JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 33, 1433-1446 (2017)
DOI: 10.6688/JISE.2017.33.6.3

An Adaptive Rule-Based Approach to Resolving Real-Time VoIP Wholesale Billing Disputes

NAZISH YAQOOB[1], SEEMAB LATIF[1], RABIA LATIF[1],
HAIDER ABBAS[1,2,3] AND ASIF YASEEN[4]
[1]*National University of Sciences and Technology*
Islamabad, 44000 Pakistan
[2]*King Saud University*
Riyadh, 11653 Saudi Arabia
[3]*Florida Institute of Technology*
Melbourne, FL 32901 USA
[4]*The University of Queensland*
Brisbane, 4072 Australia
E-mail: hsiddiqui@ksu.edu.sa; dr.h.abbas@ieee.org

The Voice over Internet Protocol (VoIP) industry has grown immensely since its inception, and is predicted to grow at double the rate in the coming years. The growth of the VoIP industry has made significant contributions to the economy, and has also increased the volume of data, which is a challenge for processing. VoIP security vulnerabilities and lack of appropriate tools and infrastructure can lead to billing disputes and fraud attacks, impacting VoIP wholesalers' profits. To reduce economic losses, these challenges need to be addressed in a comprehensive and efficient way. This study proposes an intelligent and adaptive rule based reconciliation process to resolve real-time billing disputes with minimal revenue loss. Real-time disputed Call Detail Records are analyzed to generate adaptive rules to cater for dynamic data sources. These rules are used to classify the Call Detail Record into six categories. A summarized report is generated at the end of the analysis that can be used to come to a better resolution during the billing dispute negotiation process. The complexity and volume of data affects the execution time of reconciliation processes. Spark, a distributed processing framework, is used to reduce execution times. The distributed processing solution has reduced execution times by 81.8% on average as compared to non-distributed solutions. The performance of the proposed solution is evaluated against the CALLS Dispute management system (an Aiztek Technologies solution), and the proposed solution has detected 38% more billing disputes in less time as compared to the existing solution.

Keywords: call detail record (CDR), billing dispute, VoIP wholesale, distributed processing, distributed rule-based algorithm, reconciliation process

1. INTRODUCTION

The VoIP industry has grown phenomenally in recent years. There has been a gradual market shift from analogue Plain Old Telephone System (POTS) to digital VoIP telephony recently [1]. Growth of VoIP industry has had a valuable impact on the economy, and growth is expected to quicken in coming years. Infonetics Research says that collectively $377 billion has been spent on business and residential VoIP services between 2012 and 2016[1]. Persistence Market Research has estimated the value of the

Received December 28, 2016; revised February 5, 2017; accepted March 19, 2017.
Communicated by Basit Shahzad.

[1] http://www.tmcnet.com/voip/ip-communications/articles/311105-voip-uc-markets-see-continued-strong-growth-over.htm

VoIP market at $85.9 billion for the year 2015, which is expected to rise to $194.5 billion by 2024[2].

Wholesale VoIP is a service provided by wholesale carriers to other service providers, and provides services to start-ups and extensions to their networks[3,4]. A VoIP wholesaler who buys resource traffic from one wholesaler and then sells it to another wholesaler is known as a Simple Wholesaler. The start-point of a route for traffic is termed an Origination Point and the end-point of a route for traffic is termed a Termination Point. The terminator may be a wholesaler or a retailer. Thus, a wholesaler who acts in the role of termination point is termed a Termination Wholesaler. Software that is used to transit VoIP traffic between the carriers is known as a softswitch. The main function of a softswitch is to maintain the continuous routing of the large volumes of long-distance VoIP calls. Wholesaler profits are generated from the difference of buying prices and selling prices. Thus, their main objective is to find the cheapest price for the required route and then to sell it on at a higher price.

Billing system is crucial to VoIP services; it should be reliable and accurate. VoIP service providers rely on billing to charge their customers for every billable service. Incorrect billing of services results in avoidable loss of revenue, which is obviously unacceptable in today's competitive business climate. On the other hand, VoIP subscribers believe that they should be charged only for calls they have made and based on accurate call duration. VoIP billing should be reliable and free from any inconsistencies, which result in billing disputes between the service provider and the subscriber.

Session Initiation Protocol (SIP) and Real-time Transport Protocol (RTP) are integral parts of VoIP systems. These protocols are vulnerable to billing attacks such as call establishment hijacking, call termination hijacking, call forwarding hijacking, inviter replay, fake busy, bye delay, and bye drop. These attacks cause unauthorized calls and longer than authorised call durations. As a result, subscribers are overcharged or they are billed for calls not made by them [1, 2]. Other reasons for inconsistencies are network latency, slow VoIP switches and time-zone differences that cause discrepancies in call time. All these discrepancies can lead to billing disputes, and to overcome these discrepancies, some reconciliation mechanism is required.

