مجلة العلوم البحثة والتطبيقية
Journal of Pure & Applied Sciences
www.Suj.sebhau.edu.ly ISSN 2521-9200
Received 27/06/2019 Revised 19/08/2019 Published online 11/12/2019

A Study Of Using Big Data And Call Detail Records For Criminal Investigation

N. Abuhamoud[1] , *E. Geepalla[2]
[1]Department of Electrical and Electronic Engineering, College of Engineering and Technology, Sebha University, Libya
[2] Department of Biomedical Engineering, College of Engineering and Technology, Sebha University, Libya
*Corresponding author: ems.geepalla@sebhau.edu.ly

Abstract In this paper we propose a new model to see how graph technologies can be used to analyze Call Detail Records (CDR) in order to find potential criminals. We face the challenging task of automatically deriving meaningful information from the available data, by using an unsupervised procedure of data analysis and without including in the model and a priori knowledge on the applicative context. Therefore, in this paper a big date technology (Neo4j) used to analyze the users' behaviors in order to detect abnormal behavior which might help the investigators to find the criminals.

Keywords: Neo4j, Big data, Criminals, Call, Detail, Record.

دراسة لاستخدام البيانات الضخمة وسجلات تفاصيل المكالمات لاكتشاف الجرائم

ناصر ابوهمود[1] و *امسيب جيب الله [2]
[1]قسم الهندسة الكهربائية والالكترونية– كلية الهندسة و التقنية–جامعة سبها، ليبيا
[2]قسم الهندسة الطبية–كلية الهندسة و التقنية–جامعة سبها، ليبيا
ems.geepalla@sebhau.edu.ly

الملخص في هذه البحث نقترح نموذجًا جديداً لمعرفة كيف يمكن استخدام تقنيات البيانات الضخمة المبنية على الرسومات لتحليل سجلات تفاصيل المكالمات (CDR) للعثور على المجرمين المحتملين. الإشكالية التي تسعى هذه الورقة لإيجاد حل تكمن في تطوير آلية لاستخلاص معلومات مهمة تفيد التحقيقات على أن تتم تلك العملية أوتوماتكيا من البيانات المتاحة، وذلك دون تضمين أي نموذج ومعرفة مسبقة في السياق التطبيقي. في هذه الورقة نستخدم أحد تقنيات البيانات الضخمة لتتبع تصرفات الأشخاص المتهمين وذلك من اجل الوصول إلى التصرفات الغير طبيعية التي يمكن أن تسهم في الوصول للمجرمين.

الكلمات المفتاحية: البيانات الضخمة – سجلات تفاصيل المكالمات – اكتشاف التصرفات الغير طبيعية.

I. Introduction

A Call Data Record (CDR) is a data structure storing relevant information about a given telephonic activity involving an user of a telephonic network. A CDR usually contains spatial and temporal data and it can carry other additional useful information.

Population census have been widely used in the past for keeping track of the demography and geographical movements of the population. Nowadays, due to short term and everyday mobility, more flexible methods such as various registers and indirect databases are employed: CDRs represent an optimal candidate in this sense [1, 2]. One of their main advantage is that they offer a statistically accurate representation of the distribution of people in an area and they can be used to track large and heterogeneous groups of people. Since CDRs evolve accordingly to the changes of user's behavior, the information they carry "automatically" updates over time [3, 4]. Telecom operators continuously gather a huge quantity of CDRs, from which it is possible to extract additional information with low additional costs and generate valuable datasets. Analyses of CDR data can be successfully employed in many different fields, like monitoring the network, adaptation of supplied services (e.g., customers' billing, network planning), understanding of the economic level of a certain area [5, 6].

The fact that a mobile phone can be a dangerous thing to have for a professional criminal has entered the popular culture a while ago. In the wire for example, drug dealers use "burners", cheap phones they dispose of regularly. This is because the phone operator is authorized to collect information about whom you call, for how long and from where [7]. In certain circumstances, that data can be used by law enforcement. Therefore, in this paper we are going to study the use of graph technologies to analyze phone calls to find criminals.

