GEORGETOWN UNIVERSITY

SECURITY AND SOFTWARE ENGINEERING RESEARCH CENTER

Marlene H. Dortch, Secretary
Federal Communications Commission
445 12th Street, SW
Washington, DC 20554

April 18, 2017

Dear Ms. Dortch,

The Security and Software Engineering Research Center (S^2ERC) is a National Science Foundation-sponsored, industry-supported[1] research center with the mission of connecting people and bringing the world closer together through new technologies, policy, law, and economics around communication technology. The goal of our research is to enable security and software technology gains within member organizations and to protect the security and stability of our public networks.

With funding from S^2ERC affiliates, most notably Verizon, the S^2ERC embarked on a project to examine rural call completion issues. The attached paper is a technical report describing our findings, results, and our experience with a new metric we developed which we call Human Retries (HMR). We presented this paper at the Industry Workshop hosted by Verizon on March 29, 2017.

We welcome any questions or comments on our work. I can be reached by phone at 202-687-4107 or by e-mail at eric.burger@georgetown.edu.

Sincerely,

Dr. Eric Burger
Research Professor of Computer Science
Director, S^2ERC at Georgetown University

[1] The NSF supports the work of the S2ERC through grant IIP-1362046. The NSF defines 'industry' as any funding source, public or private, that is not the NSF.

GEORGETOWN UNIVERSITY

S^{2}ERC Project:	Rural Call Completion
Report:	Issues, Analysis, and Tools For Rural Call Completion Issues
Author:	Trent Stohrer, Research Staff
	Andrew Stewart, M.S. Student
	Dr. Eric Burger, Research Professor of Computer Science
Status:	For Publication
Date:	March 27, 2017

Abstract

Changes to the wireline telephone network, including the introduction of new technologies such as SIP and the gradual reduction of wireline subscribers, has led to a network environment with higher reports of issues connecting calls to rural areas than there were ten years ago. Old network performance metrics seem incapable of identifying these previously unseen or unreported problems. The Federal Communications Commission (FCC) cites three factors: uncaptured or incorrect signaling, the presence of automated call traffic, and the increase of phone numbers without subscribers, which work together to reduce the capability of older metrics to measure network health. Using data from wireline providers and our knowledge of the symptoms of the connection problems, we created a new metric, called HMR, intended to be as independent from these factors as possible, with the intent being to deploy it to identify and resolve problems with calls to rural areas on a day-to-day basis or more frequently. While we were unable to completely disentangle HMR from some issues that cause problems for the old metrics, we were able to detect anomalies that potentially indicate problems that the other metrics were not able to capture. More work needs to be done to further reduce the influence of the complicating factors and to determine whether the data anomalies represent actual problems in the network.

This material is based upon work supported by the National Science Foundation under Grant No. 1362046 and the industry affiliates of the Security and Software Engineering Research Center (S^{2}ERC). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the S^{2}ERC affiliates. Payments are made to Georgetown University and the funds are used to cover the expenses of the study and related academic and research activities of the institution.

Introduction

Rural Call Completion is a blanket term for a series of problems that have been reported to the FCC by several states as well as trade associations that represent rural carriers carriers (collectively "rural associations"). As the name suggests, these Rural Call Completion problems (RCC from here forward) involve issues with completing calls that exist specifically with calls to the rural areas of the PSTN (Public Switched Telephone Network).[1 2] RCC is not one specific technical problem but rather a series of symptoms which, according to reports of incidents at the FCC, include the following: "lengthy periods of dead air on the calling party's end after dialing a number, audible ringing tones on the calling party's end when the called party's telephone never rings at all, false busy signals, inaccurate intercept messages, the inability of one or both parties to hear the other when the call does go through, and calls simply not arriving at their destinations." The FCC has stated that, "the inability to complete calls reliably threatens public safety and contravenes the public interest." For example, the FCC reported "examples of life-threatening call failures, including a situation where an on-call surgeon was unable to receive a call from a hospital for emergency surgery and a 911 call center was unable to do emergency call backs."[3]

According to the FCC, "there appear to be multiple factors that may cause rural call completion problems. Rural associations posit that the call completion problems may arise from the manner in which originating providers set up the signaling and routing of their calls, and that many of these call routing and termination problems can be attributed to intermediate providers." Least cost routing carriers (also known as LCRs) offer terminating services at low rates, and the rural associations argue that some LCRs who provide such intermediate transport may provide inferior service to achieve their lower rates. The FCC offers LCRs could be a cause of RCC problems due to high access charges for calls to rural areas which provides an incentive to use cheaper LCRs. Another factor cited is there are fewer potential routes to terminate to each rural location, thus meaning that for universal connectivity it is difficult to impossible for any single carrier to connect to each and every rural carrier. According to an industry speaker from the First RCC Industry Workshop,[4] some of the LCRs they investigated are able to function during times of low call volume. However, during peak traffic, when their circuits are full, they begin to handle calls incorrectly. This can manifest as the LCRs holding onto calls without handing them back for re-routing, playing a ring or other treatment before any connection has been established, or simply returning a release code and releasing the call rather than attempting to complete the call.

The FCC has stated that "one key reason for the increased problems in rural areas is that a call to a rural area is often handled by numerous different providers in the call's path. Given the particularly high rates long-distance providers incur to terminate long-distance calls to rural rate-of-return carriers, long-distance providers have additional incentives to reduce the per-minute cost of calls. For example, the disparity between interstate rates can be 5-6 cents per minute for rate-of-return areas and just over half a cent per minute for price cap areas.[5] As a result, there is greater incentive for the long-distance provider to hand off the call to an intermediate provider that is offering to deliver it cheaply – and potentially less incentive to ensure that calls to rural areas are actually completed properly." This problem is potentially exacerbated by the industry's move towards Voice over Internet Protocol (VoIP) and the ease with which a party can set up a server capable of handling and routing Internet-based calls. In the First Industry Workshop, it was

[1] Federal Communications Commission, *Report and order and further notice of proposed rulemaking in the matter of rural call completion*. 2013, pp. 1–2, WC Docket no. 13-39.

[2] By "rural carrier," we mean an incumbent high-cost, rate-of-return carrier, as defined by the FCC. The FCC defines a rural carrier as those designated as such by the National Exchange Carrier Association (NECA). *See Report and Order*, ¶9.

[3] Federal Communications Commission, *Report and order and further notice of proposed rulemaking in the matter of rural call completion*. 2013, p. 8, WC Docket no. 13-39.

[4] Verizon Public Policy, *2015 rural call completion industry workshop: Panel 1*, 2015. [Online]. Available: https://www.youtube.com/watch?v=quQnIlAm3Qc (visited on 03/10/2017).

[5] Federal Communications Commission, 2017. [Online]. Available: https://www.fcc.gov/general/intercarrier-compensation-0 (visited on 03/10/2017).

mentioned that in an investigation into an RCC problem one of the panelists' companies found a LCR that had purchased a SIM box and was acting as an alternate vendor for terminating calls.[6] In other words, with some relatively cheap hardware, free software, and a flat-rate plan these actors were able to insert themselves into the PSTN. We have also heard of a user on a flat rate plan routing calls through his plan and acting as a "least cost router."