Reconciliation is the processes of matching the Call Detail Records (CDRs) of the provider and of the subscriber call-by-call to find inconsistencies. CDRs provided by different operators are in different formats, which raise a problem for analysers/researchers trying to interpret different data fields. CDRs do not have unique call IDs, which leads to uncertainty while attempting to match corresponding calls [3]. Other difficulties faced during dispute resolution include differences in time zones and the volume of CDR data. A large volume of data requires more efficient methods than would be used with traditional dispute handling.

The reconciliation process is required to resolve the billing dispute between two parties. To perform a reconciliation process on high-volume and complex data, a framework is required that can process a large volume of heterogeneous data with accurate reconciliation. In this research, a system is proposed which uses a Big Data platform to deal with large volumes of data. To perform accurate reconciliation, an adaptive rule-based approach is proposed. The proposed system also provides an analysis summary report and reconciled CDRs. This summary report serves an evidentiary purpose, and

[2] http://wholesale-voip.tmcnet.com/articles/421746-global-voip-service-market-could-reach-1945-billion.htm
[3] https://www.voicebuy.com/what-is-wholesale-voip-or-voip-wholesale/
[4] http://blog.voipinnovations.com/blog/index.php/what-is-wholesale-voip

facilitates the negotiation process. A wholesaler can analyse CDRs and can provide evidence of disputed calls to the other party, and hence negotiate to resolve the dispute, thus making dispute resolution process transparent.

This paper is organised as follows: Section 2 describes literature review, Section 3 includes details regarding datasets used in this research and it's pre-processing, Section 4 explains the rule-based reconciliation algorithm, Section 5 presents results and analysis, and finally Section 6 concludes the work with future directions.

2. LITERATURE REVIEW

CDRs are generated by VoIP switches and contain information on call duration, calling number, called number, time and date, origin of the call, and other details. CDRs are primarily used for billing purposes. The information provided in CDRs is very useful; it has been used in many research contexts such as those pertaining to fraud detection and fraud prevention [4-10], QoS and anomaly detection [11, 12, 17, 18], business marketing [13-15], and resolving billing issues [1, 2, 16]. Research carried on CDRs has a significant contribution in improvement of overall telecommunication network.

CDRs are extensively used to detect failure in voice telecommunication systems. Breda and Mendes [11] proposed two algorithms; a real-time algorithm and a sample space algorithm for failure detection. The real-time algorithm is more efficient on detection of extensive failures whereas the sample space algorithm is effective on more sensitive and smaller proportion failure detection. A study carried out by Gaspar and Gocza used CDR for VoIP quality assessment [12]. They presented a new approach to assess VoIP quality using Bayesian networks that overcomes the issue of power consumption from intensive computation.

Many studies have been carried out using CDRs for fraud detection based on different approaches such as a rule-based approach, Neural Networks, Naïve Bayes and Fuzzy Logic. Lack of telecom equipment can result in corrupted and/or missing values in CDRs. These corrupted and missing values in CDRs may contain fraud evidence [4]. The study carried out by Baharim and Kamaruddin uses Naïve Bayes approach to explore useful information present in rejected proportion, which is then used for fraud analysis. Another study carried out by Augustin *et al.* uses a rule-based approach that performs well on real-time CDRs provided by a service provider with only a 4% false positive rate [7]. Rule-based models have also been proposed in research carried out by Padmavathamma and Rajani to detect fraud [8]. Their system not only detects fraud with minimum false alarm rate but also provides reason for the alarm. Another work by Hoffstadt *et al.* proposed multilayer SUNSHINE framework architecture for detection and prevention of VoIP fraud and misuse [9]. The techniques used in this research are profiling, clustering, neural networks, and self-organizing maps. To detect grey traffic in VoIP services, Anwar and Shabbir used CDRs [10]. Their analysis used pcap programming to extract call-log details and then used CDRs to classify VoIP traffic into legal and illegal traffic.

Growth in the telecommunication industry has increased competition. Operators seek different ways to keep subscribers on their networks. CDRs are also used to extract useful patterns for telecom promotions. The increasing volume of CDR is also a challenge

for efficient analysis. A study carried out by Jayawardhana *et al.* proposes scalable architecture of a system called Karnataka [13]. Karnataka uses a NoSQL approach to query the periodic behaviour of subscribers to decide whether the subscriber is eligible for a promotion or not. Another study carried out by Lin and Wan clusters the customer behaviour using K-mean clustering to provide practical proposals for market promotion campaigns [14]. K-mean clustering helps to achieve distinguishable groups that aid better marketing strategy.