The remainder of this paper is organized as follows. Section II provides a review of CDR AND Neo4j. Section III presents a description of the problem. In Section IV we briefly describe the running example, Section V illustrates our effort to transform the call detail records into neo4j. Section VI describes the analysis of the CDR data using Neo4j and cyphe query language. Finally, the paper ends with a conclusion in Section VIII.

II. PRELIMINARY

This section provides a brief introduction to call detail records, Neo4j and Cypher query language.

A. Call Detail Records

A call detail record (CDR) is a data record produced by a telephone exchange or other telecommunications equipment that documents the details of a telephone call or other telecommunications transaction (e.g., text messages) and any other official communications transmission. that passes through that facility or device. The record contains various attributes of the call, such as call duration, start time, completion status, calling number, and called number. [8, 9]. The call detail record simply shows that the calls or messages took place, and measures basic call properties.

B. Neo4j

Neo4j is the implementation chosen to represent graph databases. It is open source for all noncommercial uses. It has been in production for over five years. It is quickly becoming one of the foremost graph database systems. According to the Neo4j website, Neo4j is “an embedded,disk-based, fully transactional Java persistence engine that stores data structured in graphs rather than in tables” [9]. The developers claim it is exceptionally scalable (several billion nodes on a single machine), has an API that is easy to use, and supports efficient traversals. Neo4j is built using Apache’s Lucene 3 for indexing and search. Lucene is a text search engine, written in Java, geared toward high performance.

C. Cypher query

Cypher is an expressive (yet compact) graph database query language. Cypher is designed to be easily read and understood by developers, database professionals, and business stakeholders. Its ease of use derives from the fact that it is in accord with the way we intuitively describe graphs using diagrams. Cypher enables a user (or an application acting on behalf of a user) to ask the database to find data that matches a specific pattern [9].

III. Description of the Problem

CDRs are very important to all mobile phone operators. These records are important because they contain all the information related to any phone calls. For example, each of these records contains data about the caller, the called, the duration of the call, etc. Analysis of such data is very complex and hard this because the huge size of the data and the data is stored in raw format. To illustrate out approach next we describe a running example.

IV. Running example

To illustrate our use case, let’s use a common scenario. In 23 Alsahaba Avenue in Tripoli, a store robbery is committed during the day by a group of 4 criminals. The criminals are masked, use a stolen vehicle and leave no fingerprints. In that kind of case, finding an answer may take a lot of legwork. A witness noticed that one of the criminal used his phone to make a call minutes before the crime.

Equipped with a search warrant, a police officer can contact mobile phone operators to collect information about the calls made and received near the robbery when it happened.

V. Transformation of CDR into Graph

The data phone operators provide law enforcement is highly tabular. Trying to identify unique phone numbers and their relationships in tabular data is very hard. We are thus going to use the phone calls data to build a graph. That graph will show how the phone numbers are connected by phone calls. From a list of calls, we are inferring a network.

We made a model for phone calls, where everything centered around calls. A single phone call connects together 4 entities: 2 phone owners, a location (the cell site the caller was next to when he initiated the call), a city and a neighborhood.

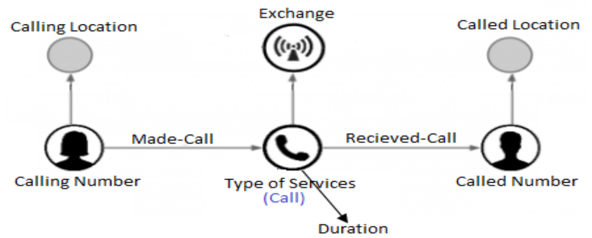

Fig. 1. Graph model to represent the phone calls.

