessed 07-06-2024].
- [33] P. Gupta, B. Srinivasan, V. Balasubramaniyan, and M. Ahamad, “Phoneypt: Data-driven Understanding of Telephony Threats,” *NDSS*, 2015.
- [34] S. Prasad, T. Dunlap, A. Ross, and B. Reaves, “Diving into robocall content with SnorCall,” in *32nd USENIX Security Symposium (USENIX Security 23)*. Anaheim, CA: USENIX Association, Aug. 2023, pp. 427–444. [Online]. Available: <https://www.usenix.org/conference/usenixsecurity23/presentation/prasad>



- [35] “patrickvonplaten/wavlm-libri-clean-100h-base-plus · Hugging Face — huggingface.co,” <https://huggingface.co/patrickvonplaten/wavlm-libri-clean-100h-base-plus>, [Accessed 07-06-2024].
- [36] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An asr corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.
- [37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf)
- [38] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [39] S. A. Leland McInnes, John Healy, “Benchmarking Performance and Scaling of Python Clustering Algorithms,” [https://hdbscan.readthedocs.io/en/latest/performance\\_and\\_scalability.html](https://hdbscan.readthedocs.io/en/latest/performance_and_scalability.html), 2016, [Accessed 06-06-2024].
- [40] “Open ASR Leaderboard - a Hugging Face Space by hf-audio — huggingface.co,” [https://huggingface.co/spaces/hf-audio/open\\_asr\\_leaderboard](https://huggingface.co/spaces/hf-audio/open_asr_leaderboard), [Accessed 08-10-2024].
- [41] “openai/whisper-medium · Hugging Face — huggingface.co,” <https://huggingface.co/openai/whisper-medium>, [Accessed 07-06-2024].
- [42] Sathvik Prasad and Bradley Reaves, “Robocall Audio from the FTC’s Project Point of No Entry,” North Carolina State University, Tech. Rep. TR-2023-1, Nov 2023.
- [43] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/0377042787901257>
- [44] T. Caliński and J. Harabasz, “A dendrite method for cluster analysis,” *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.
- [45] U. C. Bureau, “Nearly 68 Million People Spoke a Language Other Than English at Home in 2019 — census.gov,” <https://www.census.gov/library/stories/2022/12/languages-we-speak-in-united-states.html>, [Accessed 07-06-2024].
- [46] “What Is RRAPTOR? - Legal Calls Only — legalcallsonly.org,” <https://legalcallsonly.org/what-is-rraptor/>, [Accessed 07-06-2024].
- [47] “APWG | Unifying The Global Response To Cybercrime — apwg.org,” <https://apwg.org/>, 2003, [Accessed 05-06-2024].
- [48] FTC, “As Nationwide Fraud Losses Top \$10 Billion in 2023, FTC Steps Up Efforts to Protect the Public,” <https://www.ftc.gov/news-events/news/press-releases/2024/02/nationwide-fraud-losses-top-10-billion-2023-ftc-steps-efforts-protect-public>, 2024, [Accessed 04-06-2024].
- [49] TransNexus, “STIR/SHAKEN statistics from April 2024,” <https://web.archive.org/web/20240528131945/https://transnexus.com/blog/2024/shaken-statistics-april/>, [Accessed 06-06-2024].
- [50] FCC, “FCC EB Issues First Ever C-CIST Classification — fcc.gov,” <https://www.fcc.gov/document/fcc-eb-issues-first-ever-c-cist-classification>, 2024, [Accessed 06-06-2024].
- [51] “FCC Adopts Expanded Robocall Rules That Require One-Day Traceback Support from All Voice Service Providers and Expand Traffic Blocking Obligations for Originating Providers,” <https://www.fcc.gov/document/fcc-further-expands-its-robocall-blocking-rules>, [Accessed 08-10-2024].
- [52] Joshua Bercu, Jessica Thompson, “Enforcement Bureau Requests Information on the Status of Private-Led Traceback Efforts of Suspected Unlawful Robocalls,” <https://www.fcc.gov/ecfs/document/105010043202095/1>, 2024, [Accessed 08-10-2024].

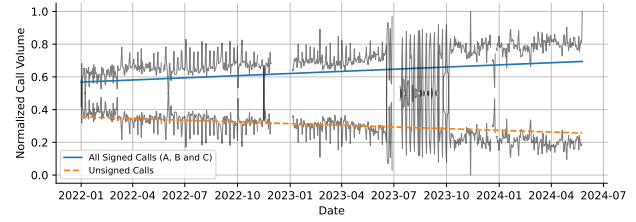


Figure 10: Calls received by Robocall Observatory Signed with STIR/SHAKEN from 2019 to 2024 (Normalized by total calls received per day)