CDRs are also used for reconciliation processes to find discrepancies in call records, billing comparisons, statistical discrepancies in CDRs, and billing disputes. Reconciliation is the process of ensuring that records in two CDRs are in agreement. This is the only process through which two CRDs are compared, the only difference lies in how this process is performed – for example, reconciliation is mostly performed by manual comparison of two CDRs by performing automated record-by-record comparison or by defining some rules for comparison.

The CALLS Dispute Management System is a software solution developed by Aiztek Technologies[5] for resolving billing disputes and is offered as a service to its customers. The CALLS Dispute Management System uses a rule-based reconciliation process to compare CDRs of different operators to identify any differences in billing. It is developed in a non-distributed framework with a fixed set of rules. Its processing time, on average, is over 70 minutes per CDR, and accuracy is 89% due to a fixed rules set. There are quite a few systems developed for resolving billing disputes: for example, Bill Reconciliation from Arcavia compares electronic bills to find differences in billing, Dispute Resolution, CDR Comparison from R&R performs automated CDR comparison to find differences and analyses statistical discrepancies. Different companies offer these systems as services. Performance analysis of these systems, in terms of execution time and accuracy, is not available or easily accessible. We selected the CALLS Dispute Management System as a benchmark and compared our solution with this system, since performance analysis of this system is available.

3. DATA USED IN RESEARCH

For this research, we have taken three real-time disputed CDR data sets from a local VoIP wholesaler, shown in Table 1. Each CDR contains 10 attributes; of which 5 were used in the analysis, as these are the common attributes present in every CDR. The other 5 attributes vary from wholesaler to wholesaler. Selected attributes are shown in Table 2. Each dataset consists of two CDR files in CSV format, one pertaining to the service provider and the other pertaining to the service subscriber. These CDRs are named as '*S*' and '*P*', where *S* represents subscriber's data and *P* represents provider's data.

3.1 Data Pre-processing

CDRs provided by different operators have heterogeneous formats. The differences of formats raise some problems while preforming analysis, which need to be resolved. The problems include missing unique IDs, differences in attributes, different formats used while recording date, and difference in names of attributes. Some fields are common to all CDRs, such as start time, end time, called number, and call duration. For

[5] http://www.aiztek.com

this research, these common attributes were selected for the analysis. A unique ID was added to each file to identify each record. The row number of a record and first letter of the file name is concatenated to make a unique ID, *e.g. P*1 is the ID for the first row of the provider CDR. Start time, end time and date has different formats in subscriber and provider CDRs. A common string format for both the CDRs (*i.e.*, "*day/Month/Year Hour*: *Min:Sec AM/PM*") was used. Both start and end times are converted to common string formats before analysis. Common attribute names are also assigned for example S.ID means subscriber's ID and P.Start_time is provider's call start time.

For this research, Spark was used for implementing the proposed algorithm. Spark performs better on parquet file format, so both files in each dataset are covered from CSV to parquet file format. Parquet is columnar storage format that consumes less memory space and follows type-specific encoding.

Table 1. Datasets used in research.

File Name (CDR)	Number of Calls (Dataset)		
	1	2	3
Provider	246,412	134,939	11,119
Subscriber	249,962	137,324	11,795

Table 2. Call detail record attributes.

CDR attribute	Description
Start time and date	Start time and date of a call
End time and date	End time and of a call
Called Number	Phone number of called party
Duration	Duration of a call
ID	Unique Call record ID

4. PROPOSED DISTRIBUTED RULE-BASED ALGORITHM

A reconciliation process involves matching of calls record by record in order to find a corresponding match. It is a time consuming and costly process due to lack of widely acceptable standards. This paper proposes a rule based reconciliation algorithm to effectively and efficiently resolve billing disputes. The aim of the algorithm is to classify all the records in both CDRs into six pre-defined categories, given in Table 3. To classify the disputed data, a wholesaler needs these categories. The Exact Call category contains those records that are exactly matched. These are non-disputed records. Call not within margin, subscriber not found, and provider not found are the categories that contain disputed records.

Table 3. Categories.

Category	Description
Exact Call	The call is perfectly matched and there is no doubt in corresponding match. Rule 2 define this category
Similar Call	The call is matched but with configurable threshold. Rule 3,4,5,6 represent these calls
Subscriber Call Not Found	The Provider's CDR call do not have corresponding match in Subscriber's CDR. Rule 7 define this category.
Provider Call Not Found	The Subscriber's CDR call do not have corresponding match in Provider's CDR. Rule 8 define this category.
Calls Not within Margin	If Provider or Subscriber's CDR deviates from allowed threshold. Rule 9 defines this category.
Duplicate Call	The record has more than one copy in same file. Rule 1 defines this category.

