Journal of AI and Data Mining
Vol 7, No 2, 2019, 239-247
DOI: 10.22044/JADM.2019.6924.1812

Behavior-Based Online Anomaly Detection for a Nationwide Short Message Service

Z. Shaeiri[1], J. Kazemitabar[2*], Sh. Bijani[3] and M. Talebi [4]

1,2. Department of Electrical and Computer Engineering, Babol Noushirvani University of Technology, Babol, Iran.
3. Department of Computer Science, Shahed University, Tehran, Iran.
4. Hamrah-e-Aval Telecom Operator, Tehran, Iran.

Received 02 April 2018; Received 29 December 2018; Accepted 23 January 2019
**Corresponding author: j.kazemitabar@nit.ac.ir (J. Kazemitabar).*

Abstract
As fraudsters understand the time windows and act fast, real-time fraud management systems becomes necessary in the Telecommunication Industry. In this work, by analyzing the traces collected from a nationwide cellular network over a period of a month, an online behavior-based anomaly detection system is provided. Over time, users' interactions with the network provide a vast amount of data usage. This data usage is modeled to profiles by which the users can be identified. A statistical model is proposed, which allocates a risk number to each upcoming record, which reveals deviation from the normal behavior stored in profiles. Based on the amount of this deviation, a decision is made to flag the record as normal or abnormal. If the activity is normal, the associated profile is updated; otherwise, the record is flagged as abnormal, and it will be considered for further investigations. For handling the big dataset and implementing the methodology, we used the Apache Spark engine, which is an open source, fast, and general-purpose cluster computing system for big data handling and analysis. The experimental results show that the proposed approach can perfectly detect deviations from the normal behavior, and can be exploited for detecting anomaly patterns.

Keywords: *Short Message Service, Behavioral Profiling, Anomaly Detection, Apache Spark.*

1. Introduction

Fraud is a major concern in the telecommunication industry. The main definition of telecommunication fraud is the abusive usage of an operator infrastructure and resources without the intention of paying them. Victims of the telecommunication fraud are the subscribers, businesses, and operators. In addition to being victims of fraud, some subscribers just do not like to get services from an operator who has been a victim of fraud attacks. Since the same services can be offered by multiple operators, the subscribers can switch between them easily. There are various types of Telecom fraud such as identity theft, internet fraud, SIMboxing, spamming, phishing, spoofing, and flooding. The number and diversity of fraud techniques continue to grow, and new technologies have led to raid new Telecom fraud strategies. The latest fraudulent tactics are difficult to track and investigate because of their frequency, their layers of anonymity, and their global nature. Each year, the suppliers and the Telecom operators lose a significant proportion of their revenues as a result of fraud. Thus explorative digging and analysis is required to get a deeper insight into the customer behavior, their preferences, usage patterns, and signs of fraudulent exploitation of the network or the services. Generally, subscribers' interactions with the Telecom services and networks produce a large amount of user-related data. Telecom operators store all of these phone traffic data in the shape of CDR (Call Detail Record) files. The CDR data analysis involves a number of challenges that are related to its large volume. These challenges are posed in collection, storage, analysis, and mining of the data. Besides, refining information or interesting data from noise is important for governments, companies, and

