Detecting Anomalies in Mobile Telecommunication Networks Using a Graph Based Approach

Cameron Chaparro and William Eberle
Department of Computer Science
Tennessee Technological University
Cookeville, Tennessee, 38505
cvchaparro42@students.tntech.edu and weberle@tntech.edu

Abstract

According to a survey conducted by the Communications Fraud Control Association an estimated $46.3 billion were lost due to telecommunications fraud in 2013. This suggests that the potential for intentional exploitation of unsuspecting users is an ongoing issue, and finding anomalies in telecommunications data can aide in the security of users, their phones, their personal information, and the companies that provide them services. Most anomaly detection approaches applied to this type of data use some type of statistical representation; however, we think that a more natural representation is to consider telecom traffic as a graph. In this paper, we specifically focus on using graph-based anomaly detection to find and report anomalies in telecom data. Up until now, little work seems to be focused on detecting and reporting anomalies in telecommunications data represented as a graph. Moreover, even less work seems to focus on detecting anomalies in phone call history with this same representation. Our goal in this application paper is to use real-world cell phone traffic to detect anomalies in user patterns based on phone call and text message history.

Introduction

According to the International Telecommunication Union (ITU), in 2014 mobile subscriptions in underdeveloped nations are estimated to be quickly growing and mobile subscriptions in developed nations are estimated to start reaching levels of saturation [ITU 2014]. This increase in the use of mobile devices can have serious implications ranging anywhere from protecting the security of user information to protecting mobile phone service providers from fraudulent usage of services such as cloning SIM cards, etc. With this abundance of mobile telecommunications data, it is possible and increasingly valuable to find and report anomalies in the data to prevent personal threats to users, financial threats to service providers, or other types of unexpected threats. One area of research that can aid in this type of potential threat is anomaly detection. In this paper we aim to show that, specifically in the case of mobile telecom data, a graph-based anomaly detection approach can provide some valuable insight into the calling patterns.

Examination of call records shows the intuitive nature of representing this data in terms of a graph. For example, Onnela et al., while not specifically focusing on the problem of anomaly detection, have success representing their large-scale phone call data as a call graph [Onnela et al. 2007]. Similarly, Eberle and Holder showed that anomalies in movements and social relationships can be detected using data from mobile devices represented as a graph [Eberle and Holder 2008]. This representation follows from the fact that we can consider phone calls as a type of transaction between individuals which indicates a relationship between them. Take for example, a phone call from person A to person B who, in turn, calls person C. We now have an indirect relationship between person A and person C. Thus, upon representing each person as a node in a graph and the phone calls between them as edges, it is straightforward to visualize the relationships between each person.

We believe that representing telecom data as a graph will provide an intuitive and efficient method for detecting anomalies. To evaluate our hypothesis, we will use the Graph-Based Anomaly Detection (GBAD) tool - provided by Eberle and Holder and discussed in their 2007 paper - in the hopes of finding anomalies in the data [Eberle and Holder 2007]. We include phone call and text message data as our primary anomaly detection features.

In the next section, we discuss what work has already been done; particularly work that has been done using anomaly detection on mobile telecommunication networks; and then we focus on relevant work that gives more insight into our reason for representing our data as a graph. In the following section we explore the structure of the data and how we combined different sets of data for use in our experiments. We also provide some information relating to how much data we use, and explain our reasons for selecting specific sections for experimentation. Then in the section that follows we discuss what experiments were run on the data set, and we present the results obtained from running our experiments. We then conclude the paper with some suggestions for future work that might be done in order to improve upon the results presented here.

Related Work

This paper makes use of two primary types of related work: (1) anomaly detection for mobile telecommunication networks, and (2) representing phone call data as a graph. The sources relating to mobile telecommunication network anomaly detection have a more direct relation to our work, since we consider the detection of anomalies in this type of network. The sources relating to representing phone call data as a graph mostly contribute to supporting our decision to use a graph-based approach for representing the data and a graph-based tool for running our experiments on the data.

