Repucall: Towards Describing the Reputability of Phonenumbers

Saran Ahluwalia
ssaluwa@ncsu.edu

Abstract

Automated calls, known as robocalls, have significantly degraded the usefulness and trustworthiness of the phone network. Individuals, enterprise organizations, and providers have no effective solution to address this problem. A significant barrier to progress in this space is the lack of visibility into the practices of automated callers.

In this paper, we are the first to apply an unsupervised retrieval technique to crowd-sourced data (COC) in order to describe the authenticity of phone numbers. We consider over 100,000 complaints collected via crowd-sourced efforts (e.g., 800notes.com). Overall, our results show that the proposed generative probabilistic model produces topics that are quantitatively and qualitatively better than matrix decomposition techniques, such as Latent Semantic Analysis (LSA). In this paper, we propose this model as a more efficacious mechanism for authenticating phone numbers. The proposed model identifies malicious phone-numbers adaptively. Additionally, the model labels and ranks numbers by their re-use across different scam campaigns. This last insight, in particular, possesses implications for research combating phishing attacks, account fraud and identity theft.

1 Introduction

The phone voice channel has many vulnerabilities and affords a very appealing attack surface for spam campaigns, robocalls, and phishing scams. Moreover, there have been over seven million complaints made to the FTC, regarding unsolicited calls made to consumers.[1]. There are numerous solutions, such as YouMail [2], and TrueCaller[3] that have surfaced in order to combat these unwanted calls.

Previous state-of-the-art work introduced the first methodology for blacklisting the following data sources: honeypot call detail records (CDR), and Federal Trade Commission public reporting [27]. In this work, the authors leverage latent semantic indexing (LSI) in order to describe their data sets. Through this procedure, the authors do not assess topic coherence [24], but instead revert to manual selection for both the topics and for the optimal number of topics to use for modeling a call transcript. This results in manual analysis of topics and the discarding of topics from a transcript's topic mixture vector. This is due to the inherent issues in LSI: that topic mixture vectors are not humanly readable [20]. Evaluation is possible through finding similar words for each word in the latent space, but are otherwise not human interpretable.

In addition, the previous state-of-the-art work decided on the number of topics based on certain heuristics that rely on domain expertise. A conventional approach is to sort the cumulative singular values in descending order and then a cut-off [21]. However, there is variability between samples; hence manual analysis is often needed for new training.

Within this context, we organized our research questions around the following:

1. What proportion of COC corpora contain a phone number, and if so, does a given complaint contain important information (wireless service provider and geographic location of the source of the call etc.)?

2. Of those numbers from the previous question, are any of these malicious numbers seen across different scam campaigns?

3. Can we use any other information retrieval techniques to measure phone reputation? If so, how do we assess these techniques?

To fill this knowledge gap, in this paper we systematically investigate COC data sources that may be leveraged to automatically learn new facets of crowd-sourced reporting.

In summary, we provide the following contributions:

- We cluster and describe phone numbers with the crowd sourced online complaints. Our analysis uncovers numerous malicious behaviors, including bulk scam and phishing campaigns that share themes with specific phone numbers. Most critically, our analysis shows that numbers, which are being used to deliver malicious scam campaigns, directly undermine mitigations from previous work [17].

- In order to associate phone numbers that are part of a long-running scam campaign, we apply a novel probabilistic generative model on user reported complaints. We then identify the top campaigns from

each data source, and introduce an evaluation metric to assess the coherency of the topic model. This has particular implications for blacklist construction and for systems-level application development that would be consumed by users. Specifically, our model enables keyword-specificity for searching for scam campaigns and for similar complaints associated with a scam campaign theme and geographical location.

- We show that our model is more resilient to threats to internal validity and to the reputability of the results that we are generating from the model. These threats are described, in detail, in Section 3.

The remainder of this paper proceeds as follows. Section 4 describes the implementation of our pipeline (presented in Figure 1.). Section 5 demonstrates various methods to assess the quality of our topic modeling methodology. Section 7 discusses implications for aggregating COC complaint data and shortcomings in our approach. Section 8 describes related work. Section 9 provides concluding remarks.

2 Data Collection

2.1 Crowd-Sourced Online Complaints

This dataset contains the online comments obtained from popular online forums, such as 800notes.com). The dataset contains over 100,000 actual raw complaints filed by users, containing the phone number that made the unwanted call, a timestamp, and a description of the content of the call. The complaints stem from May 10, 2010 until September 17, 2017. Since the comments are entered manually by anxious and frustrated users, the text describing the content of the call is typically comprised of many misspellings, grammatically inaccurate sentences, expletives, and in some cases, irrelevant information. In addition, many of the complaints contain URLs to resources and links to external fraud-prevention resources. Another externality to consider is that the phone number and timestamp provided by the users could be mistyped.

2.2 A Note on the Inclusion of Spam Messaging Feeds

Our original proposal for this research investigation stipulated that we also apply this same model to spam messaging data provided by Project HoneyPot [4]. From this we would evaluate a COC-only trained blacklist, a spam messaging-only trained blacklist and a COC and spam messaging blacklist in terms of unwanted call blocking rates. We would then conduct experiments that show whether or not spam and COC data sources provide additional insights into the longevity of spam campaigns. This would complement insights into both the origin and the reuse of phone numbers for robocalls. Finally, we would compare our blacklists' classifications to two third-party service providers.