individuals. One of the sophisticated technologies available to meet the existing requirements is the profiling technology [1-4]. Imperva's Web Application Firewall (WAF) uses dynamic profiling to learn the web users' expected behavior [5]. LightCyber's Advanced Threat Protection profiles user and device behavior and accurately detects anomalous attacks [6]. HP's TippingPoint Advanced Threat Appliance uses behavioral detection techniques to detect anomalies [7]. Damballa's Failsafe network security system uses profilers to assess the behavior [8]. RSA's Web Threat Detection constructs behavioral profiles to support the identification of anomalous behaviors [9]. Hillstone's Next Generation Firewall (NGF) associates users' behavior with their traffic, detects user-related behavioral abnormalities appearing over time by comparing traffic to a behavioral baseline, and customizes user-behavior models based on the observed traffic patterns [10]. Cisco's Traffic Anomaly Detector Module detects anomalous behaviors to prevent denial of service attacks in the network [11]. Among different approaches in the literature, we found two works that were close to our work and were appropriate for making comparisons [12, 13]. In both of these works, anomalous SMS activities have been taken into account. We provided comparisons between the proposed method and the methods presented in [12] and [13] in table 1. In these comparisons, different aspects of the models were investigated. In [13], the SMS data over a six-month period obtained from a large Asian telecommunication company (source and raw data confidential) has been used for creating the models. The power-law mixture model is used to capture community formation behaviors, and the Poisson-panel mixture model is exploited to uncover the abnormal behaviors in text messaging. It was shown that this approach could detect spammers. Meanwhile, it was applied on a small subset of the whole dataset, and it was not scalable. In another paper [12], SMS communication traffic of Machine to Machine (M2M) devices connected to the network of one of the main tier-1 providers in the US has been analyzed. Two algorithms are proposed for detecting anomalous SMS activities and attacks on the aggregate, cluster, and individual device levels. The first algorithm is contact-based and has been proven to be efficient against DDoS attacks and other anomalies characterized by an increase in the SMS load or a redirection of the SMS load aiming to saturate a specific node. The second algorithm is volumetric-based and can detect the anomalies characterized by a decrease in the traffic load. Actually, in the first algorithm, artificial anomaly is injected in the traffic and then it is detected by the contact-based method. In the second algorithm, the volumetric detection engine successfully identifies an anomaly caused by several M2M connected devices within a given cluster being down. In Table 1, the computational complexity of the three algorithms are compared. The computational complexity of both [12] and [13] is $O(N^2)$. In contrast, the computational complexity of the proposed method is $O(N)$. The results obtained show that the proposed method is considerably efficient, and it is appropriate for anomaly detection from SMS activity data. One of the key differences between the proposed method and the methodologies proposed in [12] and [13] is that in both [12] and [13], abnormal activities have been detected by comparing each new activity with a reference that is derived from an aggregated data. This reference is a combination of behaviors of all the users. If the user behavior deviates from the mentioned reference, then it is considered as an anomaly. This reference is equally used for all the individuals. In contrast to these approaches, in profiling methods, the user behavior is only compared with his/her own behavior. If a user acts differently compared with his/her own history of activities, his/her risk number increases. A user profile is a collection of explicit personal data usage associated with the user. Depending on the system with which the subscriber is interacting and his/her attitude towards the system, different types of profile datasets can be made. Profiling in e-commerce systems, social networking, recommender systems, and mobile user profiles are examples of profiling contexts. In profiling methods, the analysis is no more based on the original raw data but on an aggregated pre-processed summary of the original data. By analyzing the profiles with lower volumes compared with the original raw data, the performance will improve and also the privacy will be better preserved. In this paper, an anomaly detection method based on the profiling technology is introduced for online fraud detection from CDR data in combination with the recent big data analytical tools (Spark). For each subscriber, a profile is generated, which contains the behavior in communicating with others by sending or receiving messages. After generating profiles for a sufficiently large time interval, a statistical model is proposed and trained on the resulting profiles. For all the subsequent records, a risk number is calculated.

Table 1. Summary of results on performance and quality of methods; the first two rows are the consequences of machine to machine method [12], and statistical analysis and anomaly detection on SMS social network [13]. The new profiling method in the paper provide an online, scalable and efficient approach in row 3.s

Methods	Scalable	Computational Complexity	Based on Profiling	Online
M2M [12]	No	$O(N^2)$	No	No
Statistical Analysis [13]	No	$O(N^2)$	No	No
Proposed Method	Yes	$O(N^2)$	Yes	Yes

Based on the risk number and by exploiting a wisely measured threshold value a decision is made for labeling the record as normal or anomaly. If the record is determined as normal it is used for updating the associated subscriber's profile, whereas if the record is labeled as anomaly the associated user will be flagged for further investigation.

By exploiting the big data analytical tools, the approach reduces the cost of data processing compared to traditional data warehouse approaches since big data analysis reduces computational complexities and uses distributed processing engines. The rest of the paper is organized as follows. In Section II, we review existing fraud detection methods in this context. In Section III we define the problem in hand that was defined for us by Hamrah-e-Avval the main telecom operator in Iran. In Section IV, we propose our solution. In Section V, we provide simulation results. Finally, Section VI concludes our paper.

2. CDR data fraud detection methods

Research works on the MO-CDR or MT-CDR fraud detection is scarce, limited, and problem-specific. Many of the existing security methods are unable to provide comprehensive solutions for network-based fraud detection. Many of the existing methods reported in the literature are based on malware detection on mobile devices [14-16]. In the field of text messages (SMS), Murynets et al. have proposed a clustering-based approach to detect anomalous behaviors. Their work simulated the data by a tier-1 US-based cellular provider (AT&T) [12]. In [13], the authors have analyzed the social networking aspect of SMS users to find the anomalous behaviors. These approaches require extra computational resources on mobile devices, which lead to battery exhaustion. In general, behavioral methods for fraud detection rely on the inherent behavioral characteristics of users. The users unconsciously reveal their usage habits and patterns when they use their phones for their daily communications and conversations. The behavioral analytics solutions are designed to understand the normal behavior of each individual user, calculate the risk of each new activity, and then choose intervention methods commensurate with the risk. The key characteristics that make behavioral analytics effective are automatically monitoring all activities for all users, not just devices or records; no requirement for a prior knowledge of the specific fraud that the fraudster is attempting; and providing a detailed historical context for suspicious activities. The earlier fraudulent activity is detected, the easier and less costly it is to prevent. Behavioral analytics will detect the early stages of a fraud attack before the fraud is done. Because it is based on behavior, it will detect anomalous activities regardless of the type of attack, even newly emerging schemes. Behavioral analytics also provides context for all anomalous activities, which is extremely helpful for investigations when contacting the customer for something suspicious. Knowing a prior activity, what is normal, and the specific details of what makes the current activity high-risk make it possible to determine whether the activity is truly fraudulent or can be explained. Creating behavioral profiles involves deriving various effective and security-related features from the available usage patterns, for instance, the number of sent and received SMSs, the diversity of sent and received SMSs that means a unique number of receivers/senders a user is communicating with, and so on. To design models for distinguishing between normal and fraudulent activities, examples of both cases are required. The data relating to fraud are relatively rare, and any effort to characterize them in detail is difficult, time-consuming, and requires a specialized knowledge.

