REAL-TIME GREY CALL DETECTION SYSTEM USING COMPLEX EVENT PROCESSING

K.G.D.C. Kehelwala[1,2], H.M.N.D. Bandara[2], R.A. Yasaratne[1], P. De Almeida[1], I.K.K.S. Ilesinghe[1], P.D.K.E. Wickramasinghe[1]
[1]Dialog Axiata PLC, [2]University of Moratuwa

ABSTRACT
Illegal international voice call termination (aka. grey calls) is one of major source of revenue loss in Telco industry. Annually billions of dollars revenue is lost due to this fraud. The effects of illegal termination are felt by both the mobile operators and their end users, as it causes severe quality of service degradations. Hence, detection of Grey callers in real-time allows to nullify the impact. Existing solutions are based on Call Detail Records (CDR) analysis. These tools follow traditional database reliant store first, process then approach, Moreover, CDR collected within a large time window is considered for decision making, thus lacks real-time features and unable to effectively control the fraud. This paper presents a CDR based, real-time grey call detection system using Complex Event Processing (CEP). Our primary focus is to utilize the power of CEP to detect complex, fraudulent caller patterns in real time, in spite of high arrival rate of CDR records.

1. INTRODUCTION

Illegal international voice call termination (aka. Grey Calls, SIMbox fraud or Bypass Calls) is a major source of revenue loss for a telecommunication provider, which also deteriorates Quality of Service (QoS) offered to customers [1]-[3]. This scenario happens when fraudsters accumulate an incoming international call through the Internet via Voice over Internet Protocol (VoIP) and then injects it back to the destination country's telecom network as a local call using local SIM (Subscriber Identity Module) cards installed in a device called SIMbox. Therefore, the call reaches the called party with a local CLI (Calling Line Identity) and billed as a local call.

When the price difference between international call termination and local calls is high fraudsters can easily gain revenue that exceeds the break-even point with few fraudulent calls. So detection of such fraudulent phone numbers with minimum delay allows to regain significant portion of revenue and effectively control this kind of fraud.

Call Detail Records (CDRs) [4] are one of the most valuable data repository of a telecom operator. CDR is the data record that contains information related to a single instance of telephone call or other related transactions. Mobile Switching Center (MSC) or related telecommunication nodes create a CDR record when transaction passed through it. Based on the functionality of telecommunication nodes, level of information in CDR may be vary. In most cases it contains information such as the origination and destination addresses of the call, the time the call started and ended and the duration of the call.

CDR is generated in real time and it creates a stream of events that consists of useful patterns that reflects the user behavior. Data source of most of the bypass detection tools is CDR. But, traditional systems make detections by generating a set of features or aggregate values by querying CDR stored in static database over a large time window and make decisions based on such values. This is time consuming and by that time a fraudster can make number of successful calls before being detected. So real-time analysis of CDR streams is required to support real-time detection of such calls. Also, Grey calls create several unique patterns in CDR streams and systems can effectively use those patterns to detect such fraudulent numbers in real time.

The remainder of this paper is organized as follows. Section 2 describes the theoretical

background related to this subject including in details description about Grey Calls. Section 3 reviews the work related to CDR based grey call detection approaches available in literature. Section 4 presents our research approach by explaining novel feature set, identified complex patterns and experimental setup. The results are compared and contrasted in Section 5. Finally, Section 6 presents conclusion.

2. GREY CALLS

Grey Calls are taken place when the cost of international termination calls is considerably higher than charges for local incoming call in particular country. Therefore impact of this fraud is more severe in certain parts of the world, especially in Asia, Africa and North America.

When international carrier set fraudulent route as their least cost route, fraudster takeover those international calls and transfer it through Internet to the destination country. Then Voice over IP (VoIP) calls are injected to back to destination network via SIM cards which is installed on device called SIMbox. Since this activity bypass the legal international interconnections between Telco operators, international calls are billed as national calls and significant revenue leakage is occurred. SIMbox is the device which is capable of convert VoIP calls to GSM network call. With expansion of technology high capacity SIMboxes with advanced technologies like International Mobile Equipment Identity (IMEI) swapping and intelligent SIM swapping algorithms makes the SIMbox behavior more close to actual subscriber [8].