Anomaly Detection in Mobile Telecom Networks

Büschkes, Kesdogan, and Reich present an algorithm using a statistical approach, Bayes Decision Rule, which they use to detect anomalies in user behavior on cellular radio networks [Büchkes, Kesdogan, and Reich 1998]. Using a security focus, they apply their approach by tracking user locations through network cells and determining the probability of a user's transition from one cell into the next based on the user's prior behavior. However, as pointed out in their research, high rates of change in behavior associated with commuting adversely affects the effectiveness of their approach.

Sun et al. present two detection schemes, Lempel-Ziv and fixed-order Markov model, that they use to create user mobility profiles through cellular networks and compare the results of each approach [Sun et al. 2006]. Moreover, they dynamically update the mobility profile using the exponentially weighted moving average technique. Both of these anomaly intrusion detection techniques, similar to Büschkes, Kesdogan, and Reich, track user movements through network cells as their intrusion detection feature. In their research, Sun et al., show that their Lempel-Ziv-based method, which derives from a data compression scheme, outperforms their fixed-order Markov model-based method in both high detection rate and low false alarm rate, especially for low-speed users.

In the paper by Damopoulos et al. they explain how they evaluated four different machine learning algorithms – Bayesian network, radial basis function, K-nearest neighbors, and random Forest – for their effectiveness in the detection of anomalies in mobile devices when considering phone call history, SMS history, and web browsing history both separately and in conjunction [Damapoulos et al. 2011]. To evaluate their results, they use 10-fold cross-validation and 66% split methods. While their results are promising, they noted on several occasions that certain algorithms performed poorly, when compared to the others, due to lack of enough data values.

Representing Phone Call Data as a Graph

The research of Onnela et al. examines the structure of a very large data set and the "tie" strengths for interactions between individuals [Onnela et al. 2008]. They show that, contrary to one's intuition, in the removal of strong ties first, the network does not disintegrate, but it does shrink; whereas, upon removing weak ties first, the network quickly dissolves. They also consider the effect of tie strength on information diffusion throughout the network. On this front they find that neither strong nor weak ties have any effect on the spread of information in the network.

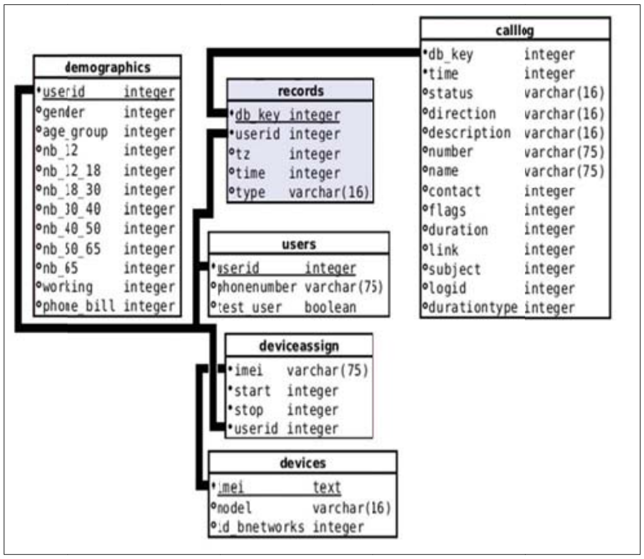

Figure 1. The condensed database schema containing only the tables we use in our research.