This classification is done through number of rules, as defined by Algorithm 1. Rules are defined after interviewing the domain experts and analysing the shortcomings of the CALLS Dispute Management System. Inputs to this algorithm are pre-processed provider and subscriber CDRs P and S respectively, empty rule set R and threshold T. Thresholds are used to classify 'Similar_Call' category in to sub-categories. If two calls do not match exactly, then an adjustable threshold value is needed to reconcile the calls. This threshold is the error margin. The maximum value of threshold or upper limit is the maximum error margin that is acceptable by both disputed parties. Similarly, the minimum value of threshold is the lower limit of error margin that is acceptable. All calls that are within range of the upper and lower limits of threshold are categorised as "Similar_Call". All those calls that are not within the margins are considered as unmatched and are categorised as "Call not within margin".

As depicted in the algorithm, all duplicate and exact calls are filtered using rules R1and R2 respectively. If two calls do not match exactly then an adjustable threshold value is needed to reconcile the calls using rule R3, R4, R5 or R6 and categorized as similar call category. In some cases, a call has multiple matches within range of threshold. For example in Table 4, when the upper limit of a threshold is 5, provider CDR's call 3P has two matches in subscriber CDR 2S and 14S. In such cases, the call is matched with the call having minimum difference of call duration and time. In Table 4, 3P is matched with 2S as it has minimum difference for both call duration and time. If a provider's call does not have a subscriber's match or a subscriber's call does not have its corresponding provider's match, it is filtered using rules R7 and R8 respectively and assigned categories accordingly.

Algorithm 1: Rule based Reconciliation Algorithm

Input: Provider CDR as P {Start_time, End_time, Called_No, Duration}, Subscriber CDR as S {Start_time, End_time, Called_No, Duration}, Threshold as T{UpperLimit, LowerLimit}, Empty Rule_Set R
Output: Category C and Subcategory SC
where Category = {'Duplicate_Call', 'Exact_Call', 'Similar_Call', 'Subscriber_Call not found', 'Provider_Call not found', 'Calls not within margin'}
Subcategory = {'Same duration but start and end time are within margin', 'Same Start time but duration and end time are within margin', 'Same end time but duration and start time are within margin', 'Duration, end time and start time are within margin'}
begin
1: R1: **if** (P.Start_time = P.Start_time ∧ P.Called_No = P.Called_No ∧ P.End_time = P.End_time ∧ P.Duration = P.Duration)
2: **then** C = 'Duplicate_Call' ∧ SC = ''
3: **end if**
4: R2: **if** (P.Start_time = S.Start_time ∧ P.End_time = S.End_time ∧ P.Called_No = S.Called_No ∧ P.Duration = S.Duration)
5: **then** C = 'Exact_Call' ∧SC = ''
6: **end if**
7: R3: **if** (P.Called_No = S.Called_No ∧P.Duration = S.Duration ∧T.UpperLimit ≥ (P.Start_time-S.Start_time) ≥ T.LowerLimit ∧ T.UpperLimit ≥(P.End_time – S.End_time) ≥ T.LowerLimit)
8: **then** C = 'Similar_Call' ∧SC = 'Same duration but start and end time are within margin'
9: **end if**
10: R4: **if** (P.Called_No = S.Called_No ∧T.UpperLimit ≥ (P.Duration – S.Duration) ≥ T.LowerLimit ∧ P.Start_time = S.Start_time ∧ T.UpperLimit ≥ (P.End_time – S.End_time) ≥ T.LowerLimit)
11: **then** C = 'Similar_Call' ∧ SC = 'Same Start time but duration and end time are within margin'

```
12:     end if
13:     R5: if (P.Called_No = S.Called_No ∧ T.UpperLimit ≥ (P.Duration – S.Duration) ≥ T.LowerLimit
              ∧T.UpperLimit ≥(P.Start_time - S.Start_time) ≥ T.LowerLimit ∧ P.End_time = S.End_time)
14:           then C = 'Similar_Call' ∧ SC = 'Same end time but duration and start time are within margin'
15:       end if
16:     R6: if (P.Called_No = S.Called_No ∧T.UpperLimit ≥ (P.Duration – S.Duration) ≥ T.LowerLimit
              ∧T.UpperLimit ≥ (P.Start_time – S.Start_time) ≥ T.LowerLimit) ∧T.UpperLimit ≥
               (P.End_time – S.End_time) ≥ T.LowerLimit)
17:           then C = 'Similar_Call' ∧ SC = 'Duration, end time and start time are within margin'
18:       end if
19:     R7: if   (for P.Called_No corresponding record in S.Called_No does not exist)
20:           then C = 'Subscriber_Call not found' ∧SC = ''
21:       end if
22:     R8: if (for S.Called_No corresponding record in P.Called_No does not exist)
23:           then C = 'Provider_Call not found' ∧SC = ''
24:       end if
25:     R9: If (P.Called_No = S.Called_No ∧ (P.Duration – S.Duration) > T.UpperLimit ∧ (P.Start_time –
              S.Start_time) > T.UpperLimit ∧(P.End_time – S.End_time) > T.UpperLimit ∧ (P.Duration –
              S.Duration) < T.LowerLimit ∧ (P.Start_time – S.Start_time) < T.LowerLimit ∧ (P.End_time –
              S.End_time) < T.LowerLimit
26:           then C = 'Call not within margin' ∧SC = ''
27:       end if
End
```

Table 4. Multiple matches of provider's call.