However, we were unable to obtain the dataset from Project HoneyPot. In addition, the length of time for constructing, training and evaluating each blacklist's Call Blocking Rates (CBR) - as detailed by S. Pandit et al. [27] - was beyond the allotted time that was available to the researchers that first implemented these experiments. Therefore, we were unable to validate many of these aforementioned hypotheses. We will leave this for future work.

3 Threat Model

Our threat model is grounded in the validity of our classification of specific phone numbers. While it is difficult to ensure the authenticity of numbers, we are assuming that an external adversary is attempting to pollute the model classification task. This is a direct threat to the internal validity of our testing through selection bias of the input samples. For example, complaints associated with a phone number could be flooded with new comments that are favorable or describe the phone number with entirely different sets of words. This causes the model to mis-classify or produce topics that are not human interpretable. This also has another assumption that the adversary understands the model's classification task. Hence we are specifically assuming that the adversary is concerned with undermining the *reputation* of the classification task. In section 5 we propose and we evaluate the resiliency of our model to word intrusion and internal validity threats.

4 Methodology and Design

As depicted in Fig. 1, the main components of our spam campaign detection, analysis and investigation framework are the following:

4.0.1 Persistent Storage

Due to the volume of corpora, a NoSQL database - particularly a document-based database - was used instead of a SQL-based relational database management system (RDBMS). Document-oriented databases were favored for their flexibility and high scalability. However, one caveat is the lack of a powerful query language like SQL. Moreover, the ability to represent relationships between

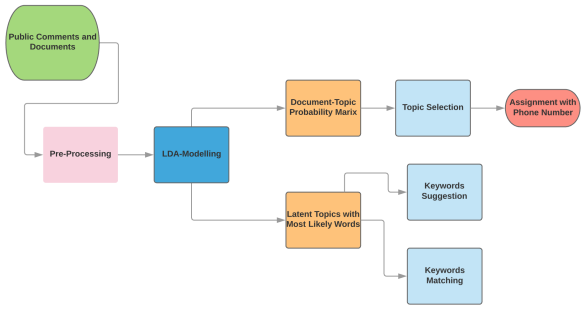

Figure 1: Architecture of pipeline

comments, scam campaigns and domain names is essential. These relationships can be stored and processed efficiently in a graph database. In our proposed system, we employ MongoDB (MongoDB, 2017), which is an open-source NoSQL database [5]. MongoDB has the flexibility of document-based databases as well as the capability of graph-based ones for managing relationships between records. It has the advantages of being performant, scalable and supports SQL-styled queries. Complaints are stored in the database as documents with different fields, such as title, textual content, embedded URLs, etc. Additionally, user information, phone-numbers, domain names and timestamps of the correspondence are also saved as different document types. For each document type, the database maintains a hash table of a specific field to eliminate duplications. The database also stores the relationships between different documents, such as between the forum correspondence, scam campaign, and phone number.

4.0.2 Elimination of Stopwords

Our implementation of Latent Dirchlet Allocation (LDA), and the pre-processing of corpora have been influenced heavily by the seminal work of Blei at al [11]. Specifically, previous work has detailed that one should remove common stop words in a document before running Latent Dirichlet Allocation [12]. Stop words are words that are very common and do not alter the semantic meaning of a sentence. Typical stop words include ”in”, ”the” and ”as” etc. Kim et al. automatically filtered out words from the vocabulary that were present in more than 50 percent or less than 5 percent of the documents [33]. Kim et al. state that this heuristic is an effective way of removing both stop words, misspelled words and non-words. Therefore, in our implementation, a static list of common stop words was used as suggested by Blei et al [12].

4.0.3 Stemming

In this project the NLTK version of the Snowball Stemmer [6] was used. This project is originally based on the stemmer proposed by Lovins [22]. Stemmers do not take into consideration polysemy. For example, the Snowball stemmer stems both ”informational” and ”informality” to the word inform. Because ”space” and ”spacial” do not possess the same semantic meaning there is a risk of losing information when using stemming [32].

However, due to prior works' common practice to use a stemmer when performing natural language processing, all experiments were conducted with the snowball stemmer.

- Phone Number: Each phone number is extracted and parsed into a format that demarcates it's country code, time zone and carrier provider. Initial parsing and validation of the phone number was conducted using Google's open sourced libphonenumber module [7].

- Complaint: We discarded complaints that contained less than 5 words or that only contained URLs. In addition, we only considered complaints that had been ”upvoted” (an indicator for how useful the complaint was to other users). Finally, we discarded any complaints that only contained profanity and/or comments that derided another user. This resulted in a 91 percent reduction in the total number of individual documents that were considered for topic modeling.

4.0.4 Online Learning for Latent Dirichlet Allocation (LDA) Models using (Online Variational Bayes) applied to COC datasets

The Latent Dirichlet Allocation can intuitively be described as a model that identifies topics of documents based on the word frequency distribution in each document. Therefore, the words in a document are dependent on the latent topic distribution. The Latent Dirichlet Allocation relies on the Dirichlet prior that the words in a document are generated based on the topic distribution of the document [12]. If one were to assume that this is the case we can use any method of posterior inference to infer the latent variables in the Latent Dirichlet Allocation model. Both Blei et al. and Kim et al. used an online version of the variational Bayes inference model [33].