The FCC has taken on a number of regulatory actions in order to alleviate RCC issues. This has included agreements with companies to report levels of monthly answer rates, measured by answer-seizure ratio (also known as ASR), and perform investigations into rural carriers[7] with unexpectedly low monthly answer rates. A number of companies have entered into Consent Decrees with the FCC whereby the companies have taken a number of steps intended to ameliorate RCC issues. These steps include investigations into Rural OCNs[8] with negative spikes of ASR[9] or if they have any other reason to believe they are having RCC issues. The Bureau has also worked on expanding the adoption of safe harbor rules whereby an Inter-Exchange Carrier (IXC) agrees to either deliver calls directly to rural carriers or only hand the call off to carriers that will directly deliver the call to the rural carrier. Essentially, the IXC agrees to use at most one LCR in the path of the call, which makes untangling any issues that do arise in the network much easier. As part of a consent decree, this study was commissioned to review current methods for detecting and resolving RCC problems, improving metrics and data collection, and to develop new tools to improve the RCC situation.[10]

Today's problem detection is either ad hoc, when a caller complains, or post hoc, from running batch reports on call detail records (CDRs). Neither method is substantially real time. This has drawn us to a line of research where we investigate whether the data is available, in the real-time signaling path, to collect meaningful metrics on call failures. The current reporting statistics for RCC involve looking at network data over a period of time, usually a month, to try to identify anomalies in call completion or network performance, and then to take responsive action. Given the importance of these issues, there is a desire to try to remediate issues in as near-real time as possible. In our research, we have determined that the current ratio-based metrics are inadequate for uncovering RCC-specific problems, and as such we have worked to develop a new metric, which we refer to as Human Retries, or HMR, to help detect the various RCC symptoms relayed above. The metric can be deployed in a variety of ways and is not reliant on long time frames of data, such as a month, but rather can be used on a day-after basis, or perhaps faster, depending on how quickly a carrier generates call records. We evaluated the feasibility of calculating and analyzing HMR as call detail records are added to the database and have also attempted, with limited success, to calculate HMR in real time.

1 Description and Shortcomings of ASR and NER

Two of the more well-known measures of a network's ability to deliver calls are Answer-Seizure Ratio (ASR) and Network Effectiveness Ratio (NER). ASR is a metric developed by industry and mandated by the FCC for reporting. Carriers report on a month-to-month basis to the FCC. As its name suggests, ASR is the ratio of answered calls to total line seizures. In practice this can be calculated as answered calls with normal call clearing over all calls not receiving SS7 release cause

[6] See video in note 4, all other references to the industry workshop refer to this video.

[7] An ILEC is an "Incumbent Local Exchange Carrier." In this paper, we refer to high-cost carriers as "rural carriers." For the most part, high cost, rate-of-return carriers are rural carriers, even though a high cost carrier may not be in a rural region and a carrier in a rural region may not be a high cost carrier.

[8] OCN is an acronym for an Operating Company Number, used to designate different telephone companies.

[9] A negative spike is a sharp decrease from prior measurements over a short time. The exact parameters of what is a "sharp decline" varies. The important aspect is the sudden short-lived aspect of the change.

[10] Federal Communications Commission, *Order in the matter of Verizon*. 2015, 18(b)[2], DA 15-74.

code[11] 1 for unallocated and unassigned numbers.[12] Some methods for determining ASR include all calls in the denominator, including the call attempts receiving code 1, but we will get into why that is a bad idea in a bit. NER is a related metric to ASR. NER is also a ratio and has the same denominator as ASR, excluding calls receiving code 1, but it counts some calls as successes that ASR does not. The general idea behind NER is to not count the behavior of the terminating users against the network. For example, under ASR a call with a busy signal is considered a failure but in NER it is considered a success; the network performed as expected and only received a busy signal because the end user was already on the phone. The exact implementation of NER can differ slightly depending on what behavior you're trying to capture. The most forgiving of these methods is to count all calls receiving code 16, 17, 18, or 31 as successes. In our method, the numerator for NER includes calls that receive release codes for user busy (code 17) and no user responding (code 18) in addition to the answered calls with normal clearing counted in ASR. If signaling were completely reliable in all cases NER would also count as successes calls which are connected but not ever answered, perhaps because the end user is not home. This is also known as a "ring no answer." unfortunately, as we shall see in a bit, in practice counting ring no answers as successes is not reliable, as they are often indistinguishable from other types of call failure. In theory, we could count every call receiving code 16 (normal call clearing) as a success, and some carriers do this, but in practice this method ignores and obscures certain facts about signaling in the current networking environment.

ASR and NER are fundamentally similar metrics in that they are ratios of calls considered successes over all calls (ASR), or all calls that have a chance to succeed (NER). In our research, the difference between the two is in many cases negligible, especially when excluding the code 1 calls from the denominator. We will elaborate on this shortly. First, though, we will describe the data set and approach we used to analyze the RCC situation and to create RCC tools.

2 Call Data and Approach

The major goal of this project was to review current metrics, methods, and data for detecting RCC and other network problems in order to develop real-time tools for detecting and resolving RCC problems. Our aim was to create metrics and tools that could at worst be deployed on a day after basis, as that would be a massive improvement over using monthly metrics to trigger investigations, with the ideal being more real-time, on the order of five to fifteen minutes. In order to help us review existing metrics and to create new ones, Verizon provided us with one month's rural call data.[13] Subsequently, Verizon responded to our request for more records with over five months-worth of additional data. We also received call data from inContact as part of their consent decree.[14] Level 3 provided rural call data at the request of FCC staff to assist in evaluating the tools for general applicability.[15] Both inContact and Level 3 provided us with two months of call data.[16]

Of the 50 Terminating OCNs[17] in our main dataset with the lowest ASR (where the code 1 calls are included), 23 of them also rank in the 100 worst OCNs in terms of NER. Not surprisingly,

[11] ITU-T, "Digital subscriber signalling system no. 1 and the signalling system no. 7 isdn user part," International Telecommunications Union Telecommunications Sector, Tech. Rep. Q.850, May 1998.

[12] We will be discussing many release codes in this paper. A release code is a standardized code that explains why a line was released. In the example above, we include all release codes except for code 1, 16, and 17, representing unassigned number, normal call clearing, and user busy respectively, all being the most prominent. *See Q.850* for more details.

[13] We will refer to this as the main or primary dataset.

[14] Federal Communications Commission, *Order in the matter of inContact, Inc.* 2016, 19(b), DA 16-466.

[15] The first month of data from Level 3 is what we refer to as our secondary dataset.

[16] All three carriers entered into strict non-disclosure agreements with Georgetown; there was no intermingling of the data; the machines were not connected to the Internet; the data was striped and stored on encrypted partitions; and only the three Georgetown researchers had access to the data.

[17] OCN is not the most granular measure and in some instances a carrier will use different routes to get to different End Offices (where customers' lines physically connect) within that OCN. That said, the FCC asks for reports on an OCN-by-OCN basis and in our main dataset there are a more manageable 1216 OCNs compared to 7959 End Offices.

the 21 remaining OCNs that are outside of the 200 worst in NER have 34% or more of their calls receiving code 1 and many of them have over 50% of their calls receiving code 1.[18] Therefore, when we calculate both ASR and NER we exclude calls receiving code 1 from our calculations as these calls introduce a tremendous amount of statistical noise and make network performance appear much worse than it is. The preponderance of code 1 calls arises from the current network environment and the large presence of robo-callers and other auto-dialers. There are cases of 'smart' auto-dialers that only call numbers believed to belong to people or businesses. However, many automated systems just call a block of numbers in sequence, indiscriminately. There is no incentive for an LCR to send a false code 1, as these calls do not play into inter-carrier compensation. So, we can be relatively certain that if a call receives a code 1 that it is to a disconnected number and thus never had a chance of being successful. On the other hand, when humans place calls they tend to call people and businesses. Except in the relatively rare cases of a mis-dialed, changed, or a recently disconnected line, calls to people from people do not typically receive code 1. As such, a large percent of calls receiving code 1 tells us much more about the profile of those placing calls to a particular area and the allocation of the numbers in that area,[19] than it does about a network's ability to deliver calls there.