Since manually detecting abnormal activities among a huge number of records is inefficient and sometimes inaccurate, in most cases, exploiting supervised models, a.k.a. classifiers is not common or even possible. On the other hand, since the dataset is imbalanced, most supervised methods cannot perform very well. More precisely, their classification performance is not equal across both the fraud and non-fraud samples. In this regard, unsupervised models are a good fit for handling the problem and ranking anomalies. In our analysis, an unsupervised thresholding method was applied on the created profiles. Profiles contained probabilities of users' activities in sending or receiving messages. These probabilities were calculated from a sufficiently large volume of data from the past. Using these profiles for all subsequent records, a risk value was computed, which showed the level of abnormality of the associated users' activity. Finally, users were ranked based on their risk numbers, and the top risky users were flagged for further investigations.

3. Business problem

From a Telecom operator in Iran that covers the entire country, we collected anonymized SMS logs (MO-CDR) for a period of time. The task of collecting data was not an easy one since the security department in Hamrah-e-Avval does not easily release the user data. This is due to both the user privacy and the national security concerns. Not only the dataset given to us was anonymized but also we had limitations in showing some of our results. The goal defined by the operator is to provide a methodology for online anomaly detection. We were also asked to provide monitoring tools that could be used for business intelligence purposes. First, behavioral profiles are created for subscribers. These profiles contain the users' activities and behaviors in communication with networks. A statistical ranking method is proposed and applied on the profiles. Based on the results derived from the ranking procedure, users are labeled as normal or abnormal. Normal profiles are updated, while abnormal cases are flagged for further investigation. Due to the size of the network, the analysis has provided excellent informative insights into the SMS traffic characteristics and fraudulent activities of behavioral change type.

4. Discussion and methodology

In this paper, we introduce a behavioral technique for SMS log (MO-CDR) fraud ranking and detection. We use the concept of Log Likelihood Ratio (LLR). Its usage is common in Bayesian hypothesis testing. The justification for using ratio comes from a famous theorem in detection theory titled "Neyman-Perason Lemma" [17]. In coding and specifically message passing algorithm used in LDPC, the same concept, i.e. LLR, is used for determining whether the received signal was a zero or one [18]. What we are doing in this work is testing the hypothesis whether a text message is coming from a fraudster or a genuine user. Consider the following likelihood ratio for a given text message event:

$$\frac{P(Fraudster \mid Event)}{P(User \mid Event)} = \frac{P(F \mid E)}{P(U \mid E)} \tag{1}$$

Using the Bayes rule, the above equation can be re-written as:

$$\frac{P(E \mid F).P(F)}{P(E \mid U).P(U)} \tag{2}$$

Since $P(F)/P(U)$ is a constant not related to the text message event, we will be dealing with $P(E \mid F)/P(E \mid U)$ from now on. Also for ease of calculations, we will use the logarithm of the likelihood ratio instead, i.e. $\log\{P(E \mid F)/P(E \mid U)\}$. In order to calculate the conditional probability of $P(E \mid U)$, we go over the history of the user's behavior in the past. In other words, if *E* refers to a texting event that took place on Tuesday 10:00 AM from cell number XYZ-WST-MNOP, we should see how likely it is for that number to do that. For that matter, we first determine how frequent it was for that specific number in the past to send a text message on Tuesdays or how frequent it was for him/her to send messages on or around 10:00 AM. We then impose a naïve Bayes [19] assumption and declare $P(E \mid U) = P(Tuesday \mid U)$. $P(10:00AM \mid U)$. As such, the risk score can be written as the summation of different risk components (log of product is sum of logs), namely time-of-day risk and day-of-week risk.