We have adopted a similar implementation proposed by Hoffman et al. [20]. Hoffman et al. provides an online version of Latent Dirichlet Allocation. The difference from the original Latent Dirichlet Allocation is that it requires only one pass over the data set. Hoffman et al. define online learning for Latent Dirichlet Allocation as shown in Algorithm 3; definitions of all parameters can be

Define $\rho_t \equiv (\tau_0 + t)^{-\kappa}$
initialise λ randomly
for $t = 0$ *to* ∞ **do**
  E step:
  Initialise $\gamma_{tk} = 1$. (The constant 1 is arbitrary.)
  repeat
    Set $\phi_{twk} \propto \exp E_q[\log(\theta_{tk})] + E_q[\log(\beta_{kw})]$
    Set $\gamma_{tk} = \alpha + \sum_w \phi_{twk} n_{tw}$
  until $\frac{1}{K} \sum_k |$ *change in* $\gamma_{tk}| < 0.00001$;
  M step:
  Compute $\tilde{\lambda}_{kw} = \eta + D n_{tw} \phi_{twk}$
  Set $\lambda = (1 - \rho_t)\lambda + \rho_t \tilde{\lambda}$
end

Algorithm 3: Online learning for Latent Dirichlet Allocation

Table 1: Shows what each symbol in Algorithm 3 represents.

symbol	Definition
α	The hyper-parameter for the Dirichlet distribution θ
θ	The k-dimensional Multinomial parameter that specifies the topic distribution for a document
β	a $k \times V$-matrix where each row describes the probability of words for a specific topic
γ	The variational Dirichlet parameter
ϕ	The variational Multinomial parameter
λ	The variational parameter on β
η	The hyper-parameter for the Dirichlet distribution on β
κ	Controls the rate at which old values of λ are forgotten
τ_0	Slows down early iterations of the Online variational Bayes algorithm
k	Number of topics
D	Size of corpus ($D \to \infty$) in a true online setting [5]

Figure 2: Online Variational LDA Inferential Step

found in the proceeding, sub-table, Table 1. While operating in a similar manner to the "regular" Latent Dirichlet Allocation there are several differences. κ_t is a parameter that adjusts the rate at which prior iterations are "forgotten". τ_0 varies the weight of the early iterations while κ defines the decay rate of previous λ. λ is the result of the previous mini-batches. More precisely, it is the variational parameter on the matrix with word frequency spectrum of all the topics. While Algorithm 3, above, shows that the algorithm can process each document individually, it is common practice to collect a mini-batch of documents and then run the algorithm on all documents in each mini-batch [20] [11]. Furthermore Hoffman et al. show that if and only if κ is within the range of $(0.5, 1]$ convergence is guaranteed [20].

For the COC datasets, we created 10 different Online Variational LDA models by varying the K parameter (i.e. the number of topics) from 1 to 100 and repeating this process five times. The assessment for optimizing the number of topics is included in Section 5.0.1.

In our implementation, the Dirichlet parameters are set to be symmetrical for the smoothing of words within topics, η, $\eta = 1/V$ and topics within the set of documents, K, $\alpha = 1/K$. By keeping $\alpha < 1$, the modes of the Dirichlet distribution are close to the corners, thus favoring just a few topics for every document and leaving the larger part of topic proportions very close to zero. The LDA models are created using the Python library Gensim [36].

Gensim uses variational inference from the online model in order to approximate the posterior [19]. The convergence iteration parameter for the expectation step (i.e. the E-step) is set to 1000; the part where per document parameters are fit for the variational distributions are detailed in Algorithm 2 of the original implementation [19].

Parameters	min	max	step
Mini-batch size	20	100	20
tau	0.5	1	0.1
kappa	0.5	1	0.1
alpha	0.05	0.35	0.05
eta	0.05	0.35	0.05

Figure 3: Empirical evaluation of parameters

5 Evaluation

An experiment was set up to find the parameters that yielded the best measurements on a run with 9,278 complaints. The measurements were made with perplexity - which Blei et al. and Hoffman et al. also used [12] [20].

The parameters used were found by, first, empirically testing different values for each parameter. This led to the set of parameter ranges found in Figure 3. A validation step permuted all combinations of these these parameters on the same 9,278 complaints. The algorithm assessed perplexity after each mini-batch together with the average perplexity and the standard deviation of the perplexity with a decaying window of 50 mini-batches. The best parameters were found in the permutation that yielded the lowest sum of the average perplexity and standard deviation of the perplexity.

5.0.1 Empirically Choosing the Optimal Number of Topics

One of the parameters in LDA is the number of topics K. Optimizing K can be accomplished by measuring the information gain provided in each topic compared to a baseline measures [8]. We use log likelihood, because this empirical measure evaluates how well the corpora fits the model. In this case, this is the topic space model produced by LDA. When performing parameter estimation, a common strategy is to *maximize* the log likelihood [18]. We employ this technique to measure the effectiveness of each LDA model, varying the number of topics K. The results are presented on Figure 4. Notice that the log likelihood is maximal after approximately 10 topics. Hence we chose ten topics as the optimal parameter in our subsequent experiments for classifying complaints.

5.0.2 Evaluating Topic Models

Because topic modeling is an unsupervised task the implementer is not aware of what the topics will be, after training. This makes evaluation of the topic model a very difficult problem. This dilemma is not apparent in supervised learning (for example, logistic regression, classification), in which one compares predicted labels to expected labels. There are no "expected labels" in our the pipeline for topic modeling.