With the code 1 calls removed, the correlation between ASR and NER is made even clearer. Of the 50 OCNs with the lowest ASRs (code 1's removed from denominator) in our main dataset, all but six, or 44 out of 50, are also among the 50 worst in NER. Of the remaining six, three have NERs that would rank in the 100 worse and the other three all have similar reasons for having a low ASR but a high NER. Specifically, all three OCNs are in the top five for calls receiving busy signals, having 22-32% of their calls receiving busy signals (the average is 3.7%). Beyond that, for these three OCNs the busy signals are mostly caused by calls from a single originating number in two cases[20] and to a single terminating number in the third case.[21] The metrics are even closer when we look at the worst performing OCNs in terms of NER. Of the 50 OCNs with the worst NER, 43 of 50 are also among the 50 worst in ASR, but all 50 are among the 75 worst in ASR. These examples are not to say one metric is better than the other or to say that they are exactly equivalent. Rather, they both perform similarly when it comes to identifying where, or if we were to slice up the data differently, when, the network is performing abnormally or especially poorly. However, as the dive into the low ASRs that do not have correspondingly low NERs suggests, these metrics can be skewed rather easily by single numbers on either side of the call path (origination or termination), meaning NER and ASR seem to be better at catching acute problems with call delivery rather than chronic problems.

Since ASR and NER are calculated in similar ways, the two have similar shortcomings. Due to current factors in the PSTN, such as the large presence of robo-callers and disconnected numbers, these metrics have limited utility. Both metrics can easily be calculated and analyzed on various slices of the data: month-by-month, day-by-day, hour-by-hour, OCN-by-OCN, end office-by-end office, by intermediate carrier (or LCR), etc. No matter how we have sliced the data, though, the problems that exaggerate how bad certain slices are continue to exist because they are, as far as we can tell, inherent to the network. This is not to say that NER and ASR are not useful, in fact they both perform quite well in finding times and places where an outsize number of call failures are happening. Rather, the network characteristics we keep referring to, which can compound together and be difficult to control for, skew both metrics to varying degrees.

The prime network factors that tend to skew ratio-based metrics are somewhat similar and, in terms of results, closely related: robo-callers/auto-dialers and disconnected numbers. Robo-callers (we will use this term to refer to both robo-callers and auto-dialers[22] from here on) drive an enormous

[18] The average of release code 1's for OCNs in the main dataset is 6%.

[19] The percent of numbers that are allocated differs from end office to end office and can be quite low. To be more specific, offices are given numbers in blocks of 1000 numbers. A typical rural carrier has 2400 customers. Since they will have at least three blocks of 1000 numbers (3000 in total), quite a few will not be assigned. In this example, 20%

[20] 84% and 60% of calls receiving busy signals from a single number in the cited case.

[21] 93% of the calls receiving busy signals are to a single number in the cited case.

[22] The two are similar in that they place lots of calls to lots of different numbers and do so with varying levels of

percentage of the traffic. Because their call patterns are so different from typical customers, their calls can wreak havoc on ratio-based metrics. Identifying which callers are robo-callers is currently an unsolved problem, but using a method recommended to us by one of our industry contacts,[23] we can demonstrate the sort of impact that robo-callers have on the current environment. Our main dataset has 20.3 million unique originating callers, of which only 3,449 numbers, or 0.17%, meet our somewhat simplistic criteria for robo-caller. However, those 0.17% of originating numbers account for 42% of the calls in the set. If robo-callers behaved more like human callers this would not necessarily be a problem, but unfortunately, they differ from humans in that they call disconnected numbers more frequently than humans do.[24] Additionally, with the widespread adoption of features like caller ID, even when robo-callers do call a human rather than a disconnected number, there is a higher than normal chance the receiving party will simply not pick up, because they do not recognize the number. This means calls by robo-callers have a lower chance of being answered even when the called party is human, meaning they have a higher probability of being counted as failed calls in both ASR and NER.

The other major complicating factor for the ratio-based metrics is the presence, and, in fact, preponderance of disconnected numbers. As with robo-callers, the presence of disconnected numbers is not necessarily a problem in itself, but rather the current network conditions cause various issues that result in disconnected numbers obscuring our metrics. Research with industry suggests there are both OCNs and LCRs that cause calls to disconnected numbers to receive a release code other than 1. In some cases, this may be due to misconfigured or outdated hardware returning the wrong code,[25] and in other cases caller hang-ups can cause a failure to capture the release code, defaulting the call to a non-answered, normally cleared call (code 16), which will count as a call failure in both ASR and the more conservative version of NER we described above.[26] It should be noted that from the vantage of the originating carrier, we do not believe it is possible to tell where signaling issues originate, just that they exist. Our investigations into calls to specific OCNs suggest that there are in fact problems with communicating the correct code, though our limited network view makes it impossible to determine why these problems exist and where in the call path they occur. The simplest, and crudest, way to show this impact is the overall effect of removing all terminating numbers with zero answered calls in the dataset from our metrics.[27] This one adjustment moves the overall NER of the dataset from 63.3% to 81.4%. On a more granular level, there is plenty of evidence that problems with signaling for disconnected numbers exists. In our main dataset, 11% of the OCNs have no calls receiving code 1's and 30.4% have fewer than 10 calls in the set receiving code 1, with none of those accounting for more than .4% of the calls for that OCN in the set. Many of these OCNs with very low percentages of calls getting code 1 have inflated percentages of calls receiving codes that are similar to code 1, but with subtly different meanings.[28] On the other extreme end, 10% of the OCNs in the dataset have over 95% of their calls receiving code 16 for normal call clearing. It is possible that these percentages are simply a reflection of the nature of the calls to these OCNs, since if only calls by humans are made to an OCN and those people rarely misdial it would be reasonable for very few calls to receive code 1. However, given the dropping number of landline subscribers and

discrimination, though from what we understand both types of parties are not usually very selective in who they call.

[23]This method considers any originating number that places over 1400 calls/day during the course of a month to be a robo-caller.

[24]44.8% of the calls placed by the numbers identified above receive code 1, compared to 18% of the calls placed by all other numbers receiving code 1.

[25]Such as code 3 or 34 instead of code 1, both of which will count against the denominator, and thus the 'score' of both of the ratio based metrics.

[26]To be more exact, what happens is that the release signal from the party hanging up (which will be code 16) reaches the carrier before the signal with the code indicating that the call is to a disconnected number. This is the network working as intended, but providing data that makes the call look like a failure.

[27]This is an illustrative exercise we performed, but we did not use it for analysis. We believe something like this may be useful for getting some noise out of the data but our sample is too small to definitively start ruling numbers with a single unanswered call as "disconnected numbers." This would, we believe, cast far too wide a net and miss calls to active numbers that just happen to not answer in the few calls we have for them.

[28]Most commonly we have seen inflated percentages for code 3 (no route to destination) and code 31 (normal, unspecified) in our main dataset and code 63 (Service or option not available, unspecified) in our secondary dataset.

the continued existence of robo-callers, we believe that this unlikely scenario cannot be the case for such a large number of OCNs and that there is sufficient evidence in the data that we should expect a codes other than 1 for some calls to disconnected numbers, which will inevitably result in such calls being labeled as failures by existing metrics.