P.ID	P. Start_Time	P.End_Time	P.Duration	P.Called_Number
3P	20/09/2015 11:57:33 PM	21/09/2015 12:04:35 AM	422	971507921047
2S	20/09/2015 11:57:32 PM	21/09/2015 12:04:35 AM	423	971507921047
14S	20/09/2015 11:57:32 PM	21/09/2015 12:04:37 AM	425	971507921047

For evaluation, the algorithm was first implemented using the Java language in a non-distributed framework. The execution time for the data was over 70 minutes, which is not suitable for real-time processing. A distributed approach is required to process real-time data. Spark, a big data distributed processing framework, was used to implement the proposed algorithm. Spark was used with Hive, which provides SQL based analysis. It is a distributed computing framework based on the Hadoop MapReduce algorithm and is a lazy learner that stores intermediate and final results in the memory. Data computation efficiency is improved due to in-memory analysis.

5. ANALYSIS AND RESULTS

The proposed algorithm is applied on datasets given in Table 1. Dataset 1 is more complex and bigger in size than datasets 2 and 3. These datasets are used to test and analyse the proposed algorithm with respect to execution time and accuracy in terms of dispute analysis performance.

5.1 Execution Time

The execution time of the proposed algorithm is compared for distributed and non-distributed implementation. It is also compared with the execution time of CALLS Dis-

pute management system, discussed in Section 2. For all the comparisons, the algorithm is executed 10 times and average values were considered for the comparison. The CALLS Dispute management system took 72 minutes and 47 seconds on Dataset 1, 49 minutes and 10 seconds on Dataset 2, and 15 minutes and 29 seconds for Dataset 3 execution. On the other hand, our proposed algorithm took 14 minutes and 6 seconds for Dataset 1, 8 minutes and 31 seconds for dataset 2, and 1 minute and 41 seconds on Dataset 3 execution. This comparison is shown in Table 5.

Table 5. Execution time of CALLS dispute management system and proposed solution.

Data set	Execution Time (minutes: seconds)			
	CALLS Dispute Management System	Proposed Solution		
		Non-Distributed	Distributed	Reduction in Execution time (Non-Distributed and Distributed)
Dataset 1	72:47.089	70:54.306	14:06.356	80.0%
Dataset 2	49:10.006	44:23.198	8:31.436	81.2%
Dataset 3	15:29.734	9:04.006	1:41.336	84.4%

Our proposed algorithm's non-distributed implementation has execution time less than CALLS Dispute management system. The execution time varies with the size and the complexity of the dataset. Another reason for the better execution time of the proposed solution is its distributed implementation.

5.2 Accuracy of Dispute Analysis

The dispute analysis results of the proposed system are compared with those of the CALLS Dispute management system based on categories and number of matched calls. The CALLS Dispute management system provides one file at the end of the analysis that contains categories of CDR calls. The Categories of the CALLS Dispute management system are shown in Table 6 along with the corresponding categories of the proposed system. Our system has two additional categories "Duplicate Call" and "Calls Not within Margin" which are not present in the CALLS Dispute management system. For "Different Date Time" our solution provides detailed information about differences in time duration using adaptive thresholds. This information will help both parties to negotiate effectively and bring transparency to the dispute resolution process. The threshold values are chosen with customer's agreement, it is adjustable and also helps to resolve dispute with minimum loss. This feature is missing in the CALLS Dispute management system.

Table 6. CALLS dispute management system categories.

Category	Description	Corresponding Category in Proposed Solution
Same Call	Call is perfectly matched and there is no doubt in corresponding match.	Exact Call
Different Date Time	The call date and time, either start or end is different.	Similar Call, also provides details of differences in duration using adaptive threshold
Local Not Found	Provider CDR call does not have corresponding call in subscriber CDR.	Provider Call Not Found
Foreign Not Found	Subscriber CDR call does not have corresponding call in provider CDR.	Subscriber Call Not Found

Table 7. Comparison of CALLS dispute management system and proposed solution.