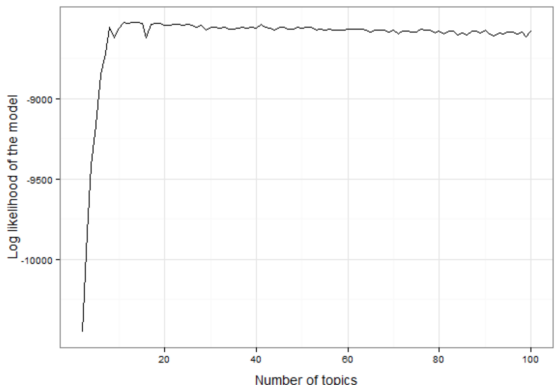

Figure 4: Empirical evaluation of parameters

Moreover, each topic modeling method - for example, LDA and LSA, possesses a unique method for measuring the "internal quality" (perplexity and reconstruction error, for example)[11]. However, what is often not cited in evaluation methods, is that these internal measures are an artifact of the particular approach taken when applying the model. These include, for example, Bayesian methods and matrix factorization. In addition, there is no evidence-based approach to compare such scores across different types of topic models. The most feasible method to assess the quality of unsupervised models is to evaluate how each model can *improve* the superordinate task of the model task for which the implementer is training each collection of models for.

For example, when our goal is to retrieve semantically similar documents, one can manually tag a set of similar documents and then assess how well a given semantic model maps those similar documents together.

Such manual tagging can be resource intensive. Wallach et al. suggest a "word intrusion" method that works well for models where the topics are meant to be "human interpretable", such as LDA [13]. For each trained topic, they obtain that topic's first ten words, then substitute one of those words with another, randomly chosen word (the intruder) and see whether a human can reliably tell which one it was.

If so, the trained topic is topically coherent; alternatively, if the topic does not have a discernible theme, the model is considered to be poorly fitted.

We performed this evaluation as follows:

- Select the 50 words for each of of the 10 LDA topics
- Collect all 50 words from all 10 topics, as one set
- For each of the 10 topics we replace a word at a different index
- We then split each document (complaint) into two parts, and verify that first, topics of the first half of the document are similar to topics of the second half of the document; and second, each of half of each respective complaint is mostly dissimilar with other complaints.

  In order to benchmark our results, we applied the aforementioned result on the same set of documents as were trained using LDA, using LSI.

  We then tested our evaluation method on 1000 documents that were not used in either LDA or LSI training. Our results are presented on Table 1 and on Table 2. We used the cosine similarity to measure first, the similarity between corresponding documents (first row) and second, a randomized selection of 10,000 halves (second row). A higher cosine similarity between corresponding documents is considered to be an attribute of a better-fitted model. Conversely, a lower cosine similarity score between randomized documents are attributes of a better-fitted topic model [13].

Table 1: LDA Results

Model	**Average cosine similarity**
Corresponding parts	0.776225069646
Randomization	0.254734527925

Table 2: LSI Results

Model	**Average cosine similarity**
Corresponding parts	0.606533434328
Randomization	0.0748434974254

We should note, at this juncture, that this does not completely address the security threat to the internal validity of our model. We understand that within a more volatile environment, there is no guarantee as to what the inter-topic distance will be, nor will we be able to control the the quality of data that is ingested by our proposed pipeline.

5.0.3 Evaluating Coherence

In order to further benchmark our online model to other topic modeling implementations, we compared the following topic models coherence measures [28]. The following models were compared:

- LSI (Latent Semantic Indexing)
- HDP (Hierarchical Dirichlet Process)
- LDA (Latent Dirichlet Allocation)

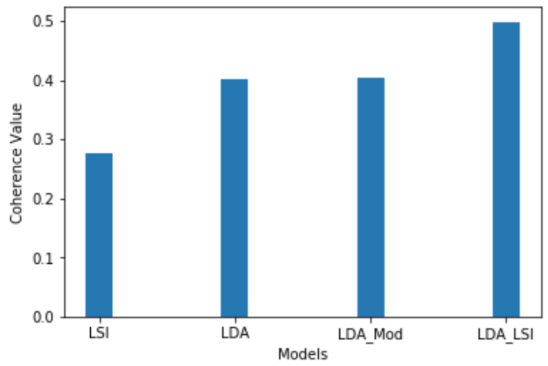

Figure 5: Coherency Analysis for Justifying choice for the use of LDA over other model variants

- LDA (optimized to find optimal number of topics - (LDA Mod))
- Projecting LDA into lower dimensional space (LDA LSI)

  Each model was parameterized to 10 topics. We used the CV measurement, due to the fact that CV has been shown to resemble most closely human measurements and ratings. The details of the CV measurement are described in detail by Both et al. [28]. It should be noted that a higher coherence value, closer to 1, is associated to more human - interpretable topics. Our results indicated that LDA, and its variants, scored above 0.4 versus LSI (0.28). This is agreement with our previous evaluation evaluating the similarity exhibited between corresponding document halves for LDA.

6 Analysis of Phone Numbers

After applying the model described in Section 4.0.4, we now have a set of topics, Z, that are labeled with a relevant campaign theme, and we aim to do two things:

1. Decide what source numbers from the online complaints should be be considered to be labeled as scam.
2. Leverage the topic assignments with their corresponding complaint (and hence the phone number(s) extracted from the complaint) to group together complaints and related source phone numbers that likely belong to the same spam campaign.