As discussed above, incorrect signaling for disconnected numbers, and in some cases *correct* signaling, will result in the network appearing to perform worse than it is performing in reality, as calls that should not be counted towards the ratios are counted as failures. Unfortunately, ambiguous signaling also brings about another phenomenon in the data that further reduces the effectiveness of the ratio-based metrics. This phenomenon results in an originating number attempting to call the same terminating number over and over again, in many cases once per second for a minute or more. We refer to this phenomena as "machine retries"[29] or "call bursts." In one particularly extreme example from our main dataset, there are 428 records of one number calling another over about six and a half minutes, all but two of which received a code 3. According to the company who provided these records, they try at most six paths before handing a call back to the previous carrier, and even then almost all calls try only one path. It is likely that either some carrier further back in the call path or the originator of the call itself, in the case of hardware designed to retry when given certain release codes, simply retried different paths for the call or the call itself over and over and over. Though a rigorous analysis of the composition of the codes that come with this phenomenon has not been done, it should be noted that nearly every investigation of OCNs with outsize percentages of abnormal release codes has come with machine retries present to differing degrees.[30] Now, these retries could possibly put some stress on the network, and we have seen cases where that is almost certainly the case, but they also act to heavily skew the ratio-based metrics. Let's take those 428 calls to one number as a simple example. The end office[31] for that terminating number had 928 total calls for the day when that burst occurred. Even if all other calls were successful, which is extremely unlikely, the office would still only have an NER of only 54%. Thus, one cannot make any qualitative judgment from ASR or NER without additional context. 54% could be a fantastically good ratio. Likewise, it could be poor. In particular, the signal of a true RCC failure could be totally lost in the noise of the robocallers. When single numbers and single pairs of numbers can and do have such an exaggerated effect on the ratio-based metrics we believe it obscures these metrics' ability to identify which parts of the network are truly having problems. Except in cases of obvious network congestion, we never saw any cases where calls to certain OCNs or end offices were affected across the board in terms of low NER or ASR, as we would expect if the network itself had problems delivering calls in RCC-type ways and these metrics captured those problems. These numbers were usually dragged down by robo-callers, terminating numbers with individual NERs of zero,[32] or machine retries.

3 Motivations for Developing HMR

These various problems and false alarms caused by complications with the ratio-based metrics led us to think of different methods for analyzing network performance. One of our first thoughts was to create black lists of potential robo-callers and disconnected numbers which we would then filter out of our calculations any call whose originating or terminating number, respectively, fell onto those lists.[33] As described in an earlier anecdote, these simple adjustments vastly improved how the network looked in terms of performance in NER. In fact, the apparent network performance improved so much that we suspected we were filtering out too many calls from our calculations. Unfortunately, we did not get more historical data to further refine our black lists for filtering so

[29]This is not an entirely accurate name, as it is unclear why exactly these retries exists, the 'machine' in the moniker is merely to denote that these retries are done so rapidly that they could not be due to a human redial.

[30]We have seen machine retries with standard codes 1 and 17 and even, in a few cases, code 16.

[31]Think of an end office as a smaller version of an OCN. As the OCN is to a state, the end office is to a city.

[32]Thus, having a higher probability than normal of being disconnected numbers

[33]The method we used for constructing both of these lists was to consider an originating number that placed at least 1400 calls/day to be a robo-caller and a terminating number that had no answered calls in the set to be unassigned.

the only tweaks we have been able to make and analyze are slight changes to the criteria already described for data we already have, such as only excluding numbers with at least 10 total calls, all unanswered, as disconnected numbers.

In any event, a static list of bad numbers will not work. This is because new robo-callers can pop up at any time. More insidiously, unscrupulous robo-callers can spoof their numbers and rotate them.[34] As such, it is important to derive dynamic identification of bad numbers.

There were other reasons beyond the data issues for us to develop new metrics. Mainly we were motivated by the idea that Rural Call Completion issues might be caused by "bad actors," like the hypothetical unscrupulous LCRs described in the introduction, and, as such, release codes and answer signals should not necessarily be trusted. Given the FCC reports that an LCR has played treatments or rings for calls when they are not supposed to and other reports of calls where one party or the other cannot hear anything from the second they connect, it is reasonable to believe that false answer signals from unscrupulous LCRs is a possibility. According to our industry contacts, they have never seen false answer domestically, though they have seen it in international calls, meaning it is technically possible, even if such false answers do not exist in the domestic PSTN. Given that possibility and known issues with signaling for disconnected numbers, we wanted to develop a metric where even if a call has a good release code and has an answer signal, that does not mean the call is automatically considered a success. Likewise, a call that has a bad release code is not automatically considered a failure. We wanted to create a metric as independent from signaling as possible.

Our first step to reaching this goal was to think of the symptoms of RCC and how callers experiencing these symptoms would respond. All of the symptoms share the trait that the call abnormally fails in a way that is obvious to the caller. A caller does not expect to hear silence when they place a call, they do not expect 10 or more rings to occur, they do not expect to hear sound so garbled they cannot hear anything, and they do not expect to hear a message saying, "your call cannot be completed as dialed" when they were able to connect just the week before. For every RCC symptom we have heard of, with the possible exception of a multitude of rings or perhaps a busy signal, we would expect the caller to try the call again in short order, perhaps waiting a few minutes or perhaps calling back immediately. As such we wanted a metric that would capture when people retried calls rather quickly due to these obvious failures without also capturing the machine retries we discussed above.

4 Description of and Results for HMR

The metric we developed, Human Retries, or HMR for short, is a flexible measure of whether a call was retried or not. The metric starts as a record-by-record feature, classifying each record by whether the call record has another call within a short time window after it. To do this, our implementation of HMR set the default value for HMR as No and then scans subsequent records for calls with the same originating and terminating numbers as the call in question. If any of the records with the same telephone number pair fall between 13 seconds and 3 minutes after the first call, we set its HMR to Yes. In our observations of the data it takes low volume callers at least 13 seconds to retry a call. This heuristic comes from the time for a person to realize that the call has failed, hang up, and then redial. Three minutes allows the caller to leave some time to "let things sort out" but hopefully excludes situations where the caller experiences an expected failure, such as a ring no answer where the caller decides "they're probably not home, I'll try again in a bit." Of course, it may be normal for people in such ring no answer situations to call twice or more before "giving it some time," but this situation is built into the aggregating of calls with Yes for HMR, which we will get to in a bit. This initial calculation is by far the most computationally expensive but once the HMR is set it can be modified very easily. For example, if a carrier wants to apply the black lists we

[34] E. Burger and J. Kieserman, "Next generation caller identification," S^{2}ERC, Tech. Rep., Jun. 2016. [Online]. Available: https://s2erc.georgetown.edu/sites/s2erc/files/files/upload/stir_status_and_analysis.pdf (visited on 03/10/2017).

described earlier they can simply set back to No any call marked with a Yes whose numbers fall into one of the black lists. Alternatively, a carrier may want to only consider unanswered calls, or calls with less than a certain amount of talk time, or only code 16, or only calls not receiving busy signals. The only place the metric is not particularly flexible is in changing the time window for what we consider retries. We are confident in the reasoning behind the 13 second to 3 minute time window[35] but re-calculating the HMR over larger datasets takes quite some time so it is possible the metric could be more useful with a wider or possibly shorter time window. More research, plus validation in an actual networking environment, is needed to determine the relative efficacy of different time windows.