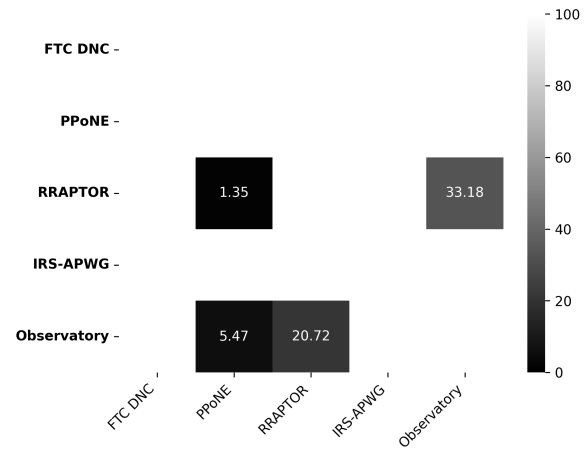


Figure 11: Heatmap of call overlaps of specific feeds relative to other feeds

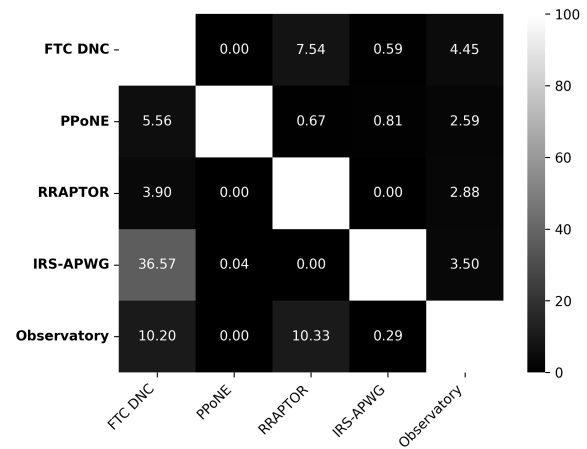


Figure 12: Heatmap of Caller ID Overlap Between Various Feeds

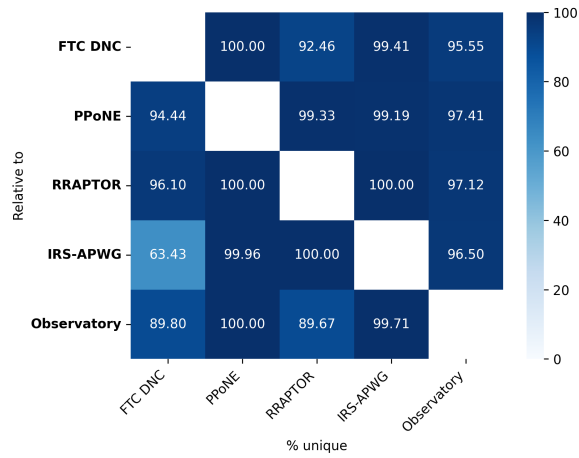


Figure 13: Heatmap of Caller ID Unique To Specific Feeds Relative to Other Feeds

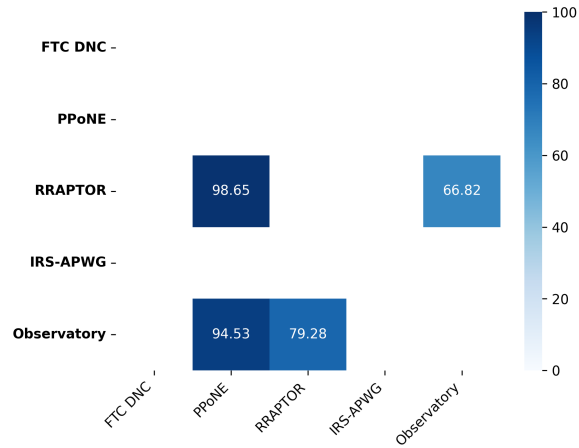


Figure 14: Heatmap of calls unique to specific feeds relative to other feeds

$$\left( \frac{\text{Clusters without a misplaced call}}{\text{Total clusters}} \right) \times 100$$

Figure 15: Equation used to calculate Cluster Perfection

$$\text{Mean} \times \left( \frac{\text{Correctly placed calls in the cluster}}{\text{Calls in the cluster}} \right)$$

Figure 16: Equation used to calculate Intra-cluster Precision

## **Appendix A. Meta-Review**

The following meta-review was prepared by the program committee for the 2025 IEEE Symposium on Security and Privacy (S&P) as part of the review process as detailed in the call for papers.

### **A.1. Summary**

This paper presents a characterization of Robocall scams. The paper collects data from multiple vantage points and presents a longitudinal analysis. As part of the analysis, the authors cluster calls using audio embeddings, which allowed them to study robocall campaigns. As part of the characterization and analysis, the authors study overlaps between vantage points across different dimensions, like caller IDs. Overall this paper presents a longitudinal analysis of 3M voice calls over 5 years.

### **A.2. Scientific Contributions**

- Provides a Valuable Step Forward in an Established Field
- Addresses a Long-Known Issue

### **A.3. Reasons for Acceptance**

- 1) This paper provides a valuable step forward in an established field by assisting researchers, regulators, industry operators, and the public in understanding the whole picture and the effectiveness of anti-robocall measures.
- 2) Addresses a Long-Known Issue. Understanding telephony scams is timely and highly important. Even though the paper shows that the amount of robocalls is going down, it is still a nuisance to anyone who owns a phone. This scourge does significant damage to trust in the telephony system, and therefore more insights into this area are important.

### **A.4. Noteworthy Concerns**

- 1) Definition of campaign. The paper defines a campaign as a "set of phone calls that play the same audio message." This notion of "campaign" is strict compared to other domains. A more versatile definition would consider calls with a common objective or from the same actor under the same campaign, even if the audio message differs. While it would be ideal to have a broader definition that accounts for same-source variance, unfortunately, there is no ground truth to evaluate against.
- 2) Having such a strict definition, the paper can guarantee that the call audio came from the same (audio) source — even without ground truth. However, the strict criteria may impact the findings obtained from the analysis of

the campaigns and the scope of this limitation cannot be determined without properly evaluating the recall performance of the clustering.