It is important to note that in real life, most of the time we would not have access to the names of the phone numbers owners. Now that we have defined a model, we are going to populate it with the data stored in the spreadsheet. To store our graph, we will use Neo4j, a popular graph database. Neo4j has a language called Cypher that makes it easy to import csv files. The following is a sample of the code that have been generated to transform the CDR data into a Neo4j graph:

```
CREATE (f:Location { name:'Location' })
SET f = ROW, f.Location = (ROW.lastCallingLocationInformation)
CREATE (g:Located { name:'Located_from' })
SET g = ROW, g.Located = (ROW.lastCalledLocationInformation)
CREATE (s:duration{ name:'DURATION'})
SET s = ROW, s.duration = (ROW.chargeableDuration)
CREATE (CALL:Call{NAME:'CALL'})
CREATE (CallinNode)-[:Made_Call]->(CALL)-[:received_call]-
>(CalledNode), (Exchange)-[:FROM]-
>(CALL), (f)- [:FRM_BTS]->(CallinNode), (g)-[:TO_BTS]-
>(CalledNode), (s)-[:DURATION]- >(CALL)
WITH CallinNode as a
MATCH (a)-[:Made_Call]->(m)-[:received_call]->(d)
WHERE a.callingPartyNumber IS NOT NULL AND d.calledPar-
tyNumber IS NOT NULL
MATCH (Exchange)-[:FROM]->(m), (g)-[:TO_BTS]->(d), (f)-
[:FRM_BTS]->(a), (s)-[:DURATION]- >(m)
WHERE f.lastCallingLocationInformation IS NOT NULL AND
g.lastCalledLocationInformation IS
NOT NULL AND s.chargeableDuration IS NOT NULL
RETURN a.callingPartyNumber as Made_Call,f.lastCallingLocation-
Information as From_Location, d.calledPartyNumber as Re-
ceiver_Call,g.lastCalledLocationInformation as To_Location, Ex
change.exchangeIdentity as Exchange ,s.chargeableDuration as
Duration_of_Call
```

The previous code transforms the original table which consists all the CDR features into a new table using Neo4j as depicted in figure 2, consisting only from usefully features which will be used to detect the abnormal behavior such as call time, calling number, called number, location of calling number, location of called number, call duration, and exchange.

From_Locationa	Made_Call	To_Location	Receiver_Call	Exchange	Duration_of_Call
06F61006415639	928184803	06F61004507D6F	912365554	MSC-BC1 BENGHAZ	0:00:17
06F61008995A1F	914042247	06F61004507E6B	918022556	MSC-BC2_GURJI_1	0:00:16
06F61004B1826F	917215147	06F610044E7F82	916731558	MSC-BC2_GURJI_1	0:00:16
06F6100001274F	919847763	06F6100DADA867	918576567	MSC-BC1 BENGHAZ	0:00:29
06F6100003EE37	917737337	06F61004507E6A	919739042	MSC-BC3_RAS HAS	0:00:17
06F6100006A189	911845729	06F6100709819D	918324459	MSC-BC3_RAS HAS	0:00:17
06F610044F7E06	916376453	06F610044E84E5	913006038	MSC-BC3_RAS HAS	0:00:38
06F6100004795F	916927921	06F61004507E92	929479604	MSC-BC3_RAS HAS	0:00:13
06F6100002EF76	917215147	06F6100C1DA5E8	916731558	MSC-BC3_RAS HAS	0:01:31
06F610000A2B17	928184803	06F610044F800F	910499001	MSC-BC3_RAS HAS	0:00:36
06F61000062A63	915887713	06F6100CE5A281	924044076	MSC-BC3_RAS HAS	0:00:36
06F610000A2AFB	916376453	06F610044E84E5	913006038	MSC-BC1 BENGHAZ	0:00:36
06F6100002277F	917215147	06F61004507E92	918576567	MSC-BC1 BENGHAZ	0:00:36
06F6100005274D	914042247	06F610044F800F	910078874	MSC-BC1 BENGHAZ	0:00:16
06F610000C29F7	911845729	06F610044F800F	910078874	MSC-BC2_GURJI_1	0:00:37
06F61000054F58	910529310	06F6100CE5A281	929479604	MSC-BC2_GURJI_1	0:00:37
06F61000044EFE	929568378	06F61004B18266	915442740	MSC-BC1 BENGHAZ	0:03:25
06F610000A2B17	924666767	06F6100DADA867	912365554	MSC-BC1 BENGHAZ	0:00:53
06F6100001283C	915454553	06F61004507D6F	919132746	MSC-BC1 BENGHAZ	0:00:53
06F6100007272F	929568378	06F61004507E6B	917649626	MSC-BC3_RAS HAS	0:01:31
06F61000049CA5	917737337	06F610044E7F82	919132746	MSC-BC3_RAS HAS	0:00:36
06F61000049C73	913876665	06F6100DADA867	916520429	MSC-BC1 BENGHAZ	0:00:36
06F610000A2B17	913876665	06F61004507E6A	917935347	MSC-BC1 BENGHAZ	0:00:29
06F61000054F58	914147014	06F6100709819D	912334786	MSC-BC1 BENGHAZ	0:06:18
06F610000528AD	911184438	06F610044E84E5	917875547	MSC-BC2_GURJI_1	0:00:29
06F610000A532D	919322145	06F61004507E6A	928505039	MSC-BC2_GURJI_1	0:01:31
06F610000A2815	924666767	06F6100709819D	918576567	MSC-BC1 BENGHAZ	0:00:36
06F610000C29F6	915454553	06F610044E84E5	910078874	MSC-BC1 BENGHAZ	0:00:36