In addition to the revenue loss SIMbox frauds causes QoS loss for customers, as these calls may reach the SIMbox through a low quality, low bandwidth IP routes. Also, a SIMbox may use 100s of SIMs simultaneously, which may overload the base stations in that area. That may cause sudden call drops and loss of QoS observed by the destination customer for IDD calls, as well as a degraded service is offered to actual customers in that area. Since SIMbox spoofs actual international Calling Line Identification (CLI) with a local CLI, SIMbox fraud may lead to privacy issues. So SIMbox fraud may severely affect customer satisfaction and operator's brand name.

We can divide Bypass fraud into two major categories as Onnet bypass and Offnet Bypass. Onnet bypass means fraudsters use the SIM cards of same network of destination number. This is the common case in most of the countries as calls within same network cost much lower than charges for calls to other operators or international calls. Figure 1 depicts the Onnet scenario. Instead of using the costly high-quality routes that goes through destination networks International Switching enter (ISC), some wholesale carriers' route calls through SIMbox operators connected though the Internet. SIMbox dials that calls via a SIM card of the same network of the destination number.

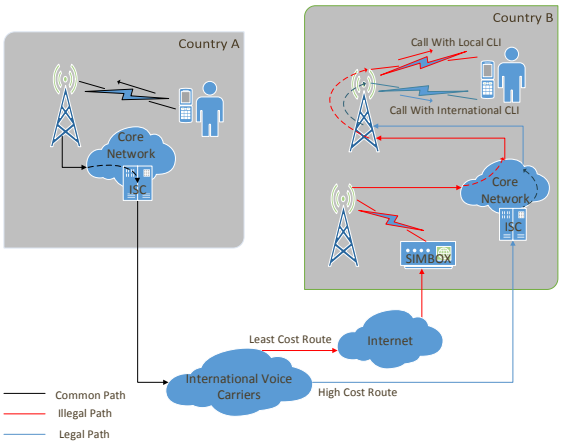

Figure 10: Onnet Bypass

In Offnet bypass fraudsters use the SIM card of a different operator in same country of destination number. Figure 2 depicts this scenario. This happens when local interconnection charges are much lower than international interconnection charges. In some countries local interconnection charges are much closer to the international interconnection charges and Offnet bypass do not take place.

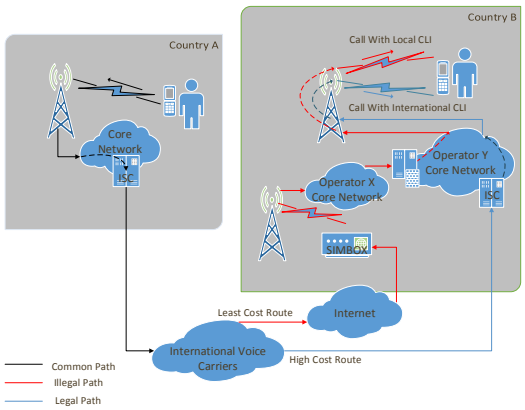

Figure 11: Offnet Bypass

Detection of SIM cards used for SIMbox fraud is a challenging task for mobile operators. There are two major approaches. First approach is actively originating calls to the target network via test units installed in several parts of the world and scan the CLI of those calls. Even though this process detects SIMbox numbers in real time, it is not capable of capturing majority of numbers. Second approach is CDR based analysis. CDRs are loaded into relational database and queries are used to identify fraudulent numbers. Rich set of attributes are needed to derive by summarizing CDR in order to achieve effective detection. In context of Onnet bypass, operator have more details including Location and Owner Information. But detection of Offnet bypass have to be performed with limited details and strong pattern mining techniques are required.

Efficient detection process must consist minimum false positive (Detect Genuine customers as Fraudulent) and false negative (Classify Fraudulent numbers as Genuine) values. In addition to that, number of attempts made by fraudulent SIM card before detection is another important factor. If this value is very high, Fraudster can cover the cost before disconnection of SIM card. Since Telecom industry is highly competitive, in most of the countries fraudsters can buy new SIM cards for very little cost. So traditional analysis methods fails here as that is based on past CDR analysis. Therefore, when operator disconnects, fraudsters use new set of SIM cards as they are able to cover profit margin. This process continues and actual task of detection process becomes damage control function.