Category		Number of Calls		
		CALLS Dispute Management System	Proposed Solution	Actual (Manual Analysis)
Dataset 1				
Same Call/ Exact Call		128687	128674	128674
Different Date Time/ Similar Call *Threshold *T* ± 1.0 sec	Total	105693	100,948	100,950
	Same duration but start and end time are within *T**	NA	18721	18723
	Same start time but duration and end time are within *T**	NA	70214	70214
	Same end time but duration and start time are within *T**	NA	11848	11848
	Duration, end and start time are within *T**	NA	165	165
Local Not Found / Provider Call Not Found		5601	5404	5403
Foreign Not Found / Subscriber Call Not Found		9981	2866	2865
Duplicate Call	Provider CDR	NA	278	278
	Subscriber CDR	NA	289	289
Calls Not within Margin	Provider CDR	NA	10543	10543
	Subscriber CDR	NA	11781	11781
Dataset 2				
Same Call/ Exact Call		40708	40707	40707
Different Date Time/ Similar Call *Threshold *T* ± 1.0 sec	Total	94228	89121	89124
	Same duration but start and end time are within *T**	NA	16380	16380
	Same start time but duration and end time are within *T**	NA	55016	55016
	Same end time but duration and start time are within *T**	NA	17725	17727
	Duration, end and start time are within *T**	NA	0	1
Local Not Found / Provider Call Not Found		2308	2385	2385
Foreign Not Found / Subscriber Call Not Found		80	42	38
Duplicate Call	Provider CDR	NA	47	47
	Subscriber CDR	NA	53	53
Calls Not within Margin	Provider CDR	NA	5023	5023
	Subscriber CDR	NA	5062	5062
Dataset 3				
Same Call/ Exact Call		5075	5073	5073
Different Date Time/ Similar Call *Threshold *T* ± 1.0 sec	Total	5016	4216	4214
	Same duration but start and end time are within *T**	NA	2005	2004
	Same start time but duration and end time are within *T**	NA	1275	1276
	Same end time but duration and start time are within *T**	NA	872	870
	Duration, end and start time are within *T**	NA	64	64
Local Not Found / Provider Call Not Found		203	956	958
Foreign Not Found / Subscriber Call Not Found		93	1052	1052
Duplicate Call	Provider CDR	NA	223	223
	Subscriber CDR	NA	256	256
Calls Not within Margin	Provider CDR	NA	307	307
	Subscriber CDR	NA	242	242

Table 8. Confusion matrix and % error.

Dataset 1											
Category		CALLS Dispute Management System					Proposed Solution				
		TP	TN	FP	FN	% Error	TP	TN	FP	FN	% Error
Same Call/ Exact Call		128674	121275	13	0	0.0	128674	121288	0	0	0.0
Different Date Time/ Similar Call		88642	131961	17051	12308	0.12	100762	148826	186	188	0.001
Local / Provider Call Not Found		5284	244242	317	119	0.001	5386	244541	18	17	0.0
Foreign / Subscriber Call Not Found		2865	239981	7116	0	0.03	2865	247096	1	0	0.0
Duplicate Call	Provider	N/A					278	249684	0	0	0.0
	Subscriber						289	249673	0	0	0.0
Calls Not in Margin	Provider CDR	N/A					10543	239419	0	0	0.0
	Subscriber						11781	238181	0	0	0.0
Dataset 2											
Category		CALLS Dispute Management System					Proposed Solution				
		TP	TN	FP	FN	% Error	TP	TN	FP	FN	% Error
Same Call/ Exact Call		40707	96616	1	0	0.0	40707	96617	0	0	0.0
Different Date Time/ Similar Call		86104	40076	8124	3020	0.08	89119	48198	2	5	0.0
Local / Provider Call Not Found		2198	134829	110	187	0.002	2381	134935	4	4	0.0
Foreign / Subscriber Call Not Found		2	137285	1	36	0.0	38	137282	4	0	0.0
Duplicate Call	Provider	N/A					47	134892	0	0	0.0
	Subscriber						53	137271	0	0	0.0
Calls Not in Margin	Provider	N/A					5022	129917	0	0	0.0
	Subscriber						5062	132262	0	0	0.0
Dataset 3											
Category		CALLS Dispute Management System					Proposed Solution				
		TP	TN	FP	FN	% Error	TP	TN	FP	FN	% Error
Same Call / Exact Call		5073	6720	2	0	0.0	5073	6722	0	0	0.0
Different Date Time/ Similar Call		4118	6683	898	96	0.08	4214	7579	2	0	0.0
Local / Provider Call Not Found		203	10082	755	1510	0.19	956	10837	0	2	0.0
Foreign / Subscriber Call Not Found		93	9784	959	1918	0.24	1052	10743	0	0	0.0
Duplicate Call	Provider	N/A					223	11572	0	0	0.0
	Subscriber						256	11539	0	0	0.0
Calls Not n Margin	Provider	N/A					307	11488	0	0	0.0
	Subscriber						242	11553	0	0	0.0

Calls that come under the "similar call" category are negotiable *i.e.* both subscriber and provider negotiate on the error margin they are going to consider to match these calls. The proposed solution provides detailed analysis of these calls by further sub-categorizing them. This feature is missing in the existing solution. The Accuracy of the CALLS Dispute Management System for this category is as low as 89.7% as compared to the proposed solution accuracy, which is 99.9%.