   To this end, one can tune the LDA model to perform assignments to estimate two things:

```
Topic 0: company service spam read operation spoofed day free protection system
Topic 1: contact address process charge amount receive office security transaction copy
Topic 2: send fraud check company loan agency fee pay western union
Topic 3: card gift code buy received won enter claim free bottom
Topic 4: country city website service code domain cell server site address
Topic 5: federal complaint trade tax commission contact irs file pay card
Topic 6: won email mobile received pound nokia claim send country prize
Topic 7: moved email reply sale product address difficult client item mail
Topic 8: correct site box time posted helpful start story empty proper
Topic 9: received time accent day hung answer guy irs computer pay
```

Figure 6: Top 10 thematic topics for 9728 complaints

3. The topic mixtures of each document (by counting the proportion of words assigned to each topic within that document).
4. The words associated to each topic (by counting the proportion of words assigned to each topic overall).

From the aforementioned heuristic, a distribution of topics, for each document, is constructed. From this distribution, we select the topic with the highest probability for a given complaint. The calls with similar textual context can be clustered together to discover spam campaigns (hence, the phone numbers related to the campaigns). Figure 6. illustrates a generated set of topics after convergence of the model. After associating each topic with each complaint we achieved the following discrete statistics (displayed on Table 3).

Table 3: Statistics (Post-LDA)

Variable	**Count**
Complaints	9278
Unique Phone Numbers	1851
Phone numbers with $>$ 1 scam association	123

In particular, we observed that only 6 percent of the phone numbers were reused for another scam campaign (Table 3). This observation supports earlier hypotheses that the phone numbers used by scammers in their campaigns are less diverse [15]. In particular purchasing new phone numbers is more expensive than the purchase of other mediums of deploying scam campaigns, such as email.

6.0.1 Scam Campaign Distributions

We clustered phone numbers by shared campaigns. Figure 7 displays the varying distributions across all scam campaign themes generated by the online LDA model.

In the event that a phone number was associated with more than two scam campaigns we selected the scam campaign topic that was most often associated with the phone number. 25.6 percent of the phone numbers were associated with technical support campaigns. 18 percent of numbers were associated with gift card scams. It should be noted that the content of these complaints for this group were associated with Nokia and Best Buy Gift Card

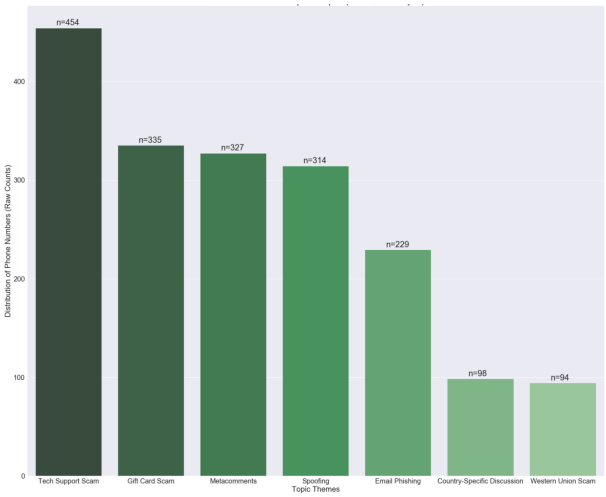

Figure 7: Top 10 thematic topics for 9728 complaints

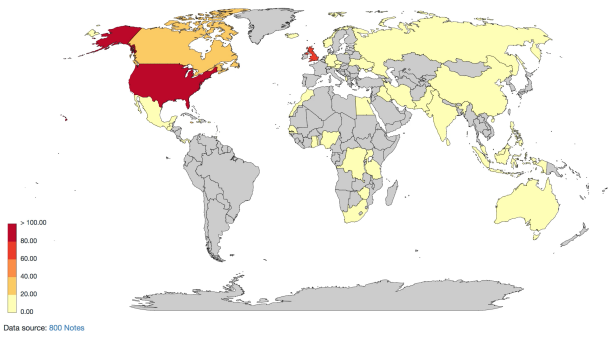

Figure 8: Distribution of Source Phone Numbers by Country

scams. All scam campaigns associated with Nokia were cited by victims in the United Kingdom. Surprisingly, we discovered that many themes were associated with what we define as meta comments: comments discussing legal policy and discussion regarding spoofing and challenges associated with discovering scammers (spoofing numbers, for example).

6.0.2 Geographic Distribution of Source Phone Numbers

We observed that over 95 percent of source phone numbers that were extracted were sourced to the United States, Canada and the United Kingdom (Figure 8.). This characterization, noted as sampling bias, is elaborated on in section 7.

Within the United States (Figure 9.), our analysis discovered that the normalized distribution of complaints that cited malicious phone numbers were primarily represented by North Carolina, Oregon, Nevada, New Mexico,

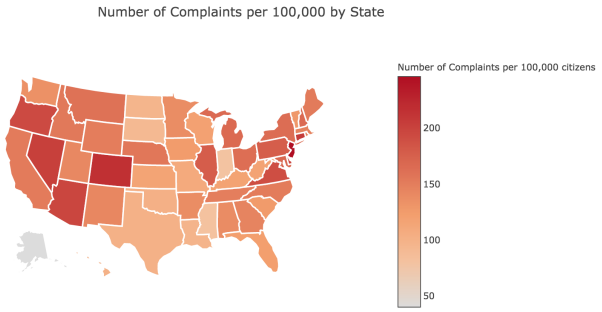

Figure 9: Distribution of Source Phone Numbers within the United States

Illinois and the District of Columbia. We discuss further implications for further research in section 7.