3. RELATED WORK

This section summarize the prominent research work related to CDR based grey call detection while critically evaluating those approaches.

Elmi et al. [5] proposed an Artificial Neural Network (ANN) method to address SIMbox fraud. They have used supervised learning method with Multi-Layer Perception (MLP) as classifier. ANN is used because of its generalizing capabilities, ability to learn complex patterns and trends within noisy data and better performance records in this domain. This system derives nine features using CDR in dataset and calculated corresponding values for each calling subscriber.

Two years later the same authors came up with a Support Vector Machine (SVM) based approach [6] for the same dataset and compared its results with the ANN based approach. They have used 10 fold cross validation technique while using same features in [5].

Even though authors were able to achieve high accuracies and lower running time in both the cases [5], [6] sustainability of this approach in practical scenario is questionable due to many reasons. When we consider the dataset its size is much smaller than typical mobile network operator. Also, they have considered CDRs from one cell id only and ratio between legitimate to fraudulent subscribers is approximately 2:1. But in real worlds more than 20,000 cell IDs [9] need to be considered and percentage of SIMbox numbers out of total customer base is very low. So we can conclude that cardinality of dataset is inferior to actual cases.

Also, they have considered CDR for two months when calculating features. But the actual requirement in SIMbox detection need to perform as early as possible. So there is no point of performing calculations within a few seconds as long time window is used for feature calculation. Additionally, scalability of these methods with large datasets were not evaluated in both the papers.

Murynets et al. [7] have presented a novel classifier for fraudulent SIMbox detection. They have selected CDR fields that accounts important details including origination and termination party details, origination time, duration, location details

and device details. Calling party's account age and corresponding customer segment are the novel fields which were absent in the previous cases [5], [6]. So this paper moved a positive step by selecting better set of CDR fields relative to [5], [6].

By considering IMEI details it is possible to block the confirmed IMEIs or detect new SIM cards that are inserted into a SIMbox with a particular IMEI. But the detection logic cannot too much depend on that since advanced SIMboxes allow changing IMEIs [8].

Authors have derived 48 features using CDR. Some of those attributes are taken directly from CDR fields while others are derived by subjecting CDR fields to simple mathematical operations. It is important to note that the feature set is per IMEI basis and they have targeted to identify SIMbox rather than SIM cards used for IMEI. This approach may incur problems as advanced SIMboxes can replace IMEIs with dummy values or other genuine IMEIs. So blocking IMEI numbers may block some genuine customers. Additionally IMEI to MSISDN mapping give false values. Since choice of device is customer's right, operator has no control on IMEI. So applicability of this system directly in practical environment can be questioned. Better option is feature calculation per MSISDN basis.

When we consider dataset, majority of the features calculated for data collected over one week period from tier-1 cellular operator in United States. So dataset is considerably larger than the previous cases. But one week period is still higher as operator loses considerable amount of revenue over that period.

Critical evaluation of above approaches reveals many areas that were not focused and thus opens up new research topics. Those facts are summarized below:

- Existing solution have only targeted the accuracy and running time of classification algorithm while considering large time window for feature calculation. Those solutions did not interpreted SIMbox detection as time sensitive operation. So these solutions are incapable of preventing financial losses as fraudsters can make profits easily by operating safely within that time window before disconnection. So in order to make near real-time detections rich set of features should be generated for short time window and classification algorithm need to be optimized according to that.
- These approaches are only capable of detecting Onnet Bypass and did not pay any attention for Offnet Bypass. Both makes similar kind of financial losses for many telecom operators. Features like Location, IMEI, IMSI, and Account type details may not be available for Offnet SIMbox numbers. So a rich set of novel features with additional measures is required to detect Offnet Bypass.
- Features are generated based on calling party behavior only. But by considering called party behavior a valuable set of attributes can be derived. For example, counting the subset of called party numbers which has received IDD calls and also belongs to set of called party numbers dialed by the considered calling party will be valuable feature in context of grey call detection.
- Previous cases have targeted CDR data stored in static databases. But CDRs records are generated at real time in telecommunication nodes and flows as streams of data. So in order to gain maximum advantage in competitive business world, the new mechanism that is capable of directly processing the multiple streams is required. Also, that mechanism should support multiple CDR streams generated by Telco nodes, as well as some static data simultaneously.
- A typical mobile service providers has a customer base of more than ten million. So data streams with very high transaction rate are generated at Telco nodes. So highly scalable and fast feature generation method is required in order to cope with current industry requirements.
- Any of the discussed methods are not capable of identifying complex events masked inside CDR. Detection of complex events allow to exploit maximum situational value. This can be effectively used for SIMbox fraud detection.