As a trade-off for being rather simple to calculate, HMR counts a lot of things as having human retries that it, as far as the theory goes, should not. For an example, we will go back to the 428 calls in 7 minutes. While it is fairly obvious that these calls are machine rather than human retries, the sheer length of the burst of calls defeats our simple calculation. With about one call per second, the very first call does not initially consider itself to have a retry until the scan hits a call with the same phone numbers outside of the 13 second range. There are so many calls though that all but the final 13 or so calls will be marked as Yes. One way to overcome this problem would be to only set HMR as Yes if there does not also exist what we would classify as a machine retry (i.e. within 5 or so seconds), but this method would make an already somewhat time consuming calculation even more time consuming. Alternatively, using number black lists would rule these out as well, since the terminating number with the myriad retries is on the list using our simple method for determining disconnected numbers.

The larger fix for this problem gets to how we think HMR should be aggregated and how it should be used to evaluate network performance. Once the individual HMRs have been calculated and filtered, they can be aggregated in different ways, much like how the ratio-based metrics can slice data along different lines of both time and location. We have been talking about HMR so far as a sort of a raw count, e.g. there were about 400 retries for this one number pair and we are going to count all of them towards the aggregate for whatever slice (OCN, hour, office, day) we are considering. We could also aggregate along each pair of originating and terminating numbers which are experiencing retries. This way, the 400 retries would only count once towards the aggregation. The idea here is that if the network or an LCR is having RCC-like issues delivering calls to a certain area then multiple different pairs of numbers are going to experience those symptoms and thus have retries. We expect that in some cases the raw number of retries will not tell us as much about performance as how many different calls are experiencing conditions that cause retries, especially given situations like burst calls. We have seen cases of 'machine' retries which are not caught by either of the black lists we have built and thus would need one of the more complicated methods to be filtered out, but aggregating on pairs of numbers with retries rather than total retries reduces their influence on the metric. Without testing, deploying, and refining in an actual network environment, though, it is hard to say what exact method of aggregating retried calls would be most informative.

Though giving a full evaluation of the usefulness of the metric and its different methods of aggregation is not possible without deployment, our initial investigations have revealed that while there is some overlap between when HMR and NER perform poorly,[36] HMR is capable of uncovering phenomena that NER cannot. Before going forward with these comparisons we should note that the form of HMR being used for these comparisons is the percent of originating/terminating number pairs having retries. Raw counts of retries do not make sense for a comparison of NER and HMR, as counts vary much more with call volume than ratios do. Using the ratio of calls that are retries to all calls has some obvious risks of overlap with NER, as in the example discussed above, where a large number of retries to the same number receiving any code other than 1 is going to drive NER down and HMR up. The ratio of number pairs receiving retries to total pairs (we will refer to this as pair HMR from here out) does seem to still have some entanglement with NER, as we shall see soon, but it also seems capable of uncovering issues invisible elsewhere. We will now go into some

[35] Though the 3 minutes limit is based somewhat on conjecture and might make sense to be refined.
[36] I.e., high HMR and low NER.

further detail on how NER and pair HMR compare to each other in our data and then follow that up with examples of pair HMR finding issues where NER would have sufficed and finding a change that NER could not have captured.

An examination of monthly NER vs pair HMR for each OCN reveals that, on average, a 1 unit increase in NER ratio results in a 9.35% decrease in the odds[37] of retries to non-retries. For context, the mean pair HMR of among all OCNs is 7.35% and the mean NER is 63.62%. A change in NER from 63.62% to 64.62% would result in a change in pair HMR from 7.35% to 6.71%. Since one could consider pair HMR as a symptom of network ineffectiveness, it is not surprising to see that pair HMR decreases as the network effectiveness ratio increases. However, as the following plots illustrate, HMR can detect phenomena which are masked by NER. The plots below are time-series of daily pair HMR and NER calculations for a single OCN located in north-central Indiana. One can clearly see large spikes in the number of unique number pairs with retries in the month of October, while there is relatively small effect on NER. In fact, the *AnomalyDetection*[38] tool does not classify the decrease in NER corresponding to the pair HMR spikes as anomalous. Additionally, there are negative spikes in NER which do not coincide with positive spikes in pair HMR, as can be seen at the end of August.

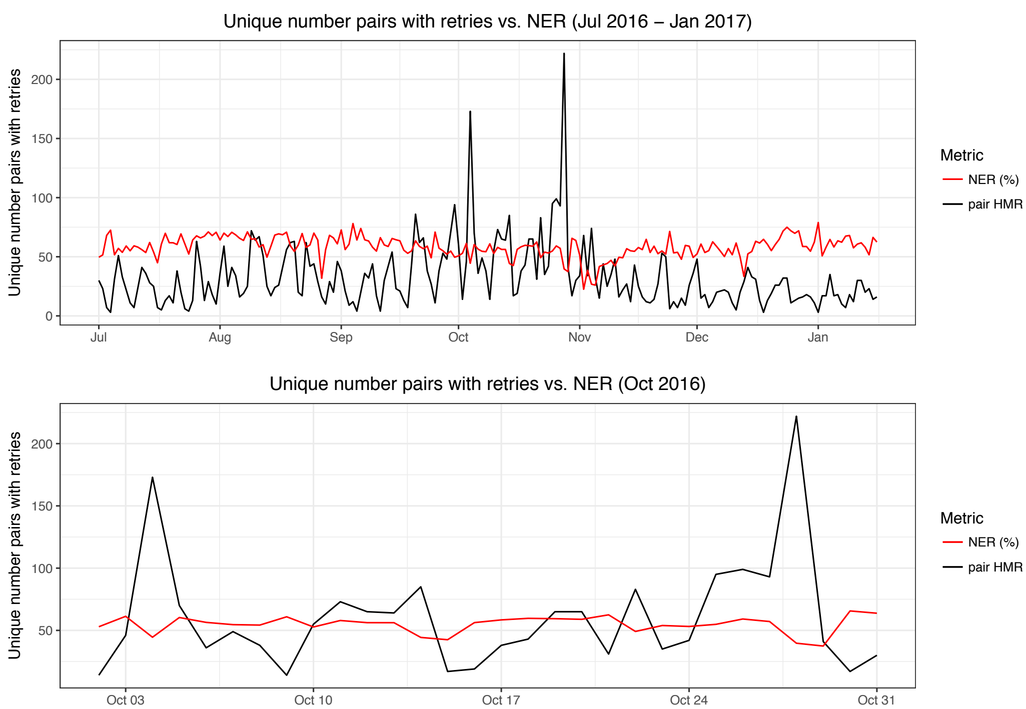

Figure 1: Time-series of daily pair HMR and NER for a single OCN.

[37] If $p = 0.05$ is the probability of being a retry, then the odds of retries to non-retries is $\frac{p}{1-p} = \frac{.05}{.95} = \frac{1}{19}$. Colloquially, we would say that "the odds of being a retry are one in nineteen."

[38] See Section 5 for discussion of the *AnomalyDetection* tool.