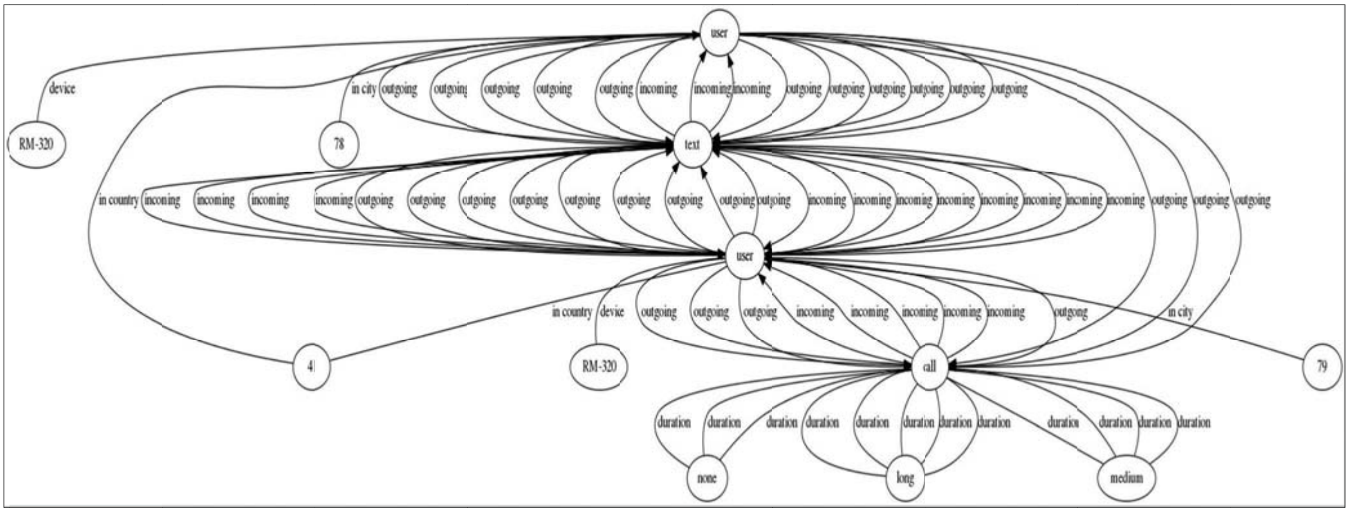

Figure 2. A visual representation of a subgraph (consisting of data for 2 users).

Data Set

Our data set comes from actual, anonymized cellular phone data provided by Nokia through the 2012 Mobile Data Challenge (MDC). We provide a condensed diagram of the database schema containing only the tables from which data was extracted in Figure 1. (For interested readers, information on requesting a full diagram of the database schema can be found at the Idiap Data Distribution website – https://www.idiap.ch/dataset/mdc/download). Note that not all the possible data is used. Instead, we focus on a subset of it including telephone calls and text messages, which are used in the experiments, and the user's demographic data for providing insight while interpreting the results of our experiments. Both topics will be discussed in more depth in the next section.

Some general statistics about the data are in order: first, we take data from 113 unique users, each with an average of about 38 calls and text messages to other users among those 113 already in the data set. While each of the users were involved in several interactions with people not contained in the data set, only those interactions between users were considered. Our research primarily incorporates data extracted from both the “calllog” and the “devices” tables which are depicted in Figure 1. The “calllog” table contains a list of phone calls and text messages between users and their contacts, and the “devices” table contains a list of phone models corresponding to each user. Also, if a user had multiple phone models, only one was used. Indirectly, we use the rest of the tables from Figure 1 for the purposes of joining data as necessary. The exception to this rule, however, is the “demographics” table which, as already mentioned, is used for accessing demographic data about certain users, particularly in the case that an anomaly was detected regarding data from their calls or messages.

We now provide a brief description of the attributes used in our experiments. First, the direction attribute from the “calllog” table has possible values of “Incoming”, “Outgoing”, or “Missed Call”. For our purposes, we used the first two types of directions since we want to only consider calls that connect. However, in the next section we describe how and why we partially account for missed calls. The description attribute is used to determine whether the transaction was a call or a text message. The duration attribute was used after being bucketized; that is, we separated the integral duration values into 4 “buckets” namely, “none”, “short”, “medium”, “long”, that were used to give more context to the duration values as well as provide a more consistent and useful normative pattern for the purposes of using GBAD. Our buckets were calculated as interquartile ranges of the integer duration values of valid calls. The model attribute was used to provide extra information about the user in our graph. Finally, the number and phone number attributes were used for determining who the other party in the call or text message was; however, since phone numbers were anonymized, we chose to represent the user's phone number by the country and city codes from the phone number.