The fundamental principal of the proposed method is that each subscriber has a unique usage pattern. By modeling these usage patterns into profiles, the users can be uniquely identified. First, the anonymized dataset is pre-processed. For performance concerns, the original raw dataset is stored by an efficient compact format using gzip. The resulting dataset is then analyzed to find and fix faults that possibly have been produced during the data gathering stages. From the resulting, the clean and compact dataset

aggregated profiles are generated. These profiles contain statistics about the users' activities such as:

- Probability of sending/receiving message in a specific time-bin of the day (we divide a full day to four 6-hour time-bin)
- Probability of sending/receiving message in a specific day of week
- Probability of sending/receiving message form a specific place
- Probability of sending/receiving message to/from a specific number

The mentioned statistics were collected for a sufficiently large time interval and gathered into the users' profiles. A number of records generated after this time span were investigated for finding the anomalous behavior. To each newly generated trace, a risk value is assigned, which is calculated as:

$$risk = \sum_i \log\left(\frac{P(F)}{P(u_i)}\right) \quad (3)$$

$P(u_i)$ refers to the probability of feature number i when a genuine user does it, as opposed to $P(F)$ that refers to the probability of the same feature when a fraudster does it. In many of the attacks performed in cellular networks, a phone is hacked by fraudsters and is then used to conduct a denial of service type of attack. At this time, the behavior seen from the victim phone number is different from his history, and thus we can detect it before it is too late. For example, a user who was not used to text during mid-night with a high volume if starts texting, all of a sudden can be a sign of an attack. In other words, attacks can be modeled by behavior change. i stands for one of the following cases:

- time-bin of sending/receiving message
- day of week of sending/receiving message
- place of sending/receiving message
- a specific phone number to/from which a message is sent/received

This risk value or anomaly score is calculated for each upcoming record. The more user's behavior deviates from his/her behavioral history or the profile the higher his/her score will be, users are ordered based on their behavior using the derived ranking scores. A threshold is required for mapping the numeric scores into labels, normal and abnormal. This threshold is calculated as follows:

$$Threshold = Range(S) \times \Theta \quad (4)$$

where S is a vector containing scores and Θ is derived by experiment.

4.1. Big data solution

Every single second, a tremendous amount of data is generated in a Telecom network. The data with this volume and complexity has been moved around the networks for years. Meanwhile, fraudsters are getting increasingly sophisticated. Since fraud impacts company revenue and brand significantly, they are seeking more real-time analytics in their analysis [22]. The global experiments in using big data-based solutions in the Telecom fraud detection reveals that big data technology gives operators the ability to detect frauds that were not detectable previously. Besides, it enables operators to detect any type of fraud in three to four minutes, what used to take 24 hours by previous systems. Telecom service providers believe that big data and advanced analytics play a certain role in supporting them to meet their business objectives. In this paper, we have used Apache Spark for the data analytics [20]. Spark was initially started by Matei Zaharia at UC Berkeley's AMPLab in 2009, and open sourced in 2010 under a BSD license. In 2013, the project was donated to the Apache Software Foundation and switched its license to Apache 2.0. In February 2014, Spark became a Top-Level Apache Project [21]. Apache Spark is an open-source project that provides an API (Application Programming Interface) centered on the RDD (Resilient Distributed Dataset), a read-only multi-set of data items distributed over a cluster of machines that is maintained in a fault-tolerant way. It provides high-level APIs in Java, Scala, Python, and R, and an optimized engine that supports general execution graphs. It can access divers data sources such as HDFS, Hive, and HBase, and is a fast and general engine for large-scale data processing.

5. Experimental rsults

Several experiments are performed on the MO-CDR (SMS) dataset. For security reasons, the CDR dataset is anonymized, and the number of sent or received messages is not shown. In figure 1, the number of sent messages is depicted versus a-number for some users with high activities in three days. Figure 1.a shows the result for a holiday, while figures 1.b and 1.c are obtained for two normal working days. It can be inferred that activity trend of these users are similar in the shown two working days, while it is somehow different in the holiday. Figure 2 shows the number of received SMSs versus b-numbers for

some recipients during the whole time interval. The results are shown for each week-day and each time-bin in figures 2.a and 2.b, respectively. From the risk analysis, users with abnormal behaviors are detected. Figures 3 and 4 show the behavior of two risky a-numbers. These illustrative results show that the proposed method perfectly detects users whose behavior has significantly changed over a very short period of time.

The cellular company provided us with a set of data that included some confirmed attacks, and asked us to detect those. After assigning a risk score to the messages, we ranked them based on the corresponding risk, and presented them to the cellular provider. They confirmed that we correctly found the attacks, albeit there were a few false alarms as well.

Discussion

One of the main aspects of the solution is monitoring the behavioral changes of all individuals/users and detecting abnormal changes. The experimental results show that the system appropriately fulfills this expectation. The system track the users' information into profiles for a sufficiently large period of time, and the profiles are updated when new records are generated. Thus profiles contain the normal behavior of users. An individual will be flagged as risky if its behavior deviates from its normal behavior in various ways. For instance, as it is illustrated in figures 3 and 4, the user is flagged as anomaly if s/he acts abnormally in a number of different ways:

- Sending an unusual number of messages,
- Sending messages from unusual places,
- Sending messages to unusual receivers,
- Sending messages in unusual weekdays, and
- Sending messages in unusual times.

In other words, if the user acts unusual with respect to the profile, his/her risk number increases. As it can be seen in figure 3, this user shows no activity for a large period of time. Then suddenly, s/he starts to send messages in a short time interval, and after that again s/he stops. In addition to deviation in the number of sent messages, we cannot observe a steady behavior from the other three plots. The same is true for the other risky sender whose behavior is depicted in figure 4, in which the number of sent messages versus time shows an impulsive behavior, although in a short time interval, the other three plots show that the user behavior has deviated from its normal state.