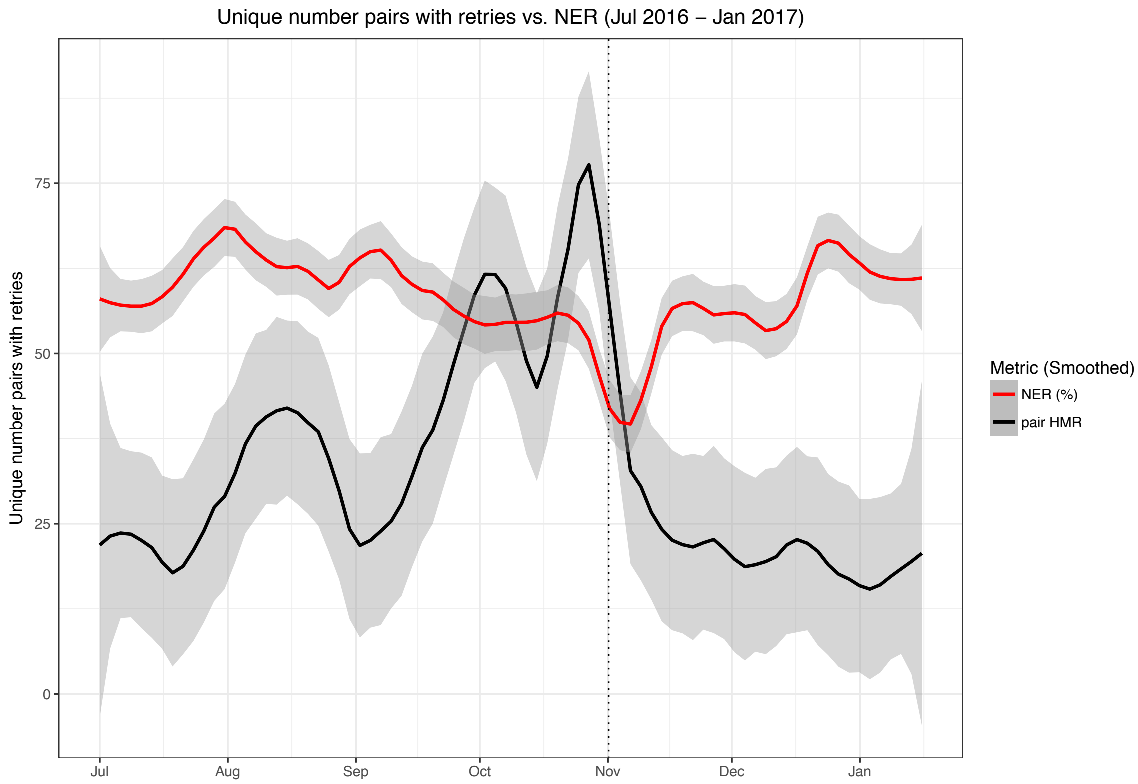

Figure 2: Smoothed time-series of daily pair HMR and NER for a single OCN.

Applying loess-smoothing to the time-series reveals the approximately inverse relationship between the two metrics more plainly. However, the relationship is not perfectly inverse, so there is clearly information which each captures that the other does not. The dotted line in figure (2) indicates where we detected a change in routing, which coincided with an immediate and sustained recovery in both pair HMR and NER from the degraded performance throughout the month of October. However, when we consulted the carrier and dug a little deeper into the data, we found that the degradation in the metrics was driven by a drastic decrease in call volume from one particular number. This number appears to be owned by a company in the political sphere and, unsurprisingly, had its call volume to this OCN drop after November 10. Though we believed we had uncovered an unexpected trend and recovery in HMR fixed with a routing change, the issue was once again seems to have been due to the composition of the callers rather than any other factor.[39] We also examined the proportion of unique number pairs with retries among all calls. Similar to NER, the proportion accounts for traffic volume, and provides a more interpretable measure of retry occurrence, though it too can be sensitive to the effect of robo-callers.[40] For example, one of our OCNs looked rather typical for all of the months that we have records except for one. In those months the OCN had an average pair HMR of 7.29%, which is quite close to the mean pair HMR for all months, but for the outlier month the OCN had a pair HMR of 19.83%. Zooming in, this OCN had a pair HMR of 48.76% on the 23rd of the month in question. However, when we decompose how the individual call pairs with retries break down by originating number, we see that a huge percentage of the calls are coming from just two numbers.

[39] The carrier in question did in fact change its routing just prior to the spike in pair HMR / negative spike in NER. However, the focus here is on the change in the values of pair HMR and NER, not on the routing change.

[40] A robo-caller which retries a few numbers many times will cause a low pair HMR proportion. Conversely, a robo-caller which retries many numbers only a few times will cause a high pair HMR.

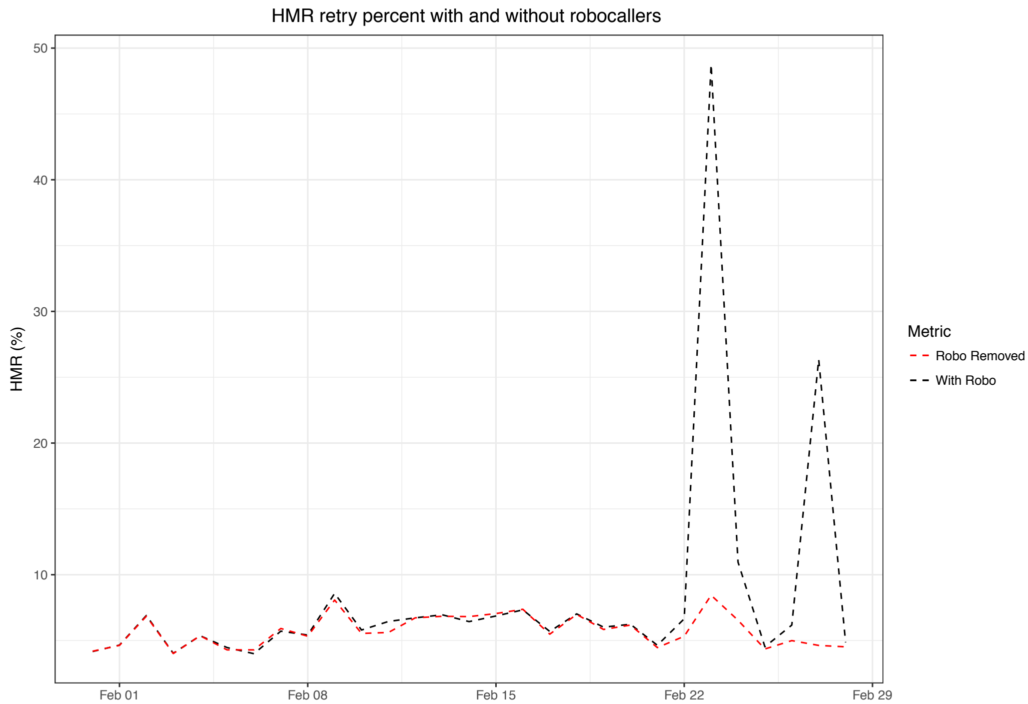

Figure 3: Time-series of daily pair HMR with and without robocallers for a single OCN.

In fact, 4247 of 5933 total call pairs for the day come from those two numbers, and 2754 of 2893 call pairs with retries on those days. Removing the call pairs including these two numbers, which we should note are both classified as robo-callers by the methods described above, improves the pair HMR for the day from 48.76% to 8.24%, much more in line with the average for the other months. There is only one other day in that month with a pair HMR of greater than 9% and, unsurprisingly, the same two robo-callers have a huge impact on the OCN for that day as well. So even pair HMR is affected by robo-callers, and so building good methods for filtering robo-callers is still going to be necessary to get the noise out of the data so carriers can focus on real network impairments.