Experimental Setup and Results

Our experimental setup consists of extracting the required data from the database, combining it to contain all the required information for each of the users, creating a multiuser graph from the data for all users, and running it through one of the anomaly detection algorithms in GBAD. In the following sections, we expand more on the main steps involved in preparing and running our experiments and then we conclude the section by providing our results and how we interpreted them. Also, following is a brief description of GBAD, the tool used to run our experiments.

The Graph

Figure 2 we provide a visual representation (consisting of data from only 2 users) of the graph topology used for one example of a subgraph. The complete graph for all users had a total of 966 vertices and 5602 edges. From this diagram, some simple observations can be made, but some clarifications are also necessary. First, we would like to point out that for interactions between two users we only have one "call" or "text" node and we use the number of edges out of that node to the "user" node to represent the number of calls or text messages from one user to the other. On that note, we should mention that the number of edges into a transaction (a call or text) node need not be equal to the number of edges out of the transaction node. This is likely a result of the fact that we did not include an edge for missed calls, but we did include the attempted call whether it was missed or not. Or, for example, in the case of text messages, one user might have sent many text messages to another but the other did not necessarily answer each text message. Another observation that can be made is that the country code (41 in this example) is shared amongst all users in the same country, yet city codes (78 and 79 in this example) are not shared. We chose to share country codes so as to potentially discover patterns associated with individual countries. However, city codes are only unique within a country, such that the same city code could be used by multiple countries.

The Graph-Based Anomaly Detection Tool

There are three general categories of anomalies in a graph: insertions, modifications and deletions. Insertions would constitute the presence of an unexpected vertex or edge. Modifications would consist of an unexpected label on a vertex or edge. Deletions would constitute the unexpected absence of a vertex or edge. The graph-based anomaly detection tool that we decided to use, GBAD, discovers each of these types of anomalies. Using a greedy beam search and a minimum description length (MDL) heuristic, GBAD first discovers the best substructure, or normative pattern, in an input graph. The minimum description length (MDL) approach is used to determine the best substructure(s) (i.e., normative pattern) as the one that minimizes the following:

$$M(S,G) = DL(G|S) + DL(S)$$

where G is the entire graph, S is the substructure, DL(G|S) is the description length of G after compressing it using S, and DL(S) is the description length of the substructure. Using a beam search (a limited length queue of the best few patterns that have been found so far), the algorithm grows patterns one edge at a time, continually discovering what substructures best compress the description length of the input graph. The strategy implemented is that after extending each substructure by one edge, it evaluates each extended substructure based upon its compression value (the higher the better). A list is maintained of the best substructures, and this process is continually repeated until either there are no more substructures to consider or a user-specified limit is reached.

In summary, the GBAD approach is based on the exploitation of structure in data represented as a graph. GBAD discovers anomalous instances of structural patterns in data that represent entities, relationships and actions. GBAD uncovers the relational nature of the problem, rather than solely the traditional statistical deviation of individual data attributes. Attribute deviations are evaluated in the context of the relationships between structurally similar entities. In addition, most anomaly detection methods use a supervised approach, requiring labeled data in advance (e.g., illicit versus legitimate) in order to train their system. GBAD is an unsupervised approach, which does not require any baseline information about relevant or known anomalies. To summarize, GBAD looks for those activities that appear to match normal / legitimate / expected transactions, but in fact are structurally different. For more information regarding the GBAD algorithms, the readers should refer to [Eberle and Holder 2007].

Finally, GBAD has two potential evaluation metrics for discovering the normative patterns: MDL and size. MDL, or Minimum Description Length, is based upon the work of [Rissanen 1989] and the idea of compression. The size metric makes a trade-off between the size and frequency of a substructure.