6.0.3 Service Providers

We further characterized numbers by their wireless service provider. An interesting observation is that certain operators are used more often than others to register scam numbers. Figure 10 shows the distribution of phone numbers used by scammers among the providers. We observe that, in our dataset, the top 4 operators (out of 32) provide more than 65 percent of fraud-related numbers. These four cell providers, in particular, represent one of the largest telecom service providers in Africa (MTN), and the Caribbean and Asia-Pacific (Digicel), respectively. After further investigation we discovered that these service providers offer one or more of the following services:

- Have an online registration service
- Offer low-cost or free international call forwarding
- Provide bundled packages for wireless services, along with subsidized cash-back call services

Although we did not confirm this, independently, we hypothesize, based on prior evidence, that communities of scammers find these service providers appealing [10].

7 Discussion and Limitations of Findings

Our analysis alone is not sufficient to draw complete conclusions. For instance, we are still uncertain how prevalent phone number re-usability is. Given that 6 percent of phone numbers are used across different scam campaigns begs the following question:

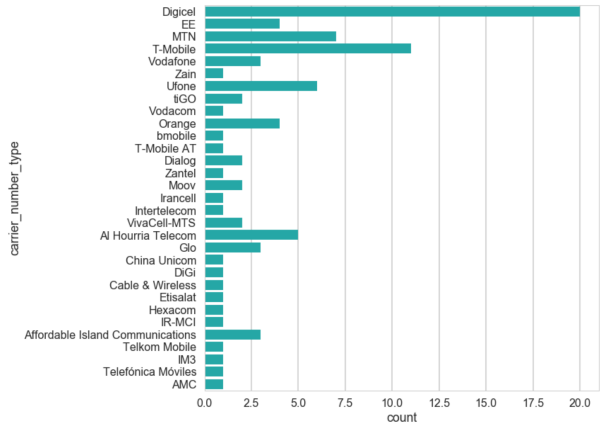

Figure 10: Wireless Service Providers

1. Does this observation imply that the remaining numbers are discarded or does this observation imply that they are simply not reported by users on that particular forum?

This question lends itself to further scrutiny into our results. Most saliently, we note that there is inherent bias in our sampling methods. 800notes.com, in particular, represents user complaints from North America and the United Kingdom. Therefore, our sample data set does not capture the full breadth of source phone numbers used by communities of scammers. In future work, we would like to diversify our data sources in order to include other public complaint forums.

Moreover, our methods for selecting valid complaints will also need to be revised. For example, due to the fact that English is predominantly represented in these forums, we do not account for other languages in our analysis. Therefore, countries where scam campaigns are ubiquitous, but under-reported are not representative of our sample. In addition, it is possible that the discarded complaints were provided by non-English speakers that do not possess English fluency to elaborate or provide detailed accounts of their experiences. Future work will include countries that our initial analysis of numbers did not account for. Namely, we will include COC datasets from China, India, the United Kingdom, and from Central America.

We also understand that during training, only a subset of source phone numbers in the COC datasets can be attributed to distinguishable campaigns. However it is a worthwhile experiment to see how effectively the COC-only model that has been trained can accurately describe a test sample.

Therefore, a more thorough evaluation of our model, which we will leave for future work, is an evaluation experiment to further assess the effectiveness of our models' label of the scam type, by creating a blacklist. We will then assess the effectiveness of that blacklist based on the blacklists' ability to block future unwanted calls. We will, subsequently, compare our blacklists against two commercial third-party carriers. We then will assess how well the blacklists formed from different data sources, namely spam and COC, can complement each other. Finally, we will describe how the blacklists we construct can block future spam messages and phone calls. Through this, we will demonstrate how effective the blacklists will be in blocking these campaigns.

To estimate the overlap between our blacklists and third party blacklists, we will select a random sample of 10,000 source phone numbers from all of our datasets, and perform reverse phone lookup queries. This will be performed using Twilio and YouMail service providers. The goal will be to verify that our numbers, each labeled with a topical theme, are described similarly by the third-party service provider. We will provide specific proportions and determine to see if our blacklists can be leveraged as a proxy to estimate how effective third-party phone blacklists are in defending users from malicious calls. We will also use that same random sample of numbers in order to validate false positives.

8 Related Work

8.1 Stability of Phone numbers

Telephony abuse has emerged as a nefarious mechanism for deploying a multitude of fraudulent and malicious privacy violations [30], [29].

The exploration of such cross-channel abuse and specific scams, has included the analysis of global operators and the origins of mobile operators. For example, Costin et. al. [15] investigated how malicious actors from various geographic regions, particularly Nigeria, deploy various scams, based on analysis of crowd sourced web listings. In this investigation, the researchers also provided a detailed analysis of the economic implications of the business models of such scams, along with the information exchanged between actors across telephony channels. Furthermore, their analysis provided detailed insights into the online and phone infrastructure used by malicious actors. Finally, Costin et. al. also examined tactics, leveraged by scammers, that actively conducted live analysis of phone numbers through calls. However, these previous results focus on demonstrating that phone numbers are a stable indicator of the identification of scam campaigns. Moreover, a more detailed analysis, mostly considered phone numbers related to outgoing calls, whereby the user contacted a given tech-support number [25].