Fig. 2. New table using Neo4j.

VI. Analysis Data

A. Analysis of the phone records

In this section we aim to explore the phone call records in order to identify the criminal who made the phone call. Therefore, for the story described above we to assume that the robbery was perpetrated at 23 Alsahaba Avenue in Tripoli on the 25th of November, 2016 around 10:40 am.

B. Find the potential suspect

In that case, the police officers usually ask the phone operators for the phone calls made 10 minutes before and after 10:40am near 23 Alsahaba Avenue. getting the answer for such question may take long time from the phone operator this is because of the size of the data and its complexity. Therefore, we have created several cypher queries to answer the question that police officers might ask. The following is one of the cypher query that we have created:

```
match (a:call)-[:located_in]->(b:location) where b.cell_site =
'0101' or b.cell_site = '0102' and 10:29:00 <toint(a.start) and
toint(a.start) < 10:49:00 with a, b
match (c:person)-[:made_call]->(a)-[:received_call]->(d:person)
return c.full_name as caller, d.full_name as called, a.start as
time, a.duration as duration, b.address as address
```

The previous query searches for the phone calls which made from the two nearest towers from 23 Alsahaba Street, when the call started between 10:29 and 10:49. The results of executing the query is illustrated in table 1.

Table 1.potential Suspects

Caller	Called	Time	Duration	Address
A1	B1	10:30:24	12	23 Alsahaba
A2	B2	10:37:25	9	23 Alsahaba
A3	B3	10:47:36	43	23 Alsahaba

This list provides us three potential suspects. They have made phone calls in the vicinity of our crime location. The only problem is that we have multiple names. So, is one of them our perpetrator?

Let's say that as a police investigator the names is the list of suspects do not ring any bells. Then we need further digging to identify our perpetrator. We could interview the different suspects and check their background but we are going to use data to speed up our investigation:

```
MATCH (a:CALL)-[:LOCATED_IN]->(b:LOCATION)
WHERE b.cell_site = '0101' OR b.cell_site = '0102' AND
10:29:00 <toInt(a.start) AND toInt(a.start) < 10:49:00
 WITH a, b
MATCH (c:PERSON)-[:MADE_CALL]->(a)-
[:RECEIVED_CALL]->(d:PERSON)
WITH c, d
OPTIONAL MATCH (c:PERSON)-[:MADE_CALL]->(a)-
[:RECEIVED_CALL]->(d:PERSON)
RETURN e, c, d
```

We are reusing the query we build to find potential suspects by adding the last part that gives us the names of the people they are in contact with. These are the second degree contacts of our suspects.

We simply have to type the suspect names in the search bar and then visually query their relationships. The result is illustrated in table 2.