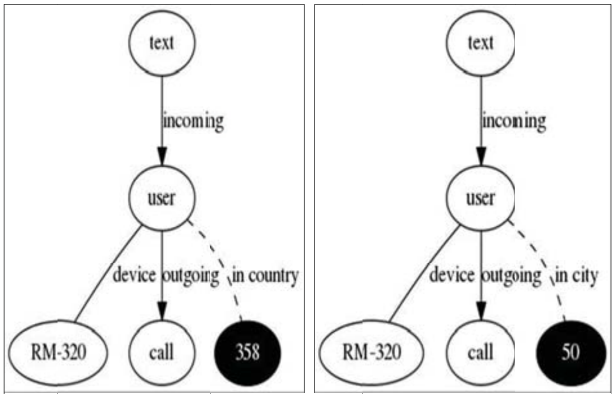

Figure 3. Anomalies detected using the size evaluation metric. (a) The anomalous insertion of a country node, "358". (b) The anomalous insertion of a city node, "50".

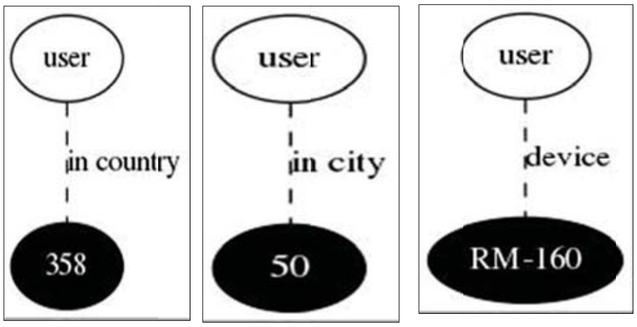

Figure 4. Anomalies detected using the MDL evaluation metric with minimum normative pattern size of 1. (a)Anomalous insertion of a country code, "358". (b) Anomalous insertion of a city code, "50", (c) Anomalous insertion of a device node, "RM-160".

The Results

Now that we have discussed the setup of our experiments and provided some background information on the GBAD tool, we will present the anomalies detected with our approach.

Running the probabilistic algorithm, used for detecting anomalous insertions, with the size evaluation metric, successfully detected two anomalies each of which are depicted in Figure 3; and using the MDL evaluation metric GBAD was able to detect the three anomalies depicted in Figure 4. The anomalies in each of the figures are depicted using a black vertex with white text to represent the anomalous insertion of a vertex and a dashed line to represent the anomalous insertion of an edge.

Further inspection of the data seems to confirm that the anomalies in Figure 3 (a) and 3 (b), are, in fact, anomalies due to the fact that of the 113 users, only one user has the country code 358, and similarly, the same user is the only one to have the city code 50 in their phone number.

When using the MDL evaluation metric, since the normative patterns were smaller, we chose to try two different normative patterns: first, the default normative pattern (single "user" vertex) and, second, the next-best normative pattern which had a minimum size of 2 vertices and an edge (the "user" and "RM-159" vertices).

The anomalies from Figure 4 (a) and 4 (b) are actually the same as the anomalies from Figure 3 (a) and 3 (b) even with a quite different normative pattern, due to using the different evaluation metric, and as such, we won't re-discuss them. The anomaly in Figure 4 (c), however, shows that an anomalous phone model node was inserted with label "RM-160". The data, again, supports this result since for the 113 users, only one user had a device with that model.

Finally, we tested, with the MDL evaluation metric, to see what, if any, anomalies would be detected with a normative pattern having a minimum size of two vertices. The result, depicted in Figure 5, was the anomalous insertion of a city node "77" which, in fact, is supported by the data since, from the 45 users with a device model of "RM-159", again, a single user was in city "77".

One final note is that while GBAD uses three distinct algorithms for detecting the three different types of anomalies in graphs, the only one that yielded results was the one for detecting anomalous insertions (the probabilistic one mentioned above). We think that a different graph topology than the one used here could potentially lead to the discovery of other types of anomalies.