4. APPROACH

This section discuss about the novel feature set derived used in our approach and the complex

patterns in CDR stream that consist decisive power in context of grey call detection.

We have considered three types of CDR, Local, National and International in order to capture the three main entry points for telecom's voice network. Those data are typically generated at different nodes of the telecommunication network, thus different mediation techniques were used to extract relevant information. Local CDR means transaction logs for calls within operator and CDR records are generated at MSCs (Main Switching Centers). National CDR means transaction logs for calls originated to or terminated from other operators within the same country. Those data is acquired from the Call Screening Server which is anode as a firewall between home network and other operators. This node is able to blacklist numbers which need to be blocked and generates transaction logs in real time with the corresponding action (Blocked or Passed). Inclusion of action field makes these logs more useful and allows us to exploit CEP to locate the patterns described in subsection 4.1. So, three CDR streams were considered for processing.

After considering available CEP engines we have selected Siddhi Complex event processor developed by Suhothayan et al. [10], [11] and later WSO2 – Open source SOA Company – developed it and made it freely available under Apache Software License v2.0 [12]. Its performance in complex pattern detection and considerable amount of successful applications were primary reasons for this decision.

4.1. **Complex Events in CDR**

It is important to mention about complex events inside CDR to understand their power in detection. Figure 3 sketches one such example. Imagine the situation where we have three different CDR streams for Local Calls, National Calls and International calls separately. At time *t* a call attempt to Subscriber *B1* from other operator number *A1* is recorded in National CDR stream. But the telecom operator has already identified *A1* as fraudulent number and it is blocked. After a few seconds at $t + \Delta t$ the same subscriber *B1* gets a call attempt from *A2* in the local CDR stream. Also, we can find a call attempt by subscriber *B1* to International number *I1* at $t - \Delta t$ in International CDR stream. In such situation there is a high probability that *A2* be a fraudulent number and we can use other attributes in order to verify it. So it is clear that complex events allows us to narrow down analysis domain.

Figure 4 shows the real-world CDR entries that corresponds to the above scenario. All telephone numbers have replaced with non-existing numbers in order to preserve privacy.

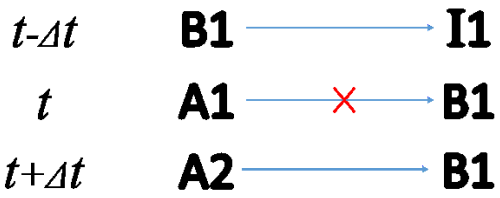

Figure 12: Complex Events in CDR created by SIMbox

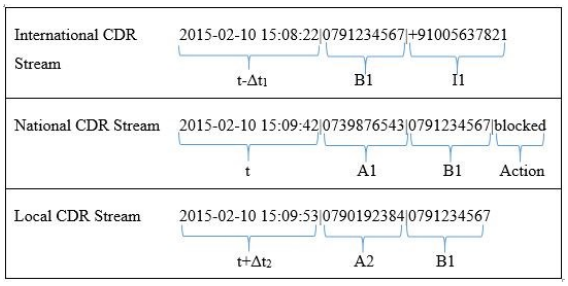

Figure 13: Complex Events in CDR created by SIMbox - Example.