Table 2.relations of suspects

Full-Name	a.Made_Call	Receiver Call	F-name	Numbers_
A1	917aaaaaa	91BBBBBB	B1.0	6
		910BBBBB	B1.1	3
		918bbbbb	B1.2	3
		929bbbbb	B1.3	3
		9185bbbb	B1.4	3
A2	915aaaaaa	91bbbbbb	B2.0	9
		910bbbbb	B2.1	7
		918bbbbb	B2.2	2
		929bbbbb	B2.3	3
		9185bbbb	B2.4	1
A3	916aaaaaa	91bbbbbb	B3.0	8
		910bbbbb	B3.1	5
		918bbbbb	B3.2	5
		929bbbbb	B3.3	1
		9185bbbb	B3.4	1

The three suspects and the calls they made. Note that there is no connections between the different suspects. From the previous table we could notice that the phone calls made by each of our suspects. If we want to see the persons our suspects are in contact with, we have to display the persons connected to the calls.

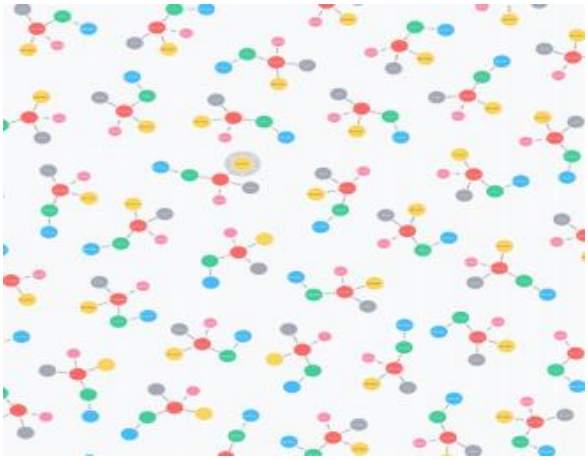

Fig. 3. The 3 suspects, the calls they made and who they made it to.

Graph visualization makes it easy to search and understand connected data. The picture above sums up the network of our suspects. That information would have required a long investigation with Excel or with traditional BI solutions.

To make the visualization more useful let's modify the data. Instead of displaying the people, the calls and the locations, we are going to focus on the people. To do that, let's create a direct relationship called "KNOWS" between everyone who share a phone call. This way we will display less data and it will be easier to analyses what is left.

```
MATCH (c:PERSON)-[:MADE_CALL]->(a)-
[:RECEIVED_CALL]->(d:PERSON)
CREATE (c)-[:KNOWS]->(d);
MATCH (a)-[r]-() WHERE NOT a:PERSON DELETE a, r;
```

The new graph schema is represented in Figure 4.

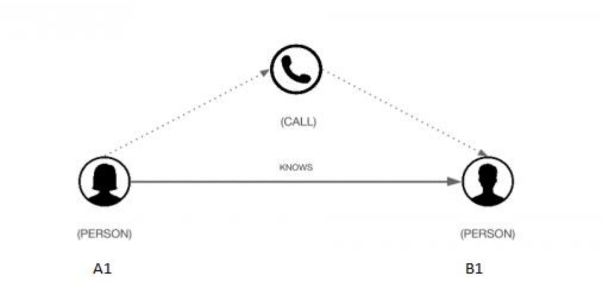

Fig. 3. New graph schema

C. Visual analysis of the network

The result of executing the query is illustrated in figure 5. From the graph depicted in figure 5 we could understand the network of our 3 suspects, A1, A2 and A3. However, such graph is fairly dense and thus hard to read. This is because it consists of 34 nodes and 150 relationships which represent the 3 suspects and the people they know. In order to simplify the graph, we can select one of the suspects to see his connections highlighted.

We assume that the police officer recognized a few names from the names that have connection with the suspect that has been selected because these names have already appeared in other investigations (i.e. B1.7 and B1.8). These names may not be directly tied to the crime we are investigating but they might be in contact with someone who is. Visually we can investigate that connection.