In real-world problems, rarely labeled or supervised data is available. The proposed method is unsupervised in its nature, and it uses only the raw data, users' interactions with the network, to evolve and to produce risk numbers. It automatically detects deviation from the normal profile even when expert knowledge is not accessible.

In addition to being simple and straightforward, the system can be used in online anomaly detection, thanks to using distributed processing engines.

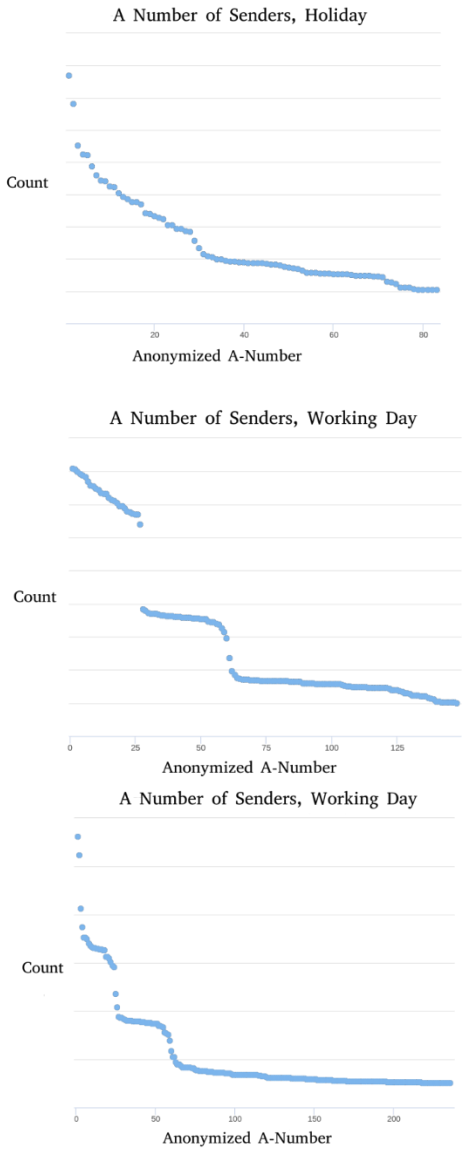

Figure 1. Number of sent messages versus users' anonymized A-numbers.

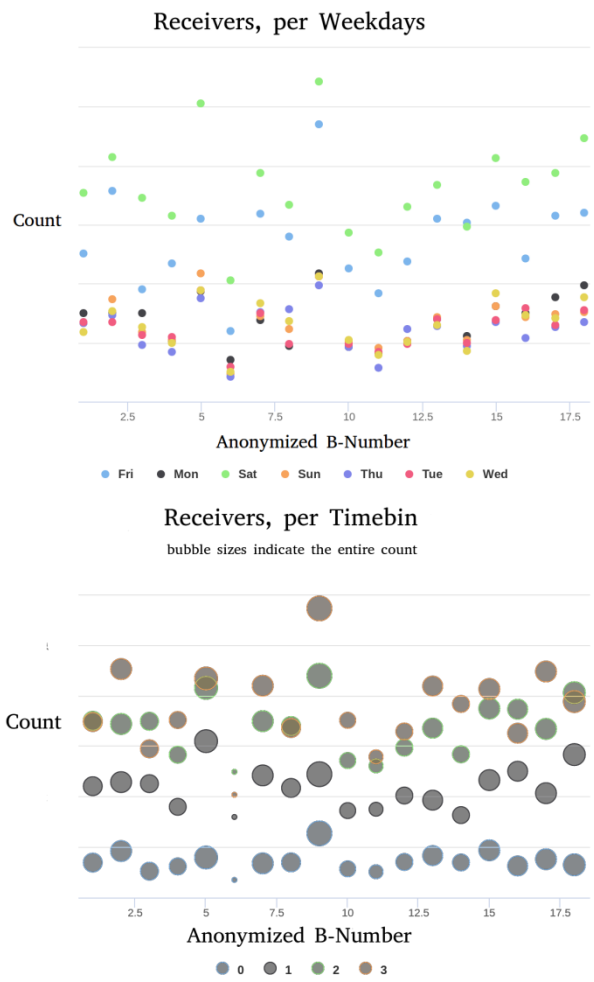

Figure 2. Number of received messages versus users' anonymized A-numbers.