In contrast to their prior work, our scope of our analysis include a broader range of campaigns. Furthermore, our

analysis extends to incorporate crowd-sourced data sets. Finally, our study instead focuses on incoming call reports and their subsequent responses from communities. From our longitudinal analysis we are able to capture and to describe the authenticity and associated scam campaign of the phone number.

8.2 Spam and Email Classification

Forensic analysis of spam messages has been a very thoroughly interrogated area of research [23]. In this study Ma et al. overcame the sparsity problem in SMS message classification by, first, using K-means to group messages into disparate classes and then, second, aggregate all the spam messages of a class into a single document. Symbol semantics was accounted for this in this model. The authors designed specific rules and introduced specific background terms in order to make the model appropriate to fully represent SMS spam. Wei et al. [34] propose an approach based on the agglomerative hierarchical clustering algorithm and the connected components with weighted edges model to cluster spam emails. Only spam emails used for advertising are tested by the authors.

In contrast, our motivation is to describe the context of a phone-number, and it's potential reuse across different spam campaigns. Our target documents are context-rich forums that provide nuanced details that are enriched with polysemy and ambiguity. Finally, the documents that we analyze, and our model evaluation account for nuances in a victim's regional differences. For example, the reporting of Nokia scam campaigns are reported differently than a Canadian lottery scam. Our procedure for describing the complaint, and for the evaluation of these complaints, account for these differences.

8.3 Classifying Using (Call Detail Records) CDR Lookup

Phoneypot, [17], is a phone honeypot that is comprised of a set of phone numbers - along with infrastructure to collect CDRs when honeypot numbers are called by external parties. Phoneypot has demonstrated the utility of honey pots in telephony research.

Although the data collected from the honeypot was analyzed to source various types of calls (attributes of the call type), the longitudinal analysis of the phone number was not studied. Moreover, the features mentioned in Phoneypot did not distinguish the specificity of the geographical origins of the phone number. In contrast, our analysis includes geographical attributes in order to identify the geography of the call, along with the service provider.

8.4 VoIP Call Detail Records

Similarly, Chiapetta et al. [14] present an analysis of a large dataset of VoIP CDRs. Subsequently, their investigation afforded insights into different call patterns and groups callers using unsupervised techniques. The features mentioned in this work are appealing for population-level uses but are not designed to differentiate between spam numbers and legitimate callers. A similar heuristic for clustering of transcripts, recorded by a phone honeypot, have also provided similar findings in identifying and, subsequently blocking actors [29], [26], [16].

However, one caveat in this previous approach lies in the fact that since the transcript is the only source of information, it can *only* serve as the single indicator for whether or not to block calls from the campaigns that are seen at the source honeypot. This cannot be generalized to other domains in the telephone network. Applying a generative and probabilistic approach based on feature sets, as in our study, allows for greater generalization beyond a specific dataset. Moreover, our analysis provides more human-interpretable output features for both the identity of and the details of a given call record, as documented by a given victim. Finally, our model continuously undergoes training in order to be flexibly deployed to novel contexts in the phone network.

8.5 Domain and IP Address Analyses

Other reputation systems have been investigated for domain names, IP addresses and other online resources such as URLs. These systems have offered robust defenses to email spam and malware infections. However, these investigations typically utilize network, and other application-specific features to label such resources which differ significantly from information available in user-facing devices.

Finally, caller profiling has been investigated by several researchers [31]. However, these previous studies are predicated on access to large volumes of private call data records. These often possess accompanying privacy concerns from service providers that do not readily make such datasets available. Yardi et al. [35] characterized macro-level behavioral patterns in order to differentiate malicious users from authentic users. Call duration and social network connections were used to separate legitimate callers from spam/scam callers in [9] by developing a global reputation based system for callers. The caveat, however, is duration information and social connections of phone users are not available. Furthermore, both the full evaluation for evaluating a human-interpretable origin of a spam campaign's numbers, and a given number's cross-validation with other ground truth data sets is also not addressed by this work. To the best of our knowledge,

our work is the first one to systematically explore crowd-sourced datasets using Online-Variational LDA sampling for modeling topics associated with phone numbers. From this we are able to both describe the origins of and authenticity of a phone number. In addition, our methods provide insights into the effectiveness of such generative models as effective instruments that can be used in classifying scam campaigns and phishing attacks.

9 Conclusion

Robocalls act as a pivotal instrument for criminal activities. In this paper, we propose an Online Variational LDA model, based on the probability theory of Latent Dirichlet Allocation, for describing and for classifying malicious scam-campaigns deployed through the telephone network. The online variant can eliminate the sparse representation issues in document classification. It is more suitable for the task of describing and for the task of classifying robocall scam campaigns. When compared with the existing state of the art techniques, namely LSA, we show that it is more suitable for analysis, and for describing the malicious content of robocall complaints, as reported on online forums. We will continue to build upon this model towards a systems solution that incorporates a client interface and keyword search capabilities. The democratization of this system will empower both everyday users, and law enforcement in identifying and preventing scam campaigns that are specific to identity theft, account fraud, and phishing attacks.

References

[1] `https://bit.ly/2KlhJyt`. [Online; accessed 12-March-2018].

[2] `https://www.youmail.com/`. [Online; accessed 20-January-2018].