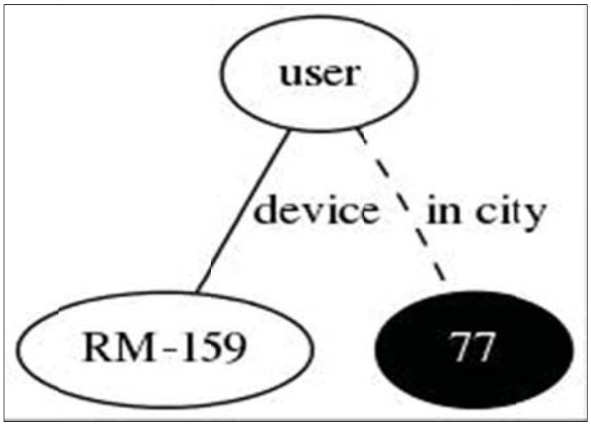

Figure 5. An anomaly detected using the MDL evaluation metric with a minimum normative pattern size of 2.

Conclusions and Future Work

In this paper, we have claimed that it can prove beneficial to put an emphasis towards using graphs for detecting anomalies in mobile telecommunications networks. We show, with real-world data, that a graph representation for the data allows for the detection of 5 (but 3 unique) anomalous substructures in a mobile call graph, two of which were detected using distinct evaluation metrics each with different normative patterns. In future work, we will attempt to apply other anomaly detection algorithms on the MDC data set, to provide a more complete picture of the effectiveness of the graph-based anomaly detection approach. Another focus of future work could be to find more graph topologies to potentially speed up the detection process, which would be essential if this approach were to be used in real-time. We also intend to further investigate the issues associated with "concept drift". Concept drift is the idea that patterns can "drift" over time causing the normative pattern for a graph at one time to potentially be different than its normative pattern at a different time. As we attempt to apply this approach to “big data”, or streaming data, we will need to evaluate the optimization of techniques that will allow for a graph-based anomaly detection approach to be used in real-time.

References

Büchkes, R.; Kesdogan, D.; Reich, P. 1998. How to Increase Security in Mobile Networks by Anomaly Detection. In Proceedings of the 14th Annual Computer Security Applications Conference, 1998, 3-12. Pheonix, AZ: IEEE.

Communications Fraud Control Association, 2013 Global Fraud Loss Survey, http://www.cfca.org/fraudlosssurvey/.

Damopoulos, D.; Menesidou, S. A.; Kambourakis, G.; Papadaki, M.; Clarke, N.; Gritzalis, S. 2012. Evaluation of Anomaly-Based IDS for Mobile Devices Using Machine Learning Classifiers. Security and Communication Networks. 5:3-14.

Eberle, W.; Holder, L. 2007. Anomaly Detection in Data Represented as Graphs. Intelligent Data Analysis 11:663- 689.

Eberle, W.; Holder, L. 2008. Analyzing Catalano/Vidro Social Structure Using GBAD. IEEE Symposium on Visual Analytics Science and Technology.

International Telecommunication Union, The World in 2014 ICT Facts Figures, http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2014-e.pdf.

Onnela, J.-P.; Saramäki, J.; Hyvönen, J.; Szabó, G.; Lazer, D; Kaski, K; Kertész, J.; Barabási, A.-L. 2007. Structure and Tie Strengths in Mobile Communication Networks. Proceedings of the National Academy of Sciences of the United States of America 104:7332-7336.

Rissanen, J. 1998. Stochastic Complexity in Statistical Inquiry. World Scientific Publishing Company.

Seshadri, M.; Machiraju, S.; Sridharan, A.; Bolot, J.; Faloutsos, C.; Leskovec, J. 2008. Mobile Call Graphs: Beyond Power-Law and Lognormal Distributions. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 596-604. New York, NY: ACM.

Sun B.; Yu, F.; Wu, K.; Xiao, Y.; Leung, V. C. M. 2006. Enhancing Security Using Mobility-Based Anomaly Detection in Cellular Mobile Networks. IEEE Transactions on Vehicular Technology 55:1385-1396.