As discussed earlier, call not within margin, subscriber not found, and provider not found are the categories that contain disputed records. CALLS Dispute management system does not take into account "call not within margin" and "Duplicate Call" category, thus missing an important aspect of billing disputes. According to the confusion matrix given in Table 8, False Positives (FP) for subscriber not found and provider not found categories is high for the CALLS Dispute management system as compared to the proposed solution's FP rate. False Negative (FN) is also very high for the CALLS Dispute management system as compared to the proposed solution. In a billing dispute, FP and FN both are very important as in the former case the subscriber will be charged for calls that are not actually made and in the latter case provider will be facing loss as they will not be paid for the services that have already been provided.

Reports generated by the proposed solution were given to the research partner organization for evaluation. Evaluation shows that our report gives better results than the CALLS Dispute management system. The output report and compared CDR files will increase trust and satisfaction of customers as it provides evidence of discrepancy with detailed information, and brings transparency to the dispute analysis.

6. CONCLUSION AND FUTURE DIRECTIONS

VoIP industry has witnessed the rapid growth and it is prediction for the coming years seems more promising. The impact of VoIP growth on economy is treasured. The vulnerabilities and challenges of VoIP need to be addressed to maximise economic benefits and provide people with cheap communication. Billing disputes and fraudulent attacks can affect the VoIP wholesaler's business profits considerably. The proposed solution supports VoIP wholesalers in resolving billing disputes with minimum economic loss. The proposed solution not only resolves billing disputes, but also overcomes the issues of large data volumes and complexity challenges, which affect execution times. The proposed system has improved on the performance of the CALLS Dispute management system on real-time claimed disputes. The proposed solution is based on distributed processing and handles the large volume of data effectively, yet it needs to be tested for scalability to fulfil future needs. One possible future research direction would be to implement and test the proposed system in real-time analysis, thus evaluating if it is practically useable in the VoIP industry.

ACKNOWLEDGEMENTS

The authors would like to extend their sincere appreciation to the Deanship of Scientific Research at King Saud University for its funding of this research through the Research Group Project no. RG-1435-048. We are also grateful to Mr. Irfan Ahmed, CEO of Aiztek Technologies, for providing us with data, guidance, and access to the company's systems for this research.

REFERENCES

1. R. Zhang, X. Wang, X. Yang, and X. Jiang, "On the billing vulnerabilities of SIP-

based VoIP systems," *Computer Networks*, Vol. 54, 2010, pp. 1837-184.
2. R. Zhang, X. Wang, X. Yang, and X. Jiang, "Billing attacks on SIP-based VoIP systems," in *Proceedings of the 1st USENIX Workshop on Offensive Technologies*, 2007, p. 18.
3. S. Tartarelli, N. d'Heureuse, and S. Niccolini, "Lessons learned on the usage of call logs for security and management in IP telephony," in *Proceedings of IEEE Communications Magazine*, Vol. 48, 2010, pp. 76-82.
4. K. N. Baharim, M. S. Kamaruddin, and F. Jusof, "Leveraging missing values in call detail record using naive bayes for fraud analysis," in *Proceedings of International Conference on Information Networking*, 2008, pp. 1-5.
5. S. Hofbauer, K. Beckers, G. Quirchmayr, and C. Sorge, "A lightweight privacy preserving approach for analyzing communication records to prevent VoIP attacks using toll fraud as an example," in *Proceedings of the 11th IEEE International Conference on Trust*, *Security and Privacy in Computing and Communications*, 2012, pp. 992-997.
6. S. Augustin, C. Gaißer, J. Knauer, M. Massoth, K. Piejko, D. Rihm, and T. Wiens, "Telephony fraud detection in next generation networks," in *Proceedings of the 8th Advanced International Conference on Telecommunications*, 2012, pp. 203-207.
7. S. Rajani and P. Padmavathamma, "A model for rule based fraud detection in telecommunications," *International Journal of Engineering Research and Technology*, Vol. 1, 2012, pp. 2278-0181.
8. D. Hoffstadt, E. Rathgeb, M. Liebig, R. Meister, Y. Rebahi, and T. Q. Thanh, "A comprehensive framework for detecting and preventing VoIP fraud and misuse," in *Proceedings of International Conference on Computing*, *Networking and Communications*, 2014, pp. 807-813.
9. U. Anwar, G. Shabbir, and M. A. Ali, "Data analysis and summarization to detect illegal VoIP traffic with call detail records," *International Journal of Computer Applications*, Vol. 89, 2014, pp. 1-7.
10. G. D. Breda and L. de S. Mendes, "QRP08-4: Failures detection in voice communication systems," in *Proceedings of IEEE Globecom*, 2006, pp. 1-5.
11. Z. Gáspár and I. Gócza, "Assessment of VoIP quality using Bayesian networks," in *Proceedings of the 15th IEEE Mediterranean Electro Technical Conference*, 2010, pp. 1389-1393.
12. P. Jayawardhana, D. Perera, A. Kumara, and A. Paranawithana, "Kanthaka: Big data caller detail record (CDR) analyzer for near real time telecom promotions," in *Proceedings of the 4th International Conference on Intelligent Systems*, *Modelling and Simulation*, 2013, pp. 534-538.
13. Q. Lin and Y. Wan, "Mobile customer clustering based on call detail records for marketing campaigns," in *Proceedings of International Conference on Management and Service Science*, 2009, pp. 1-4.
14. O. Jukić and I. Heđi, "The use of call detail records and data mart dimensioning for telecommunication companies," in *Proceedings of the 20th Telecommunications Forum*, 2012, pp. 292-295.
15. D. Kar, P. Misra, P. Bhattacharjee, and A. Mukherjee, "Real time telecom revenue assurance," in *Proceedings of the 7th International Conference on Digital Telecommunication*, 2012, pp. 130-135.