[3] `https://www.truecaller.com/`. [Online; accessed 19-January-2018].

[4] `https://www.projecthoneypot.org/`. [Online; accessed 21-March-2018].

[5] `https://www.mongodb.com/`. [Online; accessed 17-September-2017].

[6] `http://www.nltk.org/api/nltk.stem.html`. [Online; accessed 21-November-2017].

[7] `https://github.com/googlei18n/libphonenumber`. [Online; accessed 21-November-2017].

[8] L. Alsumait, D. Barbará, J. Gentle, and C. Domeniconi. Topic significance ranking of lda generative models. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part I*, ECML PKDD '09, pages 67–82, Berlin, Heidelberg, 2009. Springer-Verlag.

[9] V. Balasubramaniyan, M. Ahamad, and H. Park. Callrank: Combating spit using call duration. In *In CEAS (2007)*, CEAS '07, 2007.

[10] M. Bidgoli and J. Grossklags. Hello. this is the irs calling.: A case study on scams, extortion, impersonation, and phone spoofing. In *In Proceedings of the Symposium on Electronic Crime Research (eCrime)*, 2017.

[11] D. M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, 2012.

[12] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[13] J. Chang, J. L. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada.*, pages 288–296, 2009.

[14] S. Chiappetta, R. P. Claudio Mazzariello, and S. P. Romano. An anomaly-based approach to the analysis of the social behavior of voip users. Number 6, pages 1545–1559, 2013.

[15] A. Costin, J. Isacenkova, M. Balduzzi, A. Francillon, and D. Balzarotti. "The role of phone numbers in understanding cyber-crime schemes.". In *Privacy, Security and Trust (PST), 2013 Eleventh Annual International Conference on, pp. 213-220.*, 2013.

[16] Y. Gao and G. Zhao. Knowledge-based information extraction: a case study of recognizing emails of nigerian frauds. In *In Proceedings of the 2005, NLDB conference*, pages 161–171, 2005.

[17] P. Gupta, B. Srinivasan, V. Balasubramaniyan, and M. Ahamad. Phoneypot: Data-driven understanding of telephony threats. In *In NDSS*, 2015.

[18] G. Heinrich. Parameter estimation for text analysis. *Web: http://www. arbylon. net/publications/text-est. pdf*, 2005.

[19] M. Hoffman, F. R. Bach, and D. M. Blei. Online learning for latent dirichlet allocation. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 856–864. Curran Associates, Inc., 2010.

[20] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 50–57, New York, NY, USA, 1999. ACM.

[21] T. K. Landauer, P. W. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3):259–284, 1998.

[22] J. B. Lovins. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11:22–31, 1968.

[23] J. Ma, Y. Zhang, J. Liu, K. Yu, and X. Wang. Intelligent sms spam filtering using topic model. In *2016 International Conference on Intelligent Networking and Collaborative Systems (INCoS)*, pages 380–383, Sept 2016.

[24] D. M. Mimno, H. M. Wallach, E. M. Talley, M. Leenders, and A. D. McCallum. Optimizing semantic coherence in topic models. In *EMNLP*, 2011.

[25] N. Miramirkhani, O. Starov, and N. Nikiforakis. Dial One for Scam: A Large-Scale Analysis of Technical Support Scams. In *Proceedings of the 24th Network and Distributed System Security Symposium (NDSS)*, 2017.

[26] I. Murynets and R. P. Jover. Crime scene investigation: Sms spam data analysis. In *In Proceedings of the 2012 ACM conference on Internet measurement conference*, pages 441–442, 2012.

[27] S. Pandit, R. Perdisci, M. Ahamad, and P. Gupta. Towards measuring the effectiveness of telephony blacklists. In *In NDSS*, 2018.

[28] M. Röder, A. Both, and A. Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 399–408, New York, NY, USA, 2015. ACM.

[29] M. Sahin and A. Francillon. Over-the-top bypass: Study of a recent telephony fraud. In *Proceedings of the 23rd ACM conference on Computer and communications security (CCS)*, CCS '16. ACM, October 2016.

[30] M. Sahin, A. Francillon, P. Gupta, and M. Ahamad. Sok: Fraud in telephony networks. In *Proceedings of the 2nd IEEE European Symposium on Security and Privacy*, April 2017.

[31] V. S. Tseng, J.-C. Ying, C.-W. Huang, Y. Kao, and K.-T. Chen. Fraudetector: A graph-mining-based framework for fraudulent phone call detection. In *In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015.

[32] P. D. Turney and P. Pantel. From frequency to meaning: Vector space models of semantics. *J. Artif. Int. Res.*, 37(1):141–188, Jan. 2010.

[33] K. Wedenberg and A. Sjöberg. Online inference of topics : Implementation of the topic model latent dirichlet allocation using an online variational bayes inference algorithm to sort news articles, 2014.

[34] C. Wei, A. Sprague, G. Warner, and A. Skjellum. Mining spam email to identify common origins for forensic application. In *Proceedings of the 2008 ACM Symposium on Applied Computing*, SAC '08, pages 1433–1437, New York, NY, USA, 2008. ACM.

[35] S. Yardi, D. Romero, and G. Schoenebeck. Detecting spam in a twitter network. Number 1, 2009.

[36] R. ehek and P. Sojka. Software framework for topic modelling with large corpora. In *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks*, pages 46–50, Valletta, Malta, 2010. University of Malta.