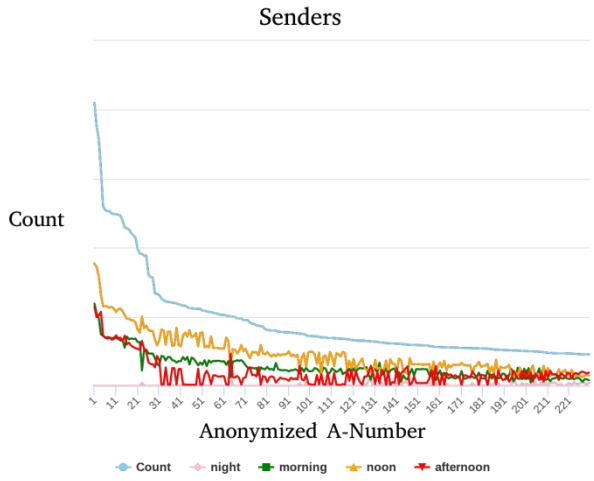

Figure 3. Number of sent messages versus sender anonymized A-numbers within time-bins of the day.

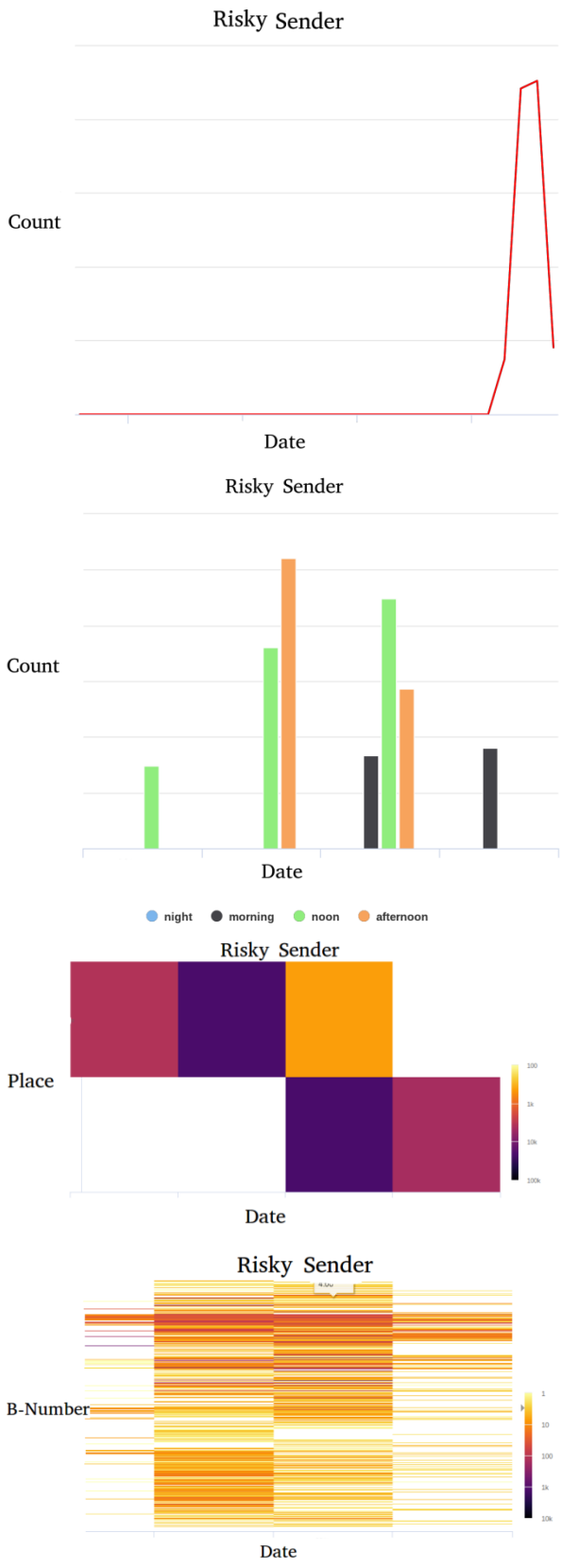

Figure 4. Top risky user.