Siddhi Query snippet shown below was used to capture this complex event.

```
from every
( a1 = internationalCDRStream
[duration == 0.0]
-> a2 = nationalCDRStream
[action=="blocked" and
called_party_id==a1.called_party_id]
-> a3 = localCDRStream
[called_party_id==a2.called_party_id
] )
select a3.calling_party_id as
blockPasscaller,"logic1" as logic
```

During our research we have identified 7 complex patterns including this example which reflects real decisive power.

4.2. **Feature SET**

In order to make decisions within small time window, the feature set used in decision making must provide maximum amount of details. By studying context details and behavior of labeled data we have derived the novel feature set which is presented in Table 1.

As described in earlier we have used Local, National and International CDR streams to derive those features. Calling party id, called party id, call duration , origination date & time, call disconnect party, call direction w.r.t home network are common features used in those three streams. For local calls we have considered cell IDs. For local CDR stream origination and destination switch details were considered.

We have deployed the set of CEP queries on to derive features in real time. We have not presented CEP queries used for feature deriving in order to save the space. But Queries were comparable complex than SQL and we have to used Event windows and unique windows extensively to achieve our targets

4.3. **Experimental Setup**

In order to evaluate our approach, the real data set acquired from major telecommunication service provider was used. Those data is animalized and encrypted in order to preserve subscriber privacy. CDR for 1 day period is collected from ISC and it contains 2,962,068 records in total. Local CDR taken from MSC contains data for 4 Hours period and consists of 2,541,630 records. Logs collected from Call Screening Server also contains data for 4 hours and has 3,198,801 records. We have used previously identified cell ids of grey callers and first call details as context details. We have stored those context data in CEP event tables in order to use them in CEP queries. The remaining 4 hours of ISC CDR, which overlaps with local and national CDR is used to generate CDR stream and remaining 20 hours was also used to generate context details. We have normalized the timestamp of 4 hours CDR with respect to fixed based value and inserted to CDR stream simulator.

The Stream Simulator was developed to convert CDR stored in files to CDR stream. The normalized timestamp value was used as hash keyfor retrieving CDR originated at given second. When trimer triggered the stream data simulator retrieves the CDR from data structure and send data through stream using data publisher. Figure 6 depicts the experimental setup.

Resultant CDR Stream goes through CEP event flow as depicted in Figure 5 and relevant standing queries are posed on it. Finally resultant feature set grouped w.r.t calling party ID is stored in SQL database. Detected complex patterns are also stored in database. Finally, rule engine decides the suspected callers by correlating these data. Different weights were assigned to types of complex events and derived the features inside rule engine. At the moment those weights are derived analyzing past labeled data manually.

Table 5: Feature set for each scenario.

Attribute	Description	Onnet	Offnet
calling_party_id	Caller ID (Primary Key)	✓	✓
og_cnt	Total outgoing call count by Subscriber	✓	✓
og_dcnt	Different numbers dialed by Subscriber	✓	✓
og_tot_dur	Total outgoing call duration by subscriber	✓	
og_ans_cnt	Number of outgoing calls answered by B party	✓	
og_max_dur	Maximum outgoing call duration	✓	
cell_count	Total number of cell ids subscriber moved within	✓	
max_cell	The cell id subscriber was their while taking most number of calls	✓	
max_cell_count	Call count subscriber has taken while in abovementioned cell	✓	
ic_cnt	Incoming call count received to this number	✓	✓

ic_dcnt	Different numbers called to this number	✓	✓
ic_tot_dur	Total call duration of subscriber for incoming answered calls from network under study	✓	✓
ic_ans_cnt	Total answered calls from subscriber for incoming calls from network under study	✓	✓
ic_max_dur	Maximum call duration of subscriber for incoming answered calls from network under study	✓	✓
og_idd_cnt	Outgoing IDD call count by subscriber	✓	
og_idd_dcnt	Distinct IDD numbers dialed by subscriber	✓	
og_idd_ans_cnt	Total IDD calls dialed by subscriber which was answered	✓	
og_idd_tot_dur	Total outgoing IDD call duration	✓	
ic_idd_cnt	Number of Incoming IDD calls to subscriber	✓	
ic_idd_dcnt	Number of distinct ID numbers dialed this subscriber	✓	
ic_idd_ans_cnt	Number of incoming IDD calls answered by subscriber	✓	
ic_idd_tot_dur	Total incoming IDD call duration of subscriber	✓	
iddb_cnt_in	Incoming IDD call count to called party numbers dialed by this number	✓	✓
iddb_dcnt_in	Distinct called party numbers dialed by this number who have received IDD calls	✓	✓
iddb_cnt_out	Outgoing IDD call count from called party numbers dialed by this number	✓	✓
iddb_dcnt_out	Distinct called party numbers dialed by this number who have dialed IDD calls	✓	✓
first_call	The date of the first call originated by subscriber (in Offnet scenario consider the date of first incoming call to network under study)	✓	✓