5 Analysis Methods and Tools

5.1 Anomaly Detection

The temporal nature of CDRs lends them to time series analysis. Business cycles and the human circadian rhythm impose weekly and daily patterns, referred to as seasonality, to call volume and frequency. In turn, any metric based upon call volumes or frequencies will also display seasonality. Like a heartbeat or the stock market, these metrics can be examined for long-term and short-term trends. Through a process called time series decomposition,[41] we break down a monthly or weekly series of measurements to identify patterns, and develop de-trended mean values against which to judge deviation. Once we have identified normal patterns, we can then identify abnormal local and global deviations, or anomalies.

This is the underlying idea of Twitter's open-source *AnomalyDetection*[42] package that we use

[41] R. Cleveland, W. Cleveland, and I. Terpenning, "STL: A Seasonal-Trend Decomposition Procedure Based on Loess," *Journal of Official Statistics*, vol. 6, no. 1, p. 3, 1990.

[42] A. Kejariwal, *Introducing practical and robust anomaly detection in a time series*, 2015. [Online]. Available: https://blog.twitter.com/2015/introducing-practical-and-robust-anomaly-detection-in-a-time-series (visited on

in order to detect anomalies in network metrics. Within this algorithmic framework, anomalies are positive or negative deviations from de-trended means. An example of a positive anomaly might be a burst in call volume from a particular OCN driven by an auto-dialer. Likewise, a negative anomaly might be the sudden decrease in NER caused by the same auto-dialer.

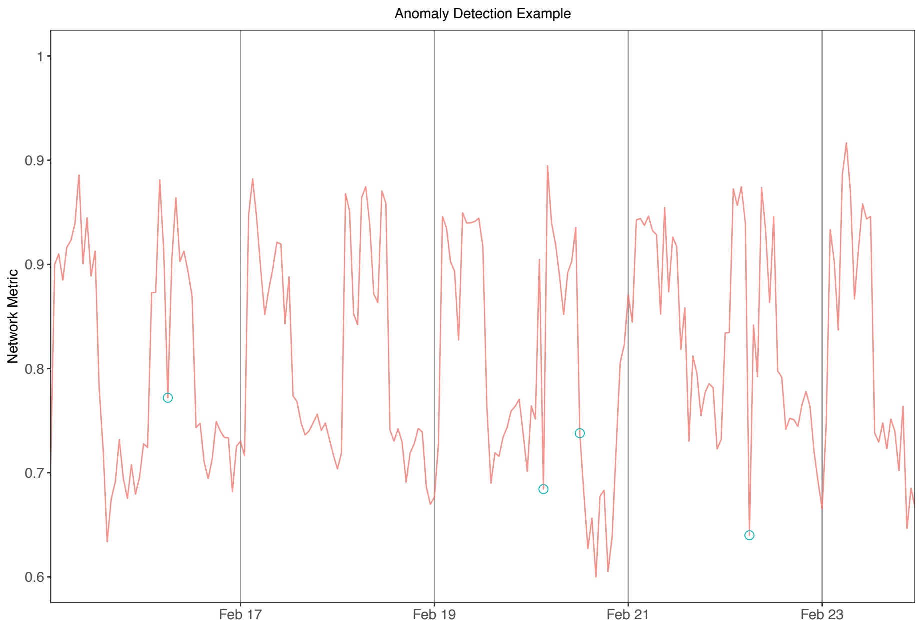

Figure 4: Example plot of *AnomalyDetection* output.

As-is, Twitter's *AnomalyDetection* package is only useful for analyzing historic, windowed data, and not as a real-time detection tool on streaming network data. Suppose that a network is collecting records and batch-processing the CDRs on an hourly basis. One possible ad-hoc method to adapt the algorithm to "real-time" analysis is to append the new batch of CDRs to the previous hour of records and perform the analysis. While hourly is certainly not real-time by most accounts, given the current pace of records analysis in telecommunications networks, usually on a monthly or daily basis, this ad-hoc method is certainly an improvement, and can be integrated into existing analytics regimes such that it operates automatically. We performed an experiment on our hardware to test the general feasibility of calculating HMR as records come in, assuming 5 minute intervals of data. We built a randomized sample of records sized as large as the largest 5-minute interval from the 5 month dataset, as well as an actual 5-minute sample of approximately the same size. Creating a index on the number pairs (necessary to optimize HMR calculations) and calculating the HMR for the records took 16 seconds for the randomized sample and 4 seconds for the real sample.[43] While the time to carry out these calculations will vary greatly from carrier to carrier based on hardware capability and database setup, our experiment at least shows that HMR could theoretically be calculated as records come in, at least in some setups. Given appropriate and informative metrics, AnomalyDetection, or a similar algorithm, can alert engineers to problems in the network as they occur, or within an acceptable window of time. We believe HMR, in its various forms, to be a useful metric to which anomaly detection may be applied, and would fit into the ad-hoc framework.

03/10/2017).

[43]We used a virtual machine running over VMware on a Cisco UCS B200 M3 blade with 2 vCPUs allocated to the virtual machine. The underlying, shared hardware is a dual Intel Xeon E5-2680 running at 2.7 GHz with 8GB RAM allocated to the virtual machine.

5.2 Retry Prediction

As discussed in the sections covering HMR, the retry metric is hypothesized to be useful for uncovering instances of rural call completion problems. In order for this metric to be more useful in a production environment, we would like to move from examination of historical records to real-time analysis or prediction. Similar to the problems with anomaly detection, we required a window size of 3 minutes to classify retries. Calculating this for each number pair on an ongoing basis would require a significant amount of computational power dedicated to the task. A more useful and economical alternative is prediction. Once we had classified the retries based on historical data, we then used machine learning methods to train a predictor model. Several methods were explored, each giving varying degrees of predictive power. In selecting the best model, a number of factors were considered. First, the model needed to be generally applicable. That is, it needed to be designed in such a way that it could be deployed on any network with little modification. Next, it should be designed in a way that it can be deployed at any level within a network's infrastructure. Finally, given the sheer volume of data flowing through a network at any given time, the model needed to have a high level of specificity. Suppose a predictor were able to catch 90% of the retries as they occur, with a 1% false-alarm rate. Further suppose we are classifying 100 calls-per-minute. With a false-alarm rate of 1%, that amounts to 60 false-alarms every hour. In the grand scheme, we deemed such an error rate to be ultimately useless, and would likely be ignored.

Thus, the model that we settled on is based on gradient-boosted decision trees[44] using logistic regression. This predictor was able to achieve a sensitivity of 13-15% with predictor variables that were not specific to any network infrastructure, while maintaining specificity at or above 99.7%. We found that the most predictive features are time of day and terminating OCN, with the two largest OCN's in our dataset by call volume being most predictive.

6 Future Research

Having recently acquired a large amount of historical data, we would like to see how HMR and the various ways of aggregating it change over time. Does it change predictably over time or does it stay relatively stable. Are there unexpected shifts? Are there other complicating factors like there are with our ratio-based metrics or is it possible that HMR gets perturbed in ways that can reveal underlying RCC problems? Many of these questions would be best answered by carriers in the industry implementing the metric, but we have enough historical data now that we should be able to reach some more definitive conclusions concerning HMR with just a little more research. We have also begun experimenting with a method to identify and filter robo-callers based on a technique discussed in a 2013 journal article in the Proceedings of the National Academy of Sciences.[45] We deeply believe that improving methods for filtering noise in the form of calls from robo-callers and to improperly signaled disconnected numbers out of the data will go a long way to improving HMR and other existing metrics. Another idea we are in the early stages of exploring is to monitor fluctuations in the percentages of different result codes on smaller slices, such as LCR-end office pairs or looking for times when carriers have sudden changes in the code distribution during periods of high traffic. If unscrupulous carriers are falsifying signaling or playing the wrong treatments, especially during times of high traffic, we would expect to see changes in the codes for calls going through that carrier. One thing that makes real-time tracing of RCC problems difficult is when a call never reaches the rural carriers they are destined for, the rural carrier has no idea they did not get the call. Likewise, the originating carrier, if it receives any signaling or treatment, thinks the call has been properly handled. If an LCR is dropping calls or playing improper treatments,

[44] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *22nd SIGKDD Conference on Knowledge Discovery and Data Mining*, 2016.