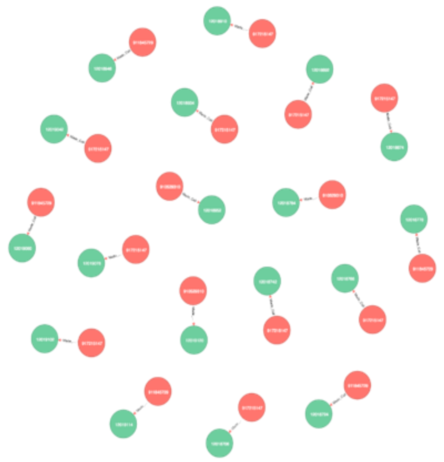

Fig. 4. The 3 suspects and the people they know.

VII. RESULTS

34 nodes and 150 relationships represent the three suspects and people they know. we can select one of the suspects to see his connections highlighted. As a police investigator we are going to assume that we recognize a few names that have already appeared in other investigations: B1.7 and B1.8. These people are not directly tied to the crime we are investigating but they are in contact with someone who is. Visually we can investigate that connection.

The phone call analysis shows that A1is connected to two known criminals: B1.7and B1.8. They are part of a small community within the larger graph. Among our initial suspects, A1 is the most likely to be a criminal. We should focus our investigation on him. In a few steps, we turned lines and lines of call records into one specific insight: A1 is the likeliest suspect in our criminal investigation. In order to achieve that result, we simply used the power of graph analysis.

VIII. CONCLUSIONS

Our proposed idea through this research paper is to develop a system which takes as input mobile number/s and extracts corresponding CDRs, thereby generating multiple databases of CDRs. After this it analyzes these databases and finds various links between various suspects (mobile numbers) and generates as output, conclusions of its analysis. This conclusionconsists of phone numbers, names of suspects. With proper analysis of the CDRs of the various suspects, the Anti - Crime team can move forward on multiple fronts simultaneously.

References

[1]- Geepalla E., Abuhamoud N., Abouda A. (2018) Analysis of Call Detail Records for Understanding Users Behavior and Anomaly Detection Using Neo4j. In: Alenezi M., Qureshi B. (eds) 5th International Symposium on Data Mining Applications. Advances in Intelligent Systems and Computing, vol 753. Springer, Cham

[2]- Kumar, Manish, M. Hanumanthappa, and TV Suresh Kumar. "Crime investigation and criminal network analysis using archive call detail records." In Advanced Computing (ICoAC), 2016 Eighth International Conference on, pp. 46-50. IEEE, 2017.

[3]- Sara B. Elagib A. Aisha-Hassan Hashim R. F. Olanrewaju "CDR analysis using Big Data technology" International Conference on Computing Control Networking Electronics and Embedded Systems Engineering (ICCNEEE) pp. 467-471 September 2015

[4]- Vukotic, Aleksa, Nicki Watt, Tareq Abedrabbo, Dominic Fox, and Jonas Partner. Neo4j in action. Manning Publications Co., 2014..

[5]- Gabi Kedma Mordehai Guri Tom Sela Yuval Elovici "Analyzing users' web surfing patterns to trace terrorists and criminals" Intelligence

and Security Informatics (ISI) pp. 143-145 June 2013.

[6]- Fiadino, Pierdomenico, Victor Ponce-Lopez, Juan Antonio, Marc Torrent-Moreno, and Alessandro D'Alconzo. "Call Detail Records for Human Mobility Studies: Taking Stock of the Situation in the "Always Connected Era"." In Proceedings of the Workshop on Big Data Analytics and Machine Learning for Data Communication Networks, pp. 43-48. ACM, 2017.

[7]- Hung, Shin-Yuan, David C. Yen, and Hsiu-Yu Wang. "Applying data mining to telecom churn management." Expert Systems with Applications 31, no. 3. 51 pp: 5-524. 2006.

[8]- Hoteit, Sahar, Guangshuo Chen, Aline Viana, and Marco Fiore. "Filling the gaps: On the completion of sparse call detail records for mobility analysis." The Eleventh ACM Workshop on Challenged Networks, pp. 45-50. 2016.

[9]- N. Abuhamoud, E. Geepalla. Using Neo4j For Fraud Detection In Banks Sector. In the Libyan International Conference on Electrical Engineering and Technologies (LICEET2018). 2018