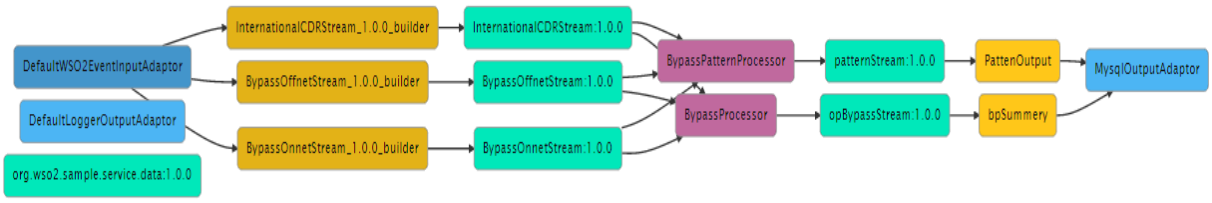

Figure 14: Event flow inside WSO2 Siddhi CEP.

So weighted sum of normalized feature values is calculated and rule engine triggers the alarm once that summation exceeds the defined threshold.

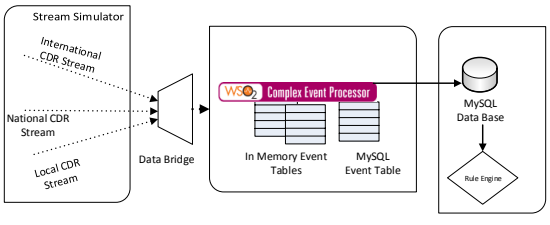

Figure 15: Experimental setup

5. RESULTS

This section compares the accuracy and efficiency of the proposed system with respect to classification provided by operator. Since we have focused efficiency of CDR based feature generation and effective use of complex event patterns for grey number detection, simple rule based tool was used for decision making. Data records of 1,573,680 distinct callers were included in Onnet data set and system has detected 31 grey numbers. Form 913,876 of Offnet callers system has identified 18 grey callers. Operator has detected 25 Onnet and 15 Offnet bypass numbers. But when we compare with operator's data 9 out

of 31 Onnet bypass numbers were genuine. Table 2 and 3 presents the confusion matrix for Onnet and Offnet bypass detections performed by the proposed system.

Table 6 : Confusion Matrix of Proposed System for Onnet Numbers		System Detected Class	
		Genuine	**Fraud**
Actual Class	**Genuine**	1573646	9
	Fraud	3	22

Table 7: Confusion Matrix of Proposed System for Offnet Numbers

		System Detected Class	
		Genuine	**Fraud**
Actual Class	**Genuine**	913856	5
	Fraud	2	13

Inherently grey call detection problem has uneven class distribution as few grey callers are operating within massive customer base. So accuracy value easily exceeds 99.999% as majority of genuine customers were correctly classified by our system in both Onnet and Offnet scenarios. Therefore *F-measure* or *F1-value* gives more reasonable view. Achieved F1-value for Onnet and Offnet scenarios are 0.786 and 0.789 respectively. [1], [2] and [3] has also achieved above 98% accuracies and F-measure of those instances was in 0.9 range due to more even distribution of grey callers and normal customers. But proposed system was evaluated against real data set without performing any filtering on CDRs.

But important observation is that operator have made detections with average answered attempt count of 7.4, but our system were able to reduce it to 5.7. So Proposed system had made detections 1.7 answered attempts quicker than operator detections. This is important parameter as increasing answered attempts means increasing revenue loss. [1], [2] and [3] did not used that important parameter when measuring performance. So proposed system perform detections with considerable accuracy without compromising detection speed.