[45] J. Zhi-Qiang, W.-J. Xie, M.-X. Li, B. Podobnik, W.-X. Zhou, and H. E. Stanley, "Calling patterns in human communication dynamics," *Proceedings of the National Academy of Sciences*, vol. 110, no. 5, pp. 1600–1605, 2013. DOI: 10.1073/pnas.1220433110.

neither the originating nor terminating side has a definitive way to be privy to the fact that calls did not go through without either complaints from customers or access to the other company's records. Better communication between the IXCs and rural carriers could also lead to tracking down and eliminating the machine retries we see all over the place in the data. We believe what we have so far with HMR is a promising, new sort of metric, one that could uncover issues invisible to metrics that rely on result codes and answer signals. Further refinement is needed, especially in filtering out numbers on both ends without over-filtering, but the metric measures calls in ways that traditional ratio-based ones do not, and attempts to identify RCC-specific problems in ways they cannot.

One question for research is whether the deployment of the all-IP telecommunications network will help or harm the RCC problem. One of the issues uncovered by our research is that some signaling anomalies stem from legacy class 4 and access tandem trunk configurations. It is not uncommon to find these configurations are running on 20-year-old equipment and provisioned as long ago. We believe a regulatory mandate to rural carriers or access tandem operators to harmonize signaling would be ineffective as it is likely that neither the personnel nor manufacturer support is available on this legacy equipment. We do believe that a move to the all-IP public network, using SIP in particular, affords an opportunity to apply upgrades and harmonization as needed.

Consider the common situation where a person calls a vacant number. The rural carrier returns a release code 1, but the access tandem starts to play SIT and a vacant number announcement. At this point in time, the caller hangs up the phone, sending a result code 16 towards the rural carrier. The result code 1 from the rural carrier is lost. In a full SIP interworking environment, if all intermediate paths return the 404 (Not Found), a 404 can go from the rural carrier to the originating carrier.

SIP is not necessarily a panacea. For example, just as there are literally hundreds of SS7/ISUP release codes, one area of research needs to be to develop uniform guidance on SIP configuration. For example, what if the ILEC configures a vacant number as 410 (Gone)? On the one hand, the originating carrier knows calls to that destination will fail just as they would for a 404. However, having two codes meaning 'vacant' leaves open the opportunity for interworking errors. Ongoing work in forums such as the NNI Work Group[46] needs to be done to ensure we do not repeat the mistakes of the SS7 network in SIP.

7 Summary

We described a novel network performance metric, Human Retries (HMR). HMR captures RCC issues missed by ASR and NER. We validated the operation of HMR on static data from three carriers, but we were only able to compare two full months of data.

As for next steps, the industry should ensure some of the research mentioned above is done. In particular, we are close to real-time HMR validation. It would be valuable to finish that work. As well, identifying and, better yet, dealing with robo-callers is important, valuable research as well. A step in that direction is the S^{2}ERC project *STIR Implementation*.[47]

Our task was to look at the RCC problem form the originating and IXC transport carrier perspective. However, as shown above, originating carriers have no idea if calls are truly being mishandled once they leave the carrier's network. We propose that we instrument the network at the point where we can detect call failure: at the rural carrier's interfaces with the rest of the public network. We can create tooling for rural carriers to detect deviations between normal call volumes and reduced call volumes. Note such tooling is not as straightforward as it sounds. Our expectation is again robo-calls will have a profound impact on any network health metric. In addition, such metrics will need to be cognizant of the general decline of wireline voice minutes and subscribers. We see this as interesting, important work to address a problem the FCC says impacts rural carriers

[46] *Atis/sip forum ip-nni task force*. [Online]. Available: http://www.atis.org/01_strat_init/IP-NNI/index.asp (visited on 03/10/2017).

[47] https://s2erc.georgetown.edu/projects/PSTNtransition/STIR

and thus rural Americans.

8 Acknowledgements

In addition to the support provided by the NSF under grant IIP-1362046, we would like to acknowledge the support and advice provided by the FCC Enforcement Bureau. In particular, the FCC assisted us with contacts at inter-exchange carriers that proved invaluable for us to get the data from which we built and verified our tools. We would also like to thank numerous individuals from Verizon, Level 3, and inContact for both providing data as well as making their engineers available whenever we had questions. We also appreciate the assistance provided by the NECA for giving us depth and context on Rural Call Completion problems and making some introductions to rural carriers for us. As stated above, the opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and not of any of the named companies, organizations, or government entities.

References

[1] Federal Communications Commission, *Report and order and further notice of proposed rulemaking in the matter of rural call completion*. 2013, pp. 1–2, WC Docket no. 13-39.

[2] ——, *Report and order and further notice of proposed rulemaking in the matter of rural call completion*. 2013, p. 8, WC Docket no. 13-39.

[3] Verizon Public Policy, *2015 rural call completion industry workshop: Panel 1*, 2015. [Online]. Available: https://www.youtube.com/watch?v=quQnIlAm3Qc (visited on 03/10/2017).

[4] Federal Communications Commission, 2017. [Online]. Available: https://www.fcc.gov/general/intercarrier-compensation-0 (visited on 03/10/2017).

[5] ——, *Order in the matter of Verizon*. 2015, 18(b)[2], DA 15-74.

[6] ITU-T, "Digital subscriber signalling system no. 1 and the signalling system no. 7 isdn user part," International Telecommunications Union Telecommunications Sector, Tech. Rep. Q.850, May 1998.

[7] Federal Communications Commission, *Order in the matter of inContact, Inc.* 2016, 19(b), DA 16-466.

[8] E. Burger and J. Kieserman, "Next generation caller identification," S^{2}ERC, Tech. Rep., Jun. 2016. [Online]. Available: https://s2erc.georgetown.edu/sites/s2erc/files/files/upload/stir_status_and_analysis.pdf (visited on 03/10/2017).

[9] R. Cleveland, W. Cleveland, and I. Terpenning, "STL: A Seasonal-Trend Decomposition Procedure Based on Loess," *Journal of Official Statistics*, vol. 6, no. 1, p. 3, 1990.

[10] A. Kejariwal, *Introducing practical and robust anomaly detection in a time series*, 2015. [Online]. Available: https://blog.twitter.com/2015/introducing-practical-and-robust-anomaly-detection-in-a-time-series (visited on 03/10/2017).

[11] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *22nd SIGKDD Conference on Knowledge Discovery and Data Mining*, 2016.

[12] J. Zhi-Qiang, W.-J. Xie, M.-X. Li, B. Podobnik, W.-X. Zhou, and H. E. Stanley, "Calling patterns in human communication dynamics," *Proceedings of the National Academy of Sciences*, vol. 110, no. 5, pp. 1600–1605, 2013. DOI: 10.1073/pnas.1220433110.

[13] *Atis/sip forum ip-nni task force*. [Online]. Available: http://www.atis.org/01_strat_init/IP-NNI/index.asp (visited on 03/10/2017).