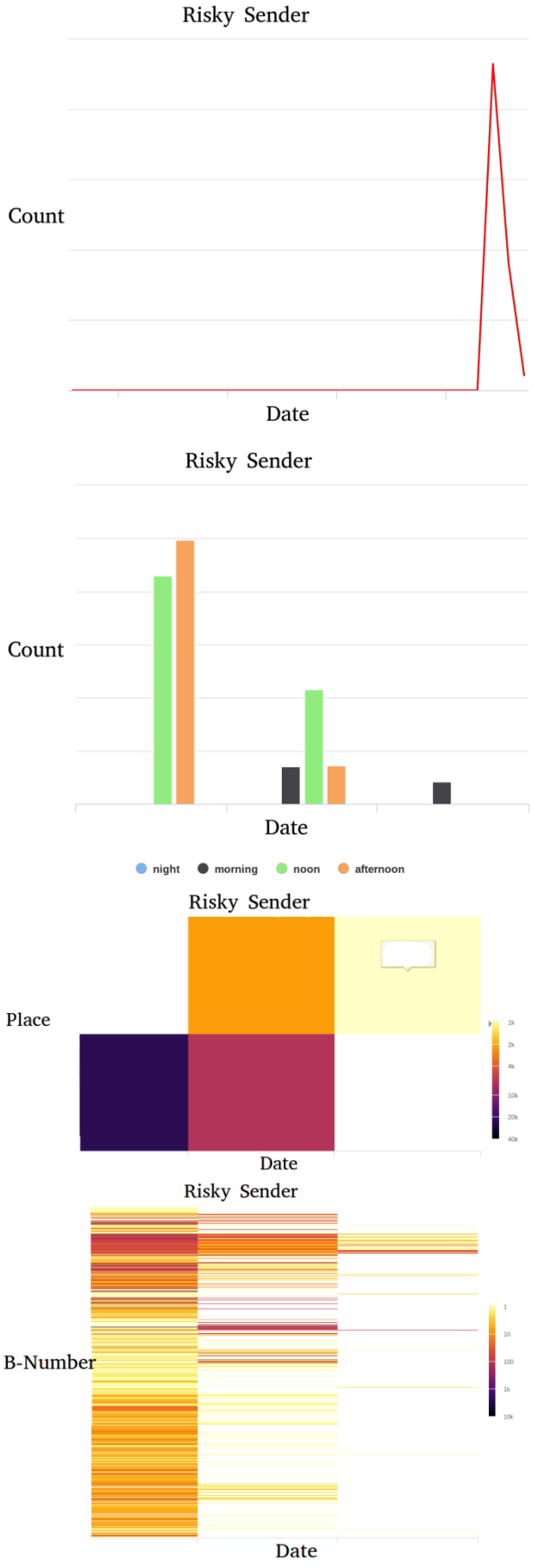

Figure 5. Second risky user.

6. Conclusion

In the age of communication, fraud is a major concern. As fraud tactics are evolved every day, the Telecom operators must increasingly fight newly emerged security threats. To this end, Telecom companies are adopting identity-based fraud detection techniques to improve networks' security. In this paper, an online behaviour-based anomaly ranking and detection method was proposed for fraud detection from CDR dataset (SMS logs - MO-CDR). First, raw CDR dataset was transformed to user profiles, which characterized users' behaviour in the form of statistical information. Modelling user behaviour in Telecom could be used to improve network security, improve services, provide personalized applications, and optimize the operation of electronic equipment and/or communication protocols. In generating profiles, users' privacy is respected, that is except from some coarse user behaviour characteristics, all private data, e.g. called number, calling location are hidden from the analyst. Next, an online anomaly ranking and detection technique is proposed, in which users are scored based on their risk level. As it is evident from the experiments, the proposed method perfectly identifies change in behaviour or deviations from the normal behaviour. We believe that this method can be a general framework that can be used in any IT-based company.

7. Acknowledgments

This research work was supported by Hamrah-e-Avval and Faraconesh Co. We thank our colleagues Dr Narges Arastoei and Mr. Hadi Portavassoli form Hamrah-e-Avval and Mr. Saeed Karimi and Mr. Ahmad Reza Khorramshekouh from Faraconesh Co. who provided expertise that assisted the research work.

References

[1] Rosas, E. & Analide, C. (2009). Telecommunications Fraud: Problem Analysis-an Agent-based KDD Perspective. epia2009.web.ua.pt.

[2] Hilas, C. & Sahalos, J. (2007). An application of decision trees for rule extraction towards telecommunications fraud detection. International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, pp. 1112-1121.

[3] Kou, Y. & Lu, C. (2004). Survey of fraud detection techniques. Sensing and Control, pp. 749–754.

[4] Fawcett, T. & Provost, F. (1996). Combining Data Mining and Machine Learning for Effective User Profiling. International Conference on Knowledge Discovery and Data Mining KDD, pp. 8-13.

[5] Imperva's Web Application Firewall data sheet. Available: https://www.imperva.com/docs/TB_Dynamic_Profiling.pdf

[6] LightCyber and Check Point Advanced Threat Protection solution brief. Available: https://www.checkpoint.com/download/downloads/products/solution-brief/SB_LightCyber.pdf

[7] HP Unifies Network Security Detection to Identify, Contain and Neutralize "Patient Zero" Infections. Available: http://www.hp.com/hpinfo/newsroom/press_kits/2014/HPProtect2014/HPTippingPoint_Advisory.pdf

[8] Finding Advanced Threats Before They Strike: A Review of Damballa Failsafe Advanced Threat Protection and Containment. Available: http://www.sans.org/reading-room/whitepapers/analyst/finding-advanced-threats-strike-review-damballa-failsafe-advanced-threat-protecti-34705

[9] How RSA web threat detection identifies account takeover. Available: http://www.emc.com/collateral/solution-overview/h11847-so-rsa-web-threat-detection.pdf

[10] Hillstone T-Series Intelligent Next-Generation Firewall Whitepaper: Abnormal Behavior Analysis. Available: http://www.hillstonenet.com/wp-content/uploads/Abnormal-Behavior-Analysis-Whitepaper-042415.pdf

[11] Cisco Traffic Anomaly Detector Module. Available: https://www.cisco.com/c/en/us/products/collateral/interfaces-modules/catalyst-6500-7600-router-traffic-anomaly-detector-module/product_data_sheet0900aecd80220a6e.html

[12] Murynets, I & Jover, R. P. (2013). Anomaly detection in cellular Machine-to-Machine communications. IEEE International Conference on Communications (ICC), Budapest, pp. 2138-2143.