6. CONCLUSION

We have started with the goal of detecting grey call numbers in real-time. The solutions available in research literature lacks these real-time features due to unsuitability of traditional database reliant store first process then approach for latency sensitive applications, large time windows for feature generation, shallow feature set, less awareness about context and ignoring complex patterns in CDR in decision making.

So we have primarily focused on generating rich set of features in real time and using power of complex events using CEP engine. Although we were able to detect the values with higher true positive true negative detection rate, false positive value is comparably higher. So we are working on integrate this real time system with past data and use more context details and feedback to increase accuracy.

Additionally inclusion of machine learning techniques and in-cooperating the manually defined rule engine with neural network or Tree based classifier can be used to achieve better F-measure value.

Due to privacy concerns operators not willing to expose these data to outside parties, so we identify the limitation of reproducibility. Also we expect to perform this experiment with larger data set spanned in more interval inside distributed computing environment in order to obtain better results. However, this system landmarks the good initiative real-time grey call detection by deviating from traditional database reliant approach.

7. ACKNOWLEDGEMENT

Special appreciation to IET Sri Lanka Network for providing this great opportunity.

8. REFERENCES

1) R. Aronoff. (2013, September 4). *Global Fraud Loss Survey 2013 by Communications Fraud Control Association*. [Online]. Available: http://www.cvidya.com/media/62059/global-fraud_loss_survey2013.pdf

2) E. Okutoyi. (2012, March 13). *SIM Box Fraud - New Headache for Africa's Mobile*

Operators. [Online]. Available: http://www.humanipo.com/news/142/sim-box-fraud-new-headache-for-africas-mobile-operators/

3) F. Kornbo. (2012, July). Carrier bypass: No drastic surgery required to protect revenue. [Online]. Available: http://www.telecomasia.net/pdf/CSGI/CSG_AsiaConnectionsJuly2012_CarrierBypass.pdf

4) Wikipedia Contributors. (2015, June 5*). Call detail record* [Online]. Available: http://en.wikipedia.org/wiki/Call_detail_record

5) A.H. Elmi, S. Ibrahim, and R. Sallehuddin, "Detecting SIM Box fraud using neural network," IT Convergence and Security 2012. Springer, vol. 215, pp. 575-582, 2013.

6) R. Sallehuddin, S. Ibrahim, A.M. Zain, and A.H. Elmi, "Classification of SIMbox fraud detection using support vector machine and artificial neural network," International Journal of Innovative Computing, Universiti Teknologi Malaysia, vol. 4, no. 2, 2014.

7) I. Murynets, M. Zabarankin, R.P. Jover, and A. Panagia, "Analysis and detection of SIMbox fraud in mobility networks," in Proc. IEEE INFOCOM '14, pp. 1519-1526, Apr. 2014.

8) Etross Telecom Co. Ltd. (2013). *GSM Modem Pool 8 Ports 32Sims ETS-8132 with SIM Rotation* [Online]. Available:http://www.etross.com/products_ys/&productId=30&comp_stats=comp-FrontProducts_list01-1364547681813.html

9) OpenCellID Community. (2015, June 7). *OpenCellID Database* [Online]. Available: http://opencellid.org/#action=statistics.cells&type=2&dateFrom=&dateTo=&mcc=&mnc=&sortBy=1

10) S. Suhothayan, K. Gajasinghe, I.L. Narangoda, and S Chaturanga, "Siddhi: A second look at complex event processing architectures," ACM GCE Workshop, 2011.

11) S. Suhothayan, K. Gajasinghe, I.L. Narangoda, and S. Chaturanga, "Siddhi-CEP, " B.Sc. Project Report, Dept. of Computer Sci. and Eng., Univ. of Moratuwa, Moratuwa, Sri Lanka,2011.

12) *WSO2 Complex Event Processor Documentation Version 3.1.0*, WSO2 Inc., 2014.

13) C. G. Weng and J. Poon, "A new evaluaton measure for imbalanced datasets," 7th Australian Data Mining Conference.