16. Z. Han and Y. Zhang, "Spark: A big data processing platform based on memory computing," in *Proceedings of the 7th International Symposium on Parallel Architectures*, *Algorithms and Programming*, 2015, pp. 172-176.
17. H. Abbas, C. Magnusson, L. Yngstrom, and A. Hemani, "Addressing dynamic issues in information security management," *Information Management and Computer Security*, Vol. 19, 2011, pp. 5-24.
18. M. Hinkka, T. Lehto, and K. Heljanko, "Assessing big data SQL frameworks for analyzing event logs," in *Proceedings of the 24th Euromicro International Conference on Parallel*, *Distributed*, *and Network-Based Processing*, 2016, pp. 101-108.

Nazish Yaqoob received BE degree in Software Engineering from Fatima Jinnah Women University, Rawalpindi, Pakistan and the MS degree in Computer Software Engineering from National University of Science and Technology, Islamabad, Pakistan. Her research interests include big data analytics and machine learning.

Seemab Latif received Bachelor of Software Engineering from Fatimah Jinnah Women University, Pakistan, Master of Software Engineering from National University of Sciences and Technology, Pakistan and Ph.D. from Manchester University, UK. She is working as Assistant Professor in the Department of Computing, SEECS, National University of Sciences and Technology, Pakistan. Her research interest includes artificial intelligence, machine learning, data mining, NLP. Her professional services include Industry Consultations, Conference Chair, Technical Program Committee Member and reviewer for several international journals and conferences.

Rabia Latif is currently working as an Assistant Professor at Information Security Department, National University of Sciences and Technology, Pakistan. She received her Bachelor degree in Computer Science from COMSATS Institute of Information Technology, Islamabad and Master degree in Information Security from National University of Sciences and Technology, Pakistan. She has completed her Ph.D. program in the area of cloud-assisted wireless body area networks at National University of Sciences and Technology, Pakistan. Her research interests include wireless body area networks, cloud computing and information security.

Haider Abbas is a Senior Member IEEE and Cyber Security Professional who took professional trainings and certifications from Massachusetts Institute of Technology (MIT), USA, Stockholm University, Sweden, IBM and EC-Council. He received his MS in Engineering and Management of Information Systems (2006) and Ph.D. in Information Security (2010) from KTH, Sweden. He is an associate editor or on the editorial board of a number of international journals. Dr. Abbas also won many awards and received several research grants for ICT related projects from various research funding authorities. He is the principal advisor for several graduate and doctoral students at King Saud University, KSA, National University of Sciences and Technology, Pakistan and Florida Institute of Technology, United States.

Asif Yaseen is a Postdoctoral Research Fellow at the ARC-Industry Transformation Training Centre, The University of Queensland, Australia. He received a Ph.D. in Agribusiness from University of Queensland in February 2015. He earned his master degree in ICT entrepreneurship from KTH Royal Institute of Technology (Stockholm), Sweden. His major research interests are Entrepreneurship and innovation in agri-food sectors and developing intelligent agri-food value chains.