[13] Zhang, B., Ma, L. & Krishnan, R. (2011). Statistical Analysis and Anomaly Detection of SMS Social Networks. Thirty Second International Conference on Information Systems, Shanghai, pp. 3007-3015.

[14] Zhang, X. & Jin, Zh. (2016). A new semantics-based Android malware detection. 2nd IEEE International Conference on Computer and Computations, pp. 1412-1416.

[15] Arora, A., Garg, Sh. & Peddoju, S. K. (2014). Malware detection using network traffic analysis in Android based mobile devices. 8th International Conference on Next Generation Mobile Apps, Services and Technologies, pp. 66-71.

[16] Ariyapala, K., GiangDo, H., Anh, H. N., NG, W. K. & Conti, M. (2016). A host and network-based intrusion detection for Android smartphones. 30th International Conference on Advanced Information Networking and Applications Workshops, pp. 849-854.

[17] Neyman, J. & Pearson, E. S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. Phil. Trans., pp. 694–706.

[18] Lin, S. & Costello, D. J. (2004). Error Control Coding (2nd Edition), Pearson.

[19] Russell, S. & Norvig, P. (2003). Artificial Intelligence: A Modern Approach (2nd ed) Prentice Hall.

[20] The Apache website (2019), Available: https://spark.apache.org/docs/latest.

[21] The Apache blog (2014), Available: https://blogs.apache.org/foundation/entry/the_apache_software_foundation_announces50.

[22] Rezvani, M. (2017) Assessment Methodology for Anomaly-Based Intrusion Detection in Cloud Computing. Journal of AI and Data Mining, Shahrood University, vol. 6, pp. 387-397.

نشریه هوش مصنوعی و داده کاوی

کشف ناهنجاری بلادرنگ مبتنی بر رفتار برای سرویس پیام کوتاه کشوری

زهرا شعیری[۱]، سیدجواد کاظمی‌تبار امیرکلایی[۲*]، شهریار بیژنی[۳] و محمد طالبی[۴]

[۱و۲] دانشکده مهندسی برق، دانشگاه صنعتی نوشیروانی بابل، مازندران، ایران.

[۳] گروه آموزشی علوم کامپیوتر، دانشکده علوم پایه، دانشگاه شاهد، تهران، ایران.

[۴] اپراتور همراه اول، تهران، ایران.

ارسال ۲۰۱۸/۰۴/۰۲؛ بازنگری ۲۰۱۸/۱۲/۲۹؛ پذیرش ۲۰۱۹/۰۱/۲۳

چکیده:

با توجه به اینکه امروزه کلاهبرداران با مفهوم مهم پنجره زمانی آشنا هستند و سرعت بالایی در انجام تقلب دارند، طراحی سیستم‌های بلادرنگ مدیریت ریسک در صنعت مخابرات امری ضروری است. در این مقاله، یک سیستم تشخیص ناهنجاری برخط مبتنی بر رفتار، جهت اعمال بر داده‌های جمع‌آوری شده از یکی از اپراتورهای مخابراتی در مدت یک ماه ارائه می‌شود. در طول زمان، تعامل کاربران با شبکه مخابراتی منجر به تولید داده‌ای حجیم می‌گردد. با استفاده از این داده (خام) برای هر کاربر پروفایلی تشکیل می‌شود که در بردارنده رفتار آن کاربر است و کاربران توسط پروفایل‌های خود از یکدیگر بازشناخته می‌شوند. یک مدل آماری ارائه شده است که در آن به هر فعالیت کاربر، براساس میزان اعوجاج رفتار او از رفتار نرمالش که توسط پروفایل او قابل درک است، یک نمره ریسک اختصاص داده می‌شود. براساس عدد ریسک محاسبه شده و قانونی برای تصمیم‌گیری، به هر کاربر پرچم نرمال و یا ناهنجار تخصیص داده می‌شود. اگر فعالیت کاربر طبیعی باشد پروفایل او به روزرسانی می‌شود؛ در غیر اینصورت جهت بررسی بیشتر برچسب ناهنجار به آن اختصاص می‌یابد. جهت پیاده‌سازی روش و اعمال آن بر روی کلان داده، تکنولوژی اسپارک آپاچی[۱] مورد استفاده قرار گرفته است. نتایج پیاده‌سازی‌ها نشان می‌دهد که روش پیشنهادی دارای توانایی بالایی در تشخیص ناهنجاری از نوع تغییر ناگهانی رفتار می‌باشد.

کلمات کلیدی: سرویس پیام کوتاه، پروفایلینگ مبتنی بر رفتار، تشخیص ناهنجاری، اسپارک آپاچی.

[۱] Apache Spark Engine